Benin Journal of
Statistics
Vol. 1, pp. 61– 67 (2018)

Identification of influential observations in a data set on economic variables

Ifeyinwa C. Ezeilo[1]* and Chinwendu A. Uzuke[2]

[1,2]*Department of Statistics, Nnamdi Azikiwe University, P.M.B. 5025, Awka, Nigeria*

Abstract. In this study, different measures of identifying influential observations were discussed. The measures were applied to a dataset on economic variables. Influential observations in the dataset were identified without a control for multicollinearity.

Keywords: influential observation, multicollinearity, Cook's distance, DEFFITs and DFBETAs.

Published by: **Department of Statistics, University of Benin, Nigeria**

1. Introduction

it is well known that not all observations in a data set play equal roles when fitting a regression model. Occasionally a single or a subset of a data set exerts a disproportionate influence on the fitted regression model. that is, the parameter estimates may depend more on the influential sub set than the majority of the data. Belslyey et al. (1980) defined an influential observation as one which either individually or together with several other observations has demonstrably large impact on the calculated values of of various estimates, than is the case of most of the observations. Outlier may exist in a data set. Although it is not all outliers that need to have undue influence on the estimation of the parameters in a regression model (Andrew and Pregibon, 1978). Since not all outliers matter, examining residual alone might not lead to the detection of influential observations. Other ways of detecting influential observations are needed. Stevens (1984) discussed the four diagnostics that are useful in identifying outliers namely: the studentized residual, the hat element, the Cook's distance and the Mahalanobis distance. An important fact is that outliers may not necessarily be influential on the regression coefficients.

Regression diagnostics may be used in the identification of influential data points and multicollinearity (Besley et al.,1980). This approach includes methods of exploratory data analysis. Multicollinearity can be detected using several diagnostics as follows:

- Eigen-structure of $X^{'}X$: let λ_1, λ_2, λ_3, ..., λ_p be the eigenvalues of $X^{'}X$ (in correlation form), where X is the independent variables. If at least one eigenvalue is close to zero, then multicollinearity exists (Greene, 1993, Walker, 1999).
- Condition Number (CN): Condition number is given by:

$$CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{1}$$

where λ_{max} is the largest eigenvalue and λ_{min} is the smallest eigenvalue. If CN lies between5 and 30, it is considered that multicollinearity exists, (Vinod and Ullah, 1981).

*Corresponding author. Email: prisca087@yahoo.com

- Variance Inflation Factor(VIF): This is computed as:

$$VIF_i = \frac{1}{1 - R_j^2}; \tag{2}$$

$j = 1, 2, \cdots, p$, where R_j^2 is the coefficient of multiple determination in the regression produced by regressing an explanatory variable the x_i , against the remaning variables $x_j (j, \neq i)$.

Several diagnostic methods have been developed to detect influential observations. Firstly, Cook (1977) introduced Cook's Distance, (D_i), which is based on deleting the observations one after the other and measuring their effect on the linear regression model. Other measures developed on the idea of Cook's Distance include: Modified Cook's Distance (D_i^*), Difference in Fits (DFFITs) by Welsch and Kuh (1977), Difference in Betas (DFBETAS) by Besley et al. (1980).

The plan of this paper is as follows. Section 2 discusses influential observations. Section3 contains the methodology used for for the study. Section 4 presents the illustration, while section 5 discusses the results. Finally, Section 6 concludes the paper and suggests area for further studies.

2. Influential Observation

In regression analysis, unusual observations are generally either outlier or influential data point. Hawkins (1980), pointed out that an outlier is an observation that deviates so much from other observations as to arouse suspicion. On the other hand, an influential observation is 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 other observations (Belsley et al, 1980). The effect of influential observations on estimated values has been studied by several authors, Cooks (1977), Cook and Weisbery (1980), Hampel (1985), Brade (1997), Flores (2015) and Genton and Hall (2016). There are several sources of influential observations. It could be as a result of improper record of data either at the source or in the transcription to computer readable form. The influence of an observation is measured by the effect it produces on the fit when it is deleted in the fitting process. The deletion is always done one point at a time. Let $\hat{\beta}_{1(i)}$, $\hat{\beta}_{2(i)}$ ... $\hat{\beta}_{p(i)}$ denote the regression coefficients obtained when the ith observation is deleted $(i = 1, 2, \cdots, n)$. Let $\hat{y}_{1(i)}$, $\hat{y}_{2(i)}$, ... , $\hat{y}_{n(i)}$ and σ_i^2 be the predicted values and residual mean square respectively when the ith observation is dropped. Note that

$$y_{m(i)} = \beta_{0(i)} + \hat{\beta}_{1(i)} x_{m1} + \hat{\beta}_{2(i)} x_{m2} + ... + \hat{\beta}_{p(i)} x_{mp} \tag{3}$$

is the fitted value for observation m when the fitted equation is obtained with ith observation deleted. The differences produced in quantities such as $(\hat{\beta}_j - \hat{\beta}_{j(i)})$ and $(\hat{y}_j - \hat{y}_{j(i)})$ is usually considered. Examination of the data and the ability to find influential observations can be beneficial to reveal spurious observations that might be as a result of error during data collection or the processing of data. It makes the researcher aware of the possibility that some part of the data might come from another regime or sub population that have very different features compared to the population under study as well as help uncover features of the data that could cause difficulty in fitting a regression model.

3. Methodology

To identify influential data points, we adopt the following single-case influential measures.

3.1 *Cook's distance (Cook, 1977):*

It measures the difference between the fitted values obtained by deleting the ith observation. The Cook's Distance is defined as;

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})}{\hat{\sigma}^2(p+1)} \tag{4}$$

$$D_i = \frac{r_i^2}{p+1} * \frac{h_{ii}}{1-h_{ii}} \tag{5}$$

where $r_i = \frac{e_i}{\sigma\sqrt{1-h}}$ and h_{ii} is the leverage of the ith observation estimated by $h_{ii} = X(X`X)^{-1}X`$.

3.2 *$DFFITs_i$ (Welsch and Kuh,* 1977*)*

The accronym DFFITs stands for the difference in Fits. DFFITs is used to identify influential data points. it quantifies the number of standard deviations that the fitted value changes when the ith data point is omitted. It is the scaled distance between the ith fitted value obtained from the full data and the ith fitted value obtained by deleting the ith observation.

$$DFFITS_i = \frac{x_i(\hat{\beta} - \hat{\beta}_i)}{SE(x_i\hat{\beta})} \tag{6}$$

This can be expressed as;

$$DFFITs_i = r_i^*\sqrt{\frac{h_{ii}}{1-h_{ii}}}, \tag{7}$$

(i = 1, 2, ... n), where r_i^* is the standardized residual defined as $r_i^* = \frac{e_i}{\sigma_(i)\sqrt{1-h_{ii}}}$

3.3 *Hadi's influence measure (Hadi, 1992)*

This is a measure of the influence of the ith observation. The measure is constructed on the fact that influential observations are outliers in the response variables or in the prdictors or both. Accordingly, the influence of the ith observation is measured by

$$H_i = \frac{h_{ii}}{1-h_{ii}} + \frac{p+1}{1-h_{ii}}\frac{d_i^2}{1-1d_i^2} \tag{8}$$

i = 1, 2, ... n where $d_i = \frac{e_i}{\sqrt{SSE}}$ is called normalized residual.

3.4 *$DFBETAS_{ij}$ (Belsley et al.,* 1989*)*

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

$$DFBETAS_{ij} = \frac{b_i - b_{i(i)}}{\sqrt{S_{(i)}^2(X`X)_{ij}^{-1}}} \tag{9}$$

3.5 *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 (Belsley et al., 1980)

$$COVRATIO = \frac{det(s_i^2(X_i^{'}X)^{-1}}{det(s^2(X^{'}X)^{-1}} \tag{10}$$

This measure can also be expressed as;

$$COVRATIO = \frac{(\frac{n-p^{'}-r_i^2}{n-p^{'}-1})}{1-h_{ii}} \tag{11}$$

where n is the sample size, p is the number of independent variables.

Table 1: The measure criteria for the listed influential measures.

Method	Criteria
Cook's Distance	$D > F_{\alpha(p,n-p)}$
Welsch and Kuh Distance	$WK > 2(\frac{p}{n})^{\frac{1}{2}}$
Kuh and Welsch ratio	$KWR > 3(\frac{p}{n})$
Hadi's Meadure	$H > (mean(h) + 3(sd(h)))$
Dfbetas	$Df > \frac{2}{(n)^{\frac{1}{2}}}$

source: Ullah and Pashal (2009)

4. Illustration

Economic indicator data from the Central Bank of Nigeria (CBN) Statistical Bulletin 2010 will be used. The data consist of Gross Domestic Product as dependent variable (y) and ten (10) independent variable namely: Money Supply (x_1), Credit to Private Sectors (x_2), Exchange Rate (x_3), External Reserve (x_4), Agricultural Loan (x_5), Foreign Reserve (x_6), Oil Import (x_7), Non Oil Import (x_8), Oil Export (x_9) and Non Oil Export (x_{10}). The data set is subjected to test for multicollinearity using Variance Infation Factor (VIF), eigen-value of the independent variable (λ), tolerance value (T) and condition number (CN). The data used for the study shown in Table 2.

Table 2: GDP and other economic variables

S/No	(y)	(x_1)	(x_2)	(x_3)	(x_4)	(x_5)	(x_6)	(x_7)	(x_8)	(x_9)	(x_{10})
1	0	14471	8570	0.61	56195	35642	23863	120	12720	10681	343
2	53659	15787	10668	0.673	12324	31764	18977	226	10545	8003	203
3	57963	17688	11668	0.724	7171	36308	16406	172	8732	7201	301
4	64326	20106	12463	0.765	5480	24655	16266	282	6896	8841	247
5	73542	22299	13070	0.894	10998	44244	18783	52	7011	11224	497
6	74542	23806	15247	2.021	18922	68417	14904	914	5070	8369	552
7	111913	27574	21083	4.018	62554	102153	48222	3170	14692	28209	2152
8	147941	38357	27326	4.537	72267	118611	52639	3803	17643	28435	2757
9	228451	45903	30403	7.392	43953	129300	88831	4672	26189	55017	2954
10	281550	52857	334584	8.038	40293	98494	155604	6073	39645	106627	3260
11	329071	75401	41352	9.909	48620	82107	211024	7772	81716	116858	4677
12	555446	111112	58123	17.298	33392	88032	348763	19562	123590	201384	4227
13	715242	165339	127118	22.051	58824	80846	384400	41136	124493	213779	4991
14	945557	230293	143424	21.886	95329	103186	3688480	42350	120439	200710	5349
15	2008564	289091	180005	21.886	32345	164165	1705789	155826	599302	927565	23096
16	2799036	345854	238597	21.886	25896	225503	1872170	162179	400448	1286216	23328
17	2906625	413280	316207	21.886	73492	242038	2087379	166903	678814	1212499	29163
18	2816406	488146	351956	21.886	93777	215697	1589275	175854	661565	717787	34070
19	3312241	628952	431168	92.693	63709	246083	2051486	211662	650854	1169477	19493
20	4717338	878457	530373	102.105	91089	361450	2930746	220818	764205	1920900	24823
21	4909526	12699322	764962	111.943	123330	728545	3226134	237107	1121074	1839945	28009
22	7128203	1508173	930494	120.97	103104	1051590	3256873	361710	1150985	1649446	94732
23	8742647	1952933	1096536	129.356	91702	1164460	5168122	398922	1681313	2993110	94776
24	1673602	2131820	421664	131.5	144753	2083745	6589827	318115	1668931	4489472	113309
25	14800000	263714	1838390	132.147	291841	3046739	10047391	797299	2003557	7140579	105056
26	18709786	3799538	2290618	128.651	449473	4261060	10433200	710683	2397836	7191086	133595
27	20874172	5138701	3680090	125.833	544732	4425862	12221711	768227	3143726	8110500	199258
28	25424948	8029089	6941383	118.566	701675	6721075	15357293	1386730	3803073	9913651	247839

The data in Table 2 were sujected to multicollinearity test using Variance Inflation Factor (VIF), eigen-value, tolerance (T) and condition number (CN) and the result is presented in Table 3

Table 3: Test for multicollinearity results

Independent Variable(x_i)	VIF	Eigenvalues (λ)	Tolerance(T)	Condition Number
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	49.4148	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 3, detected the presence of multicollinearity in the data. This is because most of the independent variables have $VIF > 10$, the eigenvalues close to zero (0), $T < 0.1$ and CN> 5. From Table 2, influential measure criteria were obtained.

Table 4: Measure criteria

Influential measures	calculated measure criteria
Cook's Distance	2.3479
DFFITs	1.1547
Hadi's Measure	6.2463
DFBETAS	± 0.3651
COVRATIO	> 0

Using R - Programme,to analyse the observation in Table 2, the influential data points were indentified and presenteed in Table 5.

Table 5: Influential data points

Influential measures	Influential data points
Cook's distance	$21, 22, 24, 25, 26, 27, 28, 29, 30$
DFFITs	25
Hadi measure	$14, 21, 22, 24, 26, 28, 29, 30$
DFBETAs	$14, 20, 21, 22, 24, 26, 27, 28, 29, 30$
COVRATIO	$14, 21, 22, 27, 28, 29, 30$

5. Discussion of results

GDP and other economic variables in Nigeria were analysed using different influential measures with a view to identify influential data poits when multicollinearity is present in a data set. The Cook's Distance identified thye data points $21, 22, 24, 25, 26, 27, 28, 29, 30$, DFFITs identified only one data point 25, Hadi Maesure identified points $14, 21, 22, 24, 26, 28, 29, 30$, DFBETAs identified data points $14, 20, 21, 22, 24, 26, 27, 28, 29, 30$ while COVRATIO identified data points $14, 21, 22, 27, 28, 29, 30$. An observation whose calculated influence measure is greater than its measure criteria is identified as being influential. Deletion of any of these identified influential data points from the fitting process will have a large effect on the parameter estimate.

6. Conclusion

This paper discussed the identification of influential observations in a data set. This was achieved using some influential measures with out controlling multicollinearity. This approach is different from that of Belsley et al (1989) where in multicollinearity was controlled before detecting influential observations. Data points $21, 22, 27, 28$ and 29 were detected by all the measures used except in DFFITs that identified only one observation. These points flagged as influential should be examined

carefully to determine whether they should be deleted from the analysis. Further, one can decide to control multicollinearity on the used and goes ahead to re-identify influential observation to check whether the same data point will be identified.

References

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. Journal of Royal Statistical Soceity, Series B, **40**: 85 – 93

Belsley, D. A., Edwin, K. and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York.

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

Brade, D. (1997). Indentification of outliers by means of LI regression: safe and unsafe configurations. Comp. Stat. & Data Analysis, **24**: 271 – 281

Cook, R. D. (1977). Detection of influential observations in linear regression. Technometrics, **19**: 15 – 18

Cook, R. D. and Weisberg, S. (1980). Characterisation of an emperical influence function for detecting influential cases in regression. Technometrics, **22**: 495 – 508.

Flores, S. (2015). Sharp non-assyptotic performance bound for II and Huber robust regression estimators. Test, **24**: 796 – 812.

Genton, M. G. and Hall, P. (2016). A tilting approach to ranking influence. Journal of Royal Statistical Society, B, **78**: 77 – 79.

Greene W. H. (1993). Econometric Analysis. MacMillan, New York.

Hadi, A, S. (1992). A new measure of overall potential influence in linear regression. Computational Statistics & Data Analysis, **14**: 1 – 27.

Hawkins, D. M. (1980). Identification of Outliers. Chapman & Hall, London.

Hampel, F. R. (1985). The breakdown point of the mean combined with some rejection rules. Technometrics, **27**: 95 – 107.

Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. Pearson Education, India.

Sheather, S. (2009). A Modern Approach to Regression with R. New York, NY: Springer.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. Psychologial Bulletin, **95**(2): 334 – 344.

Ullah, M. A. and Pasha, G. R. (2009). The origin and development of influence measures of regression. Pakistan Journal of Statistics, **25**(3): 295 – 307.

Vinod, H. D. and Ullah, A., (1981) Recent Advances in Regression Model. Marcel Dekker, New York.

Walker, E. and Birch, J. B. (1988). Influence measures in ridge regression. Technometrics,

30, 221 – 227.

Welsch, R. E. and Kuh, E. (1977). Linear regression diagnostics. Technical Report, Solan School of Management, Massachusetts Institute of Technology.

Appendix

R codes for identifying influential observations

SSR = $sum(r^2)$
h = $matrix(hatr(VI\ .,d))[[1]], 28, 10)$
ss = $(sqrt(h[i,i]/(1-h[i,i])))$
C = NULL
DF9 = NULL
H = NULL
DFB = NULL
COV = NULL
for i(in 1 : 28)
b1= cofficients($lm(VI\ .,d[-i,])$)
r1=c(residuals($lm(VI\ .,d[-i,])$))
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/$((sig) * (1-h[i,i])))))/(1,1)) * (h[i,i]/(1-h[i,i]))$)
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])))

res1 = cbind(C,DF,H,DFB,COV)
res2 = cbind(C2,DF2,H2,DFB2,COV2)
res3 = cbind(C3,DF3,H3,DFB3,COV3)
res4 = cbind(C4,DF4,H4,DFB4,COV4)
res5 = cbind(C5,DF5,H5,DFB5,COV5)
res6 = cbind(C6,DF6,H6,DFB6,COV6)
res7 = cbind(C7,DF7,H7,DFB7,COV7)
res8 = cbind(C8,DF8,H8,DFB8,COV8)
res9 = cbind(C9,DF9,H9,DFB9,COV9)