Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (1): 26-35, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.26.35

Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study

[1]S.K. Sarkar, [2]Habshah Midi and [1]Sohel Rana
[1]Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: Logistic regression is one of the most frequently used statistical methods as a standard method of data analysis in many fields over the last decade. However, analysis of residuals and identification of influential outliers are not studied so frequently to check the adequacy of the fitted logistic regression model. Detection of outliers and influential cases and corresponding treatment is very crucial task of any modeling exercise. A failure to detect influential cases can have severe distortion on the validity of the inferences drawn from such modeling. The aim of this study is to evaluate different measures of standardized residuals and diagnostic statistics by graphical methods to identify potential outliers. Evaluation of diagnostic statistics and their graphical display detected 25 cases as outliers but they did not play notable effect on parameter estimates and summary measures of fits. It is recommended to use residual analysis and note outlying cases that can frequently lead to valuable insights for strengthening the model.

Key words: Logit, covariate pattern, residual, outlier, leverage, bubble plot

INTRODUCTION

Often the outcome variable in the social data is in general not a continuous value instead a binary one. In such a case, binary logistic regression is a useful way of describing the relationship between one or more independent variables and a binary outcome variable, expressed as a probability scale that has only two possible values. Indeed, a generalized linear model is used for binary logistic regression. The most attractive feature of a logistic regression model is neither assumes the linearity in the relationship between the covariates and the outcome variable, nor does it require normally distributed variables. It also does not assume homoscedasticity and in general has less stringent requirements than linear regression models. Thus logistic regression is used in a wide range of applications leading to binary dependent data analysis (Hilbe, 2009; Agresti, 2002).

The vast majority of the work related to the logistic regression appears in the experimental epidemiological research but during the last decade it is evident that the technique is frequently used in observational studies. But analysis of residuals and the identification of outliers and influential cases are not studied so frequently to check the adequacy of the fitted model. Data obtained from observational studies sometimes can be considered as bad from the point of view of outlying responses. The traditional method of fitting logistic regression models with maximum likelihood, has good optimality properties in ideals settings, but is extremely sensitive to bad data obtained from observational studies (Pregibon, 1981). Frequently in logistic regression analysis applications, the real data set contains some cases that are outlier; that is the observations for these cases are well separated from the remainder of the data. These outlying cases may involve large residuals and often have dramatic effects on the fitted maximum likelihood linear predictor. It is therefore, important to study the outlying cases carefully and decide whether they should be retained or eliminated and if retained, whether their influence should be reduced in the fitting process and/ or the logistic regression model should be revised (Menard, 2002; Hosmer and Lemeshow, 2000).

For logistic regression with one or two predictor variables, it is relatively simple to identify outlying cases with respect to their X or Y values by means of scatter plots of residuals and to study whether they are influential in affecting the fitted linear predictor. When more than two predictor variables are included in the logistic regression model, however, the identification of outlying cases by simple graphical methods becomes

Corresponding Author: S.K. Sarkar, Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

difficult. In such a case, traditional standardized residual plots can highlight little regarding outliers and some derived statistics and their plots from basic building blocks with lowess smooth and bubble plots are potential to detect outliers and influential cases (Kutner *et al*., 2005; Hosmer and Lemeshow, 2000).

There are three ways that an observation can be considered as unusual, namely outlier, influence and leverage. In logistic regression, a set of observations whose values deviate from the expected range and produce extremely large residuals and may indicate a sample peculiarity is called outliers. These outliers can unduly influence the results of the analysis and lead to incorrect inferences. An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outliers. An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. In fact, the leverage indicates the geometric extremeness of an observation in the multi-dimensional covariate space. These leverage points can have an unusually large effect on the estimate of logistic regression coefficients (Cook, 1998).

Christensen (1997) suggested that if the residuals in binary logistic regression have been standardized in some fashion, then one would expect most of them to have values within ±2. Standardized residuals outside of this range are potential outliers. Thus studentized residuals less than -2 and greater than +2 definitely deserve closer inspection. In that situation, the lack of fit can be attributed to outliers and the large residuals will be easy to find in the plot. But analysts may attempt to find group of points that are not well fit by the model rather than concentrating on individual points. Techniques for judging the influence of a point on a particular aspect of the fit such as those developed by Pregibon (1981) seem more justified than outlier detection (Jennings, 1986).

Detection of outliers and influential cases and corresponding treatment is very crucial task of any modeling exercise. A failure to detect outliers and hence influential cases can have severe distortion on the validity of the inferences drawn from such modeling exercise. It would be reasonable to use diagnostics to check if the model can be improved in case of Correct Classification Rate (CCR) is smaller than 100. The main focus in this study is to detect outliers and influential cases that have a substantial impact on the fitted logistic regression model through appropriate graphical method including smoothing technique.

MATERIALS AND METHODS

The Bangladesh Demographic and Health Survey is part of the worldwide Demographic and Health Surveys program, which is designed to collect data on fertility, family planning, maternal and child health. The BDHS is a source of population and health data for policymakers and the research community. BDHS-2004 is the fourth survey conducted in Bangladesh and preparations for the survey started in mid-2003 and field work was carried out between January and May 2004. We have been using the women's data file. A total of 11,440 eligible women were furnished their responses. But in this analysis there are only 2,212 eligible women those are able to bear and desire more children are considered. The women under sterilization, declared in fecund, divorced, widowed, having more than and less than two living children are not involved in the analysis. Those women who have two living children and able to bear and desire more children are only considered here during the period of global two children campaign.

The variable age of the respondent, fertility preference, place of residence, highest year of education, working status and expected number of children are considered in the analysis. The variable fertility preference involving responses corresponding to the question, would you like to have (a/another) child? The responses are coded 0 for no more and 1 for have another is considered as desire for children which is the binary response variable (Y) in the analysis. The age of the respondent (X_1), place of residence (X_2) is coded 0 for urban and 1 for rural, highest year of education (X_3), working status of respondent (X_4) is coded 0 for not working and 1 for working and expected number of children (X_5) is coded 0 for two and 1 for more than two are considered as covariates in the logistic regression model. Several standardized residual plots, lowess smooth and diagnostic plots are used to detect influential outliers.

FORMULATION OF THE BINARY RESPONSE MODEL

The binary logistic regression model computes the probability of the selected response as a function of the values of the explanatory variables. A major problem with the linear probability model is that probabilities are bounded by 0 and 1, but linear functions are inherently unbounded. The solution is to transform the probability so that it is no longer bounded. Transforming the probability to odds removes the upper bound and natural logarithm of odds also removes the lower bound. Thus,

setting the result equal to a linear function of the explanatory variables yields logit or binary response model (Allison, 1999).

Suppose in a multiple logistic regression case, a collection of k explanatory variables be denoted by the vector X' = (X_1, X_2,... X_k). Let the conditional probability that the outcome is present be denoted by P (Y = 1|X) = $\theta(X)$. It is evident that Sigmoidal-shape curve configuration has been found to be appropriate in many applications for which the outcome variable is binary and the corresponding model having more than one explanatory variable can be written as:

$$Y_i = \theta_i(X) + \varepsilon_i;\ i = 1, 2... n \quad (1)$$

Where:

$$\theta_i(X) = \frac{\exp(Z_i)}{1+\exp(Z_i)} \quad (2)$$

with $Z_i = \beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_kX_{ki} = X\beta$. Here Y is n×1 vector of response having y_i=0 if the ith case does not possess the characteristic and y_i = 1 if the case does possess the characteristic under study, X is an n × (k+1) design matrix of explanatory variables, β is a (k+1) ×1 vector of parameters, ϵ is also an n×1 vector of unobserved random errors. The quantity θ_i is the probability for the ith covariate satisfying the important requirement $0 \le \theta_i \le 1$. Then the log-odds of having Y = 1 for given X is modeled as a linear function of the explanatory variables as:

$$E(Y|X) = \hat{\theta}' = \ln\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \cdots + \beta_kX_{ki} \quad (3)$$

The function:

$$\theta_i = \frac{\exp(X\beta)}{1+\exp(X\beta)}$$

is known as logistic function. The most commonly used method of estimating the parameters of a logistic regression model is the method of Maximum Likelihood (ML) instead of Ordinary Least Square (OLS) method. Mainly for this reason the ML method based on Newton-Raphson iteratively reweighted least square algorithm becomes more popular with the researchers (Ryan, 1997). The sample likelihood function is, in general defined as the joint probability function of the random variables whose realizations constitute the sample. Specifically, for a sample of size n whose observations are (y_1, y_2, ... y_n), the corresponding random variables are (Y_1, Y_2, ... Y_n). Since the Y_i is a Bernoulli random variable, the probability mass function of Y_i is

$$f_i(Y_i) = \theta_i^{Y_i}(1-\theta_i)^{1-Y_i};\ Y_i = 0 \text{ or } 1 \text{ and } i = 1, 2... n \quad (4)$$

Since Y's are assumed to be independent, the log-likelihood function L (β) is defined as:

$$L(\beta) = \sum_{i=1}^{n} Y_i \ln\left(\frac{\theta_i}{1-\theta_i}\right) + \sum_{i=1}^{n} \ln(1-\theta_i) \quad (5)$$

For convenience in multiple logistic regression models, the likelihood equations can be written in matrix notation as

$$\frac{\partial L(\beta)}{\partial \beta} = X'(Y-\theta) \quad (6)$$

Where $X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n\times(k+1)}$, $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times1}$ and $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}_{n\times1}$

Now, theoretically putting:

$$\frac{\partial L(\beta)}{\partial \beta} = 0$$

produces $\hat{Y} = \hat{\theta}$, satisfying $X'(Y-\hat{Y}) = 0$. In fact, the maximum likelihood estimates of β in the multiple binary logistic regression models are those values of β that maximize the log-likelihood function given in Eq. 5. No closed form solution exists for the values of $\hat{\beta}$ that maximize the log-likelihood function. Computer-intensive numerical search procedures are therefore required to find the maximum likelihood estimates $\hat{\beta}$ and hence $\hat{\theta}$, because the multiple logistic regression model computes the probability of the selected response as a function of the values of the predictor variables. There are several widely used numerical search procedures, one of these employs iteratively reweighted least squares algorithm. In this study, we shall rely on standard statistical software programs specifically designed for logistic regression to obtain the maximum likelihood estimates of parameters.

GOODNESS-OF-FIT OF THE MODEL

In order to check the goodness-of-fit of an estimated multiple logistic regression model one should assume that the model contains those variables that should be in the model and have been entered in the correct functional form. The goodness-of-fit measures how effectively the

model describes the response variable. The distribution of the goodness-of-fit statistics is obtained by letting the sample size n become large. If the number of covariate patterns increases with n then size of each covariate pattern tends to be small. Generally, the term covariate pattern is used to describe a single set of values for the covariates in the model. Distributional results obtained under the condition that only n become large are said to be based on n-asymptotic. The case most frequently encountered in practice that the model contains one or more continuous covariates. In such a situation the number of covariate patterns is approximately equal to the sample size and the current study contains two continuous covariates and the number of covariate patterns may not be an issue when the fit of the model is assessed. To assess the goodness-of-fit of the model, researcher should have some specific idea about what it means to say that a model fits. Suppose we denote the observed sample values of the response variable in vector form as Y where, $Y' = (y_1, y_2, \ldots y_n)$ and the corresponding predicted or fitted values by the model as $\hat{Y}' = (\hat{y}_1, \hat{y}_2, \cdots \hat{y}_n)$. We may conclude that the model fits if summary measures of the distance between Y and $\hat{Y}$ are small and the contribution of each pair $(y_i, \hat{y}_i)$, i = 1, 2, …, n to the summary measures is unsystematic and is small relative to the error structure of the model. Thus, a complete assessment of the fitted model involves both the calculation of summary measures of the distance between Y and $\hat{Y}$ and a thorough examination of the individual components of these measures. When model building stage has been completed, a series of logical steps should be used to assess the fit of the model. The components of proposed approach are: (1) computation and evaluation of overall summary measures of fit, (2) examination of the individual components of the summary statistics with appropriate graphics and (3) examination of other measures of the distance between the components of Y and $\hat{Y}$ (Hosmer and Lemeshow, 2000). The summary measures of goodness-of-fit, as they are routinely provided as program output with any fitted model and give an overall indication of the fit of the model. The different summary measures like likelihood ratio test, (Hosmer and Lemeshow, 1980) goodness-of-fit test, (Osius and Rojek, 1992) normal approximation test, (Stukel, 1988) test and other supplementary statistics indicate that the model seems to fit quite well. It is also evident that the individual predictors in the fitted model have significant contribution to predict the response variable through likelihood ratio test as well as Wald test (Sarkar and Midi, 2010). The elaboration of these measures is beyond the scope of the study. Before concluding that the model fits, it is crucial that other measures be examined to see if fit is supported over the entire set of covariate patterns. This is accomplished through a series of specialized measures falling under the general heading of residual analysis and regression diagnostics (Cook and Weisberg, 1982).

RESIDUAL ANALYSIS AND RESIDUAL PLOTS

Residual analysis for logistic regression is more difficult than the linear regression models because the responses take on only the values 0 and 1. Thus the ith ordinary residual will assume one of the two values as:

$$\hat{\varepsilon}_i = \begin{cases} 1-\hat{\theta}_i & \text{if } Y_i = 1 \\ -\theta_i & \text{if } Y_i = 0 \end{cases} \quad (7)$$

The ordinary residuals will not be normally distributed and, indeed their distribution under the assumption that the fitted model is correct is unknown. Plots of ordinary residuals against fitted values will generally be uninformative. In linear regression a key assumption is that the error variance does not depend on the conditional mean E (Y|X). However, in logistic regression, there are binomial errors and, as a result, the error variance is a function of the conditional mean as V (Y|X) = θ (1-θ). Hence, the ordinary residual can be made more comparable by dividing them by the estimated standard error of Y_i which is known as Pearson residual denoted by pr_i and defined as:

$$pr_i = \frac{\hat{\varepsilon}}{\sqrt{\hat{\theta}_i\left(1-\hat{\theta}_i\right)}} = \frac{\left(Y_i - \hat{\theta}_i\right)}{\sqrt{\hat{\theta}_i\left(1-\hat{\theta}_i\right)}} \quad (8)$$

The Pearson residuals are directly related to the Pearson chi-square goodness-of-fit statistic. The square of Pearson residual measures the contribution of each binary response to the Pearson chi-square test statistic but the test statistic does not follow an approximate chi-square distribution for binary data without replicates. The Pearson residuals do not have unit variance since no allowance has been made for the inherent variation in the fitted value $\hat{\theta}_i$. A better procedure is to further standardize the ordinary residuals by their estimated standard deviation that is called studentized Pearson residuals. The standard deviation is approximated by:

$$\sqrt{\hat{\theta}_i\left(1-\hat{\theta}_i\right)\left(1-h_{ii}\right)}$$

where:

$$H = \hat{W}^{1/2} X \left(X'\hat{W}X\right)^{-1} X'\hat{W}^{1/2}$$

h_{ii} is the ith diagonal element of the n×n estimated hat matrix H, whereby in logistic regression it is called hat diagonal or Pregibon leverage and measures the leverage of an observation. More clearly leverage is a measure of the importance of an observation to the fit of the model. Here, $\hat{W}$ is the n×n diagonal matrix with elements $\hat{\theta}_i(1-\hat{\theta}_i)$, X is the n× (k+1) design matrix defined earlier. The hat matrix for logistic regression satisfies approximately the expression $\hat{\theta}' = HY$ where, $\hat{\theta}'$ is the n×1 vector of linear predictors. Then studentized Pearson residuals spr_i are defined as:

$$spr_i = \frac{(Y_i - \hat{\theta}_i)}{\sqrt{\hat{\theta}_i(1-\hat{\theta}_i)(1-h_{ii})}} = \frac{pr_i}{\sqrt{1-h_{ii}}} \quad (9)$$

Studentized Pearson residuals are primarily helpful in identifying influential observations and those build in information about the influence of a case, whereas Pearson residuals do not. More influential cases with high leverages result in high studentized Pearson residuals. Studentized Pearson residuals approximately follow the standard normal distribution for large (n≥30) sample and it can be used as an approximate chi-square distribution.

Deviance residual is another type of residual. It measures the disagreement between any component of the log likelihood of the fitted model and the corresponding component of the log likelihood that would result if each point were fitted exactly. Since, the logistic regression uses the maximum likelihood principle, the goal in logistic regression is to minimize the sum of the deviance residuals. Deviance residuals can also be useful for identifying potential outliers or misspecified cases in the model. The deviance residual for the ith case is defined as the signed square root of the contribution of that case to the sum for the model deviance as:

$$dr_i = sign(Y_i - \hat{\theta}_i)\left\{-2\left[Y_i \ln(\hat{\theta}_i) + (1-Y_i)\ln(1-\hat{\theta}_i)\right]\right\}^{1/2} \quad (10)$$

McCullagh and Nelder (1989) expressed a preference for the deviance residuals because they are closer to being normally distributed than are the Pearson residuals. Approximate normality is certainly a desirable property of residuals, but it is also desirable to use some type of residual that will detect influential cases for necessary modifications to a logistic regression model so as to improve CCR. Like Pearson residual the square of each deviance residual measures the contribution of each binary response to the deviance good ness-of-fit statistic. Studentized Pearson residuals, deviance residuals and Pregibon leverage are considered to be the three basic building blocks for logistic regression diagnostics in detection of influential outliers and shown in Table 1.

Table 1: Binary logistic regression residuals and hat matrix diagonal elements for BDHS-2004 data

	Case						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	Y_i	$\hat{\theta}_i$	$\hat{\varepsilon}_i$	pr_i	spr_i	dr_i	h_{ii}
1	0	0.2579	-0.2579	-0.5896	-0.5903	-0.7724	0.0025
2	0	0.2889	-0.2889	-0.6374	-0.6378	-0.8258	0.0011
3	1	0.7418	0.2582	0.5900	0.5907	0.7729	0.0024
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2210	0	0.1918	-0.1918	-0.4872	-0.4879	-0.6527	0.0029
2211	0	0.2507	-0.2507	-0.5784	-0.5790	-0.7597	0.0023
2212	1	0.1336	0.8664	2.5466	2.5494	2.0065	0.0022

A good way of looking at the impact of various residuals is to graph them against either the predicted probabilities or simply case numbers. Since the sample size of the current study is large enough, the various residuals are plotted against the predicted mean response or estimated logistic probability instead of case numbers in Fig. 1. Different residual plots exhibited in Fig. 1a-d indicate two trends of decreasing residuals with slope-1. These two linear trends result from the fact that the residuals take on just one of two values at a point X_i, $1-\hat{\theta}_i$ or $0-\hat{\theta}_i$. Plotting these values against estimated logistic probability will always produce two linear trends with slope -1. The remaining plots lead to similar patterns. It is visualized from Fig. 1c and d, a few residuals appear with magnitude less than -2 and greater than +2 and beyond of this range definitely deserve closer inspection because standardized residuals outside of this range are potential outliers. If the logistic regression model were in fact true, one would expect to observe a horizontal band with most of the residuals falling within ±2 (Christensen, 1997). Under the existing 2-σ rule, the standardized residuals outside of ±2 may be considered as potential outliers and those are clearly visualized in Fig. 1c and d.

It is well known phenomena that in ordinary linear regression, residual plots are useful for diagnosing model inadequacy, non constant variance and the presence of potential outliers in response as well as in covariate space. Non constant variance is always present in the logistic regression setting and response outliers are difficult to diagnose. So, the current study focused on the detection of model inadequacy and potential outliers in the covariate space only. If the logistic regression model is correct, then E (Y_i) = θ_i and it follows asymptotically that $E(Y_i - \hat{\theta}_i) = E(\hat{\varepsilon}_i) = 0$.

This suggests that if the model is correct and no significant incorporation of potential outliers, a lowess smooth of the plot of the residuals against the estimated logistic probability or linear predictor should result approximately in a horizontal line with zero intercept. Any

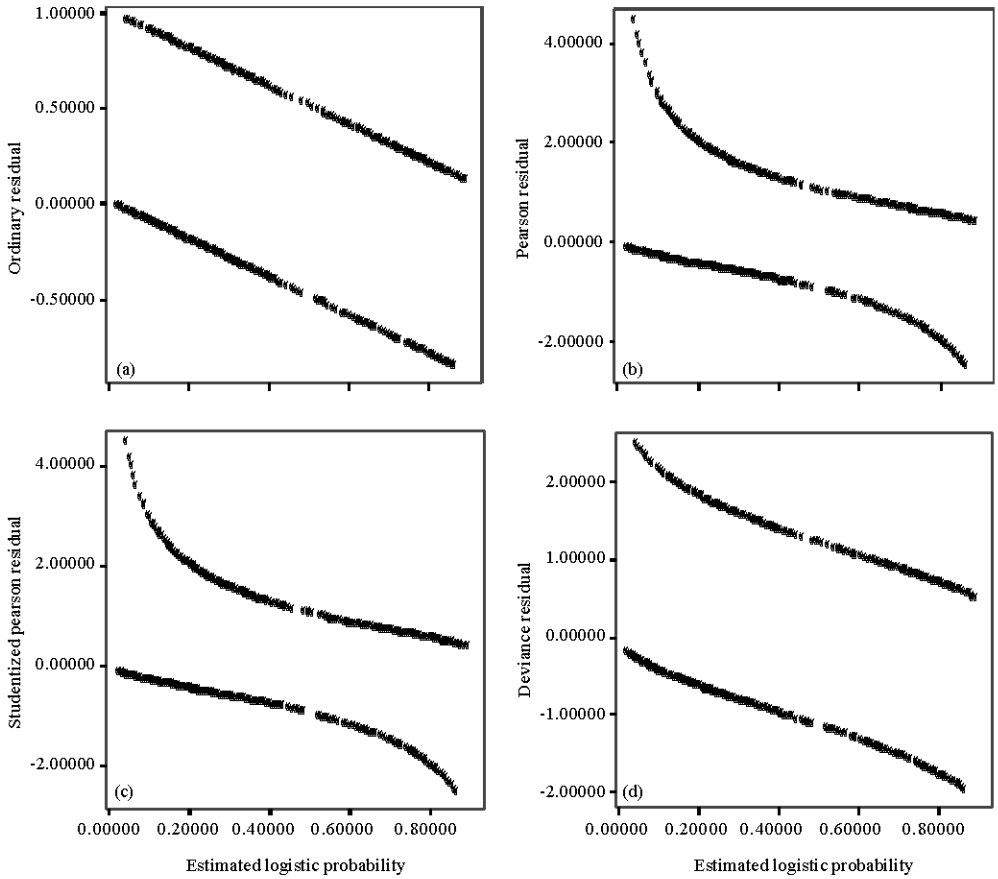

Fig. 1: Selected Residuals plotted against Estimated Logistic Probability for BDHS-2004 Data

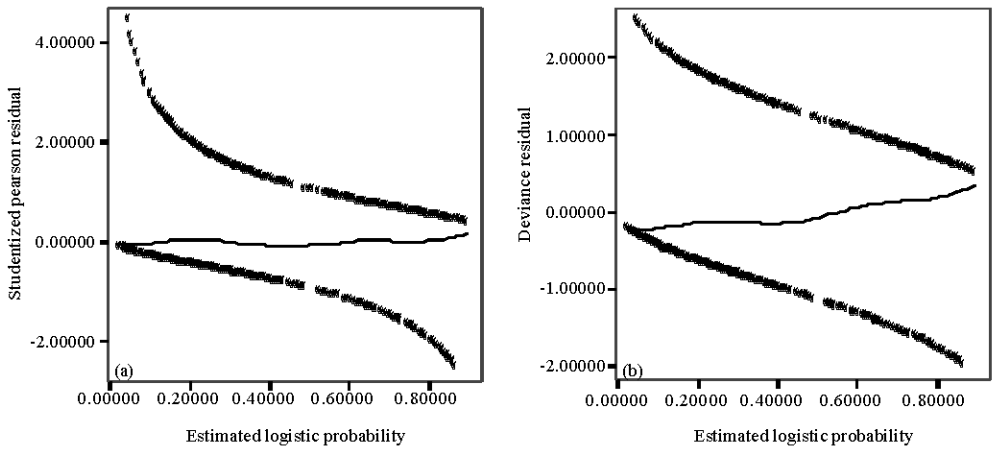

Fig. 2: Standardized Residual Plots with lowess Smooth for BDHS-2004 Data

significant departure from this suggests that the model may be inadequate and potential outliers may have dramatic impact on the fit of the model. The lowess smooth of the studentized Pearson residuals and deviance residuals are demonstrated in Fig. 2. In Fig. 2a and b, the studentized Pearson residuals and deviance residuals are plotted against the estimated logistic probability respectively and in both case, the lowess smooth approximates a line having zero slope and intercept. Hence, it can be concluded that no significant model

inadequacy and presence of influential outliers are observed in the covariate space. Thus the existing outliers detected by the residual plots are not so influential.

DIAGNOSTIC STATISTICS AND DIAGNOSTIC PLOTS

In case of more than two covariates in the logistic regression setting, the standardized residual plots can highlight little regarding influential outliers. In such a situation, some derived diagnostic statistics like change in Pearson chi-square, change in deviance, change in parameter estimates from basic building blocks and their plots including proportional influence or bubble plots are potential to detect outliers and influential cases. Several measures of influence for logistic regression have been suggested. These measures have been developed for the purpose of identifying observations, which are influential relative to the estimation of the logistic regression coefficients (Midi *et al*., 2009). Such a useful diagnostic statistic is one that examines the effect of deleting single subject on the value of the estimated coefficients (β) and the overall summary measures of fit, like Pearson chi-square (χ^2) statistic and deviance (D) statistic. Let, χ^2 denotes the Pearson chi-square statistic based on full data set and $\chi^2_{(-i)}$ denotes that statistic when case i is deleted. Using one-step linear approximations given by Pregibon (1981), it can be shown that the decrease in the value of the Pearson chi-square statistic due to deletion of the ith subject is:

$$\Delta\chi_i^2 = \chi^2 - \chi^2_{(-i)} = \frac{pr_i^2}{1-h_{ii}} = spr_i^2 \quad (11)$$

The one-step linear approximation for change in deviance when the ith case is deleted is as:

$$\Delta D_i = D - D_{(-i)} = \frac{dr_i^2}{1-h_{ii}} \quad (12)$$

The change in the value of the estimated coefficients is analogous to the measure proposed by Cook (1977) for linear regression. It is obtained as the standardized difference between $\hat{\beta}$ and $\hat{\beta}_{(-i)}$, where these represent the maximum likelihood estimates based on full data set and excluding the ith case respectively and standardizing via the estimated covariance matrix of $\hat{\beta}$. Thus, one step linear approximation is given as:

$$\Delta\hat{\beta}_i = \left(\hat{\beta}-\hat{\beta}_{(-i)}\right)' \left(X'\hat{W}X\right)\left(\hat{\beta}-\hat{\beta}_{(-i)}\right) = \frac{pr_i^2 h_{ii}}{(1-h_{ii})^2} = \frac{spr_i^2 h_{ii}}{(1-h_{ii})} \quad (13)$$

Table 2: Pearson residuals, studentized residuals, hat diagonals, deviance residuals, delta chi-square, delta deviance and delta beta statistics for the BDHS-2004 data

	Case						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	pr_i	spr_i	h_{ii}	dr_i	$\Delta\chi_i^2$	ΔD_i	$\Delta\hat{\beta}_i$
1	-0.5896	-0.5903	0.0025	-0.7724	0.3485	0.5975	0.0009
2	-0.6374	-0.6378	0.0011	-0.8258	0.4068	0.6824	0.0005
3	0.5900	0.5907	0.0024	0.7729	0.3489	0.5982	0.0008
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2210	-0.4872	-0.4879	0.0029	-0.6527	0.2381	0.4267	0.0007
2211	-0.5784	-0.5790	0.0023	-0.7597	0.3353	0.5779	0.0008
2212	2.5466	2.5494	0.0022	2.0065	6.4996	4.0401	0.0142

The derived influence statistics are listed in Table 2. These diagnostic statistics are conceptually quite appealing, as they allow us to identify those cases that are poorly fit (large values of $\Delta\chi_i^2$ and ΔD_i) and those that have a great deal of influence on the values of the estimated parameters (large values of $\Delta\hat{\beta}_i$).

A number of different types of diagnostic plots have been suggested to detect outliers and influential cases. It is impractical to consider all possible suggested plots, so we restrict our attention to a few of the more easily obtained ones that are meaningful in logistic regression analysis. These consist of plotting $\Delta\chi_i^2$, ΔD_i and $\Delta\hat{\beta}_i$ against the estimated logistic probability and plotting ΔD_i versus estimated logistic probability where the size of the plotting symbol is proportional to the size of $\Delta\hat{\beta}_i$, where it is usually called proportional influence plot or bubble plot. The derived diagnostic statistics $\Delta\chi^2$ and ΔD plotted against estimated logistic probability are shown in Fig. 3a and b, respectively.

The shapes of the plots are similar and show quadratic like curves. Cases that are poorly fit will generally be represented by points falling in the top left or top right corners of the plots. Assessment of this distance is partly based on numerical value and partly based on visual impression. Since, the current fitted model contains two continuous covariates, the number of covariate patterns is of the same order as sample size. Under n-asymptotic the value of upper ninety-fifth percentile of chi-square distribution with 1 degree of freedom is 3.84 and may provide some guidance as to whether an observation is an outlier or influential point. Thus the cases having numerical values larger than this cut-off point which is based on $\Delta\chi^2$ or ΔD can be considered as outlying observations. It can be observed from Table 3 that 25 observations are detected as outliers and these points fell at the top left corner of the plots displayed in Fig. 3. The range of $\Delta\chi^2$ is much larger than ΔD. This is a property of Pearson versus deviance residuals. Figure 3c shows the plot of the derived influence statistic $\Delta\hat{\beta}$

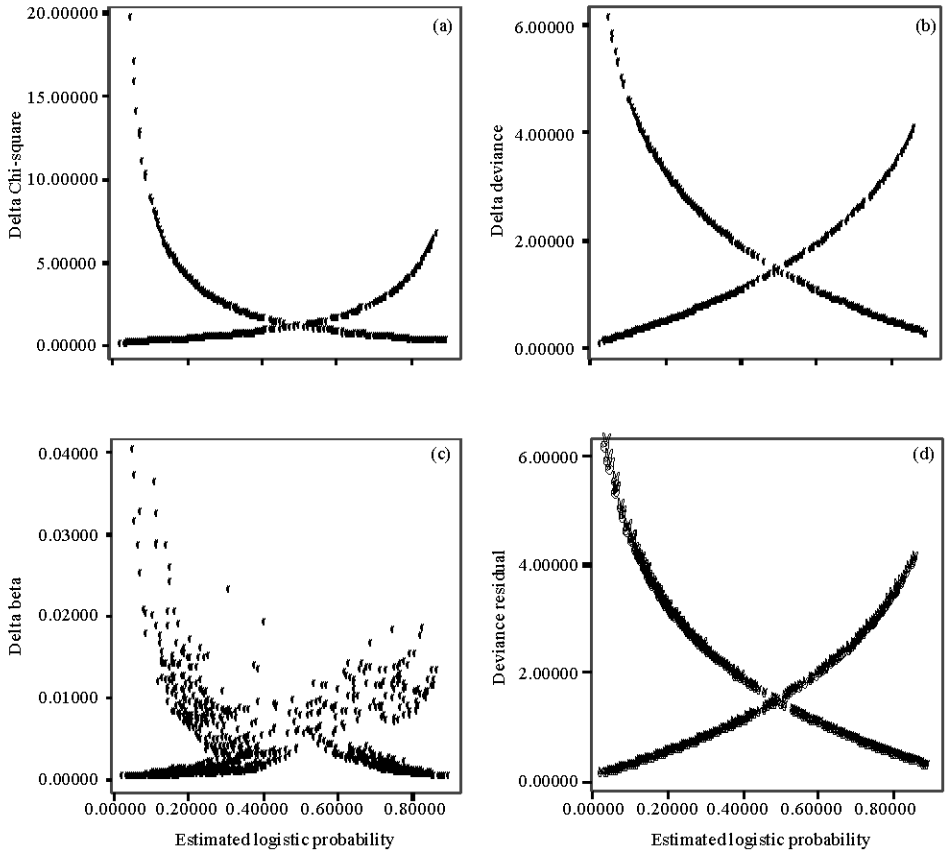

Fig. 3: Delta Chi-square, delta deviance, delta beta and proportional influence plots

Table 3: Outlying cases and their impact on influence statistics for BDHS-2004 data

	Original case No.								
Sr No.	i	(1) y_i	(2) $\hat{\theta}_i$	(3) dr_i	(4) spr_i	(5) h_{ii}	(6) ΔD_i	(7) $\Delta\chi_i^2$	(8) $\Delta\hat{\beta}_i$
1	2052	1	0.0487	2.4583	4.4233	0.0020	6.0833	19.5652	0.0401
2	2047	1	0.0562	2.3995	4.1026	0.0022	5.7946	16.8314	0.0369
3	1607	1	0.0602	2.3708	3.9555	0.0020	5.6517	15.6461	0.0311
4	1441	1	0.0670	2.3252	3.7357	0.0020	5.4345	13.9554	0.0282
5	1684	1	0.0734	2.2853	3.5555	0.0020	5.2475	12.6415	0.0249
6	2170	1	0.0749	2.2768	3.5197	0.0026	5.2164	12.3882	0.0326
7	1419	1	0.0844	2.2237	3.2974	0.0019	4.9655	10.8726	0.0205
8	1399	1	0.0898	2.1958	3.1877	0.0020	4.8412	10.1616	0.0199
9	1679	1	0.0919	2.1850	3.1464	0.0018	4.7918	9.8999	0.0175
10	1793	1	0.0919	2.1850	3.1464	0.0018	4.7918	9.8999	0.0175
11	2153	1	0.1033	2.1306	2.9491	0.0022	4.5591	8.6969	0.0195
12	140	1	0.1065	2.1166	2.9034	0.0043	4.5162	8.4299	0.0362
13	918	1	0.1136	2.0858	2.7988	0.0037	4.3792	7.8333	0.0287
14	446	1	0.1152	2.0792	2.7771	0.0036	4.3511	7.7121	0.0282
15	667	1	0.1162	2.0747	2.7609	0.0024	4.3230	7.6225	0.0186
16	2169	1	0.1182	2.0667	2.7377	0.0043	4.3033	7.4949	0.0321
17	1830	1	0.1214	2.0535	2.6919	0.0016	4.2283	7.2462	0.0117
18	149	1	0.1251	2.0392	2.6485	0.0025	4.1754	7.0143	0.0174
19	2160	1	0.1283	2.0267	2.6103	0.0024	4.1238	6.8135	0.0163
20	2064	1	0.1294	2.0222	2.5960	0.0018	4.1015	6.7389	0.0121
21	1694	1	0.1318	2.0133	2.5695	0.0021	4.0671	6.6023	0.0139
22	2095	1	0.1318	2.0133	2.5695	0.0021	4.0671	6.6023	0.0139
23	545	0	0.8672	-2.0096	-2.5584	0.0020	4.0513	6.5453	0.0130
24	2212	1	0.1336	2.0065	2.5494	0.0022	4.0401	6.4996	0.0142
25	1989	1	0.1348	2.0020	2.5365	0.0024	4.0231	6.4337	0.0152

against the estimated logistic probability. This plot is known as influence plot. We observe that few points lie somewhat away from the rest of the data. The values themselves are not large enough, as all are less than 0.040. The value of such influence statistic for an individual case must be larger than 1 to have an effect on the estimated coefficients. The largest values of $\Delta\hat{\beta}$ are most likely to occur when both $\Delta\chi^2$ and leverage are at least moderately large. However large values can also occur when either component is large. This is the case in influence plot.

The proportional influence plot or bubble plot is exhibited in Fig. 3d. The actual influence of each case on the estimated coefficient can be shown in this plot. This plot allows us to ascertain the contributions of residual and leverage to $\Delta\hat{\beta}$. The large circles in the top left corner correspond to the largest value of ΔD. No such large circles are visualized within $0.1 \leq \hat{\theta}_i \leq 0.9$ which indicates insignificant contribution of leverage on the estimates, because within the said range of estimated probability leverage gives a value that may be thought of distance.

DISCUSSION AND CONCLUSION

Logistic regression is a special case of generalized linear modeling, where the usual approach to outlier detection is based on large sample normal approximations for the deviance and studentized Pearson residuals. It is important to note that deviance residuals are valuable tool for identifying cases that are outlying with respect to covariate space. Global tests of model adequacy use the corresponding chi-squared approximations for the deviance and Pearson Statistics. Although normal approximations to the deviance and studentized Pearson residuals are often reasonable they are questionable for logistic regression with sparse data and with small sample (Hosmer and Lemeshow, 2000). Under the normality assumption with sufficiently large sample, deviance residuals or studentized Pearson residuals follow the chi-square distribution with single degree of freedom. Thus, the upper ninety-fifth percentile value of chi-square distribution which is approximately 4 may be considered as crude cut-off point to detect outlying cases. Crude in the sense, that the distribution of the delta statistics is unknown except under certain restrictive assumptions. Examination of Fig. 3 and numerical values of column 6 and (7) presented in Table 3 identifies 25 ill-fitted cases with outlying values on the basis of diagnostics statistics ΔD and $\Delta\chi^2$. These cases contribute heavily to the disagreement between the data and the predicted values of the fitted model on the basis of observed response y_i and estimated logistic probability $\hat{\theta}_i$ shown at column 1 and 2 in Table 3. Detected outlying cases are one type of observations that has a large value of ΔD and $\Delta\chi^2$ correspond to the misclassified observations. The fitted model predicts that it is unlikely for the subjects to respond when in fact they do ($\hat{\theta}_i$ is small and $y_i = 1$), while the opposite type of poor fit ($\hat{\theta}_i$ is large and $y_i = 0$) also present in the model.

On the other hand, high leverage values are bad. The leverage value varies from 0 to 1. A leverage value of 1 means, the model is being forced or levered to fit the corresponding case exactly. Thus the leverage can be used to detect influential outliers. The leverage of any given case may be compared to the average leverage which equals (k+1)/n, where k is the number of covariates in the model and n is the sample size. The average leverage is inversely proportional to the size of the sample. If the sample is sufficiently large, the leverage value h_{ii} tends to be smaller. Cases are declared influential having $h_{ii}>2(k+1)/n$ (Belsley *et al*., 1980; Bagheri *et al*., 2010). Two times of the average leverage of current study is approximately 0.0054. The leverage h_{ii} values listed at column 5 in Table 3 corresponding to the outlying cases is smaller than that cut-off point. Thus, it may be concluded that the outlying cases are not so influential due to sufficiently large sample.

The effect on the set of parameter estimates when any specific observation is excluded can be computed with the derived statistic based on the distance known as Cook's distance proposed by Cook (1977) in linear regression. The analogous to the measure of one step linear approximation proposed by Pregibon (1981) is $\Delta\hat{\beta}$ in logistic regression. Since an observation is called influential if it has notable effect on parameter estimates, Cook (1977) proposed that the influence diagnostic must be larger than 1 for an individual case to have an effect on the estimated coefficients. Influence diagnostic $\Delta\hat{\beta}$ corresponding to the outlying cases is tabulated in column (8) of Table 3. The values themselves are not especially large with respect to 1 and suggest that none of the outlying cases are influential in the fitting process. One problem with the influence diagnostic $\Delta\hat{\beta}$ is that it is a summary measure of change over all coefficients in the model simultaneously. For this reason it is important to examine the changes in the individual coefficients due to specific cases identified as influential. In this regard, change in individual coefficients can be obtained under the option DfBeta and observed that all the changes are very small relative to 1 (Sarkar *et al*., 2010).

Generally, deleting cases with the largest residuals or more extreme values almost always improves the fit of the model. Since the outlying cases are not influential, it is justified that there were no substantial changes in the model fit or estimated parameters when we delete each cases. The collective effect is also not substantial. So, we decided the outlying cases should be retained in the analysis.

In summary, scientists frequently have primary interest in the outlying cases because they deviate from the currently accepted model. Examination of these outlying cases may provide important clues as to how the model needs to be modified. Outlying cases may also lead to the finding of other types of model inadequacies such as the omission of an important variable or the choice of an incorrect functional form. The analysis of outlying cases can frequently lead to valuable insights for strengthening the model such that the outlying case is no longer an outlier but is accounted for by the model. Finally, it may be concluded that incase of small sample the influential outliers can be detected easily by the leverage value but as sample size increases, the detected outliers do not play any significant influence on the parameter estimates.

REFERENCES

Agresti, A., 2002. Categorical Data Analysis. 2nd Edn. John Wiley and Sons, Inc., Publication, New Jersey, ISBN: 9780471360933.

Allison, P.D., 1999. Comparing logit and probit coefficients across groups. Socio. Meth. Res., 28: 186-208.

Bagheri, A., Habshah Midi and A.H.M.R. Imon, 2010. The effect of collinearity-influential observations on collinear data set: A monte carlo simulation study. J. Applied Sci., 10: 2086-2093.

Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity. John Willey and Sons Inc., New York.

Christensen, R., 1997. Log-Linear Models and Logistic Regression. 2nd Edn., Springer-Verlag Inc., New York, ISBN-10: 0387982477, pp: 508.

Cook, R.D. and S. Weisberg, 1982. Residuals and Influence in Regression. Chapman and Hall, New York, ISBN-10: 0412042312, pp: 456.

Cook, R.D., 1977. Detection of influential observations in linear regression. Technometrics, 19: 15-18.

Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regression through Graphics. Wiley and Sons, New York, ISBN-10: 0471008397, pp: 280.

Hilbe, J.M., 2009. Logistic Regression Models. Chapman and Hall, CRC Press, New York, ISBN-10: 1-4200-7575-6, pp: 656.

Hosmer, D.W. and S. Lemeshow, 1980. A goodness-of-fit test for the multiple logistic regression model. Commun. Statistics, 9: 1043-1069.

Hosmer, W.D. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., John Wiley and Sons, New York, pp: 392.

Jennings, D.E., 1986. Outliers and residual distribution in logistic regression. J. Am. Stat. Assoc., 81: 987-990.

Kutner, M.H., C.J. Nachtsheim, J. Neter and W. Li, 2005. Applied Linear Statistical Models. 5th Edn., McGraw-Hill, New York, ISBN: 0-07-310874-X, pp: 1424.

McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2nd Edn., Chapmann and Hall, London, ISBN: 0-412-31760-5, pp: 536.

Menard, S., 2002. Applied Logistic Regression Analysis. 2nd Edn., Sage Pub., Thousand Oaks, Calif, ISBN-13: 978-0761922087, pp: 128.

Midi, H., S. Rana and A.H.M.R. Imon, 2009. Estimation of parameters in heteroscedastic multiple regression model using leverage based near-neighbors. J. Applied Sci., 9: 4013-4019.

Osius, G. and D. Rojek, 1992. Normal goodness-of-fit tests for multinomial models with large degrees-of-freedom. J. Am. Statistical Assoc., 87: 1145-1152.

Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist., 9: 705-724.

Ryan, T., 1997. Modern Regression Methods. Har/Dis Edn., Wiley, New York, USA., ISBN-10: 0471529125, pp: 515.

Sarkar, S.K. and H. Midi, 2010. Importance of assessing the model adequacy of binary logistic regression. J. Applied Sci., 10: 479-486.

Sarkar, S.K., H. Midi and R. Imon, 2010. Diagnostics of fitted binary logistic regression model based on individual subjects and covariate patterns. Int. J. App. Math., 23: 63-81.

Stukel, T.A., 1988. Generalized logistic models. J. Am. Statistical Assoc., 83: 426-431.