Applied Statistics

Article

Deletion Diagnostics in Logistic Regression

Soham Ghosh

Department of Statistics, University of Wisconsin, Madison, USA; sghosh39@wisc.edu

Article History
Received: 25 June, 2022
Accepted: 10 July, 2022
Published: 15 July, 2022

Citation
Ghosh, S. (2022). Deletion diagnostics in logistic regression. *Journal of Applied Statistics, 1*(1), 1-13. https://doi.org/10.56388/as220715

Abstract: Today, there are not many good measures for detecting influential observations in case of fitting a logistic regression model. So, the purpose of this article is to extrapolate from the pre-existing deletion diagnostics defined for detecting influential points for multiple linear regression, i.e. the DFFITS, DFBETAS and Cook's Distance to the scenario of a binary logistic regression model and then view the multinomial model as a special case of the same. The threshold for determining whether an observation is an influential observation or not is judged using the asymptotic distribution of the Cook's Distance in the multinomial setting, both for the single and the group deleted case. The results are examined under various simulation scenarios as well as over the modified Kyphosis data-set.

Keywords: Logistic Regression; Multinomial Logistic Model; Group deletion; Asymptotic Distribution; Influential Observations; Cook's D; DFFITS

1. Introduction

In fitting the regression model $Y_i = x_i'\beta + \epsilon_i$ where $i = 1,2,\dots,n$, where x_i is the i^{th} row of the design matrix X, β being the vector of parameters and ε_i being the random error component, we can think of the data point $(x_i, Y_i)'$ as a point in the p-dimensional space. Some of these points may arouse suspicions as they are 'discordant' with the other points. Such points are usually referred to vaguely as *outliers*, and they may or may not have an effect on estimation and inference using the prescribed regression model (Seber & Lee, 2003). If too many outliers remain undetected, they can really hamper the analysis. Thus, we must scrutinize these points closely and decide whether they should be eliminated from our sample. In the context of linear regression models, we mainly speak about two types of outliers: leverage points, or points that are remote in the X-coordinate space, and influential observations, being points with high leverage as well as having unusual Y-coordinate values. They are quite distant from the cloud of data points. With a view to discuss about outliers in logistic regression, it is purposeful to state the model of a binary logistic regression.

The general form of a binary logistic regression model is

$$y_i = E(y_i) + \epsilon_i$$

where the observations y_i are independent Bernoulli random variables with expected values:

$$E(y_i) = \pi_i = \frac{e^{\{x_i'\beta\}}}{1 + e^{\{x_i'\beta\}}}, \quad i = 1,2,\dots,n$$

Here $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ik})'$ is the vector of explanatory variables and $\beta' = (\beta_0, \beta_1, \dots, \beta_k)'$ is the vector of parameters.

In a logistic regression model, outliers generally refer to mis-classified observations. Drawing analogy from the linear regression model, we can define residuals in the case of a binary logistic regression model. The standardized residuals for logistic regression are given by:

$$e_i = \frac{y_i - \widehat{\pi_i}}{\sqrt{\widehat{\pi_i}(1-\widehat{\pi_i})}}, \quad i = 1, 2, \dots, n$$

where $\widehat{\pi_i}$ is the estimated probability for the i^{th} observation (Agresti, 2012).

As a matter of fact, the standardized residuals of the outlying observations deviate quite significantly from the expected range. According to (Christensen, 1997), in binary logistic regression, the standardized residuals are expected to be within ± 2. He has identified the points for which the residuals fall outside this range as potential outliers. However, analysis of residuals and identification of outliers or influential points are not studied frequently in case of logistic regression models. This might lead

to serious consequences as we traditionally fit a logistic regression model with maximum likelihood estimates of the parameters, which are extremely sensitive to outlying observations. As a result, it is worthwhile to study the problem of outlier detection in logistic regression models. For the purpose of our study, we confine ourselves to the detection of influential observations in binary and multinomial logistic regression. In both traditional and modern literature pertaining to this topic, some prominent diagnostic measures of influence have emerged with (Pregibon, 1981) using Pearsonian residuals and deviance residuals corresponding to each individual model observation; large values of these indicate that the observation is an outlier. However, he argues that these quantities cannot adequately measure the effect of the outlier on the many components of the fitted model and hence resorts to the perturbation technique (Pregibon, 1980). Moreover, there is the mean-shift outlier model (Williams, 1987) with the maximum likelihood estimate of the parameters based on the full set of observations as an initial solution. Then taking a single step of weighted least squares, he obtains an approximate relation between the estimates based on the deleted set and the full set of observations. Unlike the multiple linear regression model, deletion diagnostics like $DFFITS, DFBETAS$, and Cook's Distance are not popular in logistic regression, in the sense that they require iterative procedures and are deemed too complicated for application. However, our focus is to study the performance of these deletion diagnostics mainly $DFFITS$ and Cook's Distance. From the study and simulations, they seem to perform well in all the setups involving one regressor, multiple regressors, and also in the case of multiclass logistic regression. However, they seem to underperform in presence of multiple outliers due to the effects of *masking* and *swamping*, which we have addressed in this section.

In section 3, we have derived the explicit expressions of $DFFITS$ and Cook's Distance using one-step approximations, highlighting the fact that we can reduce much computation labor by just updating the residuals instead of fitting the entire regression model in each iteration after deletion of one observation. For the sake of improving our single deletion statistics to perform well in presence of multiple outliers, d-deletion statistics namely Generalized $DFFITS$ ($GDFFITS$) and Generalized Cook's Distance have been proposed. Section 4 demonstrates how well these diagnostics have fared in real-life modified datasets as well as simulated ones. In section 5, some further advanced research problems in this field have been mentioned along with the scope of improvement of the measures discussed in the previous sections.

Masking and swamping effects

Datasets with multiple outliers or clusters of outliers are subject to two phenomena called masking and swamping. For an intuitive understanding of these effects, we cite the following definitions from (Acuna & Rodriguez, 2004).

Masking effect: *It is said that one outlier masks a second outlier if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier, the second instance emerges as an outlier. Masking occurs when a cluster of outlying observations skews the mean and covariance estimates towards it, and the resulting distance of the outlying point from the mean is small.*

Swamping effect: *It is said that one outlier swamps a second observation if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier, the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances* skews *the mean and the covariance estimates towards it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers.*

In general, masking can occur when we specify too few outliers in the test, that is, for example, if we are testing for a single outlier when there are in fact two (or more) outliers, these additional outliers may influence the value of the test statistic enough so that no points are declared as outliers and on the other hand swamping can occur when too many outliers are specified. It is because of these effects that trying to apply a sequential single deletion diagnostic like $DFFITS$ or Cook's Distance can fail. For example, masking may cause the outlier test for the first outlier to return a conclusion of no outliers and so the testing for any additional outliers is not performed. This gives rise to a fundamental problem and therefore, we require detection techniques that are free from these problems. In an effort to tackle this problem, it is necessary to introduce a group deleted version of the deletion diagnostics discussed previously. We assume that d observations among a set of n observations are deleted. Hence, we have two cases, observations remaining in the analysis (denoted by R) and observations which have been deleted (denoted by the set D). We then modify our traditional $DFFITS$ and Cook's Distance measures for each observation lying either in R or in D. Correspondingly, the cutoff points for these measures will also change. However, the choice of d is a much deep-rooted theoretical problem that extends beyond the scope of this article. We have discussed a rough outline of an algorithm elaborating on how to choose the deletion set D for implementing $GDFFITS$ or Generalized Cook's Distance. We have also demonstrated how implementing this generalized statistic can help us detect the outliers correctly on the modified Kyphosis data set (Nurunnabi & Rahmatullah Imon, 2010) in Section 4. We now move on to the components defining each of these statistics.

2. Theoretical Background

We define the various deletion diagnostics of our interest in the context of linear regression. It is worthy to note that these are single point deletion diagnostics or also known as *leave-one-out diagnostics*. We shall discuss more generalized versions of these measures by deleting d observations instead of a single observation in the next section. For our purpose we assume that the design matrix X is of full rank.

Cook's Distance

Cook R. D. (1977) suggested a measure of influence of a point by using the squared distance between the least squares estimate based on all n points, $\hat{\beta}$ and the estimate obtained by deleting the i^{th} point, $\widehat{\beta^{(i)}}$. It can be expressed in the form:

$$D_i = \frac{\left(\widehat{\beta^{(i)}}-\hat{\beta}\right)'(X'X)\left(\widehat{\beta^{(i)}}-\hat{\beta}\right)}{pMS_{res}} \quad (1)$$

where p is the number of predictors and MS_{res} is the residual error variance for the full data-set.

Removing the i^{th} observation should keep $\widehat{\beta^{(i)}}$ close to $\hat{\beta}$ unless the i^{th} observation is an outlier. (Cook & Weisberg, 1982) indicate that the magnitude of D_i is usually assessed by comparing it to $F_{\alpha,p,n-p}$. If $D_i = F_{0.5,p,n-p}$, then deleting the point i would move $\widehat{\beta^{(i)}}$to the boundary of an approximate 50 % confidence region for β based on the complete dataset. Since $F_{0.5,p,n-p} \approx 1$, we usually consider points for which $D_i > 1$ to be influential. This cutoff of unity seems to work well in practice. Similarly, (Bollen & Jackman, 1985) suggested Cook's distance for observations is more than a cut-off of $\frac{4}{n} - p$ which is treated as the traditional approach to evaluating the influential observations. Again, the D_i statistic can be re-written as:

$$D_i = \frac{r_i^2}{p}.\frac{h_{ii}}{1-h_{ii}} \quad \text{for} \quad i \in \{1,\ 2,\ \dots,\ n\} \quad (2)$$

where r_i is the i^{th} studentized residual given by $\frac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}$ and h_{ii} is the i^{th} diagonal element of the hat matrix $X(X'X)^{-1}X'$.

The alternate form in (2) is useful to visualize the fact that Cook's Distance, being the product of a function of h_{ii} (which is a measure of leverage of a point) and the square of the studentized residual (that reflects how well the model fits the i^{th} observation y_i), it measures the joint influence on the observation being an outlier on Y -space and in the space of the predictors (X -space).

DFFITS

In order to investigate the deletion influence of thei^{th} observation on the predicted or fitted value gives rise to this diagnostic proposed by (A., Kuh, & Welsch, 1980).

$$DFFITS_i = \frac{\widehat{y_i}-\widehat{y_{(i)}}}{\sqrt{S_{(i)}^2 h_{ii}}} \quad \text{for } i \in \{1,2,\dots,n\} \quad (3)$$

Where

$\widehat{y_{(i)}}$: Fitted values of y_i obtained without the use of the i^{th} observation

$S_{(i)}^2$: Residual Sum of Squares without the i^{th} observation

The mathematical formulation of $S_{(i)}^2$ is given by $\frac{(n-p)MS_{res}-\frac{e_i^2}{1-h_{ii}}}{n-p-1}$, where e_i is the i^{th} residual $y_i - \hat{y}_i$. The denominator of $DFFITS$ is just a standardization and measures the number of standard deviation units the fitted value $\hat{y}_i$ changes if the observation i is removed. There is an empirical cutoff for $DFFITS$ that any observation with $|DFFITS| > 2\sqrt{\frac{p}{n}}$ (A., Kuh, & Welsch, 1980) warrants attention and may be labelled as an influential point.

DFBETAS

Like $DFFITS$, $DFBETAS_{ji}$ is a statistic that indicates how much the regression coefficient $\hat{\beta}_j$ changes in standard deviation units, if the i^{th} observation is deleted. It is expressed as follows:

$$DFBETAS_{ji} = \frac{\hat{\beta}_j-\widehat{\beta_j^{(i)}}}{\sqrt{S_{(i)}^2 C_{jj}}} \text{ for } j \in \{1,2,\dots,p\} \text{ and } i \in \{1,2,\dots,n\} \quad (4)$$

Where

$\hat{\beta}_j$: Estimated j^{th} regression coefficient

$\widehat{\beta_j^{(i)}}$: Estimated j^{th} regression coefficient with thei^{th} observation deleted.

C_{jj} : j^{th} diagonal element of $(X'X)^{-1}$

Thus, a large magnitude of $DFBETAS$ indicates that the observation i has considerable influence on the estimates of β_j. It is evident from the expression that $DFBETAS$ is a$n \times p$ matrix. Empirically, if $\left|DFBETAS_{ji}\right| > \sqrt{\frac{2}{n}}$ (A., Kuh, & Welsch, 1980), then the i^{th} observation warrants examination. In the next section we will define analogous measures of influence for a binary logistic regression model. Moreover, rather than relying on empirical cutoffs, we can get an idea about the approximate cutoffs for these measures by considering the critical points of their asymptotic distributions.

3. Developments

If we look at the expression in (1), (3) and (4), their distributions depend on a single quantity $\hat{\beta} - \widehat{\beta^{(i)}}$ and the rest are all standardization terms. So, if we can find the asymptotic distribution of $\hat{\beta} - \widehat{\beta^{(i)}}$, it can be utilized to find the asymptotic

distribution of all three measures for a binary logistic regression model. However, $DFBETAS$ being a matrix, it is computationally expensive to calculate the cutoff values for each of its elements and less interpretable during comparisons. Hence, we focus only on the two measures $DFFITS$ and Cook's Distance. We first derive their expressions explicitly and then comment about their asymptotic distributions.

3.1 Single deletion case

Model: Let y_i be a binary response variable taking values $\{0,1\}$ and x_i be the corresponding vector containing p covariates for the i^{th} observation, $i = 1,2,\dots,n$. Let $\beta = (\beta_1,\dots,\beta_p)'$ be the parameter vector associated with the covariate x_i and X be the design matrix of the order $n \times p$ formed by augmenting the covariates x_i as columns. The model is specified by:

$$y_i = \begin{cases} 1 \text{ with probability } \pi_i \\ 0 \text{ with probability } 1-\pi_i \end{cases}$$

Where

$$\pi_i = P[y_i = 1|x_i] = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}.$$

The likelihood function for estimation of β is given by:

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

The log-likelihood function is given by:

$$l(\beta) = \sum_{i=1}^{n}\left(y_i x_i'\beta - \ln\left(1+e^{x_i'\beta}\right)\right) \quad (5)$$

The likelihood equations $\frac{\partial l(\beta)}{\partial \beta} = 0$ implies,

$$\sum_{i=1}^{n}\left(x_i y_i - x_i \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) = 0$$

Define $s_i = y_i - \pi_i, v_i = \pi_i(1-\pi_i), S = (s_1,..,s_n)', V = \text{diag}((v_i))$ and $Z = V^{\frac{1}{2}}X$.
Thus we can rewrite the likelihood equations as:

$$\sum_{i=1}^{n} x_i(y_i - \pi_i) = 0 \quad (6)$$

Hence, we have

$$\frac{\partial l(\beta)}{\partial \beta} = X'S$$

and

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \frac{e^{x_i'\beta}}{\left(1+e^{x_i'\beta}\right)^2} x_i x_i' = -(X'VX) = -Z'Z$$

We use the Newton Raphson method with an initial solution β^* of β, in order to obtain the first order approximation of $\hat{\beta}$. Thus, we have

$$\begin{aligned} \hat{\beta} &= \beta^* - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'}\right)^{-1} \frac{\partial l(\beta)}{\partial \beta} \\ &= \beta^* + (Z'Z)^{-1}(X'S) \\ &= \beta^* + (Z'Z)^{-1}\left(Z'V^{-\frac{1}{2}}S\right) \end{aligned}$$

Where all the equalities are evaluated at $\beta = \beta^*$.

In order to find the expression for $\widehat{\beta^{(j)}}$, we consider the deleted log-likelihood:

$$l(\beta) = \sum_{i=1, i\neq j}^{n}\left(y_i x_i'\beta - \ln\left(1+e^{x_i'\beta}\right)\right)$$

Then,

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1, i\neq j} x_i(y_i - \pi_i) = X'S - x_j s_j$$

and

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1, i\neq j}^{n} \frac{e^{x_i'\beta}}{\left(1+e^{x_i'\beta}\right)^2} x_i x_i' = Z'Z - z_j z_j'$$

where $z_j = \sqrt{v_j}x_j$

Here let the initial point be $\beta_0^{(j)}$, so my first order approximation of $\widehat{\beta^{(j)}}$ becomes:

$$\widehat{\beta^{(j)}} = \beta_0^{(j)} + \left(Z'Z - z_j z_j'\right)^{-1}\left(X'S - x_j s_j\right) \quad (7)$$

We now address the issue of the choice of the initial solutions β^* and $\beta_0^{(j)}$. A clever choice of β^* would be the estimator obtained from a linear model, implying $\beta^* = (X'X)^{-1}X'y$.

For the deleted case, we start with $\beta_0^{(j)} = \left(X'X - x_j x_j'\right)^{-1}\left(X'y - x_j y_j\right)$. However, it is quite well known that:

$$\beta_0^{(j)} = \beta^* - \left(1 - h_{jj}\right)^{-1}(X'X)^{-1}x_j e_j \tag{8}$$

where $e_j = y_j - x_j'\beta^*$ and $h_{jj} = x_j'(X'X)x_j$.

From here, the difference between $\widehat{\beta^{(j)}}$ and $\hat{\beta}$ can be obtained as:

$$\hat{\beta} - \widehat{\beta^{(j)}} = \beta^* + (Z'Z)^{-1}\left(Z'V^{-\frac{1}{2}}S\right) - \left[\beta_0^{(j)} + \left(Z'Z - z_j z_j'\right)^{-1}\left(X'S - x_j s_j\right)\right]$$

Strictly, the second term on the RHS should have been evaluated at $\widehat{\beta_0^{(j)}}$, but since it is not necessary to obtain $\widehat{\beta_0^{(j)}}$ otherwise, the correction term can be evaluated at β^* instead (Sen Roy & Guria, 2008).

If we use the result:

$$\left(Z'Z - z_j z_j'\right)^{-1} = (Z'Z)^{-1} + \left(1 - z_j'(Z'Z)^{-1}z_j\right)^{-1}\left((Z'Z)^{-1}z_j z_j'(Z'Z)^{-1}\right)$$

then it follows,

$$\widehat{\beta^{(j)}} = \hat{\beta} - \left(1 - \mathrm{w_{jj}}\right)^{-1}(\mathrm{Z'Z})^{-1}\mathrm{z_j\, k_j} - \left(1 - \mathrm{h_{jj}}\right)^{-1}(\mathrm{X'X})^{-1}\mathrm{x_j\, e_j} \tag{9}$$

where $k_j = \left(\frac{s_j}{\sqrt{v_j}} - z_j'(Z'Z)^{-1}Z'V^{-\frac{1}{2}}S\right)$, evaluated at $\beta = \beta^*$ and $w_{jj} = z_j'(Z'Z)^{-1}z_j$.

The expression in (9) is useful in the sense that it allows us to calculate the difference $\widehat{\beta^{(j)}} - \hat{\beta}$ without fitting the model on the deleted dataset repeatedly, rather calculating the residual terms w_{jj}, k_j and h_{jj} will suffice for each deleted j.

3.2 Closed form of DFFITS

We know that $DFFITS_i$ is actually calculated by taking the difference of the predicted probabilities with and without deleting the i^{th} observation i.e. $\widehat{\pi_i} - \widehat{\pi_i^{(j)}}$ which is same as computing $\mathrm{logistic}\left(x_i'\hat{\beta}\right) - \mathrm{logistic}\left(x_i'\widehat{\beta^{(j)}}\right)$. If this difference is large, then we investigate the point j. However, the logistic function $f(z) = \frac{e^z}{1+e^z}$ is Lipschitz continuous, and hence it is equivalent to recognize the difference $x_i'\hat{\beta} - x_i'\widehat{\beta^{(j)}}$. Thus, a simpler version of $DFFITS$ is given by:

$$\begin{aligned} DFFITS^{(j)} &= x_j'\hat{\beta} - x_j'\widehat{\beta^{(j)}} \\ &= \left(1 - w_{jj}\right)^{-1}x_j'(Z'Z)^{-1}z_j k_j + \left(1 - h_{jj}\right)^{-1}x_j'(X'X)^{-1}x_j e_j \\ &= \left(1 - w_{jj}\right)^{-1}v_j^{-\frac{1}{2}}w_{jj}k_j + \left(1 - h_{jj}\right)^{-1}h_{jj}e_j \end{aligned}$$

A large absolute value of $DFFITS$ would mean that the j^{th} observation has a considerable impact on the fit, and hence can be declared as an outlier.

3.3 Closed form for Cook's Distance

For a binary logistic regression model, Cook's Distance is given by:

$$D^{(j)}\left(\hat{\beta}\right) = \left(\widehat{\beta^{(j)}} - \hat{\beta}\right)'(X'VX)\left(\widehat{\beta^{(j)}} - \hat{\beta}\right) \tag{10}$$

If we plug in the expressions obtained $for\widehat{\beta^{(j)}} - \hat{\beta}$, we get:

$$D^{(j)}\left(\hat{\beta}\right) = \left(\frac{k_j}{1 - w_{jj}}z_j'(Z'Z)^{-1} - \frac{e_j}{1 - h_{jj}}x_j'(X'X)^{-1}\right)Z'Z\left(\frac{k_j}{1 - w_{jj}}(Z'Z)^{-1}z_j - \frac{e_j}{1 - h_{jj}}(X'X)^{-1}x_j\right)$$

which then reduces to:

$$D^{(j)}\left(\hat{\beta}\right) = \frac{k_j^2 w_{jj}}{\left(1 - w_{jj}\right)^2} - 2\frac{k_j e_j}{\left(1 - h_{jj}\right)\left(1 - w_{jj}\right)}z_j'(X'X)^{-1}x_j + \frac{e_j^2}{\left(1 - h_{jj}\right)^2}x_j'(X'X)^{-1}(Z'Z)(X'X)^{-1}x_j \tag{11}$$

Remarks:

- w_{jj}'s can be looked upon as residuals obtained by regressing $\left(v_j\right)^{-\frac{1}{2}}s_j$ on z_j. This makes their computation for each iteration much easier.
- The most important role of this derivation is that it reduces the complexity of computation a lot, as we need not carry out the whole estimation afresh by fitting the regression model every time on the deleted dataset. Rather, it is sufficient to modify the residual terms e_j, w_{jj} and h_{jj} after each iteration.

3.4 Asymptotic distributions

We need to deduce the asymptotic distributions of these statistics in order to use their 95% critical points as reasonable cutoff values. In order to explicitly get the asymptotic distribution of $\hat{\beta} - \widehat{\beta^{(j)}}$ and consequently the Cook's Distance which is a mere quadratic form of the same, we state the following theorem.

3.4.1 Theorem 3.0.1 (Martín & Pardo, 2009)

Let $\hat{\beta}$and $\widehat{\beta^{(j)}}$ be the MLE of the parameters β based on the full set of observations and MLE based on the full set of observations minus the j^{th} observation. Then,

$$\sqrt{\mathrm{N}}\left(\hat{\beta} - \widehat{\beta^{(j)}}\right) \xrightarrow{\mathrm{L}} \mathrm{N}\left(0^{(\mathrm{k}+1)\times 1}, \Sigma^{(j)}\right) \tag{12}$$

where N *is the total number of observations in the data-set,* $\Sigma^{(j)} = \frac{v_j}{1-w_{jj}}(X'WX)^{-1}x_j x_j'(X'WX)^{-1}$, $w_{jj} = v_j^{\frac{1}{2}}x_j'(X'WX)^{-1}x_j v_j^{\frac{1}{2}}$ *and* v_j *is the* j^{th} *diagonal entry of* W *which is the limiting variance-covariance matrix. All these quantities are evaluated at* β_0*, the true value of* β.

Knowing this result, it is not difficult to establish the asymptotic distribution of Cook's Distance.

3.4.2 Theorem 3.0.2 (Proof in Appendix A)

Let $\hat{\beta}$ *and* $\widehat{\beta^{(j)}}$ *be the MLE of the parameters* β *based on the full set of observations and MLE based on the full set of observations minus the* j^{th} *observation. Then,*

$$D^{(j)}\left(\hat{\beta}\right) \xrightarrow{L} \frac{w_{jj}}{1-w_{jj}}\chi_1^2 \tag{13}$$

where w_{jj} *is evaluated at* β_0*, the true value of* β.

Thus, we can define the approximate cutoffs for our Cook's Distance statistic as the upper α point of χ_1^2 distribution.

Generally, we take $\alpha = 0.95$. For an observation j, if $D^{(j)}\left(\hat{\beta}\right)$ comes out to be significantly greater than $\chi^2_{1,0.95}\frac{w_{jj}}{1-w_{jj}}$, then we term the point j to be an influential point.

For $DFFITS$, we can derive a similar approximate cutoff by exploiting the relation between $DFFITS^2$ and Cook's Distance. (Nurunnabi & Rahmatullah Imon, 2010)

$$DFFITS_i^2 \approx pv_i^2 D_i \tag{14}$$

where Cook's Distance is given by $D_i, v_i = \pi_i(1-\pi_i)$ and p is the number of predictors.

Thus, the distribution of $|DFFITS|$ is approximately $k_i\chi_1$ where k_i is a specified constant. The cutoff values for $DFFITS$ would be inflated each point by a factor of k_i and can be compared easily.

3.5 Group deleted case

The diagnostics discussed till now are mainly proposed for the identification of a single influential observation and are ineffective when masking/swamping occur. Thus, we need measures which are free from these effects leading to the idea of group deleted observations. We assume that d observations among a set of n observations are deleted. Let us denote the set of observations remaining in the analysis by R and the set of deleted observations as D. Hence R contains $n-d$ observations. We state the generalized d-deleted measures $GDFFITS$ and Cook's Distance as follows:

$$GDFFITS_i = \begin{cases} \widehat{\pi_1^{(D)}} - \widehat{\pi_1^{(D+i)}} \text{ for i in R} \\ \widehat{\pi_i^{(D-i)}} - \widehat{\pi_i^{(D)}} \text{ for i in D} \end{cases}$$

where $\widehat{\pi_i^{(D)}}$ is the predicted probability of the i^{th} observation when the set D has been deleted.

$\widehat{\pi_i^{(D+i)}}$ is the predicted probability of the i^{th} observation when the i^{th} observation from R is incorporated into the deletion set D. That is, the deletion set now consists of $D \cup \{i\}$.

$\widehat{\pi_i^{(D-i)}}$ is the predicted probability of the i^{th} observation when the i^{th} observation which is already in the deletion set D is removed from D in order to avoid deduction twice. That is, now the deletion set D becomes $D \setminus \{i\}$.

Similarly, the generalized Cook's Distance is given by :

$$CD^{(D)} = \left(\widehat{\beta^{(D)}} - \hat{\beta}\right)'(X'VX)\left(\widehat{\beta^{(D)}} - \hat{\beta}\right) \tag{15}$$

where $\widehat{\beta^{(D)}}$ is the estimate of β when d observations have been deleted.

Although these expressions may seem compact, they are computationally quite intensive and hence we require expressions which are less tedious to compute.

We continue our quest after (6), and move on to derive the expression for $\widehat{\beta^{(D)}} - \hat{\beta}$ and then obtain a closed form for $GDFFITS$ and Cook's Distance.

From (6) we can write:

$$\hat{\beta} = \beta^* + (Z'Z)^{-1}\left(Z'V^{-\frac{1}{2}}s\right)$$

Now, we derive deletion statistics when the observations corresponding to an index set D with length m of $\{1,2,\dots,n\}$ are omitted. The deleted log likelihood becomes:

$$l^{(D)}\left(\hat{\beta}\right) = \sum_{i=1, i\notin D}^{n}\{y_i x_i'\beta - \ln\left(1+e^{x_i'\beta}\right)\}$$

Then

$$\frac{\partial l^{(D)}(\beta)}{\partial \beta} = \mathrm{X}'\mathrm{s} - \sum_{\mathrm{j}\notin \mathrm{D}} \mathrm{x_j s_j}$$

and

$$\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'} = \mathrm{Z'Z} - \sum_{\mathrm{j}\notin\mathrm{D}} \mathrm{z_j z_j'}$$

where $z_i = \sqrt{v_i}x_i$. First order Taylor approximation for the deletion of the index set D yields:

$$\widehat{\beta^{(D)}} = \beta_0^{(D)} + \left(Z'Z - \sum_{j\notin D} z_j z_j'\right)^{-1}\left(X's - \sum_{j\notin D} x_j s_j\right) \tag{16}$$

Taking $\beta_0^{(D)} = \beta^* - (X'X)^{-1}X_D'(I - H_D)^{-1}e_D$, where X_D being the the covariates of the set D clubbed together, y_D are their corresponding m responses, H_D is the $m \times m$ minor of $H = X(X'X)^{-1}X', e_D = y_D - X_D'\beta^*$.
Then it follows that:

$$\widehat{\beta^{(D)}} = \hat{\beta} - (Z'Z)^{-1}Z_D'(I - H_D^*)^{-1}e_D^* - (X'X)^{-1}X_D'(I - H_D)^{-1}e_D$$

e_D^* being $V_D^{-\frac{1}{2}}s_D - Z_D(Z'Z)^{-1}Z'V^{-\frac{1}{2}}s$ and $H_D^* = Z_D(Z'Z)^{-1}Z_D'$. (Jung, 2009)

3.6 Closed form of GDFFITS and Cook's Distance (Generalized)

The group deletion statistic GDFFITS is given by:

$$GDFFITS_i^{(D)} = v_i^{-\frac{1}{2}}z_i'(Z'Z)^{-1}Z_D'(I - H_D^*)^{-1}e_D^* + x_i'(X'X)^{-1}X_D'(I - H_D)^{-1}e_D \tag{17}$$

The explicit expression for Cook's Distance according to (Pregibon, 1981) can be obtained by noticing the change in likelihood. That is,

$$CD^{(D)} = 2\{l^{(D)}\left(\widehat{\beta^{(D)}}\right) - l\left(\hat{\beta}\right)\} \tag{18}$$

Where

$$l^{(D)}\left(\widehat{\beta^{(D)}}\right) = \sum_{i=1,i\notin D}^{n}\{y_i x_i'\widehat{\beta^{(D)}} - \ln\left(1 + e^{x_i'\hat{\beta}}\right)\}.$$

We stop our discussion on the theoretical aspects of these measures here. It is to be noted that the asymptotic distribution of $GDFFITS$ and Cook's Distance is quite far-fetched and is not covered in this article. We would refer to the empirical bound of $|GDFFITS|$ given as $3\sqrt{\frac{k}{n-d}}$, where k is the number of predictors and d is the number of deleted observations (Nurunnabi & Rahmatullah Imon, 2010). Although the expressions for this diagnostic are available for any arbitrary set of deleted cases, D, the choice of such a set is very important as the omission of this group determines the $GDFFITS$ diagnostics for the whole set.

It is quite intriguing if we consider the problem of deletion of a fraction of observations from the entire data set, that is if we delete kN observations where $k \in (0,1)$. In that case, what we can foresee is that the asymptotic confidence intervals for $DFFITS$ and Cook's Distance would change depending upon how large k is. This is basically due to the fact that on deletion of kN observations, the asymptotic normality in (12) will be affected by a constant as we no longer have the normalization factor to be $\sqrt{N}$ but rather $c\sqrt{N}$, where c is a constant which is a function of k. On changing this factor, the variance-covariance matrix $\Sigma^{(j)}$ will also change, that is all its entries would be scaled accordingly.

3.7 Generalization to the multinomial logistic model

Our statistics for binary logistic regression can be generalized to a multi-class logistic regression model, since we are interested in taking the difference in fitted values or estimates of β for two classes, so we use the philosophy that an observation either belongs to the i^{th} class or does not belong to the i^{th} class, which is analogous to binary logistic regression predicting either 1 if it is in the desired class, 0 otherwise. We first state the model of a multinomial logistic regression.

Model: For a $k-$class logistic regression problem, suppose our response Y_i can take values in the set $\{1,2,\dots,k\}$, where each index indicates a separate class.

We choose a pivot (base - class), say k, and we regress the log-odds, i.e. $\ln\frac{\pi_i}{\pi_k}$ against the set of regressors X_i. Here

$$\pi_j = P(Y = j), \text{ where } j = 1,2,\dots,k$$

So, we have:

$$\ln\frac{\pi_j}{\pi_k} = \beta_j'X_i, \text{ for j } \in \{1,2,..,k-1\}$$

So, as we can see we have different sets of weights β_j 's for each equation and can have $(p+1).(k-1)$ coefficients in total if we consider an intercept term for each regression equation, that is we have p features in all. Thus, we need to estimate $p.(k-1)$ parameters while fitting the model.

We have

$$\widehat{\pi_j} = \frac{\exp\left(\beta_j'X_i\right)}{1 + \sum_{j=1}^{k-1}\exp\left(\beta_j'X_i\right)}, \text{ for } j \in \{1,2,..k-1\}$$

Also, using $\sum_{i=1}^{k}\pi_i = 1$, we get,

$$\widehat{\pi_k} = \frac{1}{1 + \sum_{j=1}^{k-1}\exp\left(\beta_j'X_i\right)}$$

The $DFFITS$ for this model is defined similarly. It is given by the difference of the probability of any observation belonging to the j^{th} class and the predicted probability of that particular observation belonging to the j^{th} class when any observation (say i) has been deleted from the dataset. If this difference is large, it implies that the observation is actually mis-classified and does not actually belong to the j^{th} class. It is given by:

$$DFFITS_i = \hat{\pi}_i - \widehat{\pi_{(i)}} \tag{19}$$

where $\pi_{(i)}$ is the predicted probability for the i^{th} observation when it has been deleted from the data-set.
However, the expression for $DFBETAS$ is somewhat complicated to view.

$$DFBETAS_{ji} = \hat{\beta}_j - \widehat{\beta_{j(i)}} \tag{20}$$

which although might look similar to the multiple linear regression case, has a much higher order matrix $(n \times p(k-1))$. So, the number of columns has increased $k-1$ fold because of the $k-1$ classes. So, it is meaningless and tiresome to compare each and every value of this $DFBETAS$ matrix with an individual cutoff, rather $DFFITS$ being a $n \times 1$ vector promises greater interpretability. The form of Cook's Distance is analogous to that of a binary logistic regression model so, it is not worthwhile to mention it in this case.

Remark:

It should be remembered that these expressions are valid only under the assumption that there are no misclassifications in the base class. Had there been a violation of this assumption, this problem is converted to a classification problem rather than a problem of outlier detection and is beyond the reach of our discussion.

4. Simulation results

First, we present an example to show how we design our experiment. Initially, we simulated 50 observations. For the explanatory variable X, the first 25 observations are generated from $U(10, 100)$ and the last 25 observations are generated from $U(50, 500)$. To generate the original Y values, we set the first 25 values of Y at 0 and the last 25 values at 1. So, we create a general trend that Y values corresponding to bigger X values are more likely to have the value 1. To generate influential observations, we change the value of Y corresponding to the observations having large values of X to 0. This changed point should be an influential point because it appears against the pattern of the majority of the data. We see how our DFFITS measure works in this case. The data is tabulated in Table 1.

Table 1. Simulated dataset.

S NO.	X (Predictor)	Y (Response)	S NO.	X (Predictor)	Y (Response)
1	33.895	0	26	223.751	1
2	43.491	0	27	56.025	1
3	61.557	0	28	222.074	1
4	91.738	0	29	441.360	1
5	28.151	0	30	203.157	1
6	90.855	0	31	266.936	1
7	95.020	0	32	319.804	1
8	69.471	0	33	272.093	1
9	66.620	0	34	133.797	1
10	15.560	0	35	422.317	1
11	28.537	0	36	350.810	1
12	25.891	0	37	407.407	1
13	71.832	0	38	98.574	1
14	44.569	0	39	375.669	1
15	79.285	0	40	235.073	1
16	54.792	0	41	419.425	1
17	74.585	0	42	341.177	1
18	99.271	0	43	402.319	1
19	44.203	0	44	298.866	1
20	79.970	0	45	288.373	1
21	94.123	0	46	405.210	0
22	29.092	0	47	60.499	1
23	68.650	0	48	264.753	1
24	21.299	0	49	379.541	1
25	34.049	0	50	361.729	1

From Table 1, it is clear that the observations 27,38,46,47 are influential, as they are clearly going against the trend of X values and Y values. Now, we find the $DFFITS$ vector in order to validate our claim. It is indeed seen from Table 2 of $DFFITS$ that its values at these points are significantly large, especially at 47. So, our algorithm can successfully detect influential observations. $DFFITS$ is given in Table 2.

Table 2: $DFFITS$

S NO.	DFFITS	S NO.	DFFITS	S NO.	DFFITS
1	-1.453	18	-6.819	35	-1.989
2	-2.057	19	-2.105	36	-4.396
3	-3.427	20	-5.065	37	-2.380
4	-6.153	21	-6.368	38	22.060
5	-1.134	22	-1.184	39	-3.410
6	-6.073	23	-4.087	40	-6.126
7	-6.449	24	-0.794	41	-2.060
8	-4.110	25	-1.462	42	-4.809
9	-3.859	26	-5.102	43	-2.527
10	-0.542	27	22.036	44	-6.573
11	-1.154	28	-4.917	45	-6.900
12	-1.017	29	-1.569	46	817.327
13	-4.321	30	-2.207	47	22.643
14	-2.130	31	-7.211	48	-7.205
15	-5.002	32	-5.742	49	-3.270
16	-2.881	33	-7.193	50	-3.946
17	-4.570	34	14.806		

We present another interesting case this time with two predictors X_1 and X_2. We draw 25 observations out of which 20 of them come from $N(0,1)$ population, that is, $X_1, X_2 \sim N(0,1)$ for the first 20 observations. For the remaining 5 observations, we draw X_1, X_2 from a $N(2,1)$ population. We label $Y_i = 0$ for $i = 1, 2, \ldots, 20$ and $Y_i = 1$ for $i = 21, \ldots, 25$. The $DFFITS$ values for the 30 observations are presented in Table 3.

Table 3. DFFITS for $N(0,1)$ and $N(2,1)$.

S NO.	DFFITS	S NO.	DFFITS	S NO.	DFFITS
1	0.0101	10	-0.0209	19	0.2705
2	0.1098	11	-0.4985	20	0.0591
3	-0.0188	12	0.4729	21	16.1173
4	0.0738	13	0.0786	22	6.2627
5	0.0986	14	0.1012	23	20.1691
6	0.6605	15	0.0174	24	-0.0736
7	0.0040	16	0.1085	25	12.1179
8	-0.0117	17	-0.0134		
9	0.0466	18	0.2845		

As we had expected, the $DFFITS$ values for the observations $21 - 25$ stand out from the others in terms of magnitude. The only exception being observation 24. This is because the covariates X_1, X_2 for the observations $21, 22, 23, 25$ are in $(0,1)$, which closely resemble draws from $N(0,1)$, however they are labeled as 1. So, the algorithm detects these observations as outliers. On the other hand, the observation 24 having covariate values 3.0865 and 2.8355 are less likely values for a $N(0,1)$ distribution which has been rightly classified as 1. So, the $DFFITS$ value of this observation is small.

4.1 DFFITS for Multiclass logistic model

We also implement the $DFFITS$ measure for a multi-class logistic regression problem. For this purpose, we use the famous Iris data set. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant *setosa*, *versicolor* and *virginica*. There are four covariates namely *Sepal Length*, *Sepal Width*, *Petal Length* and *Petal Width*. For ease of notation, we consider *setosa* as Class 0 , *versicolor* as Class 1 and *virginica* as Class 2 with Class 0 as the base class. The first 50 observations have been classified as Class 0, the observations $51 - 100$ have been classified as Class 1 and the remaining from $100 - 150$ as Class 2. We generate three misclassifications by intentionally changing the observation 46 originally labeled as class 0, to Class 1, the observation 57 originally labeled as Class 2, to Class 3 and the observation 107 originally labelled as Class 3, to Class 1. Now we calculate the $DFFITS$ values for all the observations. Out of them the $DFFITS$ values for these three observations are given in Table 4.

Table 4. $DFFITS$ for Multiclass Logistic Model.

S NO.	DFFITS
46	0.8311
57	0.7174
107	0.5520

Thus, we see that all the three observations produce a high value of $DFFITS$ compared to the others having $DFFITS$ value of the order of 10^{-2} or less. Thus, our measure $DFFITS$ can identify mis-classifications at the multi-class level as the $DFFITS$ values get significantly inflated for these observations.

4.2 GDFFITS and Cook's Distance simulations

We now move on to check how our measures Cook's Distance and Generalized $DFFITS$ work on a standard data set- *Kyphosis* data (Nurunnabi & Rahmatullah Imon, 2010). But before we outline a general algorithm of finding $GDFFITS$, or more specifically how to choose the deletion set D and the remaining observations R.

Step 1: In the first step, we try to find out all suspect influential cases. This can be done either by graphical displays like index plot or character plot of explanatory and response variables. But these plots are not always helpful for higher dimensions of regressors and depend heavily on the experimenter's own investigation. So, it is sensible to first run a single deletion $DFFITS$ or Cook's Distance algorithm on the dataset and look for unusual values of these statistics. Hence, we list all the suspect points to form the set D. Naturally the set R is formed by the remaining observations.

Step 2: When we delete a group of observations, it is possible that some observations may be wrongly detected as suspect influential cases due to the swamping effects. So, it is to be checked for all possible subsets of D , whether our algorithm can detect all the influential points correctly or not. As a rule of thumb, the cardinality of D is usually less than $10\ \%$ of the size of the entire data set.

4.3 Modified Kyphosis Data

This data set contains information about 81 children who have had corrective spinal surgery. The response variable tells us whether a post-operative deformity (Kyphosis) is 'present' or 'absent' in the data. It is thought that Kyphosis depends on two variables, the number of vertebrae involved in the operation (number) and the beginning of the range of vertebrae involved in the operation (start). On first inspection of the original data, it seems to us that one observation (Case 43) might have a large influence in fitting the model. We now deliberately change eight observations 10,11,23,40,46,49,63,77 to make them influential and this modified data set is presented in Table 5. On implementing $DFFITS$ and Cook's Distance algorithms (single deletion case) on the data, we see that it cannot identify all the influential observations correctly, rather it has mis-identified some points as influential points. This indicates the existence of masking and swamping effects in the data- set. Thus, we need to proceed with our $GDFFITS$ measure. In order to create our deletion set D, we choose 5 suspect cases 10,11,23,43,77. We compute the $GDFFITS$ values for the entire data set based on deletion of this set D. The results have been tabulated in Table 6. From the results it is clear that some observations like 46 and 63 were *masked* by the presence of other outliers, which are now identified. This measure now correctly identifies all the influential observations. We also checked for other values of $d = |\mathrm{D}|$, and found that for $d < 5$, all outliers still could not be detected, and on taking $d > 5$, it was found that there were many harmless observations identified as influential. This is due to the fact that as we go on increasing d, the cutoff points for Cook's Distance varies a lot from the original cutoffs and hence it will be erroneous if we use the original single deletion cutoffs for these generalized values.

Table 5. Modified Kyphosis Data.

Index	Kyphosis	Number	Start	Index	Kyphosis	Number	Start
1	0	3	5	42	0	3	13
2	0	3	14	43	0	9	3
3	1	4	5	44	0	4	1
4	0	5	1	45	0	3	16
5	0	4	15	46	1	3	10
6	0	2	16	47	0	4	15
7	0	2	17	48	0	5	13
8	0	3	16	49	1	3	3
9	0	2	16	50	0	2	14
10	1	6	12	51	0	5	10
11	1	5	14	52	0	2	17
12	0	3	16	53	1	10	6
13	0	5	2	54	0	2	17

Table 5. Continued.

Index	Kyphosis	Number	Start	Index	Kyphosis	Number	Start
14	0	4	12	55	0	4	15
15	0	3	18	56	0	5	15
16	0	3	16	57	0	3	13
17	0	6	15	58	1	5	8
18	0	5	13	59	0	7	9
19	0	5	16	60	0	3	13
20	0	4	9	61	1	4	1
21	0	2	16	62	1	7	8
22	1	6	5	63	0	4	1
23	1	3	12	64	0	3	16
24	0	2	3	65	0	4	16
25	1	7	2	66	0	4	10
26	0	5	13	67	0	2	17
27	0	3	6	68	0	4	13
28	0	3	14	69	0	4	11
29	0	3	16	70	0	5	16
30	0	2	16	71	0	5	14
31	0	3	16	72	0	4	12
32	0	2	11	73	0	4	16
33	0	5	13	74	0	4	10
34	0	3	16	75	0	3	15
35	0	5	11	76	0	4	15
36	0	3	16	77	1	3	13
37	0	3	9	78	0	7	13
38	1	5	6	79	0	2	13
39	0	6	9	80	1	7	6
40	1	5	12	81	0	4	13
41	1	5	1				

Table 6. GDFFITS, DFFITS, Cook's D.

Index	Cook's D	DFFITS	GDFFITS	Index	Cook's D	DFFITS	GDFFITS
1	0.2935	-0.1844	-0.2061	42	0.0024	0.0003	-0.0009
2	0.0013	-0.0054	-0.0060	43	5.1944	-1.4948	-1.756
3	0.5088	-0.1535	-0.2445	44	0.1358	0.1301	-0.5211
4	0.5923	0.0351	0.0652	45	0.0006	-0.0132	-0.0030
5	0.0032	-0.0269	-0.0102	46	0.7778	-0.1692	1.2250
6	0.0010	-0.0065	-0.0002	47	0.0032	-0.0269	-0.1100
7	0.0006	-0.0084	-0.0001	48	0.0234	-0.0590	-0.0035
8	0.0006	-0.0132	-0.0003	49	0.5206	-0.2308	0.7498
9	0.0010	-0.0065	-0.5203	50	0.0033	-0.0004	-0.0003
10	1.2887	0.4538	0.9856	51	0.0121	-0.0366	-0.0085
11	1.0008	0.4093	1.2298	52	0.0006	-0.0084	-0.0001
12	0.0008	-0.0132	-0.0085	53	1.5894	0.4899	0.3491
13	0.3617	0.0402	0.5605	54	0.0006	-0.0008	-0.0012
14	0.0009	-0.0088	-0.0025	55	0.0032	-0.0269	-0.1100
15	0.0006	-0.0170	-0.0032	56	0.0232	-0.0653	-0.0205
16	0.0006	-0.0132	-0.0004	57	0.0024	0.0003	-0.0090
17	0.1049	-0.1434	-0.0475	58	0.0085	0.0276	0.5680
18	0.0234	-0.0590	-0.0041	59	0.6876	-0.3644	-0.5030
19	0.0214	-0.0662	-.0015	60	0.0024	0.0003	-0.0009
20	0.0087	0.0250	-0.0050	61	0.7567	-0.1091	0.3620
21	0.0010	-0.0065	-0.0020	62	0.6423	0.3572	0.4195
22	0.0432	0.0962	0.5785	63	0.3587	0.1301	0.8524
23	1.4432	0.445	1.556	64	0.0006	-0.0132	-0.0036
24	0.8535	0.1123	-0.1455	65	0.0035	-0.02999	-0.0085
25	0.0405	0.1317	0.0081	66	0.0028	0.0119	-0.0045
26	0.0234	-0.0590	-0.0035	67	0.0006	-0.0064	-0.0010

Table 6. GDFFITS, DFFITS, Cook's D.

Index	Cook's D	DFFITS	GDFFITS	Index	Cook's D	DFFITS	GDFFITS
27	0.1799	0.0847	-0.0076	68	0.0016	-0.0165	-0.0175
28	0.0013	-0.0054	-0.0040	69	0.0008	0.0006	-0.0300
29	0.0006	-0.0132	-0.0003	70	0.02144	-0.0662	-0.0152
30	0.0010	-0.0065	0.0018	71	0.0239	-0.0630	-0.0027
31	0.0006	-0.0132	-0.0033	72	0.0009	-0.0088	-0.0235
32	0.0179	0.0164	-0.0009	73	0.0035	-0.0299	-0.0074
33	0.0234	-0.590	-0.0357	74	0.0028	0.0119	-0.0395
34	0.0006	-0.0132	-0.0030	75	0.0008	-0.0009	-0.0053
35	0.0174	-0.0458	-0.0578	76	0.0032	-0.0269	-0.0012
36	0.0006	-0.0132	-0.0053	77	1.329	1.2937	1.862
37	0.0350	0.0391	0.3582	78	0.5091	-0.3254	-0.2130
38	0.0304	-0.0187	-0.1998	79	0.0059	0.0040	-0.0045
39	0.1366	-0.1416	-0.3280	80	0.2768	0.2516	0.2658
40	0.4188	0.2373	0.8952	81	0.0012	-0.0165	-0.0147
41	0.1922	-0.0240	0.1590				

Thus, we see that although Cook's Distance and $DFFITS$ could not identify the observations 40, 46 and 63 as outliers, $GDFFITS$ values for these observations are quite large compared to the cutoff 0.6, and hence are detected easily. The observation 53 was declared an outlier by $DFFITS$ and Cook's Distance, however, the value of GDFFITS shows that it is not. So, we have an instance of *swamping*.

5. Scope of Improvement and Conclusions

Although we discussed largely about how to obtain the form of $DFFITS$ and Cook's Distance for a logistic regression model throughout this article, there are many issues which we could not address due to the paucity of time. Some of these are as follows:

- We have obtained the asymptotic cutoffs of Cook's Distance and $DFFITS$ which varies with each deleted observation j. Though it is intuitively clear that the cutoffs should vary with the influence of the j^{th} point, however many authors including (Halfon, et al., 1977) have argued against this notion and has sought cutoffs which would provide a uniform bound for the Cook's Distance of all the observations using the theory of **Confidence Interval Displacement** (Martín & Pardo, 2009).
- In the extension of the binary logistic outlier detection techniques to the multinomial case, we assumed that there are no misclassifications in the base class. On dropping that assumption, the problem could have been more interesting and challenging.
- We are yet to find the asymptotic cutoffs for GDFFITS and Generalized d deleted Cook's Distance.
- We are yet to theoretically judge what should be the optimal value of d such that even after deleting those many observations, the values of the generalized measures are within the tolerance limits of the cutoffs in the single deletion case. For instance, in our example the cutoff of 0.6 for the single deletion case was valid till $d = 5$.
- We also have not investigated other techniques or procedures to detect masking or swamping outliers. Fitting a single deletion algorithm over a data set of large size just to get an idea of the suspected set of points will be quite laborious.

In this article, we have investigated how the various deletion diagnostics defined in the case of a multiple linear regression model can be generalized to the logistic regression model. The $DFFITS$ and Cook's Distance with single deletion appears to be quite useful for a dataset having very few outliers (generally one or two). However, in presence of multiple outliers, these measures under perform and thus they should be replaced by group deletion statistics $GDFFITS$ and Generalized Cook's Distance. Their performance on the Kyphosis data set with more than one outlier seems to be good enough once the optimal deletion set is chosen.

Acknowledgments:
I owe my debt of gratitude to my Master's thesis advisors at the **Indian Institute of Technology, Kanpur**: Dr. Debraj Das and Dr. Minerva Mukhopadhyay for their insightful suggestions and advice. They played an instrumental role in completing this article.

Data Availability Statement:
The modified Kyphosis data-set has adopted from Section 4.2 of the paper: https://www.researchgate.net/publication/258141262_Identification_of_multiple_influential_observations_in_logistic_regression.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Proof of Theorem 2.0.2:

Following the lines of **Matin & Pardo (2009)**, We shall use the following result: Suppose $(W_1, \dots, W_q)'$ be a $q \times 1$ random vector and is distributed as $N(0^{q\times 1}, \Sigma_W)$ and M is any real symmetric matrix of order q. Let $r = \text{rank}(\Sigma_W M \Sigma_W) \geq 1$ and let $\lambda_1, \dots, \lambda_r$ be non-zero eigenvalues of $M\Sigma_W$ $(r \leq q)$. Then $(W_1, \dots, W_q)M(W_1, \dots, W_q)' \sim \sum_{i=1}^{r} \lambda_i Z_i^2$ where $\{Z_i\}_{i=1}^{r}$ are independent random variables so that $Z_i \sim N(0,1)$. Based on Theorem 2.0.1, we need to calculate:

$$rank\left(\Sigma^{(j)}V^{-1}\Sigma^{(j)}\right) = rank\left((X'WX)^{-1}x_j x_j'(X'WX)^{-1}x_j x_j'(X'WX)^{-1}\right)$$

Let $R = (X'WX)^{-\frac{1}{2}}x_j$. Since rank($RR'RR'$)=rank($RR'$)=rank($R$) and R is a $(k+1)\times 1$ matrix, the rank of $\Sigma^{(j)}\Sigma^{-1}\Sigma^{(j)}$ is 1. On the other hand, since the nonzero eigenvalues of FG are equal to the non zero eigenvalues of GF, when the dimensions of F and G' matrices are equal, we have that the nonzero eigenvalue of

$$\Sigma^{(j)}V^{-1} = \frac{v_j}{1-w_{jj}}(X'WX)^{-1}x_j x_j' = FG$$

where $F = \frac{v_j}{1-w_{jj}}(X'WX)^{-1}x_j$ and $G = x_j'$, coincides with the eigen value of

$$GF = \frac{1}{1-w_{jj}} v_j x_j'(X'WX)^{-1}x_j = \frac{1}{1-w_{jj}} w_{jj}$$

Therefore, $\frac{w_{jj}}{1-w_{jj}}$ is the nonzero eigenvalue of $\Sigma^{(j)}V^{-1}$ and we obtain the desired result.

References

A., D., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data And Sources Of Collinearity.* New York: John Wiley and sons.

Acuna, E., & Rodriguez, C. (2004). On Detection Of Outliers And Their Effect In Supervised Classification. *IPSI 2004.* Venice.

Agresti, A. (2012). *Categorical Data Analysis.* Wiley Series in Probability and Statistics.

Bollen, K. A., & Jackman, R. W. (1985). Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. *Sociological Methods & Research*, *13*(4), 510–542. doi:https://doi.org/10.1177/0049124185013004004

Christensen, R. (1997). *Log-Linear Models and Logistic Regression.* Springer New York, NY.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics, 19*(1), 15–18. doi:https://doi.org/10.2307/1268249

Cook, R. D., & Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Halfon, E., Obenchain, R. L., Cook, R. D., Kabe, D. G., S. R., Shukla, G. K., & Guenther, W. C. (1977). Letters to the Editor. *Technometrics*, 348–351.

Jung, K. (2009). Multiple Deletions in Logistic Regression Models. *Communications for Statistical Applications and Methods, 16*, 309-315. doi:10.5351/CKSS.2009.16.2.309

Martín, N., & Pardo, L. (2009). On the asymptotic distribution of Cook's distance in logistic regression models. *Journal of Applied Statistics, 36*(10), 1119-1146. doi:10.1080/02664760802562498

Nurunnabi, A. A., & Rahmatullah Imon, A. H. (2010). Identification of multiple influential observations in logistic regression. *Journal of Applied Statistics, 37*(10), 1605-1624. doi:10.1080/02664760903104307

Pregibon, D. (1980). Goodness of Link Tests for Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 29*(1), 15-24. doi:https://doi.org/10.2307/2346405

Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics, 9*(4), 705-724. Retrieved from http://www.jstor.org/stable/2240841

Seber, G., & Lee, A. (2003). *Linear Regression Analysis.* Wiley Series in Probability and Statistics.

Sen Roy, S., & Guria, S. (2008). Diagnostics in logistic regression models. *Journal of the Korean Statistical Society, 37*(2), 89-94. doi:https://doi.org/10.1016/j.jkss.2007.03.001

Williams, D. (1987). Generalised linear model diagnostics using the deviance and single case deletions. *Journal of Applied Statistics, 36*(2), 181-191.