Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 290-302
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# On Identifying Influential Observations in the Presence of Multicollinearity

**Chinwendu Alice Uzuke, Ifeyinwa Christiana Ezeilo**

Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
Email: ca.uzuke@unizik.edu.ng, prisca087@yahoo.com

**How to cite this paper:** Uzuke, C.A. and Ezeilo, I.C. (2021) On Identifying Influential Observations in the Presence of Multicollinearity. *Open Journal of Statistics*, **11**, 290-302.
https://doi.org/10.4236/ojs.2021.112016

**Received:** December 3, 2020
**Accepted:** March 29, 2021
**Published:** April 1, 2021



## Abstract

Influential observation is one which either individually or together with several other observations has a demonstrably large impact on the values of various estimates of regression coefficient. It has been suggested by some authors that multicollinearity should be controlled before attempting to measure influence of data point. In using ridge regression to mitigate the effect of multicollinearity, there arises a problem of choosing possible of ridge parameter that guarantees stable regression coefficients in the regression model. This paper seeks to check whether the choice of ridge parameter estimator influences the identified influential data points.

## Keywords

Multicollinerity, Ridge Parameter, Influential Measures, Outliers, Leverage Point

## 1. Introduction

It is well understood that not all observations in the data set play equal role when fitting a regression model. We occasionally find that a single or small subset of the data exerts a disproportionate influence on the fitted regression model. That is, parameter estimates or prediction may depend more on the influential subset than the majority of the data. Belsley *et al.* [1] defined an influential observation as one which either individually or together with several other observations has demonstrably large impact on the calculated values of various estimates, than is the case of most of the other observations. Influential observation in either dependent or independent variable can be as a result of data error or other problem, for example, the influential data points in dependent variable can arise from skewness in the independent variable or from differences in the data generation process for small subset of sample. Obviously, outliers which are observations in

a data set which appears to be inconsistent with the remainder of other set of data [2] need not be influential observation in affecting the regression Equation [3]. Andrew and Pregibon [4] highlighted the need to find outliers that matter. They stated that it is not all outliers that need to be harmful in the way that they have undue influence on for instance, the estimation of the parameters in the regression model. If not all outliers matter, examining residual alone might not lead to the detection of influential observation. Thus, other ways of detecting influential observations are needed.

Regression diagnostic comprises of a collection of method used in the identification of influential points and multicollinearity [1]. This includes methods of exploratory data analysis for influential points and identification of violation of assumption of least squares. When the assumption of Ordinary Least Squares (OLS) method that the explanatory variables are not linearly correlated is violated, this results to multicollinearity problem and should be controlled before attempting to measure influence [1]. One of the most popular methods of controlling multicollinearity is the use of Ridge Regression (RR) suggested by Hoerl and Kennard [5]. The idea in RR method is to add small positive number ($k > 0$) to diagonal elements of the matrix $\left(X'X\right)$ in order to obtain a ridge regression estimator

$$\hat{\beta}_R = \left(X'X + kI\right)^{-1} X'Y \tag{1}$$

Though the estimator obtained is bias but it yields minimum Mean Squares Error (MSE) when compared to OLS estimator. If $k = 0$, $\hat{\beta}_R$ becomes the unbiased OLS estimator ($\hat{\beta}$).The choice of ridge parameter $k$ has always been a problem in using RR to solve for multicollinearity, hence methods of estimating the value of $k$ had been suggested by several authors. Below are some suggested methods of estimating $k$: Hoerl and Kennard [5], Hoerl *et al.* [6], Lawless and Wang [7], Nomura [8], Khalaf and Shukur [9], Dorugade [10], Al-Hassan [11], Dorugade and Kashid [12], Saleh and Kibria [13], Kibria [14], Zang and Ibrahim [15], Alkhamisi *et al.* [16], Al-Hassan [17], Muniz and Kibria [18], Khalaf and Shukur [9], Khalaf and Mohamed [19], Uzuke *et al.* [20] etc.

Several diagnostic methods have been developed to detect influential observation. Firstly, Cook [21] introduced Cook's distance ($D_i$) which is based on deleting the observations one after another and measuring their effect on linear regression model. Other measures developed on the idea of Cook's distance includes; modified cook's distance ($D_i^*$), DFFITS, Hadi's measure, Pena statistic, DFBETAS, COVRATIO, etc.

Therefore, problem of multicollinearity and influential observation affect the regression analysis or estimates remarkably. And in using Ridge Regression to mitigate multicollinearity problem, there is always a problem of the method to use to estimate the ridge parameter ($k$) to achieve reduction in variance larger than increase in bias furthermore, one may want to know whether multiticollinearity affects identification of influential observations.

## 2. Methodology

The influence of an observation is measured by the effect it produces on the fit when it is deleted in the fitting process. This deletion is always done one point at a time. Let $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}, \cdots, \hat{\beta}_{p(i)}$ denote the regression coefficients obtained when the ith observation is deleted $(i = 1, 2, \cdots, n)$. Similarly, let $\hat{y}_{1(i)}, \hat{y}_{2(i)}, \cdots, \hat{y}_{n(i)}$ and $\hat{\sigma}^2_{(i)}$ be the predicted values and residual mean square respectively when the ith observation is dropped. Note that

$$\hat{y}_{m(i)} = \beta_{0(i)} + \hat{\beta}_{1(i)} x_{m1} + \cdots + \hat{\beta}_{p(i)} x_{mp} \tag{2}$$

is the fitted value for the observations $m$ when the fitted equation is obtained with the ith observation deleted. Influential measures look at differences produced in quantities such as $\left(\hat{\beta}_j - \hat{\beta}_{j(i)}\right)$ or $\left(\hat{y}_j - \hat{y}_{j(i)}\right)$. Several diagnostic methods have been developed to detect influential observation. Firstly, Cook [21] introduced Cook's Distance ($D_i$) which is based on deleting the observations one after another and measuring their effect on linear regression model. Other measures developed on the idea of Cook's Distance includes; modified Cook's Distance ($D_i^*$), DFFITs, Hadi's influence measure, Pena statistic, DFBETAS, COVRATIO, etc. This work, adopted the following influential measures;

1) Cook's Distance

Cook [21] proposed this measure and it is widely used. Cook's distance measures the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the ith observation. Cook's distance measure is defined as,

$$C_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{\hat{\sigma}^2 (p+1)} \tag{3}$$

which can also be expressed as

$$C_i = \frac{r_i^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}} \tag{4}$$

Thus, Cook's distance is a multiplication function of two quantities. The first term in Equation (4) is the square of the standardized residual $r_i$, which is given as $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ and the second term is called potential function $\frac{h_{ii}}{1-h_{ii}}$ where $h_{ii}$ is the leverage of the ith observation given as $h_{ii} = X'\left(X'X + kI\right)^{-1} X$. If a point is influential, its deletion causes large changes and the value of $C_i$ will be large. Therefore, large value of $C_i$ indicates that the point is influential. It has also be suggested that points with $C_i$ value greater than the 50% point of the F distribution with $p + 1$ and $(n - p - 1)$ degrees of freedom be classified as influential points.

2) Welsch and Kuh Measure

Welsch and Kuh [22] developed a similar measure to Cook's Distance named DFFITs, defined as

$$\text{DFFITs}_i = \frac{\hat{y}_j - \hat{y}_{j(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} \tag{5}$$

$\text{DFFITs}_i$ is the scaled difference between the *ith* fitted value obtained from the full data and the ith fitted value obtained by deleting the *ith* observation. $\text{DFFITs}_i$ can as well be written as

$$\text{DFFITs}_i = r_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}},\ i = 1, 2, \cdots, n \tag{6}$$

where $r_i^*$ is the standardized residual defined as $r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$.

Points with $\left|\text{DFFITs}_i\right| > 2\sqrt{p+1/(n-p-1)}$ are usually classified as influential points.

3) Hadi's Influence Measure

Hadi [23] proposed a measure of the influence of ith observation based on the fact that influential observations are outliers in the response variable or in the predictors or both. Accordingly, the influence of the ith observation can be measured by

$$H_i = \frac{h_{ii}}{1-h_{ii}} + \frac{p+1}{1-h_{ii}}\frac{d_i^2}{1-d_i^2}, i = 1, 2, \cdots, n \tag{7}$$

where $d_i = \frac{e_i}{\sqrt{SSE}}$ (normalized residual). $H_i$ is an additive function. The first term of the equation is the potential function which measures outlyingness in the X-space and the second term is a function of the residual, which measures outlyingness in the response variable. Observations with large $H_i$ are influential observations in the response and/or the predictor variables. Although the measure $H_i$ does not focus on a specific regression result, but it can be thought of as an overall general measure of influence which depicts observations that are influential on at least one regression result.

4) DFBETAS [1]

DFBETAS measures the difference in each parameter estimate with and without the influential data point. It is an influential measure used to ascertain which observation influence specific regression coefficient

$$\text{DFBETAS}_{ij} = \frac{b_j - b_{j(i)}}{\sqrt{s_{(i)}^2 \left(X'X\right)_{ij}^{-1}}} \tag{8}$$

where $b_{j(i)}$ denote the regression coefficients obtained when the ith observation is deleted in fitting process $(i = 1, 2, \cdots, n)$ and $b_j$ the predicted values from the full data, when ith observation is used in the fitting process.

5) Kuh and Welsch Ratio (COVRATIO)

The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation. This influential measure is given as

$$\text{COVRATIO} = \left[\frac{\det\left(s_i^2 \left(X_i'X_i\right)^{-1}\right)}{\det\left(s^2 \left(X'X\right)^{-1}\right)}\right] \tag{9}$$

which can also be expressed as below

$$\text{COVRATIO} = \frac{\left(\frac{n-p'-r_i^2}{n-p'-1}\right)}{1-h_{ii}} \tag{10}$$

where $n$ is the sample size, $p'$ is the number of independent variable and $h_{ii}$ is the hat matrix.

The ridge parameter estimators which were selected to control multicollinearity are

a) $\hat{k}_1 = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$ Hoerl and Kennard [5]

b) $\hat{k}_2 = \frac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p} \hat{\alpha}_i^2\right)^{1/p}}$ Kibria [14]

c) $\hat{k}_3 = \max_{1\le i\le p}\left(\frac{\lambda_i S^2}{\lambda_i \hat{\beta}_i^2 + (n-p)S^2}\right)$ Alkhamisi *et al.* [16]

d) $\hat{k}_4 = \left(\prod_{i=1}^{p} \frac{\lambda_i \hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + \lambda_i \hat{\alpha}_i^2}\right)^{\frac{1}{p}}$ Muniz and Kibria [18]

e) $\hat{k}_5 = \left(\prod_{i=1}^{p} \frac{1}{m_i}\right)^{\frac{1}{p}}$ Muniz and Kibria [18]

f) $\hat{k}_6 = \left(\prod_{i=1}^{p} m_i\right)^{\frac{1}{p}}$ Muniz and Kibria [18]

g) $\hat{k}_7 = \text{median}\left(\frac{1}{m_i}\right)$ Muniz and Kibria [18]

where $m_i = \sqrt{\frac{\hat{\sigma}_i^2}{\hat{\alpha}_i^2}}$

h) $\hat{k}_8 = \frac{2p}{\lambda_{\max}} \sum_{i=1}^{p} \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$ Dorugade [10]

i) $\hat{k}_9 = \left(\prod_{j=1}^{p} w_j\right)^{1/p}$ Uzuke *et al.*, [20]

where $w_j = \frac{In2\left(\hat{\sigma}^2\right)}{(n-p)\hat{\sigma}^2 + In2\left(\hat{\alpha}_j^2\right)}$

j) $\hat{k}_{10} = \left(X'X\right)^{-1} X'Y$

## 3. Illustration

Using the Nigeria Economic indicator (1980-2010) data from the Central Bank of Nigeria (CBN) Statistical Bulletin 2010. The data consist of Gross Domestic Product as the dependent variable ($y$) and ten [10] independent variables namely Money Supply ($x_1$), Credit to Private Sector ($x_2$), Exchange Rate ($x_3$), External Reserve ($x_4$), Agricultural Loan ($x_5$), Foreign Reserve ($x_6$), Oil Import ($x_7$),

Non-oil Export ($x_8$), Oil Export ($x_9$), and Non-oil Export ($x_{10}$) shown in **Appendix III**.

Table 1 showed that there is presence of multicollinearity in the data, since most of the independent variables have VIF > 10, the eigen-value close to zero(0), T < 0.1 and CN > 5 The correlation matrix of the data set also showed the presence of multicollinearity.

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 0.7952 | 0.7218 | 0.7309 | 0.7838 | 0.7757 | 0.7789 | 0.8146 | 0.7532 | 0.7768 |
| $x_2$ | 0.7952 | 1 | 0.6813 | 0.8586 | 0.9702 | 0.9168 | 0.9420 | 0.9517 | 0.8851 | 0.9693 |
| $x_3$ | 0.7218 | 0.6813 | 1 | 0.7277 | 0.7507 | 0.8270 | 0.7650 | 0.8234 | 0.8350 | 0.7810 |
| $x_4$ | 0.7309 | 0.8586 | 0.7277 | 1 | 0.9372 | 0.9317 | 0.8657 | 0.8891 | 0.9438 | 0.8781 |
| $x_5$ | 0.7838 | 0.9702 | 0.7507 | 0.9372 | 1 | 0.9580 | 0.9365 | 0.9596 | 0.9505 | 0.9675 |
| $x_6$ | 0.7757 | 0.9168 | 0.8270 | 0.9317 | 0.9580 | 1 | 0.9660 | 0.9785 | 0.9877 | 0.9631 |
| $x_7$ | 0.7789 | 0.9420 | 0.7650 | 0.8657 | 0.9365 | 0.9660 | 1 | 0.9801 | 0.9455 | 0.9705 |
| $x_8$ | 0.8146 | 0.9517 | 0.8234 | 0.8891 | 0.9596 | 0.9785 | 0.9801 | 1 | 0.9612 | 0.9905 |
| $x_9$ | 0.7532 | 0.8851 | 0.8350 | 0.9438 | 0.9505 | 0.9877 | 0.9455 | 0.9612 | 1 | 0.9406 |
| $x_{10}$ | 0.7768 | 0.9693 | 0.7810 | 0.8781 | 0.9675 | 0.9631 | 0.9705 | 0.9905 | 0.9406 | 1 |

Identification of Influential Observations

Using five different influential measures; Cook's distance, DFFITs, Hadi influence measure, DFBETAs and COVRATIO, influential observations in the real data are identified using the criteria of Table 2 when multicolinearity is not controlled (OLS: $k = 0$) and when controlled using the selected ridge parameter estimators. The values for the measure criteria are presented in Table 2.

The influential observations identified by the five influential measures in the presence of multicollinearity and when controlled using some selected ridge parameters (k) were presented in Table 3. When compared with values of Table 2,

Table 1. Result of test for multicollinearity.

| Independent variables ($x$) | VIF | Eigen-values ($\lambda$) | Tolerance (T) | Condition Number (CN) |
|---|---|---|---|---|
| $x_1$ | 5.9983 | 8.9344 | 0.1667 | 1.00 |
| $x_2$ | 120.5980 | 0.4087 | 0.008 | 21.86 |
| $x_3$ | 6.5232 | 0.3329 | 0.1533 | 26.83 |
| $x_4$ | 18.1551 | 0.1937 | 0.0551 | 46.11 |
| $x_5$ | 155.7352 | 0.0785 | 0.0064 | 113.75 |
| $x_6$ | 84.1103 | 0.0191 | 0.0119 | 466.88 |
| $x_7$ | 49.4181 | 0.0175 | 0.0202 | 510.49 |
| $x_8$ | 282.6033 | 0.0093 | 0.0035 | 957.74 |
| $x_9$ | 131.6438 | 0.0036 | 0.0076 | 2496.18 |
| $x_{10}$ | 168.8738 | 0.0019 | 0.0059 | 4505.02 |

Table showed that there is presence of multicollinearity in the data, since most of the independent variables have VIF > 10, the eigen-value close to zero (0), T < 0.1 and CN > 5 The correlation matrix of the data set also showed the presence of multicollinearity.

**Table 2.** Influential measures, calculated measure criteria and values obtained.

| | | |
|---|---|---|
| Cook's Distance | $D_i > F_{\alpha(p',n-p')}$ | 2.3479 |
| DFFITs | $\text{GDFFITS} \geq 3\sqrt{\frac{p'}{n-d}}$ | 1.1547 |
| Hadi's Measure | $H_i^2 = \text{mean}\left(H_i^2\right) + c\sqrt{\text{var}\left(H_i^2\right)}$ | 6.2463 |
| DFBETAS | $\text{DFBETAS}_{ij} > \frac{2}{\sqrt{n}}$ | ±0.3651 |
| COVRATIO | $\lvert\text{COVRATIO} - 1\rvert > \frac{3p'}{n}$ | > 0 |

**Table 3.** Influential observations identified.

| Measures | Criteria | OLS | $k_1$ = 5.345 | $k_2$ = 5.9566 | $k_3$ = 6.3345 | $k_4$ = 10.345 | $k_5$ = 10.002 | $k_6$ = 10.984 | $k_7$ = 10.567 | $k_8$ = 4.023 | $k_9$ = 3.874 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cook's Distance | 2.3479 | 22, 24, 25, 26, 27, 28, 29, 30 | None | None | None | None | None | None | None | None | None |
| DFFITs | 1.1547 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| Hadi Measure | 6.2463 | 25, 26, 28 | None | None | None | None | None | None | None | None | None |
| Dfbetas | ±0.3651 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 |
| Covratio | ≈0 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |

any observation whose calculated influence measure is greater than the criteria value obtained is identified as an influential observation or data point. Cook's Distance and Hadi influence measure performed alike. They fail to identify influential data points when ridge estimators were used to control multicollinearity. DFFITs and COVRATIO measure identified single observation 25 in both OLS and when multicollinearity was controlled while DFBETAS identified data point 29 as well.

## 4. Summary and Conclusion

Ridge estimator affects influential observation identified. Cook's distance and Hadi influence measure were able to identify several influential data points on the data in the presence of multicollinearity but failed to identify any data points when the multicollinear effect has been controlled. DFFITs, DFBETAs and COVRATIO identified the same single data point in the presence of multicollinearity and when it has been controlled. Cook's distance and Hadi influence measure are very sensitive in the presence of multicollinearity, this made them to identify several influential data points but they are less sensitive when multicollinearity is controlled where they fail to identify and data point. DFFITs, DFBETAs and COVRATIO perform better than them and should be used when multicollinearity is controlled.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References


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

[2] Johnson, R.A. and Wichern, D.W. (2002) Applied Multivariate Statistical Analysis. Pearson Education, Delhi.

[3] Mickey, M.R., Dunn, O.J. and Clark, V. (1976) Note on the Use of Stepwise Regression in Detecting Outliers. *Computers and Biomedical Research*, **1**, 105-111. https://doi.org/10.1016/0010-4809(67)90009-2

[4] Andrews, D.F. and Pregibon, D. (1978) Finding the Outliers That Matters. *Journal of Royal Statistical Society*, *Series B*, **40**, 85-93. https://doi.org/10.1111/j.2517-6161.1978.tb01652.x

[5] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[6] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975) Ridge Regression: Some Simulations. *Communications in Statistics*, **4**, 105-123. https://doi.org/10.1080/03610917508548342

[7] Lawless, J.F. and Wang, P. (1976) A Simulation Study of Ridge and Other Regression Estimators. *Communications in Statistics A*, **5**, 307-323. https://doi.org/10.1080/03610927608827353

[8] Nomura, M. (1988) On the Almost Unbiased Ridge Regression Estimation. *Communications in Statistics—Simulation and Computation*, **17**, 729-743. https://doi.org/10.1080/03610918808812690

[9] Khalaf, G. and Shukur, G. (2005) Choosing Ridge Regression Parameters for Regression Problems. *Communications in Statistics—Simulation and Computations*, **32**, 419-435.

[10] Dorugade, A. (2014) New Ridge Parameters for Ridge Regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, **15**, 94-99. https://doi.org/10.1016/j.jaubas.2013.03.005

[11] Al-Hassan, Y.M. (2010) Performance of a New Ridge Regression Estimator. *Journal of the Association of Arab Universities for Basic and Applied Science*, **9**, 23-26. https://doi.org/10.1016/j.jaubas.2010.12.006

[12] Dorugade, A.V. and Kashid, D.N. (2010) Alternative Methods for Choosing Ridge Parameter for Regression. *Applied Mathematical Science*, **4**, 447-456.

[13] Saleh, A.K.Md. and Kibria, B.M.G. (1993) Performances of Some New Preliminary Test Ridge Regression Estimators and Their Properties. *Communication in Statistics—Theory and Methods*, **22**, 2747-2764. https://doi.org/10.1080/03610929308831183

[14] Kibria, B.M.G. (2003) Performance of Some New Ridge Regression Estimators. *Communications in Statistics—Simulation and Computation*, **32**, 417-435. https://doi.org/10.1081/SAC-120017499

[15] Zang, J. and Ibrahim, M. (2005) A Simulation Study on SPSS Ridge Regression and Ordinary Least Square Regression Procedures for Multicollinearity Data. *Journal of Applied Statistics*, **32**, 571-588. https://doi.org/10.1080/02664760500078946

[16] Alkhamisi, M., Khalaf, S. and Shukur, G. (2006) Some Modifications for Choosing Ridge Parameters. *Communications in Statistics—Theory and Methods*, **37**, 544-564. https://doi.org/10.1080/03610920701469152

[17] Al-Hassan, Y.M. (2008) A Monte Carlo Evaluation of Some Ridge Estimators. *Japan Journal of Applied Science*: *Natural Science Series*, **10**, 101-110.

[18] Muniz, G. and Kibria, B.M.G. (2009) On Some Ridge Regression Estimators: An Empirical Comparison. *Communications in Statistics—Simulations and Computation*, **38**, 621-630. https://doi.org/10.1080/03610910802592838

[19] Khalaf, G. and Iguernane, M. (2014) Ridge Regression and Ill-Conditioning. *Journal of Modern Applied Statistical Methods*, **13**, 355-363. https://doi.org/10.22237/jmasm/1414815420

[20] Uzuke, C.A., Mbegbu, J.I. and Nwosu, C.R. (2017) Performance of Kibria, Khalaf and Shukur's Methods When the Eigenvalues Are Skewed. *Communications in Statistics—Simulation and Computation*, **46**, 2071-2102. https://doi.org/10.1080/03610918.2015.1035444

[21] Cook, R. (1977) Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15-18. https://doi.org/10.1080/00401706.1977.10489493

[22] Welsch, R. and Kuh, E. (1977) Linear Regression Diagnostics. Technical Report, Solan School of Management, Massachusetts Institute of Technology, Cambridge, 923-977. https://doi.org/10.3386/w0173

[23] Hadi, A. (1992) A New Measure of Overall Potential Influence in Linear Regression. *Computational Statistics and Data Analysis*, **14**, 1-27. https://doi.org/10.1016/0167-9473(92)90078-T

## Appendix I

Algorithm for the R Programme

The model

$$Y_i = X\beta + \varepsilon_i$$

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon_i$$

Using the unit length scaling shown below:

$$\tilde{Y} = \frac{Y - \overline{y}}{L_y},$$

$$\tilde{X}_j = \frac{X_j - \overline{x}_j}{L_j},\ j = 1, 2, \cdots, p$$

where $\overline{y}$ is the mean of $Y$, $\overline{x}_j$ is the mean of $X_j$, and

$$L_y = \sqrt{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}, \text{ and } L_j = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2},\ i = 1, 2, \cdots, n$$

such that $\sum_{i=1}^{n} x_{ij}^2 = 1,\ j = 1, 2, \cdots, p$

We obtain the following model

$$\tilde{Y} = \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2 + \cdots + \beta_p \tilde{X}_p + \varepsilon'$$

Obtain $A = \tilde{X}'\tilde{X}$

Eigenvalues of $A = t_j$

Eigenvectors of $A = D$

Confirm that $DD' = I$

Confirm that $D'\tilde{X}'\tilde{X}D = t_j$

Obtain $\alpha_j = D'\beta$

Obtain $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} \varepsilon_i}{n - p}$

Methods of estimating ridge parameter k

1) $\hat{k}_1 = \dfrac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$ Hoerl and Kennard (1970)

where, $\hat{\sigma}^2 = \sum_{i=1}^{p} e_i^2 / n - p$ is the residual mean square estimate of $\sigma^2$ and $\hat{\alpha}_i$ is the ith element of $\hat{\alpha}$ which is an unbiased estimator of $\alpha = D'\beta$ where D is the eigenvectors of the matrix $X'X$

2) $\hat{k}_2 = \dfrac{\hat{\sigma}^2}{\left(\prod_{i=1}^{p} \hat{\alpha}_i^2\right)^{1/p}},\ i = 1, 2, \cdots, p$ Kibria (2003)

3) $\hat{k}_3 = \max\left(\dfrac{\lambda_i \hat{\sigma}^2}{\lambda_i \hat{\alpha}_i^2 + (n - p)\hat{\sigma}^2}\right)$ Alkhamisi *et al.* (2006)

where $\lambda_i$ is the ith eigenvalues of the matrix $X'X$ and $S^2 = \dfrac{\sum_{j=1}^{p} \varepsilon_i^2}{n - p}$

4) $\hat{k}_4 = \left( \prod_{i=1}^{p} \frac{\lambda_i \hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + \lambda_i \hat{\alpha}_i^2} \right)^{\frac{1}{p}}$ Muniz and Kibira [18]

5) $\hat{k}_5 = \left( \prod_{i=1}^{p} \frac{1}{m_i} \right)^{\frac{1}{p}}$

6) $\hat{k}_6 = \left( \prod_{i=1}^{p} m_i \right)^{\frac{1}{p}}$

7) $\hat{k}_7 = \text{median}\left( \frac{1}{m_i} \right)$

where $m_i = \sqrt{\frac{\hat{\sigma}_i^2}{\hat{\alpha}_i^2}}$

8) $\hat{k}_8 = \frac{2p}{\lambda_{\max}} \sum_{i=1}^{p} \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$, $i = 1, 2, \cdots, p$ Dorugade [10]

9) $\hat{k}_9 = \left( \prod_{j=1}^{p} w_j \right)^{1/p}$ Uzuke *et al.* [20]

where the weight $w_j = \frac{In2\left(\hat{\sigma}^2\right)}{(n-p)\hat{\sigma}^2 + In2\left(\hat{\alpha}_j^2\right)}$

10) OLS = $\left(X'X\right)^{-1} X'Y$

**Methods of detecting influential observation**

**Method 1 (cook's distance)**

$$C_i = \frac{t_i^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}},$$

The criteria is given as

$$C_i > F^{0.05}_{(p+1, n-p-1)}$$

where

$$h_{ii} = X'\left(X'X + kI\right)^{-1} X, \text{ and } t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

**Method 2 (DFFITs)**

$$\text{DFITS}_i = t_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}}, i = 1, 2, \cdots, n$$

The criteria is given as

$$\text{DFFITs} > 2\sqrt{\frac{p'+1}{n-p-1}}$$

where $r_i^*$ is the R-residual defined as $t_i^* = t_i \sqrt{\frac{n-p-1}{n-p-t_i^2}}$ and $h_{ii} = X'\left(X'X + kI\right)^{-1} X$

**Method 3 (Hadi measure)**

$$H_i = \frac{h_{ii}}{1-h_{ii}} + \frac{p+1}{1-h_{ii}} \frac{d_i^2}{1-d_i^2}, i = 1, 2, \cdots, n$$

where $d_i = \frac{e_i}{\sqrt{SSE}}$ called normalized residual.

**Method 4 (DFBETAS)**

$$\frac{b_j - b_j(i)}{\sqrt{s^2_{(i)}\left(X'X + kI\right)^{-1}_{ij}}}$$

The criteria is given as $\text{DIFBETAs} > \frac{2}{\sqrt{n}}$

**Method 5 (COVRATIO)**

$$\frac{\left(\frac{n-p'-t_i^2}{n-p'-1}\right)}{1-h_{ii}}$$

The criteria is given as

$$\left|\text{COVRATIO} - 1\right| > \frac{3p}{n}$$

where

$$h_{ii} = X'\left(X'X + kI\right)^{-1} X \text{, and } t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

## Appendix II

R Codes for Detecting Influential Observation for Different k Values

```
for(i in 1:9){
h=matrix(hatr(lmridge(V1~.,rr, k[i]],30,30)
ss=(sqrt(h[i,i]/(1-h[i,i])))
C=NULL
DF9=NULL
H=NULL
DFB=NULL
COV=NULL
for(i in 1:30){
b1=coefficients(lm(V1~.,rr[-i,]))
r1=c(residuals(lm(V1~.,rr[-i,])))
sig1=(sum(r1^2))/(n-p)
num=c[3]-b1[3]
hh=solve(t(xx[-i,])%*%(xx[-i,]))
denom=sqrt(sig1*hh[3,3])
C=rbind(C,(((r[i]^2/((sig)*(1-h[i,i])))))/(11))*(h[i,i]/(1-h[i,i])))
DF9=rbind(DF9,r[i]/(sqrt(sig1*(1-h[i,i])))*sqrt(h[i,i]/(1-h[i,i])))
H=rbind(H,(h[i,i]/(1-h[i,i]))+(11/(1-h[i,i]))*(r1[i]/sqrt(ssr)))
DFB=rbind(DFB,num/denom)
COV=rbind(COV,(sig1/sig)*(h[i,i]/(1-h[i,i])))
}
```

## Appendix III

**Table A1.** Nigerian economic indicator (1980-2010) data.

| GDP | Money Supply | Cred Priv. Sector | Exchange Rate | External Reserv | Agric Loan | Foreign Trade | Oil Import | Nonoil Import | Oil Export | Nonoil Export |
|---|---|---|---|---|---|---|---|---|---|---|
| | 14,471 | 8570 | 0.61 | 56,195 | 35,642 | 23,863 | 120 | 12,720 | 10,681 | 343 |
| 53,659 | 15,787 | 10,668 | 0.673 | 12,324 | 31,764 | 18,977 | 226 | 10,545 | 8003 | 203 |
| 57,963 | 17,688 | 11,668 | 0.724 | 7171 | 36,308 | 16,406 | 172 | 8732 | 7201 | 301 |
| 64,326 | 20,106 | 12,463 | 0.765 | 5480 | 24,655 | 16,266 | 282 | 6896 | 8841 | 247 |
| 73,542 | 22,299 | 13,070 | 0.894 | 10,998 | 44,244 | 18,783 | 52 | 7011 | 11,224 | 497 |
| 74,542 | 23,806 | 15,247 | 2.021 | 18,922 | 68,417 | 14,904 | 914 | 5070 | 8369 | 552 |
| 111,913 | 27,574 | 21,083 | 4.018 | 62,554 | 102,153 | 48,222 | 3170 | 14,692 | 28,209 | 2152 |
| 147,941 | 38,357 | 27,326 | 4.537 | 72,267 | 118,611 | 52,639 | 3803 | 17,643 | 28,435 | 2757 |
| 228,451 | 45,903 | 30,403 | 7.392 | 43,953 | 129,300 | 88,831 | 4672 | 26,189 | 55,017 | 2954 |
| 281,550 | 52,857 | 33,548 | 8.038 | 40,293 | 98,494 | 155,604 | 6073 | 39,645 | 106,627 | 3260 |
| 329,071 | 75,401 | 41,352 | 9.909 | 48,620 | 82,107 | 211,024 | 7772 | 81,716 | 116,858 | 4677 |
| 555,446 | 111112 | 58,123 | 17.298 | 33,392 | 88,032 | 348,763 | 19,562 | 123,590 | 201,384 | 4227 |
| 715,242 | 165,339 | 127,118 | 22.051 | 58,824 | 80,846 | 384,400 | 41,136 | 124,493 | 213,779 | 4991 |
| 945,557 | 230,293 | 143,424 | 21.886 | 95,329 | 103,186 | 3,688,480 | 42,350 | 120,439 | 200,710 | 5349 |
| 2,008,564 | 289,091 | 180,005 | 21.886 | 32,345 | 164,162 | 1,705,789 | 155,826 | 599,302 | 927,565 | 23,096 |
| 2,799,036 | 345,854 | 238,597 | 21.886 | 25,896 | 225,503 | 1,872,170 | 162,179 | 400,448 | 1,286,216 | 23,328 |
| 2,906,625 | 413,280 | 316,207 | 21.886 | 73,492 | 242,038 | 2,087,379 | 166,903 | 678,814 | 1,212,499 | 29,163 |
| 2,816,406 | 488,146 | 351,956 | 21.886 | 93,777 | 215,697 | 1,589,275 | 175,854 | 661,565 | 717,787 | 34,070 |
| 3,312,241 | 628,952 | 431,168 | 92.693 | 63,709 | 246,083 | 2,051,486 | 211,662 | 650,854 | 1,169,477 | 19,493 |
| 4,717,332 | 878,457 | 530,373 | 102.105 | 91,089 | 361,450 | 2,930,746 | 220,818 | 764,205 | 1,920,900 | 24,823 |
| 4,909,526 | 12,699,322 | 764,962 | 111.943 | 123,330 | 728,545 | 3,226,134 | 237,107 | 1,121,074 | 1,839,945 | 28,009 |
| 7,128,203 | 1,508,173 | 930,494 | 120.97 | 103,104 | 1,051,590 | 3,256,873 | 361,710 | 1,150,985 | 1,649,446 | 94,732 |
| 8,742,647 | 1,952,922 | 1,096,536 | 129.356 | 91,702 | 1,164,460 | 5,168,122 | 398,922 | 1,681,313 | 2,993,110 | 94,776 |
| 11,673,602 | 2,131,820 | 1,421,664 | 133.5 | 144,753 | 2,083,745 | 6,589,827 | 318,115 | 1,668,931 | 4,489,472 | 113,309 |
| 1.48E+08 | 2,637,914 | 1,838,390 | 132.147 | 291,849 | 3,046,739 | 10,047,391 | 797,299 | 2,003,557 | 7,140,579 | 105,956 |
| 18,709,786 | 3,799,538 | 2,290,618 | 128.651 | 449,473 | 4,263,060 | 10,433,200 | 710,683 | 2,397,836 | 7,191,086 | 133,595 |
| 20,874,172 | 5,138,701 | 3,680,090 | 125.833 | 544,732 | 4,425,862 | 12,221,711 | 768,227 | 3,143,726 | 8,110,500 | 199,258 |
| 25,424,948 | 8,029,089 | 6,941,383 | 118.566 | 701,675 | 6,721,075 | 15,357,293 | 1,386,730 | 3,803,073 | 9,913,651 | 247,839 |
| 2,896,746 | 9,456,480 | 9,147,417 | 148.902 | 536,428 | 8,349,509 | 13,458,920 | 1,063,544 | 4,038,990 | 8,067,233 | 289,153 |
| 3,124,539 | 11,034,941 | 10,157,021 | 150.298 | 448,268 | 7,740,508 | 19,041,169 | 2,073,579 | 5,931,795 | 10,639,417 | 396,377 |