ANALYZING AND IDENTIFYING UNUSUAL OBSERVATIONS IN MODIFIED LIU ESTIMATOR USING GLOBAL INFLUENCE TECHNIQUE

Aboobacker Jahufer and Mohamed Casim Alibuhtto

Department of Mathematical Sciences
Faculty of Applied Sciences
South Eastern University of Sri Lanka
jahufer@seu.ac.lk

Abstract

Influence concepts have an important place in linear regression models and case deletion is a useful method for assessing the influence of single case. The influence measures in the presence of multicollinearity were discussed under the linear regression models when the errors structure is uncorrelated and homoscedastic. When modified Liu estimator (MLE) is used to mitigate the effects of multicollinearity, the influence of observations can be drastically modified. In this research paper it is aimed to analyze global influence techniques to detect influential observations in MLE. To illustrate the methodologies derived in this research paper a multicollinearity real data set was used to identify influential observations using global influence techniques derived in this research paper.

Keywords: Global Influence Measures, Leverages; Residuals; Modified Liu Estimator; Multicollinearity

Introduction

In a linear regression model, the presence of multicollinearity can contribute very sensitive least squares statistical quantities. Therefore, mixed estimators and biased estimators techniques are proposed to mitigate the effect of multicollinearity. Among the biased estimation techniques, ridge type regression is very popular and extensively used measure in the presence of mulricollinearity. The multicollinearity and influential observations have come to cause substantial affect on the parametric estimation, estimation of biasing parameter in biased estimators and statistical inferences. However, many authors have attempted to prove the fact that the influence of the observations on ridge type biased estimators are different from that of the corresponding least-squares estimator and that the multicollinearity can even disguise anomalous data.

Identifying influential observations is an important step in the least squares regression model building process. Various influential measures, based on different motivational arguments have been designed to measure the influence of observations for unbiased estimators especially ordinary least squares estimator (OLSE), on the basis of a statistical analysis of the residuals and diagonal elements of a projection matrix, diagnostic plots for influential points indication are formed.

It is occasionally found that a small subset of the data exerts a disproportionate influence on the fitted regression model. That is, parameter estimates or predictions may depend more on the influential subset than on the majority of the data. It is fitting to locate these influential points and assess their impacts on the model. If these influential points are of bad values then they should be eliminated. On the other hand, there is nothing wrong with these points, if they control the key model properties, as it would be possible for them to be so because, they could affect the use of the model.

Although, many ridge type biased estimators have been proposed to fit the regression model when multicollinearity presents among the regressors very few methods were constructed or developed for detecting influential observations for ridge type biased estimators. Hence, in this research study it aimed to develop new techniques and methods based on global influential method proposed by Walker and Birch (1988) to identify influential cases ridge type Liu estimator namelymodified Liu estimator (MLE) proposed by Li and Yang (2012).

According to Belsley et al. (1980), regression diagnostics comprises a collection of methods used in the identification of influential points and multicollinearity. This includes methods of exploratory data analysis for the analysis of influential points and for the identification of violations of the assumptions of least squares. Notably, the detection of influential observations has received a great deal of attention in the statistical literature since the seminal work of Cook (1977) and many others including Belsley et al. (1980), Cook and Weisberg (1982), Walker and Birch (1988), Shi (1997) and Shi and Wang (1999).

Walker and Birch (1988) have analyzed Longly (1967) data to identify influence of observations on the biased estimator namely ordinary ridge regression estimator (ORRE) by employing approximated case deletion formulae. They derived an approximate case deletion formula to detect influential cases for ordinary ridge regression estimator.

Tsai and Wu (1990) described a diagnostic method for assessing the influence of an individual case on the transformation power estimator in the Box-Cox regression model and transform-both-sides regression model (Carroll and Ruppert 1988). They compare the method of those proposed by Cook and Wang (1983) and Hinkley and Wang (1988). The new method takes into account the deletion effect on the Jaconian of the variable's transformation. It provides a more accurate and reliable transformation power estimator. They also extend this method to analyze the case influence on the weighted parameter estimator in both the weighted regression model and the transformed and weighted regression model. Several examples are employed to illustrate their methodology.

Shi and Wang (1999) studied the above same Longly (1967) data to detect influence of observations on the ORRE using local influential analysis method by assessing influence on the selection of ORRE biasing parameter based on minor perturbations method. Therefore, the influence of the different aspects of the model can be well approached. Also they considered the perturbation of individual explanatory variables. They derived a new technique on biasing parameter of ORRE to detect influential cases in the data set.

This research paper is composed into five sections. Section 2 illustrates global influential measures in least squares. Section 3Methodology and in section 4 results and discussions are given. In the last section conclusion is given.

Global Influential Measures in Least Square Estimator

Background of Ordinary Least Squares Estimator (OLSE):

A matrix multiple linear regression model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is an n×1 response vector, $\mathbf{X}$ is an n×p centered and standardized known matrix (i.e. the length of the column of $\mathbf{X}$ is standardized to one), $\boldsymbol{\beta}$ is a p×1 vector of an unknown

parameter, $\boldsymbol{\varepsilon}$ is an n×1 error vector with $E(\boldsymbol{\varepsilon})=\mathbf{0}$ and $Var(\boldsymbol{\varepsilon})=\sigma^2\mathbf{I}_n$ and $\mathbf{I}_n$ is an identity matrix of order n. Then the OLSE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

The estimator of σ^2 is $s^2 = \mathbf{e}'\mathbf{e}/(n-p)$, where residual vector $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

Definition of influential measures in OLSE

The general purpose of influential analysis is to measure the changes induced in a given aspect of the analysis when the data set is perturbed. A particularly appealing perturbation scheme is case deletion. Note that this scheme is used throughout this article.

In general, the influence of a case can be viewed as the product of two factors: the first a function of the residual and the second a function of the position of the point in the **X** space. The position or leverage of the i-th point is measured by h_i, which is the i-th diagonal element of the "hat" matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Among the most popular single-case influential measure is the difference in fit standardized DFFITS (Belsley et al., 1980), which evaluated at the i-th case is given by

$$DFFITS(i) = \mathbf{x}_i(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))/SE(\mathbf{x}_i\hat{\boldsymbol{\beta}}), \quad (2)$$

Where $\hat{\boldsymbol{\beta}}(i)$ is the least squares estimator of $\boldsymbol{\beta}$ without the i^{th} case and $SE(x_i\hat{\boldsymbol{\beta}})$ is an estimator of the standard error (SE) of the fitted values. DFFITS is the standardized change in the fitted value of a case when it is deleted. Thus it can be considered a measure of influence on individual fitted values.

Another useful measure of influence is Cook's *D* (Cook and Weisberg, 1982), which evaluated at the i^{th} case is given by

$$D_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))/(ps^2), \quad (3)$$

Where$D_i$is a measure of the change in all of the fitted values when a case is deleted. Even though D_i is based on different theoretical consideration, it is closely related to DFFITS.

Points with large values of D_i have considerable influence on the least squares estimate $\hat{\boldsymbol{\beta}}$. In general, points for which $D_i > F_{(\alpha,p,n-p)}$ to be influential. In DFFITS measures any observation for which $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ warrants attention.

It is important to mention that these measures are useful for detecting *single* cases with an unduly high influence. These indexes, however, suffer from the problem of *masking*-that is, the presence of cases can disguise or mask the potential influence of other cases. For generalizations of equations (2) and (3) for detecting influential sets, see Cook and Weisberg (1982) and Belsley et al. (1980).

Methodology

Modified Liu Estimator (MLE)

The modified Liu estimator $\hat{\boldsymbol{\beta}}_{ML}$ proposed by Li and Yang (2012) andis defined as:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+\mathrm{dI})\hat{\boldsymbol{\beta}}+(1-\mathrm{d})\mathbf{b}_0], \tag{4}$$

Where**I** is an identity matrix, $\hat{\boldsymbol{\beta}}$ is OLSE, $\mathbf{b}_0$ is a non-stochastic vector (prior information vector) and*d* is Liu estimator biasing parameter. The MLE is very useful to mitigate the effect of near multicollinearity. Also, the recent literature, particularly in the area of econometrics, engineering and other statistical areas, the MLE has produced a number of new techniques.

Leverage and residual measures in MLE

In equation (4), if the prior information vector $\mathbf{b}_0$is assumed to be the OLSE then vector of fitted value of MLE becomes:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+\mathrm{dI})+(1-\mathrm{d})\mathbf{I}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\hat{\mathbf{Y}}_{\mathrm{ML}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = \mathbf{X}(\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+\mathrm{dI})+(1-\mathrm{d})\mathbf{I}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}_{\mathrm{d}}\mathbf{Y},$$

Where the MLE hat matrix $\mathbf{H}_{\mathrm{ML}} = \mathbf{X}(\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+\mathrm{dI})+(1-\mathrm{d})\mathbf{I}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ plays the same role as the hat matrix in OLSE. The i-thfitted value can be written in terms of elements of $\mathbf{H}_{\mathrm{ML}}$ as $\hat{y}_{\mathrm{MLi}} = \sum_{j=1}^{n} h_{\mathrm{MLij}} y_j$; consequently, $\frac{\partial \hat{y}_i}{\partial y_i} = h_{\mathrm{MLii}} \equiv h_{\mathrm{MLi}}$.

The "MLE hat diagonals" h_{MLi} can be interpreted as leverage in the same sense as the hat diagonals in OLSE. The single value decomposition (SVD) (see Mandel, 1982) allows **X** to be decomposed as **X**=**UDV**', where **D** is a p×p diagonal matrix with i-th diagonal elements $\lambda_i^{\frac{1}{2}}$ (λ_i is the i-th eigenvalue of **X**'**X**), the column of **V** are the eigenvectors of **X**'**X**. The (ij)-th element of the n×p matrix **U** (u_{ij}) is such that $u_{ij}\lambda_i^{\frac{1}{2}}$ is the projection of the i-th row, $\mathbf{x}_i$, onto the j-th principal axis (eigenvector) of **X**. Using the SVD, the Liu estimator leverage of the i-th point can be written as

$$h_{\mathrm{MLi}} = \sum_{j=1}^{p}\left[\frac{(\lambda_j+\mathrm{d})+(1-\mathrm{d})}{(\lambda_j+1)}\right]u_{ij}^2.$$

$$h_{\mathrm{MLi}} = \sum_{j=1}^{p} u_{ij}^2.$$

The i-th MLE residual is defined as

$$e_{\mathrm{MLi}} = y_i - \hat{y}_{\mathrm{MLi}} = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{\mathrm{ML}}.$$

DFFITS and Cook's Measures in Liu Estimator

The DFFITS for MLE can be written as

$$\mathrm{DFFITS}_{\mathrm{ML}}(\mathrm{i}) = \mathbf{x}_i(\hat{\boldsymbol{\beta}}_{\mathrm{ML}} - \hat{\boldsymbol{\beta}}_{\mathrm{ML}}(\mathrm{i}))/\mathrm{SE}(\mathbf{x}_i\hat{\boldsymbol{\beta}}_{\mathrm{ML}}), \tag{5}$$

Where $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}(\mathrm{i})$ is the MLE in (4) computed with the i-th case deleted and the denominator is an estimator of the standard error of the MLE fitted value. If MLE biasing parameter *d* is assumed non-stochastic, then

$$SE(\mathbf{x}_i\hat{\boldsymbol{\beta}}_{ML}) = s[\mathbf{x}_i(\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+d\mathbf{I})+(1-d)\mathbf{I}](\mathbf{X}'\mathbf{X})^{-1}$$
$$*[(\mathbf{X}'\mathbf{X}+d\mathbf{I})+(1-d)\mathbf{I}](\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}\mathbf{x}_i^{'}]^{1/2}$$

Hence the mean squared error is a function of the fitted values and the response, neither of which depends on individual eigenvalues of $\mathbf{X}'\mathbf{X}$; it is not affected by multicollinearity. For this reason, the OLSE of σ [s and s(i)] will be used as measures of scale.

At least two versions of Cook's D_i can be constructed for Liu estimator, they are

$$D_i^{*} = \frac{1}{ps^2}(\hat{\boldsymbol{\beta}}_{Ml} - \hat{\boldsymbol{\beta}}_{ML}(i))'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_{ML} - \hat{\boldsymbol{\beta}}_{ML}(i)) \quad (6)$$

and

$$D_{MLi}^{**} = \frac{1}{ps^2}(\hat{\boldsymbol{\beta}}_{ML} - \hat{\boldsymbol{\beta}}_{ML}(i))'(\mathbf{X}'\mathbf{X}+\mathbf{I})[(\mathbf{X}'\mathbf{X}+d\mathbf{I})^{-1}+(1-d)\mathbf{I}](\mathbf{X}'\mathbf{X}) \quad (7)$$
$$*[(\mathbf{X}'\mathbf{X}+d\mathbf{I})^{-1}+(1-d)\mathbf{I}](\mathbf{X}'\mathbf{X}+\mathbf{I})(\hat{\boldsymbol{\beta}}_{ML} - \hat{\boldsymbol{\beta}}_{ML}(i))$$

Where D_i^{*} is the direct generalization of Cook's D in (3) and D_i^{**} is based on the fact that

$$Var(\hat{\boldsymbol{\beta}}_{ML}) = \sigma^2(\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}[(\mathbf{X}'\mathbf{X}+d\mathbf{I})+(1-d)\mathbf{I}](\mathbf{X}'\mathbf{X})^{-1}$$
$$*[(\mathbf{X}'\mathbf{X}+d\mathbf{I})+(1-d\mathbf{I}](\mathbf{X}'\mathbf{X}+\mathbf{I})^{-1}.$$

Note that both D_i^{*} and D_i^{**} simplify to D_i in (3) when d=1.

It would be desirable to be able to write these measures as functions of leverage and residual, as was done in (2) and (3). This is not possible, however, because of the scale dependency of the Liu estimator. Since the Liu estimator is not scale invariant, **X**(i) (the **X** matrix with the i-th row deleted) has to be rescaled to unit-column length before computing $\hat{\boldsymbol{\beta}}_{ML}(i)$. In the following section some approximate deletion formulas are proposed.

Results and Discussions

The Longley data (Longley 1967) have been used to explain the effect of extreme multicollinearity on the OLSE. The scaled condition number (see Belsley et al., 1980) of this data set is 43,275. This large value suggests the presence of high level of multicollinearity among regressors. Cook (1977) used this data to identify the influential observations in OLSE using the method of Cook's D_i and found that cases 5, 16, 4, 10 and 15 (in this order) were the most influential observations (see Table 1). Walker and Birch (1988) analyzed the same data to detect anomalous observations in ORRE using case deletion method. They found that cases 16, 10, 4, 15 and 5 (in this order) were most influential observations (see Table 1). Shi and Wang (1999) also analyzed the same data to detect influential observations on the ridge regression estimator using local influence method. They detected cases 10, 4, 15, 16 and 1 (in this order) were the most five anomalous observations.

Table 1:Most first five influentialcasesusing OLSE and ORRE in Longley data

OLSE		ORRE	
Case	D_i^{*}	**Case**	D_i^{**}

5	0.614	16	0.582
16	0.467	10	0.251
4	0.244	4	0.219
10	0.235	15	0.145
15	0.170	5	0.142

In this paper, we used the same data set to assess the influential observations in modified Liu estimator using global influence methods derived in this paper such as: two versions of Cook's D_i, DFFITS, Leverage and Residual. The estimated results and the most five influential cases are given in Table 2.

Table 2:Most first five influentialcasesusing MLE in Longley data

Modified Liu Estimator									
Case	D_i^*	**Case**	D_{MLi}^{**}	**Case**	**DFFITS**	**Case**	**Leverage**	**Case**	**Residual**
10	0.652	10	0.601	10	0.452	5	0.624	10	-553.6
16	0.568	4	0.519	5	0.332	10	0.601	4	499.4
5	0.452	15	0.511	15	-0.298	4	0.542	15	456.1
4	0.410	16	0.403	4	0.212	6	0.412	5	-411.6
15	0.341	5	0.352	16	-0.126	15	0.342	6	313.3

In the above table-2 it clearly shows that the most first five observations are same as in ordinary least squares estimatorstudied by Cook (1977) and ordinary ridge regression estimator analyzed byWalker and Birch (1988),but only the order of the magnitudes are changed.

Conclusion

In this article, we showed that, when MLE is used to detect influential cases for multicollinearity data, the influence of each case changes as a function of the shrinkage parameter *d*. It is important that the MLE user not rely on influence measures obtained for least squares. Once the value of biasing parameter of MLE *d* is determined, influence measures should be computed for that *d*. Although no conventional cutoff points are introduced or developed for MLE using global influence diagnostic quantities. It is a bottleneck for cutoff values for the influence method. These are additional active issues for future research study. Case deletion approach techniques will be also accommodated for future research study.

Acknowledgements

This research work was supported by the research grant of South Eastern University of Sri Lanka. We thank the referees for the valuable and constructive comments and suggestions to improve the quality of the paper.

References

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

Carroll, R.J., and David R., (1988). *Transformation and weighting in Regression*, Ney York: Chapman & Hall.

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

Cook, R.D., and Weisberg, S., (1982). Criticism and Influence Analysis in Regression. *Sociological Methodology*, Vol. **13**, pp. 313-361.

Cook, R.D., and Wang, P.C., (1983). Transformations and Influential cases in Regression. *Technometrics*, Vol. **25**, No. 4, pp. 337-343.

Hinkley, D.V., and Wang, S., (1988), More About Transformations and Influential Cases in Regression, *Technometrics*, **30**, 435-440.

Longley, J.W., (1967). An Appraisal of Least Squares Programmes for the Electronic Computer from the Point of View of the User. *Journal of the American Statistical Association*, Vol. **62**, No. 319, pp. 819-841.

Li, Y., and Yang, H., (2012) A new Liu-type estimator in linear regression. Statistical Papers 53:427-437.

Shi, L., (1997). Local Influence in Principal Component Analysis. *Biometrika*. **84** (1), 175-186.
Shi, L., and Wang, X., (1999). Local Influence in Ridge Regression. *Computational Statistics & Data Analysis*, **31**, 341-353.

Tasi, C.L., and Wu, X., (1990). Diagnostics in Transformation and Weighted Regression. *Technometrics*, Vol. **32**, No. 30.

Walker, E., and Birch, J.B., (1988). Influence Measures in Ridge Regression. *Technometrics*, Vol. **30**, No. 2, pp. 221-227.