Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/090303

Detection of High Leverage Points in Regression Model in Apple Data

*Rizwan Yousuf** and *Manish Sharma*

Division of Statistics and Computer Science SKUAST Jammu 18009 India

Received: 6 Dec. 2021, Revised: 2 May 2022, Accepted: 3 Jun. 2022
Published online: 1 Sep. 2022

Abstract: An outlier in regression analysis is an observation with a large residual compared to the other observations in the data set. Outliers and influential points must be identified as part of the regression analysis. To identify and exclude unusual values from data, outlier detection methods have been applied. In this paper, Outliers are identified in regression models for the Apple data set. Ordinary residuals are not ideal for diagnostic purposes; rather, a modified version is recommended. Next, we have used the new approach of Modified OLS after handling HLP for detecting outliers. Real data was used to test the new approach's performance. The modified had shown better results as compared to OLS.

Keywords: Outlier, Hat matrix, Cooks Distance, Apple Data, DFFITS

1 Introduction

The most popular and effective statistical method for evaluating the relationship between a response variable (y) and explanatory variables(x) is regression. If we have dependent variable Y and independent variables X1, X2. . . Xn then the linear regression model generally can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_v X_{iv} + \mu_i \quad i = 1, 2, \ldots, n \tag{1}$$

Where

$$\beta_{0,}\beta_1, \ldots \beta_p \tag{2}$$

are regression parameters and error is normally distributed with mean zero and equal variance. Regression analysis is used in a wide range of fields, including physical and chemical sciences, engineering, economics, finance, pharmacy, life and biological sciences, social sciences, and other disciplines. It allows the mean function E(y) to be dependent on several explanatory factors and to have forms other than straight lines, and not arbitrary shapes. To this purpose, Legendre's seminal work in [1] and Gauss's seminal work in [2] proposed the least squares (LS) method, which has probably become the most popular approach to perform a regression analysis. The LS estimator may be expressed in closed form and shown to attain the minimum variance among all unbiased estimators when the underlying error distribution is normal, which justifies its popularity. It has been that LS regression may not be appropriate when the response variable differs from the regression function in an asymmetric manner, which is commonly encountered in medical data, among other venues. Regression analysis is used to predict a continuous dependent variable from a number of independent variables [3].

Given a training dataset, the goal of regression is to estimate the parameters of a model relating two sets of variables. However, because the training dataset contains outliers, this estimate is unreliable. Outliers are data that are significantly different from the rest of the dataset. Because real-world data is virtually always tainted by outliers, effective parameter estimate is critical. Outliers can occur by chance in any distribution, but they frequently suggest either measurement error or a population with a long tail. In the first case, they should be ignored or replaced by exceptional case statistics, whereas in the second, they suggest that the distribution is skewed and that utilising tools or intuitions that assume

[*] Corresponding author e-mail: rizwanwar50@gmail.com

a normal distribution should be avoided. A combination of two distributions, which may represent two distinct subpopulations or signify ’correct trial’ versus’measurement error,’ is a common cause of outliers; this is modelled using a mixing model. The ordinary least squares (OLS) method and Modified OLS after handling High Leverage Points (HLP) using interpolation techniques have been used to fit the model and to estimate the parameter values.

$$Y = x\beta + \varepsilon \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dot{y}_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1\ X_{11}\ X_{21}\ \ldots\ X_{k1} \\ 1\ X_{12}\ X_{22}\ \ldots\ X_{K2} \\ .\ \ \cdot\ \ \cdot\ \ \ \ \ \cdot \\ .\ \ .\ \ .\ \ \ \ \ . \\ 1\ X_{1n}\ X_{2n}\ \ldots\ X_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dot{\beta}_K \end{bmatrix} \tag{3}$$

$$\text{Where,} \quad \dot{\beta} = \left(X'X\right)^{-1} X'Y$$

The “goodness of fit” of the model are described by the quantity R2 and adjusted R2. In SLR model, R2 was plainly the square of the correlation coefficient between the Independent and the dependent variable. This, however, will not work with the MLR model. R2 is usually defined as the proportion of variation in the dependent variable that can be explained by regression on all of the model’s independent variables.

2 Material and Methods

Identification of bad leverages

There are several types of outliers in regression problems, which include residual outliers, vertical outliers, and high leverage points. The term ”residual outlier” refers to any observation with a big residual. Vertical outliers (VO), also known as y-outliers, are observations that are extreme or outlying in the y-axis. High leverage points (HLPs).On the other hand, high leverage points (HLPs) are those observations which are extreme or outlying in x-axis Coordinate. HLPs can be classified into good leverage points (GLPs) and bad leverage points (BLPs). GLPs are outliers in the explanatory variables that follow the majority of the data’s pattern, whereas BLPs are the absolute opposite. The computed values of various estimates are more affected by BLPs. The bad and good Leverages have been identified by using Residual analysis, Standardized Residuals, Studentized Residuals and Deleted residuals. Because the bad leverages suffer from a masking effect, these approaches may fail to detect high leverage points. As a result, in addition to the methods mentioned above, the effective steps were used to identify the bad leverages.

Weighted sum of squared distance (Wood, 1980) suggested that the weighted sum of squared distance of ith point from the center of data (WSSD) say :

$$WSSD_I = \sum_{j=1}^{K} \frac{\widehat{\beta}_j \left(\bar{X}_{ij-x_j}\right)^2}{\sqrt{MSE}} \tag{4}$$

It is used to locate the points that are remote in X-space. The general procedure to rank the points in increasing order WSSDI and concentrate on points for which the statistic is large [4].

Hat matrix (Hoaglin and Welsch, 1978) discussed the role of hat matrix

$$\mathrm{H} = \mathrm{X}\left(X'X\right)^{-1} X' \tag{5}$$

in identifying influential observation. The diagonal elements of H matrix are called the Hat values denoted by

$$\mathrm{h_{ii}\ given\ by{:}\ h_{ii}} = x_i{}^T \left(X^T X\right)^{-1} x_i, i = 1,2,3\ldots\ldots n \tag{6}$$

Since H determines the variance and covariance of

$$\acute{y}\text{and e, since v}(y) = \sigma^2 H \text{ and v(e)} = \sigma^2(1-H)\text{. The elements } h_{ij} \text{ of H may be interpreted} \tag{7}$$

as the amount of leverage exerted [5]

Cook’s distance (Cook, 1979): Cook’s distance is useful for spotting outliers in predictor variable observations. It also shows how each observation has an impact on the fitted response values. An outlier is an observation having a Cook’s distance more than three times the mean Cook’s distance. The scaled change in fitted values is known as Cook’s distance.

The normalised change in the vector of coefficients owing to the deletion of an observation is represented by each element in Cook's distance. A conventional cut-off point is 4/n, where n is the number of observations in the data set. Cook gave the method for evaluating the influential observations

$$| D_i = \frac{f_i}{p} \frac{h_{ii}}{(1-h_{ij})}, \quad \mathrm{i} = 1,2\ldots\mathrm{n} \tag{8}$$

where $[f_i]^2$ is the ith studentized $residual\ and\ hij$ is the ith diagonal element of H .In statistics ,cook's distance is a commonly used estimate of the influence of the data point when doing least square regression analysis. Cook's distance measures the effect of deleting a given observation [6].

Robust Mahalanobis distance (Rousseeuw and Leroy, 1987): In multivariate space MD is the distance between two locations in multivariate space. The independent variables in a regression equation constitute a multidimensional space in which each observation can be plotted. Construct a point that represents the means of all independent variables. The term "centroid" refers to the mean point in multidimensional space. The MD is the distance between an observation and the centroid as specified by the independent variables that are correlated. The putative HLPs are identified using this method, which is based on Rousseeuw and Leroy's minimal volume ellipsoid (MVE).The means for all independent variables. In multidimensional space this mean point is known as 'centroid'. The MD is the distance of a observation from the centroid defined by the correlated independent variables In this method the suspected HLP's are determined on the base of the minimum volume ellipsoid (MVE) developed by Rousseeuw and Leroy as

$$RMD_i = \sqrt{[X - T(X)]^T [C(X)]^{-1} [X - T(X)]} \tag{9}$$

where, T(X) and C(X) are robust locations and shape estimates of the MVE respectively.Habshah et.al (2009) suggested using the following cut-off value for the robust Mahalanobis distance Median$(RMD_i) + 3MAD([RMD]_i)$ where, MAD is the median absolute deviation [7]

DFBETAS: DFBETAS is a variable that shows how much the regression coefficient changes if the ith observation is removed. The change is expressed in standard deviation units. This statistic is

$$\mathrm{DFBETAS} = \frac{\mathrm{b_j - b_{j(i)}}}{\sqrt{\mathrm{S^2_{(i)} C_{ji}}}} \tag{10}$$

Where Cjj is the jth diagonal element of $(X'X)^-1$and $b_(j(i))$ b regression coefficient computed without the use of ith observation.

DFFITS: The impact of the ith observation's deletion on the projected or fitted value can be studied using diagnostics developed by Belsley, Kuh, and Welsch as :

$$\mathrm{DFFITS_i} = \frac{\mathrm{\hat{y}_j - \hat{y}_{(i)}}}{\sqrt{\mathrm{S^2_{(i)}\ h_{ii}}}}, \mathrm{i} = 1,2\ldots,\mathrm{n} \tag{11}$$

Where $y(i)$ is the fitted value ofy_i obtained without the use of the ith observation. The denominator is just standardization, since

$$\mathrm{Var}\left(\mathrm{\hat{y}_{(i)}}\right) = \sigma^2\ \mathrm{h_{ii}} \tag{12}$$

This $DFFITS_i$ is the number of standard deviations that the fitted value y_i changes of ith observation is removed. Computationally

$$\begin{aligned} \mathrm{DFFITS\ _i} &= \sqrt{\frac{\mathrm{1-h_{ii}}}{\mathrm{h_{ii}}} \frac{\mathrm{e_i}}{\sqrt{\mathrm{1-h_{ii}}}}} \\ &= \mathrm{t_i}\sqrt{\frac{\mathrm{1-h_{ii}}}{\mathrm{h_{ii}}}} \end{aligned} \tag{13}$$

=R-student x leverage of ith observations Where t_i is R-student

Covariance Ratio :The generalised variance is a scalar measure of precision that is defined as the determinant of the covariance matrix. The generalized variance of OLSE b is

$$\mathrm{GV(b)} = |\mathrm{v(b)}| =| \sigma^2 \left(X'X\right)^{-1}\Big) | \tag{14}$$

To express the role of ith observation on the precision of estimation, define

$$\text{COVRATIO} = \left| \frac{\left(X'_{(i)}X_{(i)}\right)^{-1} \quad \mathrm{S}^2_{(\mathrm{i})}}{\left(X'X\right)^{-1} MS_{res}} \right|, i = 1,2,\ldots,n \tag{15}$$

Studentized Deleted Residual: Outliers can be identified using studentized deleted residuals, which exhibit a T-distribution. For a reasonable size n, an SDR of magnitude 3 or greater (in abs. value) is regarded an outlier. Depending on the significance level used, any magnitude between 2 and 3 might be close. Standard deviation for this residual is

$$\begin{gathered} \mathrm{s}\{\mathrm{d_i}\} = \sqrt{\frac{\mathrm{MSE_{(i)}}}{1-\mathrm{h_{ii}}}} \\ t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{\mathrm{MSE_{(i)}}\left(1-\mathrm{h_{ii}}\right)}} \end{gathered} \tag{16}$$

It is called Studentized Deleted Residual. It Follows a T-distribution with n – p – 1 degrees of freedom allowing us to know what constitutes an extreme value.

3 Results and Discussion

In this study data have been collected through a survey in Kashmir using Snow-ball sampling technique during corona pandemic in the month of April and May, in district Budgam of Kashmir purposively on the basis of maximum growers.The occurrence of outliers in Apple data based on residuals generated from the model has been investigated in this work. The data pertains to yield (q/kanal), labour (Rs/day) and capital (Rs/kanal).

Table 1: Detection of HLP and values of distances and residuals by using various techniques

Table 1: Detection of HLP and values of distances and residuals by using various techniques.

A	B	C	D—	E	F	G	H	I	J	K	L	M
4	3.13	3.01	3.02	0.1	2.37	3.59	0.02	0.31	1.08	0.86	1.01	1.41
8	3.13	3.01	3.02	0.1	2.37	3.59	0.02	0.31	1.08	0.86	1.01	1.41
9	3.13	3.01	3.02	0.1	2.37	3.59	0.02	0.31	1.08	0.86	1.01	1.41
15	3.13	3.01	3.02	0.1	2.37	3.59	0.02	0.31	1.08	0.86	1.01	1.41
31	3.13	3.01	3.02	0.1	2.37	3.59	0.02	0.31	1.08	0.86	1.01	1.41
111	2.44	11.9	3.16	0.05	2.19	0.01	0.05	0.26	-0.11	1.51	1.02	1.45
122	1.43	5.5	4.38	0.05	3.28	0.18	0.05	0.42	-0.61	1.51	0.82	3.93
133	3.13	3.01	3.02	0.1	2.37	3.59	0.02	0.31	1.08	0.86	1.01	1.41
137	1.43	5.53	4.38	0.05	3.28	0.18	0.05	0.42	-0.61	1.51	0.82	3.93
160	2.44	11.9	3.22	0.05	2.24	0.11	0.05	0.28	-0.11	1.59	1.02	1.53
161	1.72	8.42	4.02	0.06	3.22	0.41	0.06	0.41	-1.02	2.02	0.8	4.21
164	2.44	11.9	4.47	0.12	3.27	0.41	0.01	0.42	-0.26	3.45	0.89	3.39
166	1.72	8.42	4.02	0.06	3.22	0.41	0.06	0.41	-1.02	2.02	0.8	4.2
167	1.72	8.42	4.02	0.06	3.22	0.01	0.06	0.41	-1.02	2.02	0.8	4.2

table 1 an outlier elucidation is obtained, and estimation results are presented and discussed.About seven percent of total observations(A) were found outliers w.r.t different techniques have been presented in the table which clearly indicates the outliers are HLP as their values are maximum. In mahalanobis distance(B), it has been observed the value of HLP lies between 1.43 to 3.13 whereas through Robust MCD(C) 3.01 to 11.90, For standardized robust residuals(D), cooks distance(E),Studentized(F), WSSDI(G), Hat diagonal(H), deleted residual(I), DFBETAS(J), DFFITS(K), Covariance Ratio(L), and Studentized deleted residuals(M) [8] the range of the observations are between 3.02 to 4.47, 0.05 to 0.12, 0.01 to 3.59, 0.01 to 0.06, 0.26 to 0.42, -1.02 to 1.08, 0.86 to 3.45, 0.80 to 1.02 and which are very high respectively [8] .

Table 2: Summary Statistics of apple yield (q/kanal) w.r.t labour (Rs/day) and capital (Rs/kanal.

Variable)	$Q1$	Median	$Q3$	Mean	S.D	MAD	C.V
x_1 labour (Rs/day)	12000.0	18000.0	30000.0	29389.0	75008.7	13343.4	254.6
x_2 Capital(Rs/kanal)	12272.0	18408.0	30680.0	25736.0	23739.2	13645.9	92.0
y Yield(q/kanal)	14.3	21.4	35.8	27.2	29.9	15.9	91.0

in Table 2, it has been observed the mean and standard deviation of y were found to be 27.2 (q/kanal) and 29.9 (q/kanal) which clearly indicates variation is more in the data as the mean is smaller than the standard deviation due to the presence of influential observations which will influence the model. Mean Absolute deviation have minimum value for y found to be 15.9.The first Quartile for x_1 were found to be positive 12000.0 whereas for y found to be positive 14.3. Low standard deviation means the data points are close to the mean, while high standard deviation means the data points are spread out over a wide range of values. Coefficient of variation (CV) of the y is found to be low 91.0 whereas for x_1 CV is very high indicating inconsistency in the data set which is due to High leverage points [9].

Table 3: Estimates of yield (q/kanal) w.r.t labour (Rs/day) and capital (Rs/kanal) of apple yield through Ordinary Least Square (OLS).

Variable)	Regression coefficient (S.E)	Parameters
Intercept	-6.902(0.197)	F-value 920.42(p¡0.0001) R^2 (0.90) AIC (-257.549) BIC (-244.276)
(Labour)	2.102* (0.231)	
(Capital)	-0.073NS (0.212)	

Table 3 presents the estimates results with standard error. The F-value was found to be significant indicates the model is adequate w.r.t to study variable. It also indicates the estimated coefficient may not having the correct sign due to the presence HLP. The estimates of the regression coefficient have been done through OLS. The R2 value (0.901) of the ols model is very high indicates 90 percents variation of variable is explained by this model with AIC (-257.549) and BIC (-244.276).

Table 4: Estimates of yield (q/kanal) w.r.t labour (Rs/day) and capital (Rs/kanal) of apple yield through Modified ordinary least square (OLS) after handling HLP.

Variable)	Regression coefficient (S.E)	Parameters
Intercept	-7.55 (0.1576)	F-value 920.42(p¡0.002) $R2$ (0.94) AIC (-411.315) BIC (-398.043)
(Labour)	2.36* (0.150)	
(Capital)	-0.186 (0.140)	

Table 4 presents the estimates results with standard error. The F-value was found to be significant indicates the model is adequate w.r.t to study variable. It also indicates the estimated coefficient may not having the correct sign due to the presence HLP[10]. The estimates of the regression coefficient have been done after replacing the influential observations through interpolation techniques . The R2 value (0.94) of the model is very high indicates 94 percents variation of variable is explained by this model with AIC (-411.31) and BIC (-398.043). Table 4 and 5 shows the regression coefficient estimation results using OLS and Modified OLS after dealing with HLP. The OLS estimates of labour was positive 2.10 and significant, after handling HLP, the estimates of labour were also found to be positive and significant (2.36). Furthermore, the coefficient of determination is noticeably greater in the Modified OLS after handling HLP estimate 0.943 which indicates that 94 percent variation of study variable is explained through this method. Capital comes out to be significant means that we have to decrease one unit in capital so that our yield is increased.

4 Conclusion:

In this paper, the detection of HLP in regression model has been discussed. A new approach Modified OLS after handling HLP for detecting outliers have been proposed which is quite useful in detecting outliers. The model parameters are affected by the presence of HLP. Further, when HLP was replaced with robust values, the results were more precise. Modified OLS method outperforms OLS method on the basis of high R2 value, minimum AIC and BIC. Hence, we suggest that in regression model, the new method can be used for detecting outliers. Also by removing the influential point it is found that the model adequacy has been increased (from R^2=0.90 to R^2=0.94).

References:

[1]Legendre, A.M. Nouvelles méthodes pour la détermination des orbites des comètes; par AM Legendre... chez Firmin Didot, libraire pour lew mathematiques, la marine, l'architecture, et les editions stereotypes, rue de Thionville **12(5)**,651-660,1806.

[2]Gauss, C. F. Theoria Motus Corporum Coelestium. Perthes, Hamburg. Translation reprinted as Theory of the Motions of the Heavenly Bodies Moving about the Sun in Conic Sections. Dover, New York,**23(2)**,121-132,1809.

[3]Raj, S.S. and Kannan, K.S. Detection of outliers in regression model for medical data. *International Journal of Medical Research & Health Sciences*,**6(1)**,50-56,2017.

[4]Wood, F.S. Linwood and non linwood-Linear and Nonlinear Least Squares Curve-Fitting Programs. *The American Statistician*, **34(1)**,177-179, 1980.

[5]Hoaglin, D.C. and Welsch, R.E.The Hat matrix In Regression and ANNOVA.*The American Statistician* **32(1)**,17-22,1978.

[6] Cook, R.D. Influential observations in linear regression. *Journal of the American Statistical Association*,**74(2)**, 169-174,1979.

[7]Rousseeuw, P. J. and Leroy, A. M.*Robust Regression and Outlier Detection* .Wiley Series in Probability and Statistics, Wiley &Sons, New York, NY, USA 223.214,1987

[8]Yousuf, R., Sharma, M., Bhat A.Robust Models and their Validation for maize production in Jammu Region of Jammu and Kashmir Union Territory. *Int.J. Agricult.Stat.Sci.*, **17(1)**,2065-2072,2021..

[9]Yousuf, R., Sharma, M, Bhat, M.I.J,Gupta,M 2021 Estimation and Validation of Robust Model for Productivity of Apple in UT of J&K using Cobb Douglas Method. *International Journal of Statistics and Reliability Engineering*. **8(3)**.376-380,2021.

[10]Yousuf, R., Sharma, M., Bhat, M.I.J., Rizvi, S.E.H 2021. Robust model for the quadratic production function in presence of high leverage points. *International Journal of Scientific Research in Mathematical and Statistical Sciences*.**8(2)**,8-13,2021.