The Journal of Administration & Economics / year 42/No 119/2019
ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

Selective Overview on Single Diagnostics Methods of Outliers in Logistic regression

Dr .Hassan S. Uraibi*
Email: hassan.uraibi@qu.edu.iq

Abstract

Logistic regression is an important statistical tool for modeling a set of independent variables that effect in a binary response variable. Most practitioners of statistics have used logistic regression in many scientific areas. Unfortunately, they are not aware that the method of estimation logistic regression breaks down in the presence of outlying data point(s) in the original dataset. The objective of this paper is to bring out the attention of respectful researchers in various scientific areas to single logistic regression diagnostics methods, and the effect of the presence of outliers on logistic regression estimates. Real data are considered in this paper and the results show the high performance of diagnostic methods to detect these observations that are affected on the logistic regression estimates. Some graphs are discussed and supported the results of diagnostic method the influence of outlying data point on the fitted logistic model.

1. Introduction

Regression analysis gives all observation equally role to determine the regression equation and subsequent conclusions (Chatterjee and Hadi,1988). In the real world of data, this assumption may be violated in the presence of outlying data points. The presence of such data points results in a breakdown of the traditional methods of logistic regression estimates. A huge efforts have been done in the literature to classify these data points and the remedy their influenced. Rousseeuw and Leroy (1987) pointed out that, the outlying data points have to be identified, and then either correct by using robust weight functions or trim it from the dataset. Many researches have been carried out and proven that outlying observations have unduly effect on the parameter estimates in linear regression (Midi et al., 2009). Indeed, it is undesirable that even a good quality data cannot be avoided of having 1 to 10 percent of outlying observations (Hampel et al., 1986).

* جامعة القادسية / كلية الادارة والاقتصاد .
مقبول للنشر بتأريخ 2019/5/19

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

Thus, many diagnostic methods have been developed to identify these outlying data points.

Generally, there are three types of outlying data points, outlier, leverage point and influential observation. The outliers is defined as the data points which are far from the bulk of predictors (Rousseeuw and Van Zomeren,1990). It is an extreme observation even though the extremeness can be positive or negative residual. High leverage points are the observations which are far removed from the main body of points in X space (Chatterjee and Hadi, 2006). Belsley et al. (1980) defined an observation to be influential if it is one, which either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations. In the following sections, some famous diagnostic methods for identification of outlying points in the literature of logistic regression will be discussed.

3. Logistic Regression

In several statistical applications, a set of variables of linearity is associated with a classifier that can represent a binary response variable, that by taking one value either 1 or 0. The binary response to dichotomous classification is quite common in many scientific studies. Logistic regression is often appropriate for such data because its fitted values will be inside the permitted range of binary responses that are bounded by only two values 0 and 1.

Let $\mathbf{y_i}$ be a dichotomous response is influenced by a linear combination of independent variables, $\mathbf{X_j}$, where $i = 1,2,\dots,n$ and $j = 1,2,\dots,p$, the logistic regression model is recommended to overcome for fitting this problem. It is well known that the i^{th} case of $\mathbf{y_i}$ follows Bernoulli distribution with probability $\pi_i = p(\mathrm{y_i} = 1|X_j)$ and $[1-\pi_i] = p(\mathrm{y_i} = 0|X_j)$ for success and failure classes, respectively and $0 \leq \pi_i \leq 1$ When $\mathbf{y}$ be a vector of size $n \times 1$ of binomial distribution, $\mathrm{Bin}(n, \pi_i)$, the vector of probability estimates π_i is non-linear. Berkson (1944) transformed the relationship between the response function π_i and $\mathbf{X_j}$ to linear relationship as follows,

$$\ln[\pi_i/(1-\pi_i)] = \beta_0 + \sum_{j=1}^{p} X_j'\beta_j \qquad (1)$$

where β_0 is the intercept and β_j is a $p \times 1$ vector of unknown regression coefficients and π_i is the response function can be written as follows,

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

$$\pi_i = \frac{e^{\beta_0+\sum_{j=1}^{p} x'_j\beta_j}}{1+e^{\beta_0+\sum_{j=1}^{p} x'_j\beta_j}} \qquad (2)$$

The coefficients of model (1) can be estimate iteratively by maximizing the logistic regression likelihood function which is defined as :

$$L(\beta; y_i) = \prod_{i=1}^{n} \pi_i^{\mathrm{y_i}} (1-\pi_i^{1-\mathrm{y_i}}) \qquad (3)$$

$$L(\beta; y_i) = \sum_{i=1}^{n} \{\mathrm{y_i}\, ln(\pi_i) + (m_i - \mathrm{y_i})\, ln(1-\pi_i)\} \qquad (4)$$

Differentiating equation (3) with respect to β yields, $\sum_{i=1}^{n}(\mathrm{y_i} - m_i\pi_i) = 0$ and $\sum_{i=1}^{n} X_{ij}(\mathrm{y_i} - m_i\pi_i) = 0$, these equation can be solved iteratively either using Newton-Raphson method or IRLS to get $\hat{\beta}$.

4. Single diagnostics of outlier

Consider $\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}$ is the estimated values of actual probabilities, the ordinary residuals may define as the deviation between $\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}$ and $\mathbf{y_i}$, $\boldsymbol{\epsilon_i} = \boldsymbol{y_i} - \hat{\boldsymbol{\pi}}_{\boldsymbol{i}}$ such that,

$$\boldsymbol{\epsilon} = \begin{cases} \mathbf{1} - \hat{\boldsymbol{\pi}}_{\boldsymbol{i}} & , \boldsymbol{y} = \mathbf{1} \\ -\hat{\boldsymbol{\pi}}_{\boldsymbol{i}} & , \boldsymbol{y} = \mathbf{0} \end{cases}$$

The *i*th case of $\boldsymbol{\epsilon_i}$ follows Bernoulli distribution with probability $\boldsymbol{\pi_i}$. consequently, the errors distribution is binomial and its variance is a function of the conditional mean as $\hat{\mathbf{V}}(\mathbf{Y}|\mathbf{X}) = \hat{\boldsymbol{\pi}}_{\boldsymbol{i}}(\mathbf{1} - \hat{\boldsymbol{\pi}}_{\boldsymbol{i}})$ where the values of the independent variables for each observation is unique. When it is not unique, the errors distribution is binomial and its variance is equivalent to $\hat{\mathbf{V}}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{m_i}\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}(\mathbf{1} - \hat{\boldsymbol{\pi}}_{\boldsymbol{i}})$, where $\boldsymbol{m_i}$ is the number of observations with the same values of $\boldsymbol{X_j}$ as observation. Hosmer and Lemeshow (2000) proposed person residuals as alternative to ordinary residuals by dividing them by $\sqrt{\boldsymbol{m_i}\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}(\mathbf{1} - \hat{\boldsymbol{\pi}}_{\boldsymbol{i}})}$.

The Pearson residual defined for the *i*th covariate pattern is given by $\boldsymbol{r_i} = \frac{\boldsymbol{y_i} - \boldsymbol{m_i}\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}}{\sqrt{\boldsymbol{m_i}\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}(\mathbf{1}-\hat{\boldsymbol{\pi}}_{\boldsymbol{i}})}}$, where $\boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n}$. the square of $\boldsymbol{r_i}$ has related to Pearson chi square test statistic, due to the $\boldsymbol{r_i^2}$ measures the contribution of $\mathbf{y_i}$ to the Pearson chi square goodness of fit statistic Hosmer and Lemeshow (2000). Unfortunately, with such as binary data chi-square test statistics does not follow an approximate of chi-square distribution without replicates (Sarkar et al., 2011).

A seriously problem arises where the variance of $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}$ calculated, since the $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} = \boldsymbol{y_i} - \hat{\boldsymbol{\pi}}_{\boldsymbol{i}} \approx (1 - h_{ii})\mathrm{y_i}$, hence the variance of the residual is given by $V(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}) = (1 - h_{ii})\hat{\pi}_i(1 - \hat{\pi}_i)$. It is obvious that $V(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}})$ dose not have unit

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

variance and consequently the variance of Pearson residuals is not constant.

4.1 The Studentized Pearson residuals

The Studentized Pearson residuals has been proposed in the literature by dividing r_i by the standard deviation which is approximate as $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$,where $H = \sqrt{\hat{V}}X\left(X'\hat{V}X\right)^{-1}X'\sqrt{\hat{V}}$ and h_{ii} is i^{th} diagonal element of Pregibon leverage H which is so called hat matrix. The Studentized Pearson residuals $\mathrm{spr_i}$are defined as

$\mathrm{spr_i} = r_i/\sqrt{(1-h_{ii})}$

and the i^{th} observation associated $|spr_i| > 3$ are generally identified as outlier.

Draper and John (1981) pointed out that the approach of row deletion of influential data points and then examine it effects on the estimates shows deletion the clean observation with small residual has most influential than the outlier with large residual on the parameter estimates.

5. Single diagnostics of leverage point

A large body in the literature has been done to measure the influential observation in X-direction of linear regression model. The detection high leverage point such as mahalanobise distance , twice-the-mean rule (Hoaglin and Welsch,1978), thrice-the-mean rule (Vellman and Welsch,1981) and others that are based on leverage point which measures the distance of a covariance pattern from the mean (Imon,2006). Such as these methods are impracticable in logistic regression where the most outlying data points in X-direction may have the smallest leverage.

In the setting of logistic regression, Pregibon (1981) introduced hat matrix to measure the leverage points in the independent variables as follows,

$$H = \hat{V}^{1/2}X\left(X'\hat{V}X\right)^{-1}X'\,\hat{V}^{1/2} \qquad (5)$$

Another approach is proposed by Imon (2006) based on a quantity that measures the distance of each covariance pattern from the mean. He estimated the probability of each covariate pattern for general logistic regression model $y = \pi(X) + \epsilon$ as follows,

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0+\sum_{j=1}^{p}x_j'\hat{\beta}_j}}{1+e^{\hat{\beta}_0+\sum_{j=1}^{p}x_j'\hat{\beta}_j}}$$

$\hat{\beta}$ is computed by equation (4), and then the fitted values of the i^{th} covariate pattern is calculated as $\hat{y}_i$ where,

$\hat{y}_i = \hat{\pi}_i$

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

The general element $v_i = m_i\hat{\pi}_i(x_i)[1-\hat{\pi}_i(x_i)]$ can be founded and employed with the diagonal element of the hat matrix which is defined in(5)

$$h_i = m_i\hat{\pi}_i(x_i)[1-\hat{\pi}_i(x_i)]x_i'\left(X'\hat{V}X\right)^{-1}x_i$$

The observation that possess $h_i > 2p/n$ are generally identified as high leverage point. Indemnification of high leverage values is crucial due to their responsibility for masking and swapping outliers (see, Peña and Yohai,1995) . Jennings (1986) shows that the leverage value $h_i > 2p/n$ is closely related to the $\hat{\pi}_i$, therefore, he considered the *i*th observations are high leverage points if their corresponding $\hat{\pi}_i \in [0.1,0.3]$ and/or belong to [0.7, 0.9]..

6. Cook's distance CD_i

Nurunnabi et al. (2010) defined the outliers and influential observations in logistic regression may occur as result of misclassification between the binary responses. This may occur due to meaningful deviation in explanatory variables. He rewrite the formula of Cook distance (Cook,1977) to be suitable with logistic regression setting as follows,

$$CD_i = \left[(\hat{\beta}^{(-i)}-\hat{\beta})'(X'\hat{V}X)(\hat{\beta}^{(-i)}-\hat{\beta})\right]/k\,\hat{\sigma}^2, \quad i = 1,2,\dots,3 \qquad (6)$$

where $\hat{\beta}^{(-i)}$ is the estimated parameter of β with the *i*th observation deleted, $k = p+1$.

The cutoff point that is used in this paper is formulated as follows,

$$Crit = \frac{h_{ii}}{1-h_{ii}} \times \chi^2_{0.95}(1)$$

The *i*th observation is considered as influential observation where $CD_i > Crit$

It is well known that Cook's distance measures the distance between least square estimates based on full sample size and the estimate that obtained by deletion the *i*th observation. Many researchers in robust statistics literature expressed Cook's distance in terms of the *i*th DFFITS,Standardized Pearson residual $\mathrm{spr_i}$ and leverage points h_{ii}.

Belsley et al. (1980) defined the influential observations as points that have a demonstrable impact on the various estimates and introduced DFFITS that defined as

$$\mathrm{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ii})}}, \quad i = 1, 2, \dots, n \qquad (7)$$

where $\hat{y}_i^{(-i)}$ and $\hat{\sigma}_{(i)}$ are respectively the *i*th fitted response and the estimated standard error with the *i*th observation deleted.

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

Nurunnabi et al. (2010) expressed the DFFITS formula in (8) for logistic regression in terms of Sudentized Pearson residuals $\mathrm{spr_i}$ and leverage values h_{ii} as

$$DFFITS_i = spr_i \sqrt{\frac{h_{ii}\,\hat{\pi}_i(1-\hat{\pi}_i)}{(1-h_{ii})\,[\hat{\pi}_i(1-\hat{\pi}_i)]^{(-i)}}}\ , i = 1,2,\dots,n \quad (8)$$

Where the observation possessing $DFFITS_i > c\sqrt{k/n}$ is generally identified as an influential observation.

The numerator of (6), $\left[\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)'\left(X'\hat{V}X\right)\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)\right]$ equivalents to $\left(\frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\sqrt{(1-h_{ii})}}\right)^2$ which equals to $\mathrm{DFFITS}_i^{\,2}\,\hat{\sigma}_{(i)}^2$, therefore the Cook's distance in terms of DFFITS can be expressed as

$$CD_i = \mathrm{DFFITS}_i^{\,2}\,\hat{\sigma}_{(i)}^2\,/k\,\hat{\sigma}^2 \quad (9)$$

Another relationship has been proposed in literature by using Standardized Pearson residual $\mathrm{spr_i}$ and leverage points h_{ii} as follows,

$$CD_i = \frac{1}{k}(\mathrm{spr_i})^2\left(\frac{h_{ii}}{1-h_{ii}}\right) \quad (10)$$

7. PimaIndians Diabetes

To investigate the performance of single diagnostics method, the PimaIndiansDiabetes dataset is considered. The original version of this data contains 768 observations on 12 variables, but it is corrected by removing several physical impossibilities values which are considered as missing data. For more details see Wahba et al (1995) and Ripley (1996).

First, Fig 1 shows the investigation of the difference between observed and fitted value by using marginal model plot. The dependent variable (diabetes) is plotted against independent variables. The observed data and fitted value are shown in solid and long dash lines, respectively. It is obvious that blood pressure, triceps, insulin and body mass index fit poorly. Consequently, the fit is not support the entire set of independent variables patterns. Due to that the diagnostics regression is required to detect the outlying observations that have significant impact on the model.

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

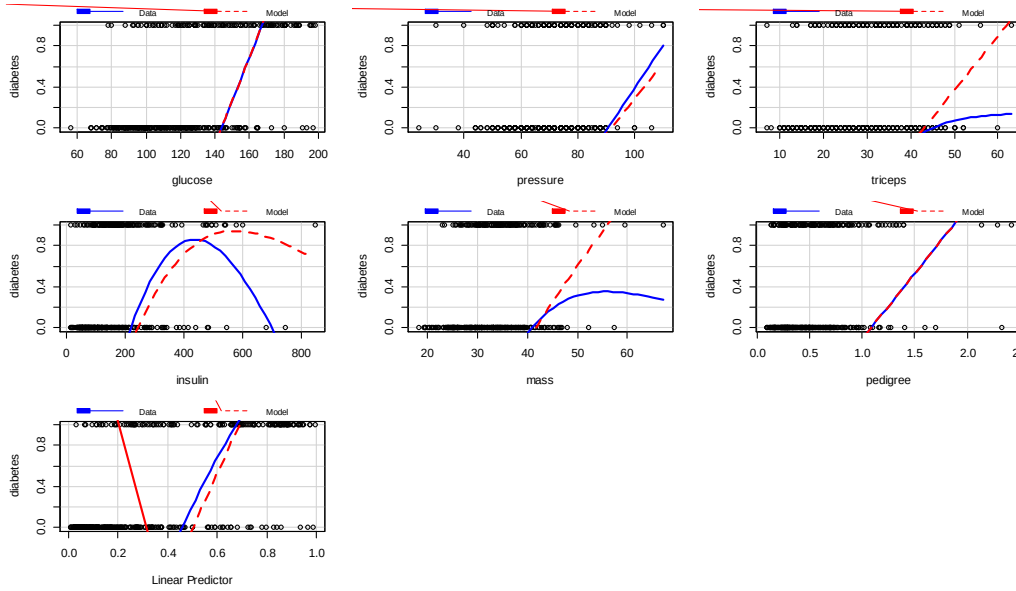

Figure 1
Marginal model plot drawing response variable against each predictors and linear predictor

Figure 2 shows the observations indexed 7 and 229 , respectevly, are most likely outliers by using studentized pearson residual method, Bonferonni P value confirms only the observation 229 is influenced outliers. Note that, sometimes the detection method of outliers identifys some observation as outliers but it is not, such as this observation is called swapping.

The results of cook distance to measure the influence of observations are present in Figure 2 and the observations 229 and 745 are identified as influential observations. The diagonal elemrents of hat matrix in the Figure 2 diagnoses two observations are leverage points which indexed 29 and 255, respectevly.

To examine the change of coefficient that may be happen as a result to the presence of outlying observations (two influential observations (229, 745) and two leverage

points (29,255) and two outliers (229,7)) two models (*Model 1* and *Model 2*) are considered with and without these outlying observation, and then estimates of each model are compared.

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

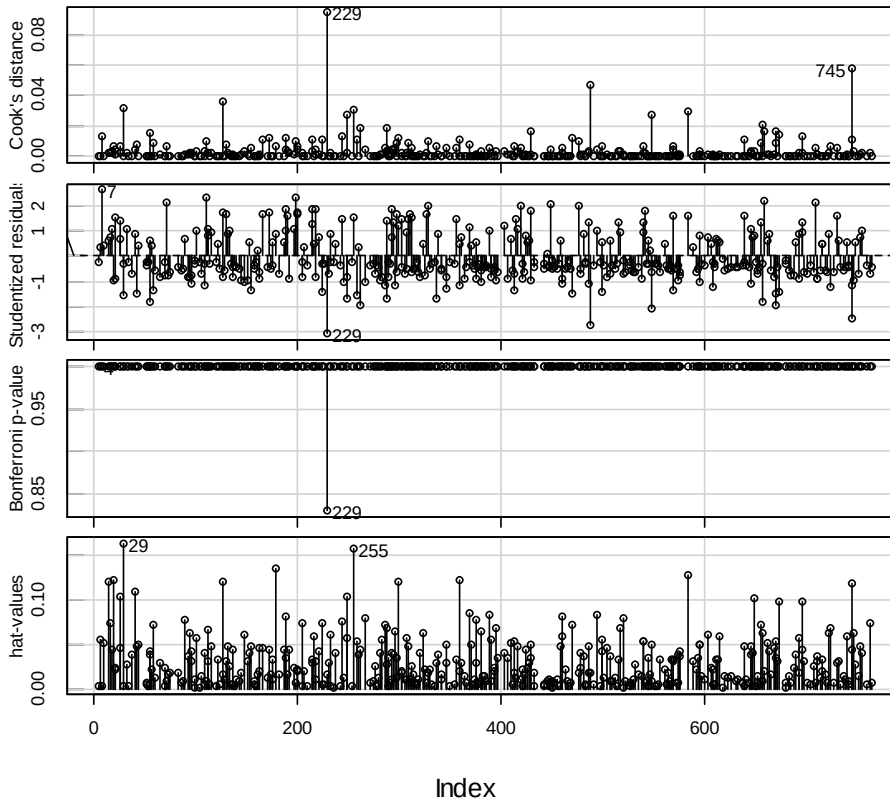

Figure 2
Diagnostic plots combining Cook's distance, Studentized residuals, Bonferonni P and hat-values.

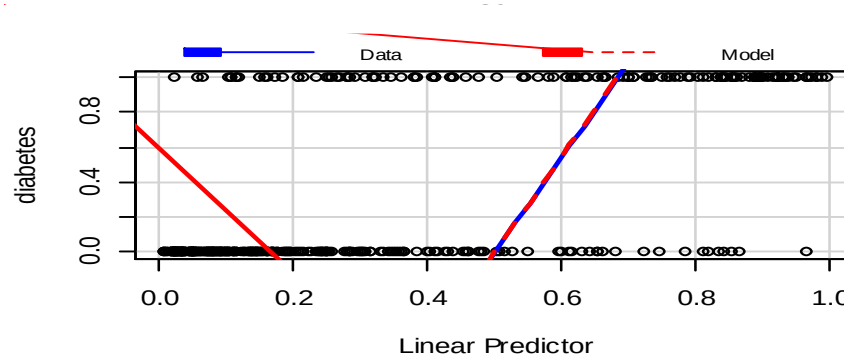

Figure 3
the diabetes observed data (solid line) and linear predictor (fitted value; long dash line) for the *Model 2*

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.5

Table 1
Displays the estimates of logistic regression of *Model 1* and *Model 2*

Variable	*Model 1*		*Model 2*	
Variable	Estimate	Standard Error	Estimate	Standard Error
Intercept	-8.87	1.17	-9.60	1.23
glucose	0.039	0.006	0.041	0.006
pressure	-0.004	0.012	-0.004	0.012
triceps	0.015	0.017	0.016	0.018
insulin	-6.48E-04	1.36E-03	-4.06E-05	1.40E-03
mass	0.063	0.027	0.064	0.028
pedigree	1.017	0.439	1.449	0.464
age_bucket31-40	0.854	0.377	0.982	0.385
age_bucket41-50	1.575	0.520	1.488	0.534
age_bucket50+	1.384	0.637	1.200	0.656
preg_bucket10+	0.828	0.767	1.411	0.838
preg_bucket6–10	-0.243	0.420	-0.243	0.428

As can be seen from Table 1 that coefficients of insulin, pedigree and preg_bucket10+are influenced after deletion the outlying observations. It is clear that the plot of diabetes (response variable) against linear predictor in Figure 3 is better than the one which has been presented in Figure 1. As Figure 3 shows the solid line of observed data of diabetes matches the long dash line of fitted values, while in Figure 1 both lines are not matches. However, the conclusion of deletion the observation 229 improves the model fitting, so it is identified an influential observation.

7. Conclusion

The main purpose of this paper is to describe the single logistic regression diagnostics methods and to give it the attention of researchers. The diagnostic method of outlying observation is presented, and then real data are considered. The results show the high performances of diagnostic methods to detect the outlying observations. The plots have been confirmed the correct identification of outliers, leverage points and influential observations by these methods. Outliers and leverage point individually are fare away from block of data. Both may be configured the influential observation but vice versa in not correct. When influential observation is dropped from the model, there will be a significant shift of the coefficient and regression line.

References

1. Belsley, D.A., Kuh, E., Welsch, R.E., (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York.
2. Berkson, J. (1944), Application of logistic function to Bio-assay. Journal of the American Statistical Association, 9,357-365
3. Chatterjee, S., Hadi, A.S.,(1988). Sensitivity Analysis in Linear Regression. Wiley, New York.
4. Cook, R.D., (1977). Detection of influential observations in linear regression, Technometrics 19, 15-18.
5. Cook, R.D., Weisberg, S., (1982). Residuals and Influence in Regression, Chapman and Hall, London.
6. Hampel, F. R., Ronchetti, E. M, Rousseeuw, P. J. and Stahel, W. A. (1986). Robust statistics, J. *Wileyand Sons, New York*.
7. Hoaglin, D. C. and Welsch, RE. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22.
8. Hosmer, D.W., Lemeshow, S., (2000). Applied Logistic Regression. 2nd ed., Wiley, New York.
9. Imon, A. H. M. R.**, (**2006). Identification of high leverage points in logistic regression, Pak. J. Statist. 22, 147 – 156.
10. Nurunnabi , A.A.M., A. H.M. Rahmatullah Imon & M. Nasser (2010) Identification of multiple influential observations in logistic regression, Journal of Applied Statistics, 37:10, 1605-1624, DOI: 10.1080/02664760903104307.
11. Midi. Habshah, M. R. Norazan and A.H.M.R. Imon (2009), "The performance of diagnostic-robust generalized potentials for the identification of mutliple high leverage points in linear regression", Journal of Applied Statistics vol. 36(5), pp. 507-520.
12. Pregibon, D. (1981). Logistic regression diagnostics. The Annals of Statistics, 9, 705-724. doi:10.1214/aos/1176345513
13. Pena, D. and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. J. R. Statist. Soc. B, 57(1): 145–156.
14. Rousseeuw, P. J, and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223.
15. Sarkar .S. K., H. Midi and S. Rana, (2011) "Detection of outliers and influential observations in binary logistic regression: An empirical study", Journal of Applied Sciences vol. 11 (1), pp.26-35.
16. Zhang Z. Residuals and regression diagnostics: focusing on logistic regression. Ann Transl Med 2016;4(10):195. doi: 10.21037/atm.2016.03.36
17. Velleman, P. F, and Welsch, RE. (1981). Efficient computing of regression diagnostics. *The American Statistician, 35*(4), 234-242.

…….
……
……