BINARY REGRESSION MODEL DIAGNOSTICS: AN APPLICATION TO CHILDHOOD DIARRHOEA DATA

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

MPHATSO LUWEMBA

UNIVERSITY OF MALAWI

DECEMBER, 2022

BINARY REGRESSION MODEL DIAGNOSTICS: AN APPLICATION TO CHILDHOOD DIARRHOEA DATA

MSc.(BIOSTATISTICS)

BY

MPHATSO LUWEMBA

BSc (Statistics)- University of Malawi

Thesis submitted to the Mathematical Sciences Department, Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

DECEMBER, 2022

DECLARATION

I, the undersigned hereby declare that this thesis/dissertation is my own original work which has not been submitted to any other institution for similar purpose. Where other people's work has been used acknowledgements have been made.

MPHATSO LUWEMBA

Full Legal Name

Signature

Date

CERTIFICATE OF APPROVAL

The undersigned certifies that this thesis represents the student's own work and effort, and has been submitted with my approval.

Signature _____ Date _____

Tsirizani Kaombe, Ph.D. (Lecturer)

Supervisor

DEDICATION

This thesis is dedicated to my late father, Peter Luwemba and my late mother, Rose Juma , had you both lived, this would have made you proud.

ACKNOWLEDGEMENTS

I would like to thank God for his love, care and guidance.

I wish to extend my deepest gratitude to my supervisor, Dr. Tsirizani Kaombe for his support, coordination , constructive corrections and criticism throughout the entire period of my study.

I am also grateful to my sisters; Aida Khuliwa, Grace Gulani and Esther Mikuti for their support and encouragement during my studies.

Special thanks to my friends, classmates and workmates; Blessings Chisambi, Ellen Gondwe, Bright Mkandawire, John Abdul, Noel Tembo and Faith Lembeno for their support and encouragement during my studies.

Finally, I am indebted to the DELTAS Sub-Saharan Africa Consortium in Advanced Biostatistics Training (SSACAB) for funding my Master's degree at University of Malawi- Chancellor college.

ABSTRACT

Binary logistic regression model is applied in many public health studies, that involve binary response variable such as presence or absence of diarrhoea in a child. In such applications, most studies in literature have focused on inferences and implications on relevant policies. There has been little effort to exhaustively understand the fit of the binary regression model to the data the at hand, before making conclusions and recommendations on policies. This study focused on utilization of the post-estimation diagnostic statistics that are available for fitting binary regression models to data, which are usually ignored in most applications of the model. This was done by applying diagnostic statistics that analyze the presence of outliers, influential observations, high leverage subjects and multicollinearity among independent variables, upon fitting binary logistic regression model to child dirrhoea data from 2015-16 Malawi demographic and health survey. The results showed that there were outliers and high leverage points in the model. Region and toilet sharing variables were mostly affected by outliers. But using Cook's distance, the individual children did not have influence on all estimated regression parameter values. The study recommends that analysts should throughly examine the fit of the logistic regression model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Regression methods have become important tools in data analysis that involves describing the relationship between a dependent variable and one or more independent variables. In some cases, the dependent variable in such relationships is categorical, hence the logistic regression model becomes a choice to characterize the relationship. The data under study can be in the field of health, agriculture, engineering and education among others. When the response variable is dichotomous, a binary logistic regression model is used while ordinal categorical responses are modeled using ordinal logistic regression model.

When the response variable has several categories that are nominal in scale of measurement, each with binary outcomes, then multinomial logistic regression model is used (OConnell, 2006; Hilbe, 2009). Multinomial logistic regression is used to predict the probability of being in a certain group compared to other groups (Hilbe, 2009).

Instead of directly predicting the response variable using a set of covariates, logistic regression models probabilities of success given the covariates(Hosmer Jr, Lemeshow, & Sturdivant, 2013; Park, 2013). When the dependent variable is ordinal then ordinal logistic regression is used. An ordinal variable is a categorical variable in which there is ordering of the category levels (Harrell, 2015). The ordinal logistic regression analyzes how much closer each predictor pushes the outcome into the next level or category of the outcome variable (Hilbe, 2009).

The term "logistic" is used because the relationship between the covariates and the probability of success resembles a logistic growth curve. Thus, binary logistic regression model does not assume linearity in the relationship between the dependent variable and independent variables. In addition, the model does not need the variables to be normally distributed (S. K. Sarkar, Midi, & Rana, 2011).

1.2 Binary regression model and estimation

Let $Y_1, Y_2, ..., Y_n$ be identically and independently distributed random variables, where each $Y_i \sim Binomial(n_i, \theta_i)$, that is, $P(Y_i = 1) = \theta_i$ and $P(Y_i = 0) = 1 - \theta_i$ giving a Bernoulli as a special case of binomial experiment . Let $X = (X_1, X_2, ..., X_P)$ be a set of explanatory variables, which can be used to estimate the parameter θ_i for data on Y_i, that is , $P(Y_i = 1|X) = \theta_i(X)$. Then, a binary regression model is given by

$$y_i = \theta_i(X) + \epsilon_i \tag{1.1}$$

where $y_i^T = \begin{bmatrix} y_1, & y_2, & \cdots, & y_n \end{bmatrix}$ is a vector of dichotomous responses, with $y_i = 1$ if the observed outcome is a success and 0 if it is a failure, $X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{12} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix}$ is a design matrix of covariates that are not linearly correlated, $\epsilon_i^T = \begin{bmatrix} \epsilon_1, & \epsilon_2, & \cdots, & \epsilon_n \end{bmatrix}$ is a vector of unknown errors that depend on $\theta_i(X)$. The errors ϵ_i have unknown probability distribution(S. K. Sarkar et al., 2011). In addition, the relationship between covariates and probabilities of success is given by the formula:

$$\theta_i(X) = \frac{exp(X\beta)}{1 + exp(X\beta)}. \tag{1.2}$$

This is the logistic function that characterizes the shape of the conditional probability of Y given $\beta^T = \begin{bmatrix} \beta_1, & \beta_2, & \cdots, & \beta_p \end{bmatrix}$, which are the regression coefficients. $\theta_i(X)$ has values between 0 and 1 (Hosmer & Lemeshow, 2002).

From above model, one can derive the link function for the model as

$$\log(\frac{\theta_i(X)}{1 - \theta_i(X)}) = X\beta. \tag{1.3}$$

This link function is called the logarithm of odds of success or simply the "logit" link, to which the model gets its name. In this case, a unit increase in the value or level of a covariate X leads to change of logarithm of odds of success for Y by a value of β (Hilbe, 2009). This interpretation of the coefficients in logistic regression is said to be at coefficient scale. Alternatively, one can exponentiate

the above equation (1.3) both sides, so that an increase in value of a covariate will lead to a change of $\exp(\beta)$ in odds of success of Y. This is called odds scale interpretation (Rawlings, Pantula, & Dickey, 2001).

1.2.1 Model estimation

Since Y_i in equation 1.1 is distributed as $Binomial(n_i, \theta_i(X))$ with probability mass function; $f(y_i, \theta_i) = \binom{n_i}{y_i}\theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$, then its conditional probability mass function can be expressed in exponential family form as:

$$f(y_i, \theta_i(X)) = \exp[y_i log(\frac{\theta_i(X)}{1-\theta_i(X)}) + n_i \log(1-\theta_i(X)) + \log \binom{n_i}{y_i}]. \tag{1.4}$$

Hence, the likelihood function for β is given by:

$$\begin{aligned} L(\beta; y_i, X) &= \prod_{i=1}^{n} f(y_i, \theta_i(X)) \\ &= \exp[\sum_{i=1}^{n}\{(y_i \log(\frac{\theta_i(X)}{1-\theta_i(X)})) + n_i \log(1-\theta_i X) + \log \binom{n_i}{y_i}\}]. \end{aligned} \tag{1.5}$$

This gives the log-likelihood as:

$$\begin{aligned} l(\beta, y_i, X) &= \log L(\beta; y_i, X) \\ &= \sum_{i=1}^{n}\{(y_i \log \frac{\theta_i(X)}{1-\theta_i(X)}) + n_i \log(1-\theta_i(X)) + \log \binom{n_i}{y_i}\} \\ &= \sum_{i=1}^{n} y_i X\beta - \sum_{i=1}^{n} n_i \log[1+\exp(X\beta)] + \sum_{i=1}^{n} log\binom{n_i}{y_i}. \end{aligned} \tag{1.6}$$

The goal is to find the value of β that maximizes the log-likelihood function $l(\beta, y_i, X)$. The critical points of a function $l(\beta, y_i, X)$ are found when its first derivative equals 0. These derivatives give score functions in respect of each

coefficient β. Maximum likelihood is one of the parameter estimation methods for binary logistic regression. The advantages of using maximum likelihood are that it gives most efficient estimators if the assumptions are satisfied and it gives unbiased estimates for large sample size (Erica, 2020). From the log-likelihood (1.6), the score function is given by:

$$\begin{aligned} U(\beta) &= \frac{\partial l(\beta)}{\partial \beta} \\ &= \sum_{i=1}^{n} X^T Y_i - \sum_{i=1}^{n} (\frac{n_i X^T \exp(X\beta)}{1+\exp(X\beta)}) \\ &= \sum_{i=1}^{n} X^T [y_i - n_i \theta_i(X)]. \end{aligned} \tag{1.7}$$

The maximum likelihood estimators, $\hat{\beta}$ for model parameters β are found by solving for β when the score functions $U(\beta)$ above are equated to zero. However, the score function equation $U(\beta) = 0$ is intractable for β. Hence, numerical techniques are used to solve for maximum likelihood estimators $\hat{\beta}$ from the equation $U(\beta) = 0$. These include iterated re-weighted generalized least squares, Newton Raphson method among others (Pregibon, 1981). With the Newton Raphson technique, the value of β such that $U(\beta) = 0$ can be obtained by observing a small step in β, that is from $\beta^{(k)}$ to $\beta^{(k+1)}$ that makes the score function, $U(\beta)$ almost static, i.e $U(\beta^{(k+1)})$ not far from $U(\beta^{(k)})$. This is essentially the slope:

$$\frac{\partial U(\beta)}{\partial \beta}|_{\beta=\beta^{(k)}} = \frac{U(\beta^{(k+1)}) - U(\beta^{(k)})}{\beta^{(k+1)} - \beta^{(k)}} = U'(\beta^{(k)}). \tag{1.8}$$

If the step $\beta^{(k+1)}$ will be the required solution, such that $U(\beta^{(k+1)}) = 0$, then one can solve for $\beta^{(k+1)}$ in the relation of middle and right most terms in the above

equation and obtain the Newton- Raphson iteration equation:

$$\begin{aligned}\beta^{(k+1)} &= \beta^{(k)} - \frac{U(\beta^{(k)})}{U'(\beta^{(k)})} \\ &= \beta^{(k)} - [U'(\beta^{(k)})]^{-1}U(\beta^{(k)}) \\ &= \beta^{(k)} + I^{-1}(\beta^{(k)})U(\beta^{(k)}),\end{aligned} \tag{1.9}$$

where $k = 0, 1, ...n$ is an iteration step, $I(\beta = -E(U(\beta))$ is the Fisher information for parameter β. Thus, to get maximum likelihood estimators $\hat{\beta}$ for logistic regression model (1) using Newton-Raphson algorithm, one need a $(p \times 1)$ vector of score functions $U \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$, the $(p \times p)$ information matrix and $(p \times 1)$ vector of initial values of estimators $\begin{bmatrix} \beta_1^{(0)} \\ \beta_2^{(0)} \\ \vdots \\ \beta_p^{(0)} \end{bmatrix}$. This is implemented in many statistical softwares(Kasza, 2015).

1.3 Diagnostic Statistics

Model validation is a very important step in building a statistical model. Model diagnostic statistics have been developed to check for and detect model mis-specifications. The validity of inferences drawn from fitting the logistic regression model to the data depends on the model satisfying the assumptions (Hosmer, Taber, & Lemeshow, 1991). It is very important to examine how well the model describes

the observed data. For example, given the observed outcomes $y^T = (y_1, y_2, ..., y_n)$ and their predicted values $\hat{y}^T = (\hat{y_1}, \hat{y_2}, ..., \hat{y_n})$ then one would expect the distances $(y_i - \hat{y}_i)$ to be small for a model that fits the data well (Menard, 2002).

Diagnostic statistics are quantities calculated from the data with the aim of finding influential data points, residuals, outliers or high leverage points so that their effect on the inferences is known and corrected for the model to remain valid (Nurunnabi & Nasser, 2011). Failure to do model diagnosis may lead to misleading or incorrect inferences (Hosmer & Lemeshow, 2002). For instance, outliers to a model cause serious problems for the application of many statistical models, especially models that assume normal distribution for the error term (Ramsey & Ramsey, 2007). An outlier is an observation whose value deviates so much from the expected range as to arouse suspicions that it was generated by a different mechanism (S. K. Sarkar et al., 2011). Outliers can be caused by human errors for example keypunch errors, recording errors and instrument errors. Outliers may cause errors in parameter estimation and misclassifying the outcomes, which can cause problems when drawing inferences (Nurunnabi & Geoff, 2012). Some of the available methods for assessing outliers include; Pearson residual , Studentized Pearson residual and Deviance residual.

Influential observations need also to be assessed when fitting binary logistic regression. Influential observations are defined as points which either individually or together with several other observations have a demonstrably larger impact on the calculated values of various estimates (Nurunnabi & Geoff, 2012). Some of the available

methods for detecting influential points in logistic regression model are Cook's distance, Dffits and Dfbetas.

Multicollinearity is another problem which is needed to be addressed when fitting logistic regression model. Multicollinearity is defined as condition where there are dependencies among the independent variables which causes parameter estimates to be unstable (Haque, Jawad, Cnaan, & Shabbout, 2002). Some of the available methods for detecting multicollinearity are variance inflation factor, condition index and variance decomposition proportions (Senaviratna & Cooray, 2019; Midi, Sarkar, & Rana, 2010).

There are also diagnostic statitistics which assess fulfillment of model assumptions of linearity and additivity. One of the assumption of binary logistic regression model is that there is linear relationship between the log odds of success outcome and the independent variables (Schreiber-Gregory, 2018). One of the methods for checking this assumption is box-tidwell transformation. Box-tidwell transformation measures linearity between the log odds and the continuous variables (Hilbe, 2009).

1.4 Logistic Regression Application in Diarrhoea Studies

Diarrhoea remains one of the major causes of morbidity and mortality in infants and children in developing countries, including Malawi. Diarrhoea is water-borne disease which is leading cause of death among children under age of five (Getachew,

Guadui, Tadie, & Gizaw, 2018). According to (Mwambete & Joseph, 2010), diarrhoea is defined as condition in which there is increase in volume of stool and number of visits to the toilet. Binary logistic regression model has been used on a number of diarrhoea studies.

For example, (Kaombe & Namangale, 2016) used Logistic regression model in comparison with Poisson and Bayesian regressions to investigate variations in risk of diarrhoea in under-five children in Malawi using 2006 Malawi Multiple Indicator (MICS) data. In their study, mother's education, toilet sharing status, place of residence and source of drinking water were among risk factors for childhood diarrhoea. While, (Getachew et al., 2018) applied binary logistic regression to model risk factors of under five child diarrhoea in rural north Gondar Zone, Northwest Ethiopia. The study found out that number of children family members in a household , mother education and age of under five child were significantly associated with diarrhoea.

Another study that applied binary logistic regression model on childhood diarrhoea was done by (Mbugua et al., 2014) using data from 2008 Kenya demographic and health survey. In that study, place of residence was found to be useful factor for determining the child diarrhoea.

In all the above cited studies, the binary logistic regression model was fitted without adequate model validation assessments. Each of these studies did not pay attention to examining outlying children, influential children and the children

who had high leverage in the fitted probabilities of diarrhoea. The estimates and conclusions were made without careful examination of the fit of model to the data. This usually leads to problematic conclusions like obtaining wrong value of coefficients.

1.5 Problem Statement

Many public health research involves dependent variable with two possible outcomes and independent variables that are categorical or continuous (Nurunnabi & Geoff, 2012). In this case, logistic regression model is mostly used for the reasons of ease of convenience, interpretation and computation which can be implemented in many available statistical softwares (Hosmer et al., 1991). The validity of inferences drawn from logistic regression model depends on assumption of the fitted model being satisfied (S. Sarkar, Midi, & Rana, 2011).

Model diagnostics are tools that are used in assessing the appropriateness of the model in fitting the data. Unfortunately, many studies which used logistic regression model on diarrheoa data did not consider examining the fit of the model as stated in previous section. When logistic regression model is applied to child diarrhoea data, most studies in literature have focused on model description and prediction (S. Sarkar et al., 2011). There has been little effort to exhaustively understand the fit of the logistic regression model.

1.6 Study Objectives

1.6.1 Main Objective

- The main objective of this study was to demonstrate utilization of the post-estimation diagnostic statistics that are available for fitting binary regression model to data which are usually ignored in most applications.

1.6.2 Specific objectives

- To fitting the binomial regression model to child diarrhoea data from 2015-16 Malawi demographic and health survey.
- To apply available methods of detecting outliers, influential points, multicollinearity and leverage values in logistic regression.

1.7 Significance of the study

Logistic regression model has received massive attention in the modeling of binary response health data. There has been little effort to exhaustively understand the fit of the model to such data. Improving child health features highly in United Nations sustainable development goals (SDGs) 2030. This includes aiming for good health and well being (goal 3) and clean water and sanitation (goal 6), which directly relates to reducing childhood diarrhoea cases (GOM, 2020). Goal number 3 of vision 2063 for Malawi is to have healthy and well-nourished citizens through reduction of diseases including diarrhoea. Moreover, in Malawi diarrhoea among under five children remain a big problem with 22% prevalence (NSO, 2017). So,

there is need to fully diagnose the binary logistic regression whenever it is fitted on diarrhoea data in order to have valid results. The findings will help in minimizing erroneous conclusions that follow from fitting logistic regression model to diarhoea which can help in improving child health.

1.8 Thesis structure

The thesis is structured as follows: Chapter 2 gives literature review of binary logistic regression model diagnostic statistics . Chapter 3 presents the methodology of the study. Chapter 4, results of the fitted model diagnostics. Chapter 5 provides discussion and conclusion.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews theoretical framework for diagnosing logistic regression model. The chapter reviews diagnostic statistics for outliers, diagnostic statistics for multicollinearity, influence statistics for the logistic model and diagnostic statistics for leverage.

2.1 Diagnostic statistics for outlier assessment

An outlier is defined as an observation that highly differ from other observations which brings suspicion that it was collected using a different way (Ahmad, Ramli, & Midi, 2012). Outliers can come from human error like making mistakes when entering the data and using values for missing observation as real observation (Ahmad et al., 2012). Outliers increase error variance which lead to decrease of power of statistical tests, alter the odds and they cause bias of estimates of the model (Osborne, 2004). There are many diagnostics for detecting outliers. Residual analysis can help in detecting the outliers. The residuals in logistic regression model are not normally distributed, since the response takes values 0 and 1 (S. K. Sarkar et al., 2011). Since the errors in logistic regression are binary in nature, so the error variance is a function of $\theta(X)$. From logistic regression

model (1.1), the residual $\hat{e}_i$ is defined as;

$$\hat{e}_i = y_i - \hat{y}_i = \begin{cases} 1 - \hat{\theta}_i, & \text{if } y_i = 1 \\ -\hat{\theta}_i, & \text{if } y_i = 0. \end{cases} \tag{2.1}$$

This type of residual is called raw or ordinary residual. Raw residual is used to measure variation between observed and fitted values, the closer the distance from y_i to $\hat{y}_i$ the better the model (Hilbe, 2009). The values of raw residual that are less than -2 and greater than +2 correspond to potential outliers to the model (S. K. Sarkar et al., 2011). There is also raw residual which is used when fitting linear regression. The raw residual of linear regression is given by

$$e_i = y_i - \hat{y}_i \tag{2.2}$$

where $\hat{y}_i = X\beta$ (Larsen & McCleary, 1972). The values of raw residual that are less than -2 and greater than +2 correspond to potential outliers to the linear regression regression(Freund, Vail, & Clunies-Ross, 1961).

2.2 The Pearson's Residual

Logistic regression is fitted with assumption that the model has important explanatory variables and these variables are entered in correct form (Hosmer & Lemeshow, 2013). After fitting logistic regression model, there is need to check if the probabilities accurately reflect the true outcome in the data (Hosmer & Lemeshow, 2013). Assuming that the observed sample values of outcome are represented by $y_i^T = (y_1, y_2, ..., y_n)$ and the estimated values by $\hat{y}_i^T = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$. Then the logistic

regression model is considered to be fitting the data if the summary measures of the distance between y_i and $\hat{y}$ are very small. The Pearson's residual is calculated by taking the difference between observed and fitted values and dividing by the estimate of standard deviation of $\hat{y_i}$ (Hosmer & Lemeshow, 2013). This is given by;

$$pr_i = \frac{\hat{e_i}}{\sqrt{\hat{\theta}_i(X)(1-\hat{\theta}_i(X))}} = \frac{y_i - \hat{\theta}_i(X)}{\sqrt{\hat{\theta}_i(X)(1-\hat{\theta}_i(X))}} \tag{2.3}$$

where $\hat{\theta}_i(X)$ are the fitted values and y_i are observed values. The observations that have higher Pearson's residuals are suspected to be outliers (Ahmad et al., 2012). Using Pearson's residuals, the observation is regarded as problematic if $|pr_i| > 3$ (LaValley & P, 2008).

2.3 Studentized Pearson Residual

Another transform of the residual (2.1) is Studentized Pearson residual. It is measured by dividing the residual by its estimate of standard deviation (Hosmer & Lemeshow, 2002). The standard of deviation of residual (2.1) is given by

$$\sqrt{\hat{\theta}_i(X)(1-\hat{\theta}_i(X))((1-h_{ii})} \tag{2.4}$$

where h_{ii} are diagonal entries of hat matrix $\hat{W}^{1/2}(X\hat{W}X)^{-1}X\hat{W}^{\frac{1}{2}}$ where X is the vector of independent variables and W is a diagonal matrix with entries $\sqrt{\theta(1-\theta)}$. The Studentized Pearson residual is given by:

$$Spr_i = \frac{y_i - \hat{\theta}_i(X)}{\sqrt{\hat{\theta}_i(X)(1-\hat{\theta}_i(X))(1-h_{ii})}} = \frac{pr_i}{\sqrt{1-h_{ii}}} \tag{2.5}$$

The model fits well, if the residuals are between -3 and $+3$ (Menard, 2002). Similarly, the studentized residual for linear regression is found by dividing the raw residual with estimated standard error. The studentized residual is given by

$$Z_i = \frac{e_i}{\sqrt{MSE}} \tag{2.6}$$

The model is regarded as a good model if the graph of Z_i resembles N(0,1) (Freund et al., 1961).

2.4 Deviance Residual

This is another residual which is based on deviance or likelihood ratio chi square statistics (Ahmad, Midi, & Ramli, 2011). The deviance residual is used to measure the difference between any component of the log likelihood of the fitted model and the corresponding component of log likelihood that will result if each point was fitted correctly (S. K. Sarkar et al., 2011). It is used to identify potential outliers (S. K. Sarkar et al., 2011). Deviance residual for i-th case in logistic regression model is defined as;

$$d_i = sign(y_i - \hat{\theta}_i(X))\{-2[y_i \log \hat{\theta}(X) + (1-y_i)log(1-\hat{\theta}_i(X))]\}^{1/2}. \tag{2.7}$$

The deviance residual is better compared to Pearson's residual because deviance residual depicts normal distribution (S. K. Sarkar et al., 2011). The model is

regarded as valid if the deviance residuals are between -2 and 2. The values of deviance residual that are less than -2 and greater than +2 correspond to potential outliers to the model. There is another type of deviance residual which is used in poisson regression model. The deviance residual for poisson regression model is given by

$$d_i = \mathtt{sgn}(y_i - \exp\{\mathbf{X}_i\hat{\beta}\})\sqrt{2\left\{y_i \log\left(\frac{y_i}{\exp\{\mathbf{X}_i\hat{\beta}\}}\right) - (y_i - \exp\{\mathbf{X}_i\hat{\beta}\})\right\}}. \quad (2.8)$$

where β's are coefficients and X_i's are independent variables (Jennings, 1986). The model is regarded as good model if d_i is between 0 and 1 (Davison, Gigli, & A, 1989).

(Chen, Yang, Chen, & Chen, 2008) applied Pearson residuals, Studentized residuals and Deviance residual on logistic regression in order to detect outliers. The study focused on detecting the outliers and influential observations of the data from experimental study. The results showed that 3 points were found to be outliers and they were removed and the logistic regression was fitted again and the comparison was made. After removing the outliers R-squared increased from 0.50 to 0.67 and also predicative estimates were found to be better.

2.5 Diagnostic Statistics for Multicollinearity

One assumption of logistic regression is that explanatory variables should be independent of each other. Multicollinearity is defined as a situation where independent

variables are associated or dependent on each other (Midi et al., 2010). Let the j-th column of the X matrix be represented by X_j, then $X = [X_1, X_2, \ldots, X_p]$. X_j contains the n values of the jth independent variable. Then multicollinearity is defined in terms of the linear dependence of the columns of X. According to (Montgomery & Peck, 2012), the vectors $X_1, X_2, \ldots, X_p$ are linearly dependent if there exist a set of constants $t1, t2, \ldots, tp$ not all zero, such that

$$\sum_{j=1}^{p} t_j X_j = \mathbf{0}. \tag{2.9}$$

Multicollinearity can be caused by ways of collecting data, model specification and putting a lot of regressors in the model (Montgomery & Peck, 2012). Some of the problems that can be caused by multicollinearity are unstable estimates and inaccurate variances which can cause errors in confidence intervals and hypothesis test (Midi et al., 2010). In other ways the p-values are decreased and confidence intervals are increased because of collinearity (Miles, 2014).

2.5.1 Variance Inflation Factor (VIF)

Variance inflation factor is defined as the reciprocal of Tolerance where Tolerance is defined as $1 - R^2$ (Miles, 2014; Midi et al., 2010). VIFs and Tolerance are so much related and they all depend on R^2 (Miles, 2014). R^2 is coefficient of determination that is obtained by regressing each independent variable as dependent variable on all other independent variables (Miles, 2014). Since Variance Inflation factor is just reciprocal of Tolerance, given by:

$$VIF = \frac{1}{Tolerance} = \frac{1}{1 - R^2}. \tag{2.10}$$

When variance inflation factor is greater than 10, it means that there is multicollinearity between the covariates under consideration (Midi et al., 2010).

Variance inflation was used in diagnosing the logistic regression in the study done by (Midi et al., 2010). The study used data from the Bangladesh Demographic and Health Survey (BDHS-2004). The dependent variable was whether the person would like to have another child and the responses were 0 for 'no more' and 1 for ' have another'. Some of the independent variables were age, education level and working status. The results showed that three independent variables were dependent to each other or there was presence of multicollinearity in the model.

2.6 Influential statistics for the logistic model

Influential observations are cases that influence the estimation of the regression coefficients vector and the deviance (Johnson, 1985). Influential observations reduce the power of the test for significance of the covariates since error variance is increased and it also causes unbiased estimates (Dhakal, 2017). Some of the methods of detecting influential points in logistic regression model are Cook's distance and Dffits (Nurunnabi, Rahmatullah Imon, & Nasser, 2009).

2.6.1 Cook's distance

Cook's distance is used to find influential observations by finding the difference between the regression parameter estimates $\hat{\beta}$ and the result if the i-th data point

is deleted (McDonald, 2002; Cook, 1977). The Cook's distance is defined as;

$$CD_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T (X^T V X)(\hat{\beta} - \hat{\beta}_{(-i)})}{p\hat{\sigma}^2} \qquad i = 1, 2..., n \tag{2.11}$$

where $\hat{\beta}_{(-i)}$ is the estimated parameter of $\hat{\beta}$ with the ith observation deleted, V is $n \times n$ diagonal containing elements $\hat{\theta}_i(X)(1 - \hat{\theta}_i(X))$, $\hat{\sigma}^2$ is the variance of response variable and p is the number of covariates in the model (McDonald, 2002; S. K. Sarkar et al., 2011). For logistic regression model, Cook's distance simplifies to;

$$CD_i = \frac{(spr_i)^2 h_{ii}}{1 - h_{ii}} \tag{2.12}$$

where spr_i is Standardized Pearson residual and h_{ii} is leverage points (S. K. Sarkar et al., 2011). The Cook's distance estimates impact of removing an observation on maximum likelihood estimator $\hat{\beta}$. If CD_i is greater that 1, then that i-th point is influential observation on $\hat{\beta}$ (Nurunnabi et al., 2009). Cook's distance is also used in linear regression to detect influential points. The Cooks distance for linear regression is given by

$$D_i = (r^2 / P * MSE) * (h_{ii}/(1 - h_{ii})^2) \tag{2.13}$$

where r_i is i-th residual, P number of independent variables in the model, MSE is mean square error and h_{ii} is the leverage (Nurunnabi, Hadi, & Imon, 2014). The rule of thumb is that when D_{ii} is greater than 4/n then that point is regarded as influential point. Cook's distance is also used in Poisson regression to detect influential points. The simplified formula for Cooks distance for poisson regression

is given by

$$D_i = \frac{\chi_i^2}{P+1} * \frac{h_{ii}}{1-h_{ii}} \tag{2.14}$$

where χ_i^2 is the value of chi-square, P is the number of covariates in the model and h_{ii} is the leverage point (Nurunnabi et al., 2014). The point is regarded as influential point if it is greater than 4/(n-1).

2.6.2 DFFITS

Another method for finding influential observations is Dffits (Uraibi, 2019), which measures impact of i-th case on fitted value $\hat{y}_i$ when i-th observation is deleted from the data, Dffits calculate the change in fit (Khan, Amanullah, Aljohani, & Mubarak, 2021). For logistic regression model, the Dffits is defined as;

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{(-i)}\sqrt{(1-h_{ii})}} \tag{2.15}$$

where $\hat{y}_{i(-i)}$ is fitted response when i-th observation is deleted, h_{ii} is leverage and $\hat{\sigma}_{(-i)}$ is estimated standard error when i-th observation is deleted (Uraibi, 2019). The Dffits can also defined as

$$DFFITS_i = spr_i\sqrt{\frac{h_{ii}\hat{\theta}_i(1-\hat{\theta}_i)}{(1-h_{ii})[\hat{\theta}_i(1-\hat{\theta}_i)]^{-i}}} \tag{2.16}$$

where spr_i is Studentized Pearson residual and h_{ii} is leverage (Uraibi, 2019). When the point has $Dffits_i > 2$ then that point is influential observation (MESTAV, 2019). Dffits are also used in poisson regression where they are used to detect

influential points. The Dffits for poisson regression are given by below formula

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{h_{ii}}} \tag{2.17}$$

$\hat{y}$ is the predicted dependent variable before deleting case i, $\hat{y}_{i(i)}$ is predicted dependent variable after deleting case i and h_{ii} is the leverage. If $DFFITS_i$ is greater than $2\sqrt{(P+1)/n}$ then that observation is influential point.

2.6.3 Dfbetas

Another method to detect influential observation on each regression coefficient is to determine how much that coefficient changes when the observation is deleted. The Dfbetas statistic is defined as the standardized difference between a regression coefficient before and after the removal of the jth observation (Dattalo, 1994). Dfbetas is expressed by;

$$DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_j^i}{\sqrt{s^2_{(i)} c_{jj}}} \tag{2.18}$$

Where $\hat{\beta}_j$ is estimated j-th regression coefficient, $\hat{\beta}_j^i$ is is estimated j-th regression with i-th observation deleted and c_{jj} is the j-th element of $(X^T X)^{-1}$. An observation is regarded as influential point if $DFBETAS_{ji} > \sqrt{\frac{2}{n}}$, where n is sample size (Ghosh, 2022).

Cook's distance and Dffits were used in the study done by (Ghosh, 2022). The main objective of the study was to extrapolate from the pre-existing deletion diagnostics defined for detecting influential points for binary logistic regression.

The study used modified kyphosis data containing 81 children who have had corrective spinal surgery. The dependent variable was whether a post-operative deformity (Kyphosis) is 'present' or 'absent' in an individual and the independent varibles were the number of vertebrae involved in the operation and the beginning of the range of vertebrae involved in the operation. The results from Cook's distance and the Dffits showed that there were no influencial points in the model since all points were less than 5.

2.7 Diagnostic statistics for Leverage

Leverage values are very important in logistic regression model. Leverage values reveals which observations in the X-space of the data are responsible for coming up with unusual fitted responses (Imon & Hadi, 2013).

2.7.1 Pregibon Leverage

Pregibon leverage, denoted by h_{ii} is i-th diagonal element of $n \times n$ estimated hat matrix H (S. K. Sarkar et al., 2011). In the logistic regression, the H matrix is defined by

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2} \tag{2.19}$$

where V is a diagonal matrix with size of $n \times n$ containing $\hat{\theta}_i(1 - \hat{\theta_i})$ and X is $n \times (p+1)$ matrix (Imon & Hadi, 2013; S. K. Sarkar et al., 2011). The i-th diagonal element of matrix H for logistic regression model is defined as

$$h_{ii} = \hat{\theta}_i(1 - \hat{\theta}_i) x_i^T (X^T V X)^{-1} x_i \tag{2.20}$$

Where x_i^T =$[1, x_{1i}, x_{2i}, ...x_{pi}]$ which is $1 \times p$ vector of observed covariates for i-th element (Imon & Hadi, 2013). When the point has $h_{ii} > 3p/n$ then that point is regarded as high leverage point. (Fitrianto & Wendy, 2016).

Pregibon leverage was used in the study by (Imon & Hadi, 2013) for detecting high leverage points in logistic regression. The main objective of the study was to identify multiple high leverage points in logistic regression. The study used artificial data where the values of Y were given in a way that the first five values were set to 0, the next five to 1 and the whole sequence was repeated once again. The results showed that there were no high leverage points in the logistic regression model.

CHAPTER 3

METHODOLOGY

3.1 Study Population and Sampling Techniques

The secondary data from 2015-2016 Malawi demographic Health Survey (MDHS) were used. The data were collected between October, 2015 and February, 2016 by National Statistical Office (NSO) of Malawi in conjunction with the Ministry of Health (MoH) and Community Health Services Unit (CHSU). Stratified sampling with selected two stages was used in 2015-2016 MDHS. In first stage, 850 Standard Enumeration Areas (SEAs), 173 SEAs from urban areas and 677 SEAs were selected with probability proportional to the SEA size and with independent selection in each sampling stratum. In the second stage, 30 households per urban cluster and 33 households per rural cluster were selected with an equal probability systematic selection from the newly created household listing.

The data for children were collected from the women aged 15-49 who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey. In this study, the sample size was 15112 children aged 0 to 59 months.

3.2 Geographic Location and Population Distribution

The data were collected in all 28 districts in Malawi. Malawi is a landlocked country in south-east Africa. Malawi is found in a land between Zambia and Mozambique and in the north shares a border with Tanzania (Kumbuyo, Yasuda, Kitamura, & Shimizu, 2014). The area of Malawi is $118,484km^2$ in which $94,276km^2$ is land and $24,208km^2$ is covered by water (Kaombe & Namangale, 2016). According to Malawi housing census survey 2018 the population of Malawi is 17, 563,749 people.

3.3 Statistical methods

The dependent variable used in this study was whether the child had diarrhoea 2 weeks before the survey. The independent variables were age of the child (0-59 months), sex of child, location (urban and rural), toilet facility (shared or not), mother's education, source of drinking water, region and wealth index. The variables were taken from the studies of (Kaombe & Namangale, 2016) and (Getachew et al., 2018) since there were found to be associated with child diarrhoea.

The binary logistic regression model given equation (1.1) was used in this study. This model was used because the dependent variable, total number of under-5 diarrhoea cases in past two weeks before the survey was following binomial distribution. Let $P(diarrhoea = Yes) = \theta_i$ be the probability of success, $P(diarrhoea = No) = 1 - \theta_i$ be the probability of failure and diarrhoea outcome Y on the ith child be Yes (coded 1) for presence of diarrhoea and No (coded 0) for absence

of diarrhoea. Then Y is Bernoulli variable. The total number of cases in one observation $\sum(Y = 1)$ was distributed as $Bernoulli(\theta)$, since the Y was Bernoulli variable with the probability of presence of diarrhoea θ. Since the number of children n was fixed, then total number of cases in n observations $\sum(Y = 1)$ was distributed as $binomial(n, \theta)$. The exponential form with natural parameter is defined by;

$$f(y; n, \theta) = \exp[\log \binom{n}{y} + n \log(1 - \theta) + y \log(\frac{\theta}{1 - \theta})] \tag{3.1}$$

θ which was the probability that a child had diarrhoea which was assumed to be dependent on some of the characteristics of the child. The assumption was made that the relationship between θ and the characteristics of the child x was non-linear, which was the logistic function given by

$$\theta(x) = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)} \tag{3.2}$$

where $\beta_0, \beta_1 ... \beta_p$ are coefficients of independent variables. Then the logistic function was put in linear form as

$$\log(\frac{\theta(x)}{1 - \theta(x)}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{3.3}$$

where

X_1 = Age of child (months)

X_2 = Region where the child is coming from

X_3 = If the household is sharing the toilet

X_4 = Sex of the child

X_5 = Breast feeding

X_6 = Wealth index

X_7 = Location

X_8 = Mothers education

X_9 = Source of drinking water

The maximum likelihood estimation was used to estimate the values of estimates of coefficients as per methods discussed in Section 1.2.

3.4 Computations of diagnosis statistics

Data analysis was done using STATA 15 and R 4.0.5. Participants with missing data were dropped and descriptive analysis was done, and percentages of each variable used in the study were reported. Pearson's chi-square test was performed to find if there was association between dependent variable and each independent variable. Binary logistic regression models were fitted in STATA to find the estimates and the AIC. The AIC was used to find the best model. The results from binary logistic regression models were reported in odds together with their correspondent 95% confidence interval.

To check for outliers, Pearson's residual, Deviance residuals and Studentized Pearson residual were used. First, estimated logistic probabilities were computed

using STATA and diagonal elements of hat matrix were computed using R. Pearson residual was calculated using this formula;

$$pr_i = \frac{y_i - \hat{\theta}_i(X)}{\sqrt{\hat{\theta}_i(X)(1 - \hat{\theta}_i(X))}} \tag{3.4}$$

where $\hat{\theta}_i(X)$ are the fitted values and y_i are observed values. Then the graph of Pearson residual against estimated logistic probability was plotted to assess if there were outliers. The points which were greater than 3 were regarded as outliers. These were dropped and the logistic regression was fitted again to data to observe the impact of outliers.

Deviance residual was also used in the study to check for outliers. Deviance residual was calculated using this formula;

$$d_i = sign(y_i - \hat{\theta}_i(X))\{-2[y_I \log \hat{\theta}(X) + (1 - y_i)log(1 - \hat{\theta}_i(X))]\}^{1/2} \tag{3.5}$$

Then the graph of deviance residuals against predicted probabilities was plotted. The cases which had values greater than 2 were regarded as outliers and were later dropped. The logistic regression was fitted again without those outliers.

Studentized residual was also used to check for outliers. Studentized residual was calculated using this formula;

$$Sp_i = \frac{y_i - \hat{\theta}_i(X)}{\sqrt{\hat{\theta}_i(X)(1 - \hat{\theta}_i(X))(1 - h_{ii})}} = \frac{pr_i}{\sqrt{1 - h_{ii}}} \tag{3.6}$$

Then graph of studentized residuals against predicted probabilities was plotted. The cases which had values above 3 were dropped because they were regarded as outliers. Then, the binary logistic regression was fitted again without those outliers.

To check for multicollinearity, Variance Inflation Factor(VIF) was used. For VIF, every independent variable was fitted as dependent against remaining independent variables to calculate R^2. For independent variables with 2 possible outcomes, binary logistic regression model was used and independent variables with more than 2 possible outcomes multi-level logistic regression model was fitted. After getting R^2s tolerance was calculated using this formula

$$Tolerance = 1 - R^2. \tag{3.7}$$

After getting Tolerance for each independent variable then VIFs were calculated using this formula:

$$VIF = \frac{1}{Tolerance} = \frac{1}{1 - R^2}. \tag{3.8}$$

After getting VIFs, then the results were interpreted.

To check for influential points on $\hat{\beta}$ Cook's distance was used. Cook's distance was calculated using this formular;

$$CD_i = \frac{(spr_i)^2 h_{ii}}{1 - h_{ii}}. \tag{3.9}$$

The graph of Cook's distance and index was plotted. Using CD_i, the cases with values greater than 1 were regarded as influential points on $\hat{\beta}$.

Dffits were also used to detect influential points. The dffits were calculated using the formula;

$$DFFITS_i = spr_i \sqrt{\frac{h_{ii}\hat{\theta}_i(1-\hat{\theta}_i)}{(1-h_{ii})[\hat{\theta}_i(1-\hat{\theta}_i)]^{-i}}} \tag{3.10}$$

Then the graph of Dffits and cases was plotted. Cases with values greater than 2 were regarded as influential points.

Diagnostic for leverage was also necessary in this study, so Pregibon's leverage was used. Pregibon's leverage was computed using this formula;

$$h_{ii} = \hat{\theta}_i(1-\hat{\theta}_i)x_i^T(X^TVX)^{-1}x_i \tag{3.11}$$

Then, the graph of Pregibon's leverage and index was plotted. Cases which had values greater than 0.0008 were regarded as high leverage points and were dropped. Then, the binary logistics regression was fitted again without those high leverage points in order to determine if the removed leverage points had effects on coefficients and the confidence intervals.

CHAPTER 4

RESULTS

4.1 Descriptive Analysis

There were 15,112 children that were included in this study, 11,976(79%) children had no diarrhoea and 3,136 (21%) children had diarrhoea in the last 2 weeks before the time of the survey. The distribution of cases is summarized in **Table 4.1**. It was shown that the cases were high in breastfeeding children, southern region and rural areas. Also, Pearson chi-square test of association results showed that breast feeding and all other factors were associated with diarrhoea except place of residence and source of drinking water. Place of residence and source of drinking water were not associated with childhood diarrhoea because their p-values were greater that 0.05. The remaining variables were associated with childhood diarrhoea because their p-values were less than 0.05.

Table 4.1: Social demographic characteristics of respondents, sample size= 15,112

VARIABLE	**CASES (%) 3136 (21 %)**	**CHI-SQUARE P-VALUE**
BreastFeeding		< 0.001
No	1,183 (17)	
Yes	1,953 (24)	
Region		< 0.001
North	475 (17)	
Central	1,221 (23)	
Southern	1,440 (21)	
ToiletShared		< 0.001
No	1,833 (19)	
Yes	1,303 (24)	
Sex		< 0.001
male	1,665 (22)	
female	1,471(19)	
Wealth Index		< 0.001
poorest	692 (23)	
poorer	710 (22)	
middle	613 (20)	
richer	604(20)	
richest	517 (18)	
Location		0.5
urban	538 (21)	
rural	2,598(21)	
Age (months)		< 0.001
0-12	917(27)	
13-24	1,013(34)	
25-36	601(20)	
37-48	385(13)	
49 +	222(8)	
Mother Education		< 0.001
no education	307 (17)	
primary	2,177(22)	
secondary	611 (20)	
tertiary	41(15)	
Source of water		0.45
piped water	693 (21)	
protected wells	2,013(24)	
unprotected wells	255(13)	
surface water	175 (12)	

4.2 Logistic regression estimation results

AIC is a mathematical method of finding a best model. The smaller the value of AIC, the better the model. The AIC is calculated from number of covariates in the model and the maximum likelihood estimate of the model. Model 1 had AIC of 14542.66, Model 2 had AIC of 14542.83, Model 3 had AIC of 14542.49 and Model 4 had AIC of 14537.63. The results from AIC showed that Model 4 was best model since it had lowest AIC among all 4 models. The table below shows the logistic regression estimation results for all 4 models. The model 1 used all independent variables for the study and model 2 source of drinking water was dropped since it was less significant. Breast feeding and location were dropped for model 3 and mother education and wealth index were dropped for model 4.

From **Table 4.2**, the results showed that the significant predictors of child diarrhoea were; region where the child is coming from, toilet shared, sex of the child and age of the child. The results showed that children from the central region and southern region were more likely to catch diarrhoea than children from the northern region. The children from the central region had 52% higher odds than children from northern region. The children from southern region had 30% higher odds than children from northern region.

The results showed that the children from families that were sharing toilet were more likely to catch diarrhoea than children from families that were not sharing the toilet. The children from the families that were sharing the toilet had 31% higher odds compared to children from families that were not sharing the toilet.

Females were less likely to catch diarrhoea compared to males. The odds of catching diarrhoea were 15% lower for females compared to males.

The odds of catching diarrhoea in children of age (13-24) months were 41 % higher compared to the children with age of (0-12) months. The odds were lower by 32% in children with age of (25-36) months, lower by 61% in children with age of (37-48) months and lower by 76% in children with age of (49+) months compared to children with age of 0-12 months.

Table 4.2: Logistic regression model estimates results

Child Characteristics	**Model 1 OR(95% CI)**	**Model 2 OR(95% CI**	**Model 3 OR(95% CI**	**Model 4 OR(95% CI**
BreastFeeding				
No				
Yes	0.93(0.83-1.06)	0.93(0.84-1.03)		
Region				
North				
Central	1.57(1.35-1.73)	1.51(1.33-1.71)	1.52(1.36-1.72)	1.52(1.36-1.72)
Southern	1.31(1.16-1.48)	1.29(1.15-1.46)	1.31(1.16-1.40)	1.31(1.16-1.40)
ToiletShared				
No				
Yes	1.27(1.18-1.39)	1.27 (1.17-1.39	1.29(1.18-1.39)	1.29(1.18-1.39)
Sex				
male				
female	0.85(0.78-0.92)	0.85(0.78-0.92	0.85(0.78-0.92)	0.85(0.78-0.92)
Wealth Index				
poorest				
poorer	0.98(0.86-1.11)	0.98(0.86-1.10)	0.98(0.86-1.10)	
middle	0.91(0.80-1.03)	0.91(0.80-1.03)	0.91(0.80-1.04)	
richer	0.98(0.85-1.12)	0.97(0.85-1.10)	0.99(0.87-1.13)	
richest	0.84(0.72-0.99)	0.80(0.67-0.94)	0.86(0.74-0.101)	
Location				
urban				
rural	0.87(0.75-1.01)	0.87(0.78-1.01)		
Age (months)				
0-12				
13-24	1.39(1.25-1.55)	1.39(1.25-1.55)	1.39(1.25-1.55)	1.39(1.25- 1.55)
25-36	0.64(0.56-0,74)	0.65(0.56-0,74)	0.65(0.56-0,74)	0.65(0.56-0,74)
37-48	0.38(0.33-0.44)	0.38(0.33-0.44)	0.38(0.33-0.44)	0.38(0.33-0.44)
49 +	0.23(0.20-0.27)	0.23(0.20-0.28)	0.23(0.20-0.28)	0.23(0.20-0.28)
Mother Education				
no education				
primary	1.34(1.17-1.54)	1.34(1.17-1.54)	1.40(0.98-1.50)	
secondary	1.25(1.06-1.47)	1.25(1.06-1.47)	1.29(0.95-1.40)	
tertiary	0.93(0.63-1.36)	0.93(0.63-1.36)	0.95(0.59-1.47)	
Source of water				
piped water				
protected wells	0.98(0.88-1.11)			
unprotected wells	0.97(0.81-1.17)			
surface water	0.95(0.78-1.17)			
AIC	14542.66	14542.83	14542.49	14537.63

4.3 Results for outliers from fitted model

The methods which were used in this study for checking the presence of outliers were; Pearson's residual, Deviance residual and Studentized Pearson residual (Table 4.3). The good method of looking at outliers is by graphing residuals against index or predicted probabilities. The residual plots showed 2 trends because the residuals are defined as $1-\hat{\theta}_i$ for Y=1 and $-\hat{\theta}_i$ for Y=0. The results are further discussed in the following subsection. Table 4.3 showed Index represented as Child i, Residual, Pearson Residual, Leverage, Studentized Residual and Deviance. The results of Table 4.3 were interpreted using Figures below.

Table 4.3: Estimates of Fitted values($\hat{\theta}_i$), Residual(ε_i), Pearson residual (pr_i), Leverage(h_{ii}), Studentized residual(spr_i) and Deviance residual (dr_i) upon fitting binary logistic regression model to diarrhoea data

Child i	Y_i	$\hat{\theta}_i$	ε_i	pr_i	h_{ii}	spr_i	dri
1	0	0.07856	−0.07856	−0.29199	0.00046	−0.29205	−0.03178
2	0	0.36951	−0.36951	−0.76554	0.00067	−0.76580	−0.35470
3	0	0.17671	−0.17671	−0.46330	0.00070	−0.46346	−0.11020
4	0	0.07856	−0.07856	−0.292	−0.292	0.00046	−0.29206
5	0	0.36951	−0.36951	−0.76554	0.00067	−0.7658	−0.3549
6	0	0.122	−0.122	−0.37274	0.00047	−0.37283	−0.06223
7	0	0.21908	−0.21908	−0.52967	0.00059	−0.52982	−0.15407
8	0	0.19249	−0.19249	−0.48823	0.00054	−0.48837	−0.12587
9	0	0.21908	−0.21908	−0.52967	0.00059	−0.52982	−0.15407
10	1	0.26071	0.73929	1.68394	0.00057	1.68442	1.21222
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
15103	0	0.21441	−0.21441	−0.52244	0.00074	−0.52263	−0.14896
15104	1	15576	0.84424	2.32815	0.00066	2.32892	1.62808
15105	0	0.0619	−0.0619	−0.25688	0.00045	−0.25694	−0.02213
15106	0	0.08372	−0.08372	−0.30227	0.00047	−0.30234	−0.035
15107	0	0.27819	−0.27819	−0.62081	0.00087	−0.62108	−0.22462
15108	0	0.13551	−0.13551	−0.39593	0.00057	−0.39604	−0.07313
15109	0	0.10709	−0.10709	−0.34631	0.00062	−0.34642	−0.05097
15110	0	0.061903	−0.0619	−0.25688	0.00045	−0.256939	−0.02213
15111	0	0.27819	−0.27819	−0.62081	0.00087	−0.62108	−0.22462
15112	0	0.0619	−0.0619	−0.25688	0.00045	−0.25694	−0.02213

4.3.1 The Pearson's Residual results

Figure 4.1 shows the graph of Pearson's residuals against predicted probabilities. The results showed that there were some potential outliers since some of the cases were found outside the range of ±3. 215 (1.42 %) cases out of 15112 cases.

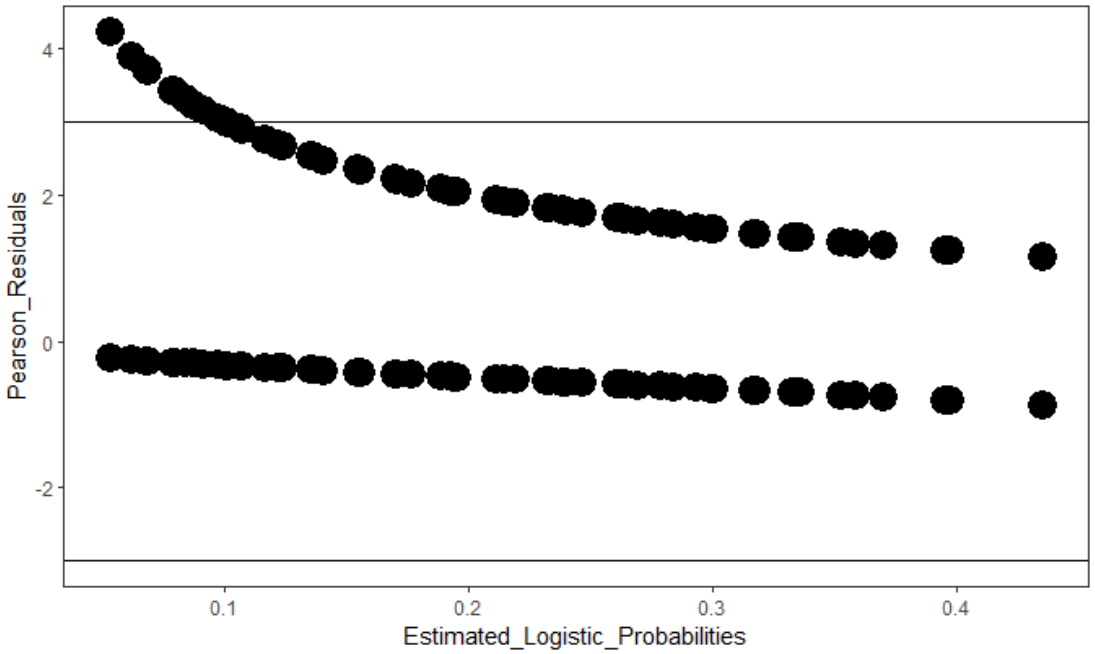

Figure 4.1: Pearson's residuals vs Estimated logistic probability

4.3.2 Re-fitted model model estimates upon removing outliers detected by Pearson residual

Table 4.4 shows the estimation of the binary logistic regression model. Comparison was made between the fitted model (with outliers) and re-fitted model (without outliers) using odds ratios and the width of 95% confidence intervals. The significance of independent variables was not affected by dropping of outliers confidence intervals did not include the value 1. The width of confidence interval is difference between the largest number and lowest number of the interval.

The odds ratio of children catching diarrhoea increased for children from central region, children from southern region, children from families that shared toilet and children with age of (13-24) months. While the odds ratio of catching diarrhoea decreased for female children, children with age of (37-48) months and children with age of (49+) months. The odds did not change for children with age of (25-36) months.

The results showed that the width of 95% CI increased for children from central region, children from southern region, children from families that were sharing toilets and children with age of (13-24) months. The width of 95% CI did not change for female children, children with age (25-36) months and children with age of (37-48) months. The width of 95% decreased for children with (49+) months.

The change of odds ratios and the width of 95% confidence intervals showed that, the outliers had affected the estimates of the model.

Table 4.4: Logistic regression model 4 estimates after removing outliers

Child characteristics	**Model 4 (with full data) OR(95% CI)**	**Model 4 (without outliers) OR(95% CI)**
Region		
North*		
Central	1.52(1.36-1.72)	1.82 (1.60-2.07)
Southern	1.31(1.16-1.40)	1.50 (1.32-1.70)
Toilet shared		
No*		
Yes	1.29(1.18-1.39)	1.44 (1.32 - 1.57)
Sex		
male		
female	0.85(0.78-0.92)	0.83 (0.77-0.91)
Age (months)		
0-12*		
13-24	1.39(1.25- 1.55)	1.42 (1.27- 1.58)
25-36	0.65(0.56-0,74)	0.68 (0.6-0.76)
37-48	0.38(0.33-0.44)	0.35 (0.30-0.40)
49+	0.23(0.20-0.28)	$0.05(0.04 - 0.07)$

4.3.3 Deviance Residual results

Deviance residuals were plotted against the predicted probabilities. **Figure 4.2** showed that there were cases that were outside the range of 2 173(1.4%) indicating the presence of outliers.

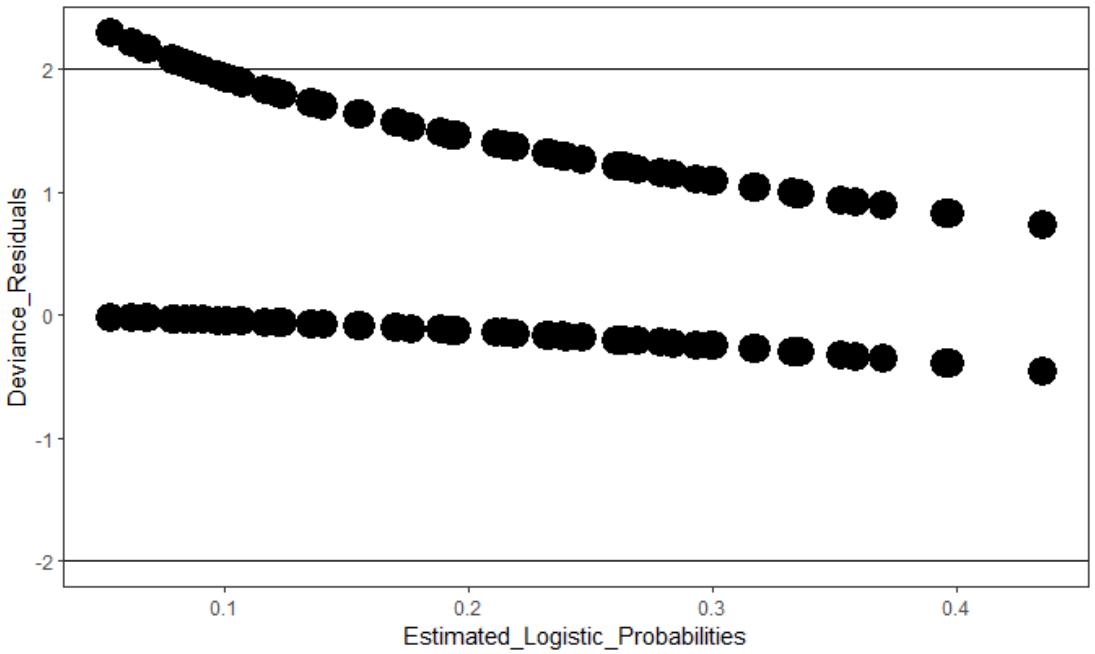

Figure 4.2: *Deviance Residual Plot*

4.3.4 Re-fitted model estimates upon removing outliers detected by deviance residual

Table 4.5 shows the estimates of the binary logistic regression after dropping outliers identified by deviance residual. The results of the re-fitted model were compared with results of the fitted model. The odds of children catching diarrhoea increased for children from central region, children from southern region, children from families that shared toilet and children with age of (13-24) months. While the odds of catching diarrhoea decreased for female children, children with age of (37-48) months and children with age of (49+) months. The odds did not change for children with age of (25-36) months.

The results showed that the width of 95% CI increased for children from central region, children from southern region, children from families that were sharing toilets and children with age of (13-24) months. The width of 95% CI did not

change for female children, children with age (25-36) months and children with age of (37-48) months. The width of 95% decreased for children with (49+) months.

The change of odds ratios when the outliers were removed meant that the outliers were affecting the estimates of the model.

Table 4.5: Logistic regression model 4 estimates after removing outliers

Child characteristics	**Model 4 (with full data) OR(95% CI)**	**Model 4 (without outliers) OR(95% CI)**
Region		
North*		
Central	1.52(1.36-1.72)	1.78 (1.56-2.02)
Southern	1.31(1.16-1.40)	1.42(1.26-1.62)
ToiletShared		
No*		
Yes	1.29(1.18-1.39)	1.40 (1.28 - 1.52)
Sex		
male		
female	0.85(0.78-0.92)	0.81 (0.74-0.88)
Age (months)		
0-12*		
13-24	1.39(1.25- 1.55)	1.42 (1.27- 1.58)
25-36	0.65(0.56-0,74)	0.68 (0.6-0.76)
37-48	0.38(0.33-0.44)	0.37 (0.32-0.42)
49+	0.23(0.20-0.28)	0.08 (0.06-0.1)

4.3.5 Studentized Pearson residual results

Figure 4.3 shows the graph of Studentized Pearson residuals against estimated logistic probabilities. The graph shows that there were some cases that were outside the range of 3 indicating the presence of outliers. From the **Table 4.3** above, 214 (1.42%) cases were outliers out of 15112 cases.

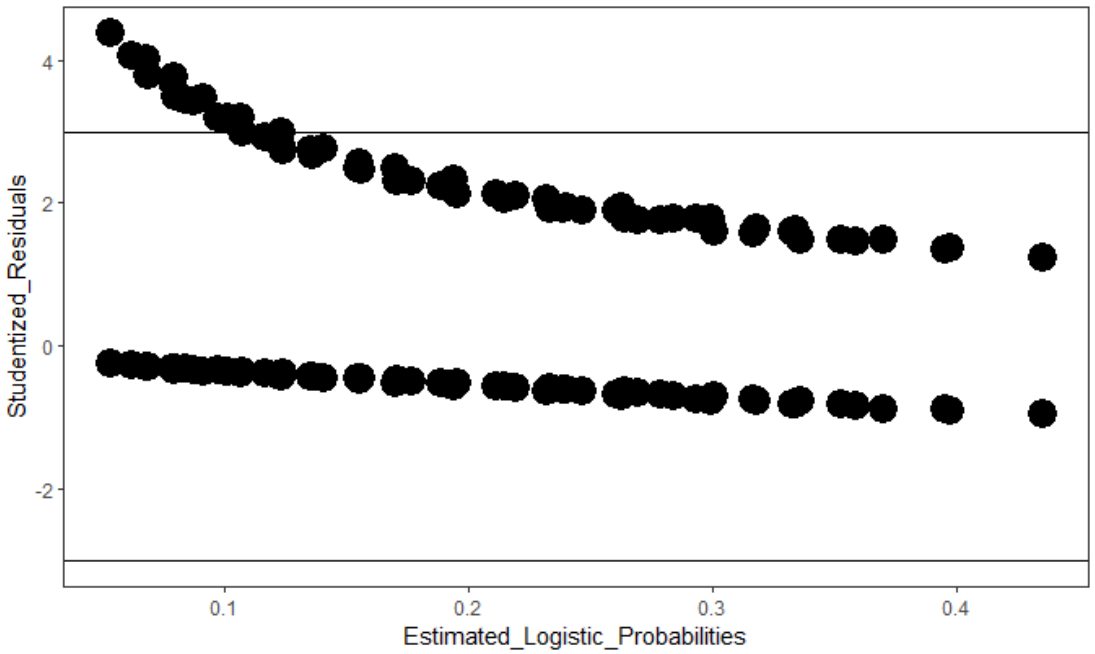

Figure 4.3: *Studentized pearson residuals*

4.3.6 Re-fitted model estimates upon removing outliers detected by studentized Pearson residual

Table 4.6 shows the estimates of the binary logistic regression. There was a change in odds ratios and the width of 95% CI upon removing outliers in the model. There was an increase of odds of catching diarrhoea for children from central region, children from southern region, children from families that shared toilet and children with age of (13-24) months. There was a decrease of odds of catching diarrhoea for female children, children with age of (37-48) months and children with age of (49+) months. The odds did not change for children with age of (25-36) months.

The results also showed that there was an increase of width of 95% CI for children from central region, children from southern region, children from families that were sharing toilets and children with age of (13-24) months. The width of 95%

CI did not change for female children, children with age of (25-36) months and children with age of (37-48) months. The width of 95% CI decreased for children with age of (49+) months.

The change of width of confidence intervals and odds ratios showed that the outliers had affected estimated odds.

Table 4.6: Logistic regression model 4 estimates after removing outliers

Child characteristics	**Model 4 (with full data) OR(95% CI)**	**Model 4 (without outliers) OR(95% CI)**
Region		
North*		
Central	1.52(1.36-1.72)	1.82 (1.59-2.02)
Southern	1.31(1.16-1.40)	1.5 (1.32-1.70)
Toilet shared		
No*		
Yes	1.29(1.18-1.39)	1.44 (1.32 - 1.57)
Sex		
male		
female	0.85(0.78-0.92)	0.83(0.77-0.91)
Age (months)		
0-12*		
13-24	1.39(1.25- 1.55)	1.42 (1.27- 1.58)
25-36	0.65(0.56-0,74)	0.68 (0.6-0.76)
37-48	0.38(0.33-0.44)	0.35 (0.30-0.40)
49+	0.23(0.20-0.28)	0.05(0.04-0.07)

In summary, the results from Pearson's residual showed that 215 cases were outliers, Deviance residual showed that 173 cases were outliers and Studentized Pearson residual showed that 214 cases were outliers. 215, 173 and 214 cases were mixture of same individuals and new individuals. The re-fitted models after removing outliers showed that there was a change in odds ratios and width of 95% confidence interval.

4.4 Results for Multicollinearity

To check if independent variables were not dependent on each other, Variance Inflation Factor (VIF) was used.

4.4.1 Variance Inflation Factor VIF

To check for multicollinearity using VIF, every independent variable was fitted as dependent against remaining independent variables to calculate R^2s. After getting R^2s, Tolerance was calculated. Then VIF was calculated by finding the reciprocal of Tolerance. Table below summarizes the results. From **Table 4.7**, VIF of every variable used in the study was 1 which is less than 10 that indicated that there was absence of multicollinearity.

Table 4.7: Results of R-squared, Tolerance and VIF

VARIABLE	R-SQUARED	TOLERANCE	VIF
Sex	0.0004	0.9996	1
age	0.0005	0.9995	1
Region	0.0016	0.9984	1
Toilet Shared	0.0025	0.9975	1

4.5 Results for influence of individual children on $\hat{\beta}$ and $\hat{y}_i$

The methods which were used in this study for checking influential points were Cook's distance, Dfbetas and Dffits. The table below summarizes the results of Cook's distance and Dffits.

Table 4.8: Influence statistics and leverage results

i	Y_i	h_{ii}	spr_i	CD_i	$Dffits_i$
1	0	0.00047	-0.29206	3.89907E-05	-0.00882
2	0	0.00067	-0.76580	0.00039	-0.02529
3	0	0.00070	-0.46346	0.00015	-0.01680
4	0	0.00046	-0.29206	3.89907E-06	-0.00882
5	0	0.00067	-0.76580	0.00039	-0.02529
6	0	0.00047	-0.37283	6.54621E-05	-0.01129
7	0	0.00059	-0.52982	0.00017	-0.01745
8	0	0.00054	-0.48837	0.00013	-0.01552
9	1	0.00059	-0.52982	0.00017	-0.01745
10	1	0.00057	1.68442	0.00017	0.04007
⋮	⋮	⋮	⋮	⋮	⋮
15103	0	0.00074	-0.52263	0.0002	-0.01933
15104	1	0.00066	2.32892	0.00359	0.05064
15105	0	0.00045	-0.25693	3.00443E-05	-0.00777
15106	0	0.00047	-0.30234	4.31522E-05	-0.00927
15107	0	0.00087	-0.62108	0.00037	-0.02432
15108	0	0.00060	-0.39604	9.36712E-05	-0.01345
15109	0	0.00062	-0.34642	7.41666E-05	-0.01207
15110	0	0.00045	-0.25694	3.00443E-05	-0.00778
15111	0	0.00087	-0.62108	0.00034	-0.02432
15112	0	0.00045	-0.25694	3.00443E-05	-0.00778

4.5.1 Cook's Distance

Cook's distance is used to assess influential points. **Table 4.8** shows Cook's distance represented by CD_i. **Figure 4.4** shows the graph of Cook's distance vs Estimated probabilities. The graph shows that there was no point that was greater than 1. This meant that there was no influential point on $\hat{\beta}$. This contradicts estimates from re-fitted model upon removing outliers in previous section. However, generalized Cook's distance estimated general influence of a child on all parameters. While the estimates from refitted model were per each independent variable. In addition, the Cook's distance values estimate influence of individual child on regression parameters, whereas the refitted models were estimating joint influence of a group of outlier children on the same regression parameters.

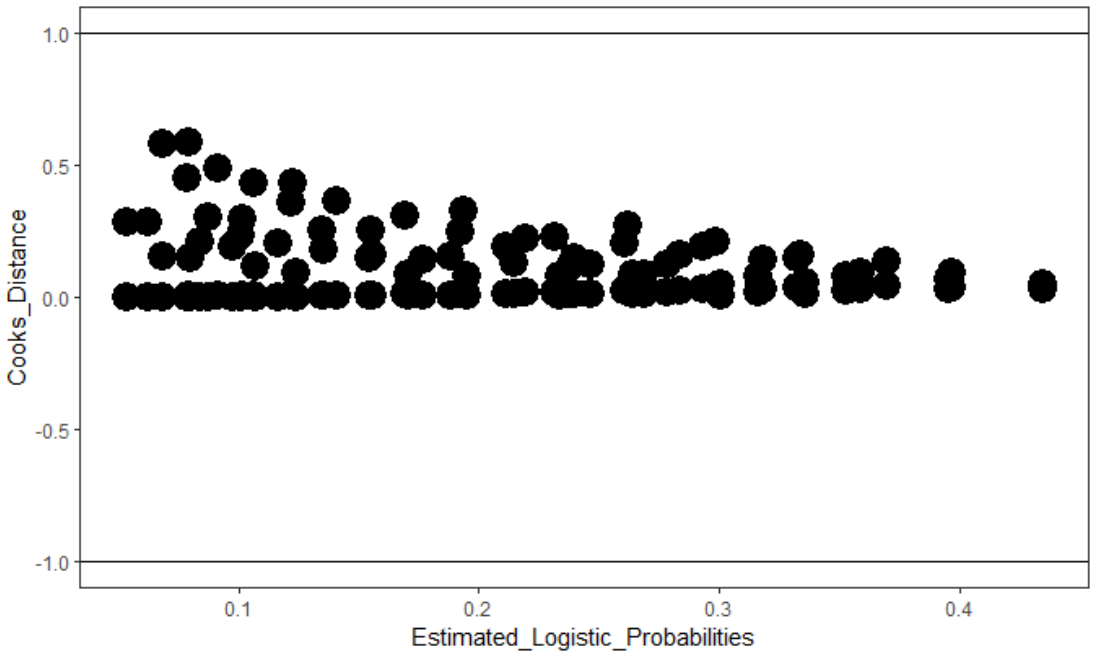

Figure 4.4: *Cook's distance vs index*

4.5.2 Dffits

Dffits were also used to detect influential points. **Table 4:8** shows the summary of dffits. The dffits were plotted against the estimated probabilities and the results are shown in **Figure 4.5** below. The graph showed that there was no point that was greater than 2. This meant that there was no influential points.

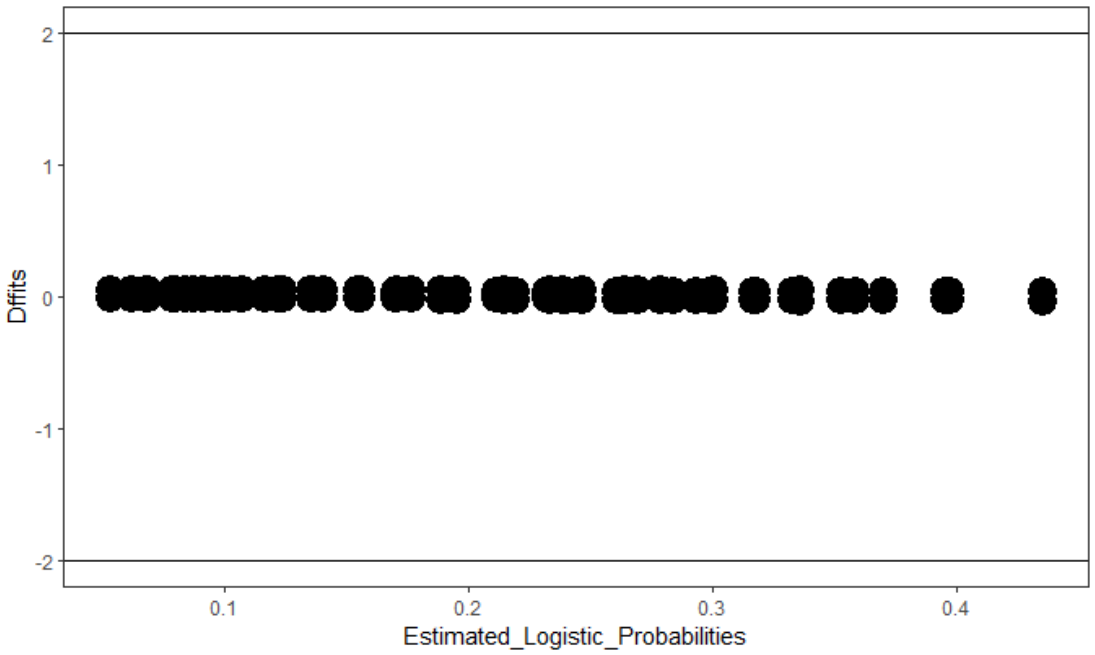

Figure 4.5: *Dffits vs index*

4.5.3 DFBetas for Covariates

Dfbetas were used to detect influential points. The Dfbetas for each independent variable were plotted against Index.

4.5.3.1 ***Sex***

The graph below shows that there was no points outside the range of 0.01 and -0.01. This meant that there was no influential points

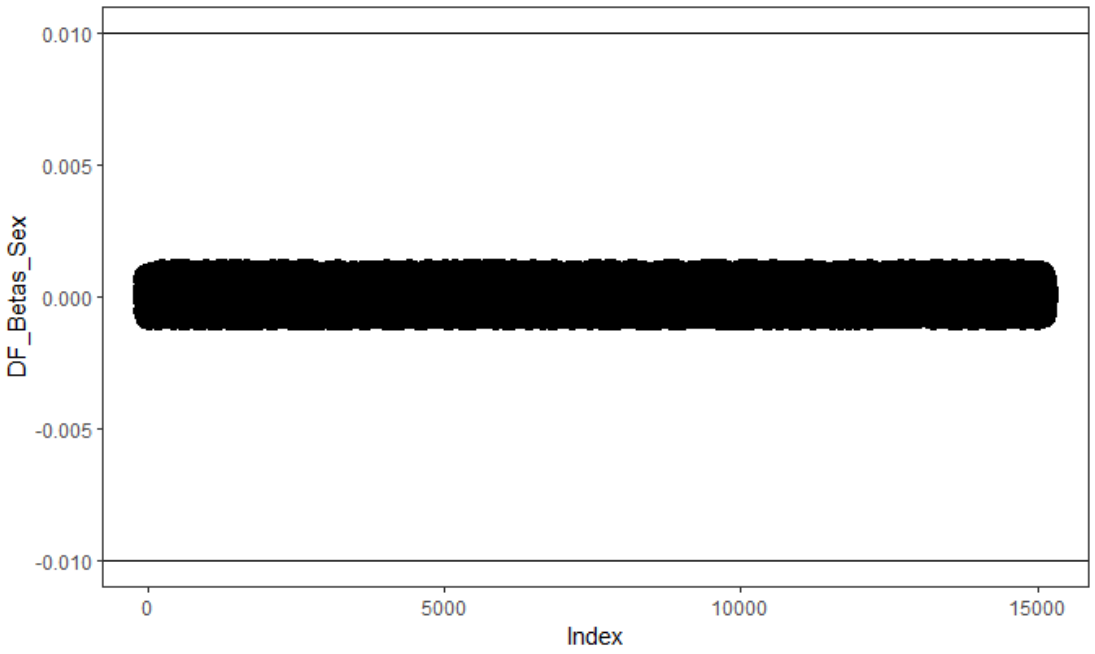

Figure 4.6: *DFbetas vs index*

4.5.3.2 ***Region***

The results were similar for Dfbetas for region. The **Figure 4.7** below showed that there were no influential points since there were no point outside the range of 0.01 and -0.01.

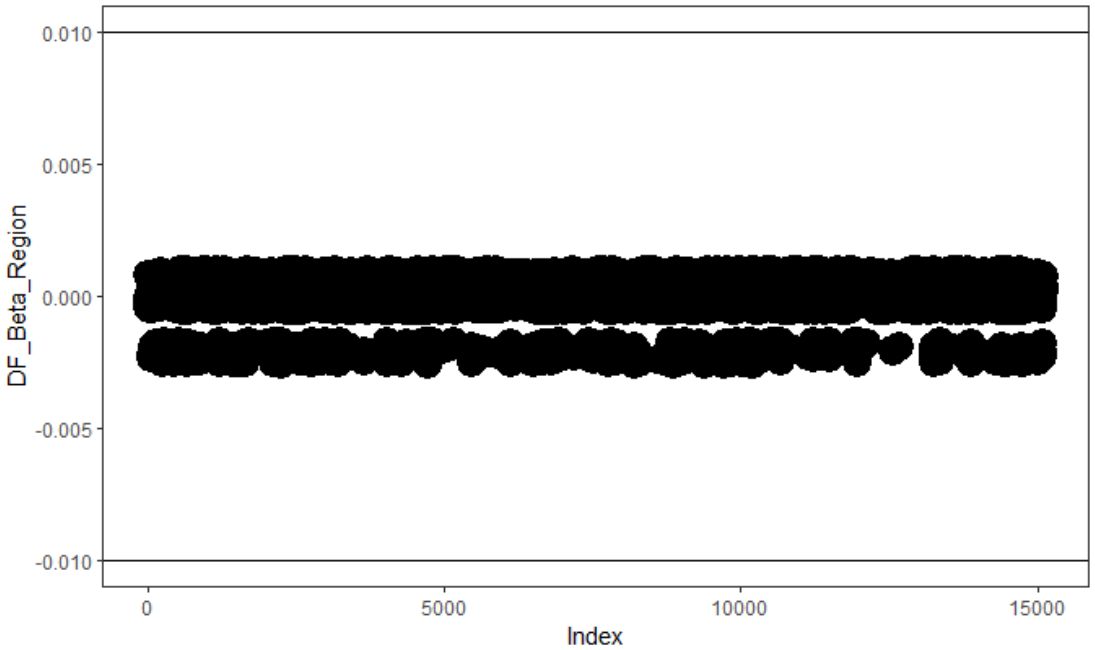

Figure 4.7: *DFbetas vs index*

4.5.3.3 ***Toilet Shared Variable***

Figure 4.8 shows that there was not influential points since there was no point which was outside the boundaries of 0.01 and -0.01.

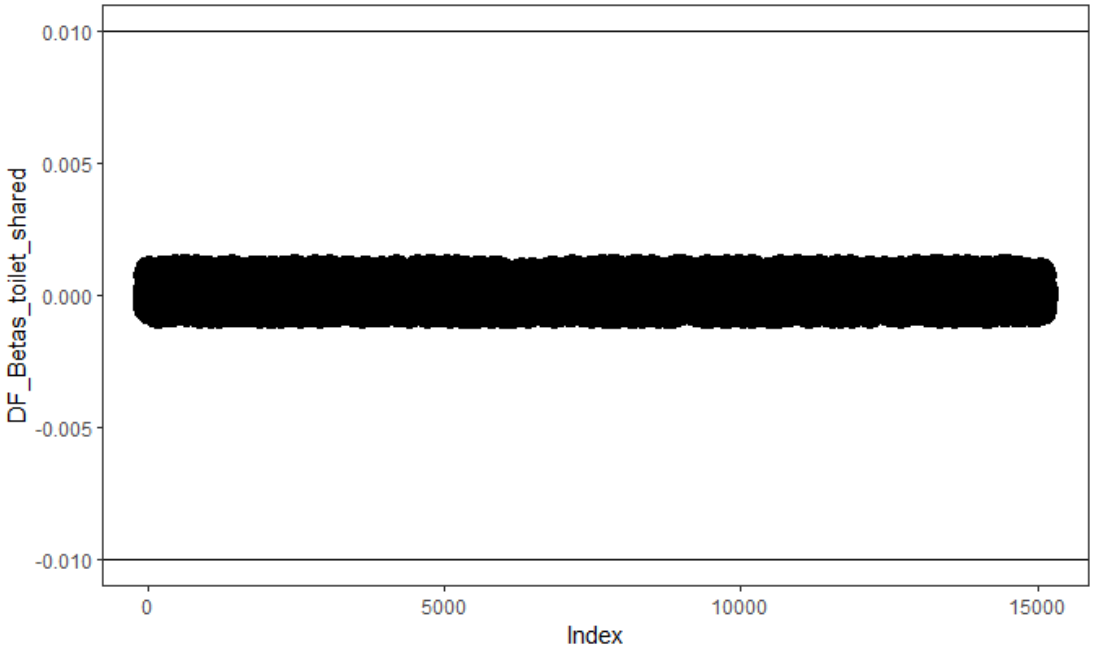

Figure 4.8: *DFbetas vs index*

4.5.3.4 ***Age***

The graph of the Dfbetas vs Index below shows that there was no point outside the boundary of 0.01 and -0.01. This meant that there were no influential points in the model.

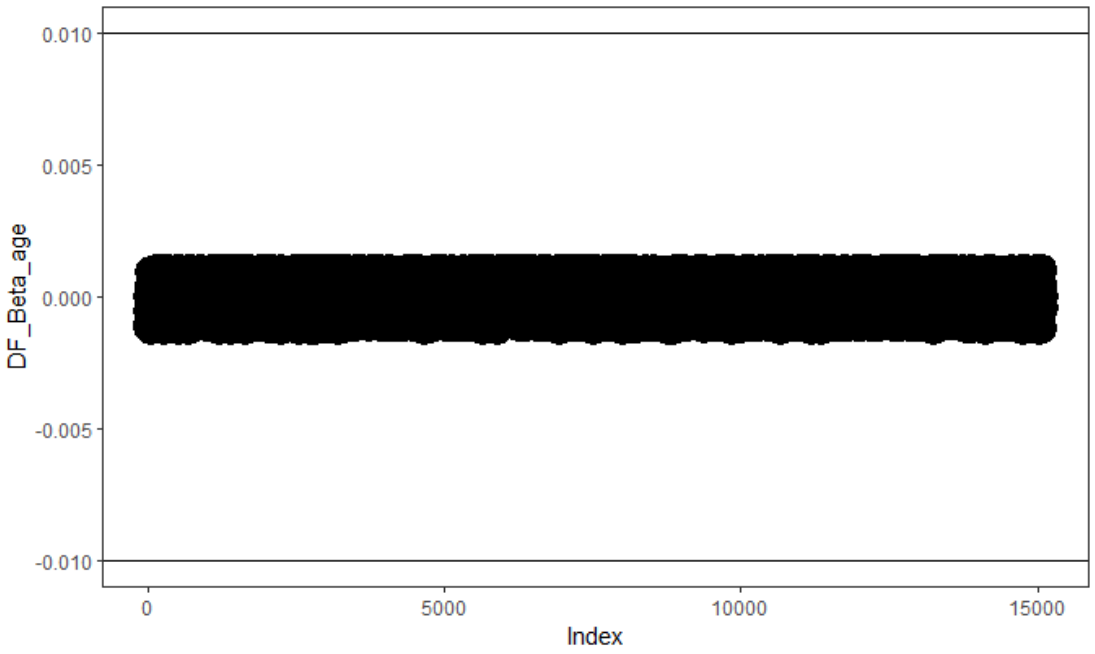

Figure 4.9: *DFbetas vs index*

In summary, results from Cook's distance, Dfbetas and Dffits showed that there were no individual influential points in the model.

4.6 Results for leverages of children on fitted values

Pregibon's leverage also known as hat diagonal was used to detect high leverage points. **Figure 4.10** below shows the Pregibon's leverage vs index. The graph showed that there were points which were above 0.0008 which meant that there were high leverage points. 1,284 points were found to be high leverage points using Pregibon's leverage. These high leverage points were dropped from the data and the logistic model was fitted again to see if these points were affecting the results.

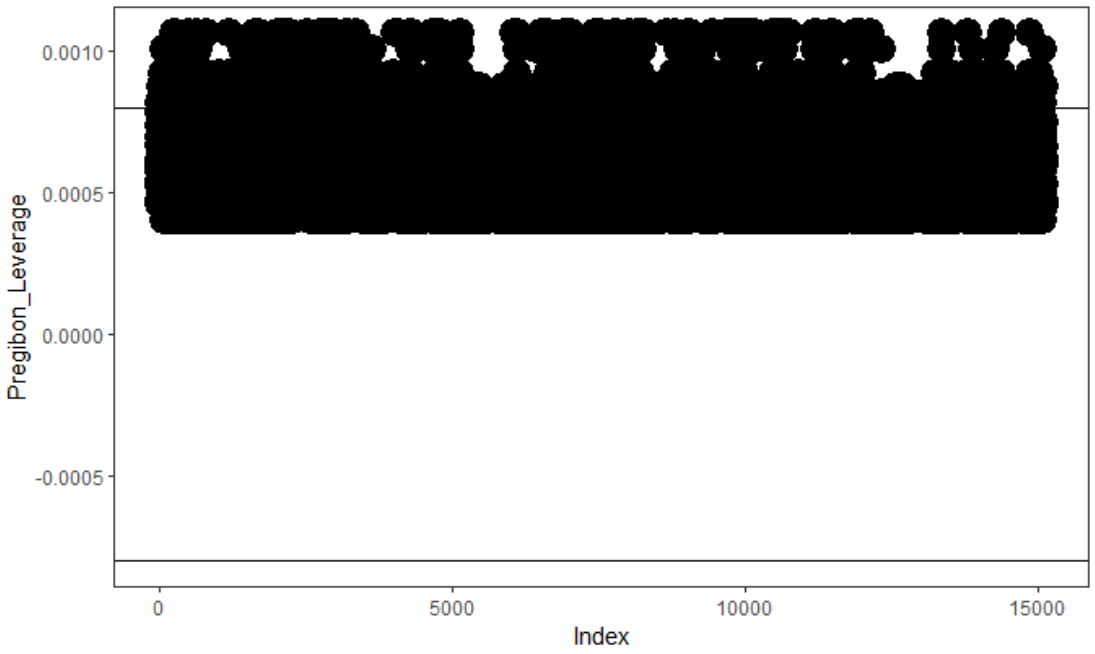

Figure 4.10: *Pregibon leverage vs index*

4.6.1 Refitted model estimates after dropping high leverage points

Table 4.9 shows the estimates of binary logistic regression after dropping high leverage points. There was a change in odds and the width of 95% CI upon removing high leverage points. The odds of catching diarrhoea decreased for children from central region, children from southern region, children from families that shared toilet and female children. There was an increase of odds of catching diarrhoea for children in all age groups.

The results shows that there was an increase of width of 95% confidence interval for children from central region, children from southern region, children from families that shared toilets, children with age of (13-24) months, children with age of (37-48) months and children with age of (49+) months. The width of 95% confidence interval did not change for female children and children with age of (25-36) months.

High leverage points had jointly greater influence on the estimated regression parameter values that why the odds and the width of 95% CI were changing upon removing those points.

Table 4.9: Logistic regression model estimates results after removing leverage points

Child characteristics	**Model 4 (with full data) OR(95% CI)**	**Model 4 (without outliers) OR(95% CI)**
Region		
North*		
Central	1.52(1.36-1.72)	1.50 (1.27-1.75)
Southern	1.31(1.16-1.40)	1.26 (1.07-1.47)
ToiletShared		
No*		
Yes	1.29(1.18-1.39)	1.30 (1.19 - 1.43)
Sex		
male		
female	0.85(0.78-0.92)	0.84 (0.77-0.91)
Age (months)		
0-12*		
13-24	1.39(1.25- 1.55)	1.52 (1.34- 1.72)
25-36	0.65(0.56-0,74)	0.70 (0.62-0.78)
37-48	0.38(0.33-0.44)	0.40 (0.35-0.46)
49+	0.23(0.20-0.28)	0.25 (0.21-0.29)

CHAPTER 5

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1 Discussion

The width of 95% confidence intervals were increasing because the removal of outliers increased the standard deviations of some of the parameter estimates and the width of 95% confidence intervals were decreasing because the removal of outliers decreased the standard deviations of some of the parameter estimates. The odds were increasing or decreasing because the removal of outliers were jointly influential on the parameter estimates. The reasons why outliers were present in this study could be due to human errors such as keypunch errors, malfunction of instruments or due to natural deviations from the population. Although the study used few independent variables, this does not cause outliers to be present in the study. The study also showed that Studentized Pearson residuals performed well compared to Deviance residuals and Pearson's residuals since the coeffiecient of determination was bigger in model where Studentized Pearson residual was used compared to models where Deviance residual and Pearson's residual were used.

This study used deleting as a method of dealing with outliers but they are many ways of treating outliers. Changing the value of outliers is another way of treating outliers. This is done by changing the values of outliers to something more representative of your data set (Cousineau & Chartier, 2010). The disadvantage of this method is that it is difficult to apply this method when you have a lot outliers. Another method of dealing with outliers is by using non maximum likelihood estimation such as Lasso (Peng, Luo, & Gao, 2022). In this method outliers are removed and replaced with interpolated values. The advantage of using this method is that there is no decrease of sample size. Another method of dealing with outliers is transformation of the regression both sides. Using methods like robust regression can help in minimising the effect of outliers by giving little weight to extremely wild outliers (Jung, Lee, & Hu, 2016).

The multicollinearity was not found in this study because there were not a lot of independent variables and also there was no independent variable which was calculated from other independent variables. The results were not agreeing with a study done by (Amare, Ahmed, & Mehari, 2019) which was looking at determinants of nutrition status among children under 5, multicollinearity was found. Some of the independent variables which were used were region, sex of child and age of child. Using variance inflation factor, multicollinearity was found among independent variables such as birth interval variable was then removed from the binary logistic regression model.

Influential points were not found in the model because individual Cook's distance

and Dffits were used to check for all coefficients used in the study. If Cook's distance and Dffits were used to check for each coefficient, one at a time, influential points could be found. The results were agreeing with a study done by (Nurunnabi et al., 2009) which looked at identification of multiple influential observations in logistic regression model. The dependent variable was post-operative deformity (Kyphosis) which had 2 levels namely present and absent in children and the independent variables were the number of vertebrae involved in the operation and the beginning of the range of vertebrae involved in the operation. Using the Cook's distance showed that there were no influential points. Dffits were used in a study done by (MESTAV, 2019) which focused on detection and diagnostic methods of multiple influential points in binary logistic regression model in animal breeding. The outcome variable was coded as 1 for lambed and 0 for un-lambed in relation to fertility rate and the independent variables were weaning weight, fleece weight and yearling weight. Dffits showed that there were no influential points since all cases were less than 2.

The model was also checked for the presence of high leverage subjects using Pregibon's leverage. The results showed that there were high leverage subjects in the model. By comparing the fitted model with high leverage subjects and re-fitted model without high leverage subjects, the results showed that there was a change in odds and width of 95% confidence intervals. High leverage subjects increased the standard deviations of some estimates whereby increasing the width of 95 % confidence intervals. The 95% confidence intervals were decreasing for some estimates because high leverage subjects decreased the standard deviation

of those estimates. The odds were increasing or decreasing because the removal of high leverage points were jointly influential on the parameter estimates. Pregibon's leverage was used in a study done by (Fitrianto & Wendy, 2016) which looked at identification of high leverage points in binary logistic regression. The dependent variable was weather the cancer patients had lymph node involvement or not and the independent variables was level of acid phosphates. Pregibon leverage showed that there were leverages.

There is other methods of diagnosing logistic regression based on predictive ability (Rufibach, 2010). One of the such methods is Brier's score. The Brier's Score is the method that measures the accuracy of probabilistic predictions. The value of the Brier's score ranges from 0.0 and 1.0, where a model with best predictive ability has a score of 0.0 and the worst predictive ability has a score of 1.0.

5.2 Conclusion

The aim of this study was to demonstrate utilization of the post-estimation diagnostic statistics that are available for fitting binary regression models to data which are usually ignored in most applications of the model. In the study, we looked at diagnostic statistics for outliers, multicollinearity, influential points and high leverage points. The findings suggest that there is need to check for outliers and leverages when fitting binary logistic regression using childhood diarrhoea MDHS 2015-2016 data.

The results showed that there is a need to fully diagnose the binary logistic

regression because outliers, influential points, leverages and multicollinearity may be present in the model which may affect the coefficients, p-vales and confidence intervals as a results wrong inferences and conclusion can be made. The conclusion made is that significant predictors of child diarrhoea are region where the child is coming from, if the family is sharing toilet with other families, sex of the child and age of the child. Therefore, the findings of the study expose the fact that use of child diarrhoea data that are collected through large surveys like DHS has to consider analysis of outliers and influential observations before making conclusions. The study also showed that there was a change in odds and width of 95% CI after removing influential points and outliers. There was a serious change when Pearson's residual was used.

5.3 Recommendations

- The study suggest that the analysts should check throughly the fit of the model to ensure the errors are removed whereby getting good models that will help in improving child health.
- Another study should be done using multivariate diagnostic statistics for logistic regression model.
- Another study should be done using Brier's score to check for predictive ability

REFERENCES

Ahmad, S., Midi, H., & Ramli, N. (2011). Diagnostic for residual outliers using deviance component in binary logistic regression. *World Applied Sciences Jounal*, *8*, 1125–1130.

Ahmad, S., Ramli, N. M., & Midi, H. (2012). Outlier detection in logistic regression and its application in medical data analysis. In *2012 ieee colloquium on humanities, science and engineering (chuser)* (pp. 503–507).

Amare, Z., Ahmed, M., & Mehari, A. (2019). Determinants of nutritional status among children under age 5 in ethiopia: further analysis of the 2016 ethiopia demographic and health survey. *Globalization and health*, *15*(1), 1–11.

Chen, C.-Y., Yang, H.-C. P., Chen, C.-W., & Chen, T.-H. (2008). Diagnosing and revising logistic regression models: effect on internal solitary wave propagation. *Engineering Computations*.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research*, *3*(1), 58–67.

Dattalo, P. (1994). A comparison of discriminant analysis and logistic regression. *Journal of Social Service Research*, *19*(3-4), 121–144.

Davison, A., Gigli, & A. (1989). Deviance residuals and normal scores plots. *Biometrika*, *76*(2), 211–221.

Dhakal, C. P. (2017). Dealing with outliers and influential points while fitting regression. *Journal of Institute of Science and Technology*, *22*(1), 61–65.

Erica. (2020). *Beginner's guide to maximum likelihood estimation.* Retrieved 2010-09-23, from `https://www.aptech.com/blog/beginners-guide-to-maximum-likelihood-estimation-in-gauss/`

Fitrianto, A., & Wendy, T. (2016). Identification of high leverage points in binary logistic regression. In *Aip conference proceedings* (Vol. 1782, p. 050006).

Freund, R. J., Vail, R. W., & Clunies-Ross, C. (1961). Residual analysis. *Journal of the American Statistical Association*, *56*(293), 98–104.

Getachew, A., Guadui, T., Tadie, A., & Gizaw, Z. (2018). Diarrhea Prevalence and Sociodemographic Factors among Under-Five Children in Rural Areas of North Gondar Zone, Northwest Ethiopia. *International Journal of Pediatrics*, *12*(1), 73–86.

Ghosh, S. (2022). Deletion diagnostics in logistic regression. *Journal of Applied Sta-tistics*, *1*(1), 1–13.

GOM. (2020). *Malawi 2020 voluntary national review report for sustainable development goals (sdgs); main report.*

Haque, A., Jawad, A., Cnaan, A., & Shabbout, M. (2002). Detecting multicollinearity in logistic regression models: an extension of bkw diagnostic. In *Proceedings of the 2002 joint statistical meeting, american statistical association* (pp. 1356–1358).

Harrell, F. E. (2015). Ordinal logistic regression. In *Regression modeling strategies* (pp. 311–325). Springer.

Hilbe, J. M. (2009). *Logistic regression models*. CRC press.

Hosmer, D., & Lemeshow, S. (2002). *Applied logistic regression.*

Hosmer, D., & Lemeshow, S. (2013). *Applied logistic regression*. New Jersey: John Willey and Sons Inc.

Hosmer, D., Taber, S., & Lemeshow, S. (1991). The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *American Journal of Public Health*, *81*(12), 1630–1635.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Imon, A. R., & Hadi, A. S. (2013). Identification of multiple high leverage points

in logistic regression. *Journal of Applied Statistics*, *40*(12), 2601–2616.

Jennings, D. E. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association*, *81*(396), 987–990.

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika*, *72*(1), 59–65.

Jung, Y., Lee, S. P., & Hu, J. (2016). Robust regression for highly corrupted response by shifting outliers. *Statistical Modelling*, *16*(1), 1–23.

Kaombe, T., & Namangale, J. (2016). Modelling Distribution of Under-Five Child Diarrhoea across Malawi. *Journal of Mathematics and System Science*, *6*, 91–101. 10.17265/2159-5291/2016.03.001

Kasza, J. (2015). Stata tip 125: Binned residual plots for assessing the fit of regression models for binary outcomes. *Stata Journal*(199-2018-3586). Retrieved from `http://ageconsearch.umn.edu/record/275965` doi: 10 .22004/ag.econ.275965

Khan, A., Amanullah, M., Aljohani, H. M., & Mubarak, S. A. (2021). Influence diagnostics for the poisson regression model using two-parameter estimator. *Alexandria Engineering Journal*, *60*(5), 4745–4759.

Kumbuyo, C. P., Yasuda, H., Kitamura, Y., & Shimizu, K. (2014). Fluctuation of rainfall time series in malawi: an analysis of selected areas. *Geofizika*, *31*(1), 13–28.

Larsen, W. A., & McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, *14*(3), 781–790.

LaValley, & P, M. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.

Mbugua, S., Musikoyo, E., Ndugi, F., Sang, R., Kamau, M., & Ngotho. (2014). Determinants of diarrhea among young children under the age of five in Kenya, evidence from KDHS 2008-09. *Operations Research*, *28*(2), 1046–1056. Retrieved from `http://aps.journals.ac.za`

McDonald, B. (2002). A teaching note on cook's distance-a guideline.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

MESTAV, B. (2019). Detection and diagnostic methods of multiple influential points in binary logistic regression model in animal breeding. *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi*, *29*(4), 677–688.

Midi, H., Sarkar, K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253–267.

Miles, J. (2014). Tolerance and variance inflation factor. *Wiley StatsRef: Statistics Reference Online*.

Montgomery, D. C., & Peck, E. A. V. (2012). *Introduction to linear regression analysis*. New Jersey: John Willey and Sons Inc.

Mwambete, K. D., & Joseph, R. (2010). Knowledge and perception of mothers and caregivers on childhood diarrhoea and its management in temeke municipality, tanzania. *Tanzania journal of health research*, *12*(1), 47–54.

NSO. (2017). *Malawi demographic health survey 2015-16.*

Nurunnabi, A., & Geoff, W. (2012). Outlier Detection in Logistic Regression:A Quest for Reliable Rnowledge from Predictive Modeling and Classification. *Conference Paper*. DOI: 10.1109/ICDMW.2012.107

Nurunnabi, A., Hadi, A. S., & Imon, A. (2014). Procedures for the identification of multiple influential observations in linear regression. *Journal of Applied Statistics*, *41*(6), 1315–1331.

Nurunnabi, A., & Nasser, M. (2011). Outlier Diagnostics in Logistic Regression: A Supervised Learning Technique. *2009 International Conference on Machine Learning and Computing*, *3*.

Nurunnabi, A., Rahmatullah Imon, A., & Nasser, M. (2009). Identification of multiple influential observations in logistic regression. *Journal of Applied*

Statistics.

OConnell, A. A. (2006). *Logistic regression models for ordinal response variables* (Vol. 146). sage.

Osborne, A., Jason W Overbay. (2004). The power of outliers (and why researchers should always check for them). In (Vol. 9, p. 6).

Park, H.-A. (2013). An Introduction to Logistic Regression: From basic concepts to Interpretation with Particular Attention to Nursing Domain. *J Korean Acad Nurs*, *43*(2), 154–164. Retrieved from `http://dx.doi.org/10.4040/jkan.2013.43.2.154`

Peng, Y., Luo, B., & Gao, X. (2022). Robust moderately clipped lasso for simultaneous outlier detection and variable selection. *Sankhya B*, 1–14.

Pregibon, D. (1981). Logistic regression diagnostics. *The annals of statistics*, *9*(4), 705–724.

Ramsey, P., & Ramsey, P. (2007). Optimal Trimming and Outlier Elimination. *Journal of Modern Applied Statistical Methods*, *6*(2), 355–360.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.

Rufibach, K. (2010). Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, *63*(8), 938–939.

Sarkar, S., Midi, H., & Rana, S. (2011). Detection of outliers and influencial observation in Binary Logistic Regression an emprical study. *Jounal of Applied Science*, *56*(4), 26–35.

Sarkar, S. K., Midi, H., & Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences*, *11*(1), 26–35.

Schreiber-Gregory, D. (2018). Logistic and linear regression assumptions:

Violation recognition and control. *Henry M Jackson Foundation*.

Senaviratna, N., & Cooray, T. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 1–9.

Uraibi, H. S. (2019). Selective overview on single diagnostics methods of outliers in logistic regression. *Journal of Administration and Economics*(119).