Zimbabwe Journal of Science & Technology pp 58-65 Vol.11 [2016]
e-ISSN 2409-0360
Zimbabwej.sci.technol

Modified Pena's Measure for Detecting Influential Observations in Biased Estimators

[1]Adewale F. Lukman, [1]Oyedeji Janet, [1]Abiola Racheal
[1] Department of Statistics, Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria.
Email: *wale3005@yahoo.com, janetatilolaoyedeji@gmail.com, abiolarachie2@yahoo.com*

Abstract

The detection of influential observations is an important aspect of model building in linear regression analysis because of their unduly large influence on regression coefficients. Different diagnostics measures on identifying these observations have been developed. In this list is the Pena's statistic diagnostic measure. In this study, Pena's approach is modified to Liu estimator (LE) and Two-parameter Liu-Ridge estimator (TPE). The performance of the proposed diagnostic for LE and TPE on detecting an influential observation is evaluated with two real life data. Results shows that that the proposed Pena measure performs very effectively in identifying influential observations and agrees with previous studies.

Keywords: Pena's statistic, Liu estimator, Two-parameter Liu-Ridge estimator, influential observations

Received:23.09.2016. Accepted:11.11.2016.

1. INTRODUCTION.

Besley *et al.* (1980) defined influential observations as one which either individually or jointly with other observations has a significant effect on particular estimates when compared to others. Influential measures are engaged to detect unusual observations that exert unduly influence on the parameter estimates of OLSE. In the last three decades, the impacts of influential observations on the regression coefficients of Ordinary Least Squares Estimator in linear regression model had received considerable attention since the seminal work of Cook (1977) and others including Besley *et al.* (1980), Cook and Weisberg (1982), Atkinson (1985), Chatterjee and Hadi (1986), Walker and Brich (1988), Belsley *et al*. (1989), Zhong, Wei and Fung (2000), Jahufer and Jianbao (2009) and Ullah and Pasha (2009) and others. The approach of most these numerical measures is to delete one data point and see how this will affects the vector of regression coefficients or the vector of forecasts. In recent years, Pena (2005) measures how the deletion of each sample point affects the forecast of a specific observation.

The objective of this article is not to propose a procedure for detecting or handling outliers but to modify Pena Statistic into Liu estimator (LE) and Two-parameter Liu-Ridge estimator (TPE).
The organization of this paper is as follows: Background information Pena statistic in OLS is given in section 2. Pena measures in LE and TPE are introduced in section 3. Section 4 covers application to two real life datasets. Conclusion is provided in the last session.

1.1 Pena Statistic for Diagnostic Analysis in Ordinary Least Squares

Consider the multiple linear regression model:

$$Y = X\beta + u \quad (1)$$

where y is an n×1 vector of response variable, X is an n×p full rank matrix of known regressors variables augmented with a column of ones. β is p×1 vector of the unknown regression coefficients and u is the nx1 vector of error terms such that $u\sim(0,\sigma^2 I_n)$ and I_n is an n x n identity matrix.

The OLS estimator is defined as:

$$\hat{\beta} = (X'X)^{-1}X'y \quad (2)$$

The residual vector is defined as e=y- $X\hat{\beta}$= (I-H) y where $H=X(X'X)^{-1}X'$ is the hat matrix.

The commonest approach in influential measure is to delete a data point and examine how this affects the regression coefficients or fitted values. The suggested measure by Cook (1977) is defined as

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_i)'(X'X)(\hat{\beta}-\hat{\beta}_i)}{ps^2} \quad (3)$$

where $\hat{\beta}_i$ is the least squares estimator of β when the ith case is deleted. This statistic can also be written as

$$D_i = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) \quad (4)$$

where h_{ii} is the diagonal elements of the hat matrix, p is the number of estimated parameters and where r_i is i^{th} internally studentized residual defined as

$$r_i^2 = \frac{e_i^2}{s^2(1-h_{ii})} \quad (5)$$

Pena (2005) introduced an alternative approach by measuring how the deletion of each sample points affects the forecast of a particular observation. This approach measure the influence of a point by the squared norm of the standardized vector s_i, that is,

$$S_i = \frac{s_i's_i}{pv(\hat{y}_i)} \quad (6)$$

where $v(\hat{y}_i) = s^2h_{ii}$ and $s_i=(\hat{y}_i - \hat{y}_{i(1)},\hat{y}_i - \hat{y}_{i(2)},\ldots\hat{y}_i - \hat{y}_{i(n)})$ such that $\hat{y}_i - \hat{y}_{i(j)}=\frac{h_{ij}e_j}{1-h_{jj}}$. This can be written equivalently as:

$$S_i = \frac{1}{ps^2h_{ii}}\sum_{j=1}^{n}\frac{h_{ij}^2e_j^2}{(1-h_{jj})^2} \quad (7)$$

S_i can also be written as a linear combination of the sample Cook's distances as:

$$S_i = \sum_{j=1}^{n}\rho_{ij}^2 D_j \quad (8)$$

where $\rho_{ij}^2 = \frac{h_{ij}^2}{h_{ii}h_{jj}} \leq 1$ is the correlation between $\hat{y}_i$ and $\hat{y}_j$.

1.2 Proposed Influential Measure

1.2.1 Pena Measure in Liu Estimator

Liu (1993) introduced the Liu estimator, $\hat{\beta}_d$, which combines the ridge estimator by Hoerl and Kennard (1970) with the stein estimator by Stein (1956). It is defined as:

$$\hat{\beta}_d = (X'X+I)^{-1}(X'y + d\hat{\beta}) \quad (9)$$

where I is an identity matrix, $\hat{\beta}$ is the least square estimator of β, d is referred to as the Liu biasing parameter

Jahufer (2013) defined Cook's Distance based on Liu estimator is as follows:

$$D_{iLIU}^{*} = \frac{1}{ps^2}\left[\frac{e_{di}}{1-m_{ii}}\right]^2 x_i'(X'X + kI)^{-1}(X'X)(X'X + kI)^{-1}x_i'$$

$$= \frac{1}{ps^2}\left[\frac{e_{di}}{1-m_{ii}}\right]^2 \sum_{j=1}^{n} h_{ij}^{2*} \quad (10)$$

Proposed Pena Statistic based on Liu Estimator is defined as:

$$S_i^{Liu} = \frac{1}{pv(\hat{y}_i)} \sum_{j=1}^{n} \frac{h_{ij}^{2*} e_j^{2*}}{\left(1-h_{jj}^{*}\right)^2} =$$

$$\frac{1}{p\sum_{j=1}^{n} h_{ij}^{2*}} \sum_{j=1}^{n} \frac{h_{ij}^{2*} e_j^{2*}}{\left(1-h_{jj}^{*}\right)^2} \quad (11)$$

where $v(\hat{y}_i)=s^2X(X'X+I)^{-1}(X'X+dI)(X'X)^{-1}(X'X+I)^{-1}(X'X+dI)X'=\sum_{j=1}^{n} h_{ij}^{2*}$

S_i^{Liu} can be written as a linear combination of D_{iTP}^{*} as

$$S_i^{Liu} = \sum_{j=1}^{n} \rho_{ij}^{2*} D_{jTP}^{*} \quad (12)$$

where $\rho_{ij}^{2*} = \frac{h_{ij}^{2*}}{h_{ii}^{2*} h_{jj}^{2*}} \leq 1$ is the correlation between $\hat{y}_{di}$ and $\hat{y}_{dj}$.

1.2.2 Pena Measure in Two-Parameter Liu-Ridge Estimator

Ozkale and Kaciranlar (2007) introduced a Two-Parameter Ridge-Liu estimator (TPE) given as:

$$\hat{\beta}_{TP} = (X'X+kI)^{-1}(X'y + kd\hat{\beta}) \quad (13)$$

where I is an identity matrix, k is the ridge parameter and d is Liu biasing parameter. $\hat{\beta}$ is the least square estimator of β. TPE is introduced for handling the problem of multicollinearity in the linear regression model. In this study, Cook's Distance based on Two-Parameter Liu-Ridge parameter is defined as follows:

$$D_{iTP}^{**} = \frac{1}{ps^2}\left[\frac{e_{tpi}}{1-m_{ii}}\right]^2 x_i'(X'X+kI)^{-1}(X'X)(X'X+kI)^{-1}x_i'$$

$$= \frac{1}{ps^2}\left[\frac{e_{tpi}}{1-m_{ii}}\right]^2 \sum_{j=1}^{n} h_{ij}^{2**} \quad (14)$$

Proposed Pena Statistic based on Two-Parameter Liu-Ridge Estimator is defined as:

$$S_i^{TP} = \frac{1}{pv(\hat{y}_i)} \sum_{j=1}^{n} \frac{h_{ij}^{2**} e_j^{2**}}{\left(1-h_{jj}^{**}\right)^2} =$$

$$\frac{1}{p\sum_{j=1}^{n} h_{ij}^{2**}} \sum_{j=1}^{n} \frac{h_{ij}^{2**} e_j^{2**}}{\left(1-h_{jj}^{**}\right)^2} \quad (15)$$

where $v(\hat{y}_i)=s^2X(X'X+kI)^{-1}(X'X+kdI)(X'X)^{-1}(X'X+kI)^{-1}(X'X+kdI)X'=\sum_{j=1}^{n} h_{ij}^{2**}$

S_i^{TP} can be written as a linear combination of D_{iTP}^{**} as

$$S_i^{TP} = \sum_{j=1}^{n} \rho_{ij}^{2**} D_{jTP}^{**} \quad (16)$$

where $\rho_{ij}^{2**} = \frac{h_{ij}^{2**}}{h_{ii}^{2**} h_{jj}^{2**}} \leq 1$ is the correlation between $\hat{y}_{tpi}$ and $\hat{y}_{tpj}$.

2. METHODS

2.1 Application to Real life Dataset

Real life data sets are used to illustrate the performance of the influential statistics.

3. RESULTS

3.1 Application to Longley Data

This was adopted from the study of Longley (1967). The datasets consist of six independent variables (X_i where i=1,2,…6) and one dependent variable, y. The Independent variables are as follows: X_1 is the gross national product implicit price deflator, X_2 is the gross national product, X_3 is unemployment, X_4 is the size of armed forces, X_5 is the non-institutional population 14 years of age and over and X_6 is the time. The scaled condition number of this data is 43,275 (Walker and Birch, 1988). Table 1 provides a summary of the different authors, methods and influential points detected. Using Pena Statistic on OLS, the influential observations identified are cases 5, 16, 6, 15 and 10. The proposed Pena Statistic based on Two-parameter Liu-Ridge estimator identified cases 4, 16, 10, 5 and 15. This agrees with the results of Cook (1977) and Ullah *et al*. (2013). The proposed Pena Statistic based on Liu estimator identified cases 16, 15, 4, 1 and 10. This agrees with the results of Walker

and Birch (1988), Shi and Wang (1999); and Jahufer and Jianbao (2009). The value of Pena Statistic based on TPE and LE are provided in Table 2 and Pena Statistic based on TPE is graphically illustrated in Figure1.

Table 1. Summary of most influential diagnostics on Longley datasets

Authors	Year	Influential points in order	Method
Cook	1977	5, 16, 4, 10, 15	Cooks distance in OLS
Walker and Birch	1988	16, 10, 4, 15, 1	Cooks distance in ridge regression
Shi and Wang	1999	10, 4, 15, 16 and 1	Local influence in ridge regression
Jahufer and Jianbao	2009	16, 4, 1, 10, 15	Modified Ridge regression of the usual diagnostics
Jahufer and Chen	2012	4, 10, 1, 5 and 6	Liu estimator in ordinary regression
Jahufer	2013	15,4 1, 6, 16	Cooks distance and DFFITs in Liu regression
Ullah et al.	2013	16, 10, 4, 6, 1 16, 10,4, 6, 5 and 16, 5, 4, 10, 15	Influential points in Liu regression for different d=0.1, 0.5 and 0.9 respectively
Emami	2015	12, 16, 2 and 5	Case deletion based on ridge semi-parametric regression model
Yasin and Murat	2016	16, 10, 6, 1 and 4	Influential points in Two–parameter ridge

Table 2. Pena Statistics Value based on Two-Parameter Liu-Ridge estimator

Cases	1	2	3	4	5	6	7	8	9	10	11
$\mathbf{S_i^{TP}}$	**0.663**	0.188	0.473	**0.930**	**0.709**	0.632	0.409	0.096	0.564	**0.814**	0.479
$\mathbf{S_i^{LIU}}$	**0.216**	0.110	0.159	**0.283**	0.102	0.114	0.114	0.012	0.136	**0.189**	0.099
$\mathbf{D^*_{iLIU}}$	**0.139**	0.036	0.004	**0.212**	0.033	0.068	0.035	0.000	0.000	**0.092**	0.000
$\mathbf{D^{**}_{iTP}}$	**0.333**	0.073	0.008	0.700	**0.486**	0.244	0.111	0.000	0.001	**0.662**	0.000
$\mathbf{D^*_{iOLS}}$	**0.141**	0.041	0.003	**0.244**	**0.614**	0.009	0.079	0.000	0.000	**0.235**	0.000

Cases	12	13	14	15	16
$\mathbf{S_i^{TP}}$	0.193	0.356	0.342	0.698	**0.845**
$\mathbf{S_i^{LIU}}$	0.084	0.128	0.170	**0.290**	**0.411**
$\mathbf{D^*_{iLIU}}$	0.004	0.039	0.006	**0.096**	**0.342**
$\mathbf{D^{**}_{iTP}}$	0.008	0.085	0.017	**0.280**	**0.659**
$\mathbf{D^*_{iOLS}}$	0.004	0.036	0.004	0.170	**0.467**

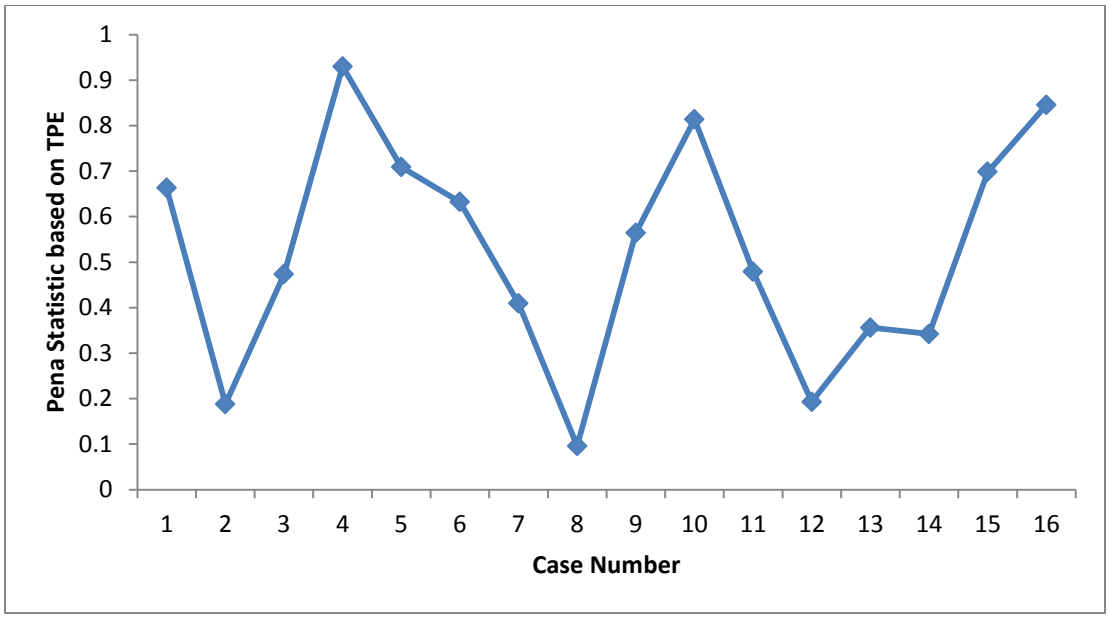

Figure 1. Plot of Pena Statistic based on TPE against its observations using Longley dataset

3.2 Application to Hald Data

Cook (1977); Yasin and Murat (2016) both adopted this dataset in the investigation of influential observations. It consists of five economic variables with thirteen (13) observations. The scaled condition number of this dataset is 249.578. This value suggests the presence of high multicollinearity. Cook (1977) applied Cooks Distance to the dataset and found that cases 8, 3, 11, 6 and 13 (in this order) were the most influential points in OLS. Yasin and Murat (2016) applied Cooks Distance based on Two-parameter ridge and found that cases 8, 11, 10, 6 and 13 (in this order) were the most influential points. In this study, cases 8, 11, 6, 13 and 7 (in this order) were identified when Cooks D in Liu estimator was used. Also, applying Cooks D based on Two-parameter Liu-Ridge cases 8, 13, 6, 11 and 2 (in this order) were identified as influential observations. Using Pena Statistic on OLS, the influential observations identified are cases 8, 11, 6, 2 and 10. The value of k and d in this study are computed to be 0.0076761 and 1.18495 respectively. The proposed Pena Statistic based on Liu estimator identified cases 8, 11, 4, 7 and 13 (in this order) as influential points. This is further illustrated in Figure 3. Cases 11, 8, 12, 6 and 13 were found with the application of Pena Statistic on Two-Parameter Liu-Ridge estimator. It was observed that the same cases given in Cook (1977) except for case 12 instead of case 3 is identified as influential case. This is further illustrated in Figure 2. Consistently, from this study and previous studies cases 8, 11 and 13 were identified more frequently even though it appears in different order. The results of the proposed Pena statistic based on Two-Parameter Liu-ridge and Liu estimators are summarized in Table 3.

Table 3. Cooks D and Pena Statistics Values based on TPE and LE using Hald dataset

Cases	1	2	3	4	5	6	7	8	9	10	11	12	13
S_i^{TP}	0.032	0.125	**0.179**	0.159	0.096	**0.257**	0.061	**0.471**	0.122	0.127	**0.645**	**0.275**	0.173
S_i^{LIU}	0.036	0.122	0.119	**0.194**	0.077	0.133	**0.143**	**0.456**	0.098	0.076	**0.441**	0.137	**0.138**
D^*_{iLIU}	0.001	0.037	0.020	0.044	0.003	**0.075**	**0.060**	**0.320**	0.015	0.021	**0.135**	0.012	**0.064**
D^{**}_{iTP}	0.015	**0.070**	0.001	0.023	0.000	**0.084**	0.070	**0.411**	0.040	0.000	**0.079**	0.002	**0.150**
D^*_{iOLS}	0.000	0.057	**0.309**	0.059	0.002	**0.083**	0.064	**0.394**	0.038	0.021	**0.171**	0.015	**0.110**

Note: The bolded values are the first five influential points**.**

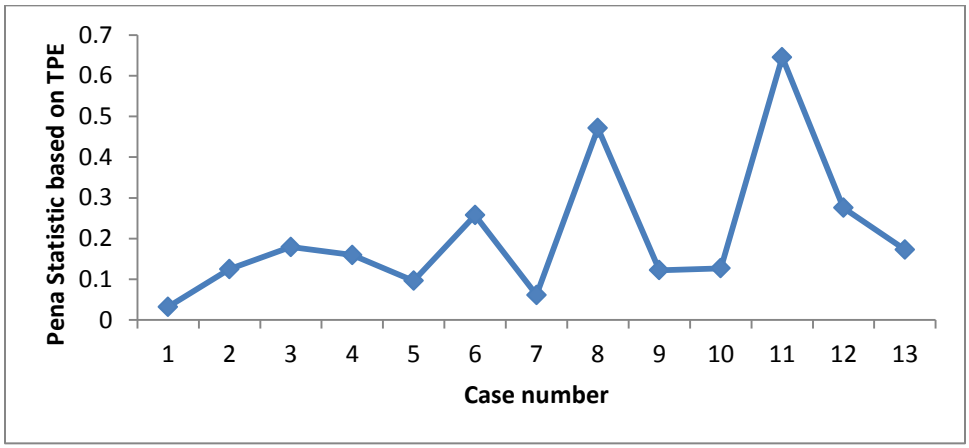

Figure 2. Plot of Pena Statistic based on TPE against its observations using Hald dataset

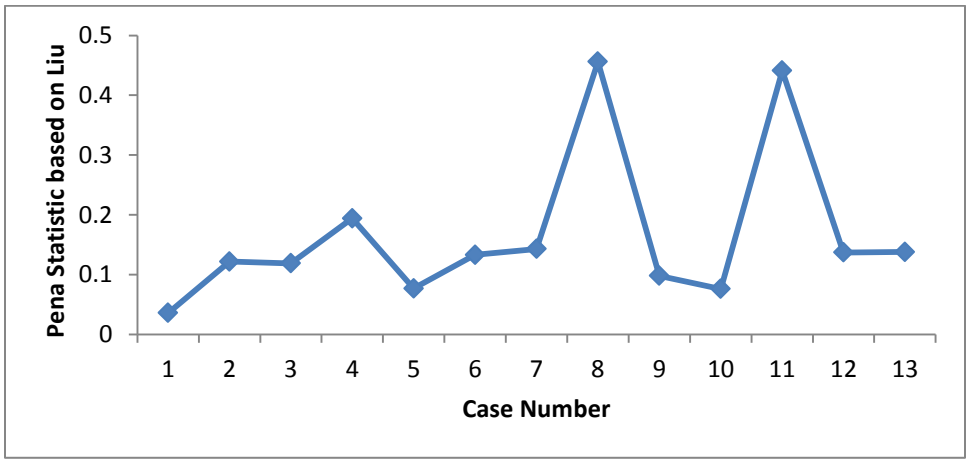

Figure 3. Plot of Pena Statistic based on Liu against its observations using Hald dataset

4. DISCUSSION AND CONCLUSIONS

In this paper, the problem of multicollinearity and influential observations were jointly considered. As suggested by Belsley *et al*. (1989), multicollinearity should be handled before attempting to detect influential points. Biased estimators such as Liu and Two-parameter Liu-ridge estimators are used to reduce the effect of multicollinearity and influence measures of Pena Statistic modified into these estimators. Application to real life data show that the proposed Pena measure S_i^{LIU} and S_i^{TP} performs very effectively in the identification influential observations. The result also agrees with previous studies.

REFERENCES

Atkinson, A. C. (1985). Plots, Transformations and Regression: Oxford, U.K. Clarendon Press.

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

Belsley, D. A., Kuh, E., and Welsch, R. E. (1989). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity New York : Wiley

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 and Hall, New York.

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

Emami, H. (2015). Influence diagnostics in ridge semiparametric regression models. Statistics and Probability Letters 105,106–115.

Hald, A. (1952). Statistical Theory with Engineering Applications. Wiley, New York.

Jahufer, A., and Jianbao, C. (2009). Assessing global influential observations in modified ridge regression. Statistics & Probability Letters 79(4), 513-518.

Jahufer, A. and Chen, J. (2012). Identifying Local Influential Observations in Liu Estimator. Journal of Metrika 75(3), 425-438.

Jahufer, A. (2013). Detecting Global Influential Observations in Liu Regression Model. Open Journal of Statistics 3(1), 5-11.

Liu, K. (1993). A new class of biased estimate in linear regression. Communications in Statistics 22(2), 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.

Ozkale, M. R. and Kaciranlar, S., (2007). The restricted and unrestricted two-parameter estimators. Communications in. Statististical Theory and Methods 36, 2707–2725.

Pena, D. A. (2005). New Statistic for Influence in Linear Regression. Journal of American Statisticians Association 47, 1-13.

Shi, L., and Wang, X. (1999). Local influence in ridge regression. Computational Statistics and Data Analysis 31(3), 341-353.

Stein, C. (1956). Inadmissibility of the usual estimator for the Mean of multivariate normal distribution. Third edition- Berkeley; California.

Ullah, M. A., Pasha, G. R. (2009). The Origin and Developments of Influence Measures in Regression. Pakistan Journal of Statistics 25, 295-307.

Ullah, M. A., Pasha, G. R., and Aslam, M. (2013). Assessing Influence on the Liu Estimates in Linear Regression Models. Communications in Statistics - Theory and Methods 42(17), 3100-3116.

Yasin, A. and Murat, E. (2016). Influence Diagnostics in Two-Parameter Ridge Regression. Journal of Data Science 14, 33-52.

Walker, E. and Birch, J. B. (1988). Influence Measures in Ridge Regression. Technometrics 30(2), 221- 227.

Zhong, X., Wei, B. and Fung, W. (2000). Influence Analysis for Measurement Error Models. Ann. Inst. Statistics Mathematics 52, 367-379.