Appendix A

1. Statistical Methods and Models

1.1 Multivariate Outliers Detection Measures for Generalized Linear Model

Considered the model

$$y = f(X) + \varepsilon \qquad (1)$$

as a generalized linear model (GLM) where the response is exponential family member and *X* is a matrix of order $n \times p$. The observation that impacts the slop of the regression line are called influential observations and considered as the bad outliers. By removing these observations, the estimated coefficients can be significantly changed. The different influence measurement statistics used to detect bad outliers for the GLM are described as follows.

Leverage is the measurement of how far the predictor variable deviate from its mean. The influential observation can be detected with the help of leverage and cases are declared to be influential if $h_{ii} > 2p/n$ [24,25]. Where h_{ii} is the diagonal element of hat matrix H. In GLM the hat matrix is defined as $H = \hat{W}^{\frac{1}{2}} X \left(X' \hat{W} X \right)^{-1} X' \hat{W}^{\frac{1}{2}}$. The diagonal elements of H is defined as $h_{ii} = diag(H) = \hat{w}_i x_i' \left(X' \hat{W} X \right)^{-1} x_i$.

Cook's Distance (CD) is widely used in the detection of influential observations in linear regression models. This method is proposed by[26]. It measures the complete change in the regression model when ith observation is removed. The observations whose $CD \geq 4/(n-1)$ are suspected to be influential observations (Hardin, 2012). For GLM the Cook's Distance is formulated as $CD_i = \left(s_{rp_i}^2 \times h_{ii} \right) / \left(p' \times 1 - h_{ii} \right)$, where $s_{rp_i}^2$ is the Standardized Pearson Residuals and defined as $s_{r\,p_i} = r_{p_i} / \sqrt{\hat{\phi}(1-h_{ii})}$ and r_{p_i} is the Pearson residual $r_{p_i} = \left(y_i - \mu_y \right) / \sigma_y$, where μ_y and σ_y are the mean and standard deviation of binomial distribution.

Modified Cook Distance (MCD) diagnose the influential observation more sharply[27]. The observation whose $MCD \geq 2\sqrt{(n-p)/n}$ are suspected to be influential observation. The MCD for GLM is defined as $MCD_i = \left[\frac{n-(p+1)}{p+1} \frac{h_{ii}}{1-h_{ii}} \right]^{1/2} |t_i|$ where $t_i = s_{rp_i} \sqrt{\frac{n-(p+1)}{n-p-s_{rp_i}^2}}$

Andrew's Pregibon (AP) gives another measure[28] to detect the influential observation and the observation are declared to be influential observation when $1 - AP > 2p/n$. The AP measures for GLM is defined as $AP_i = 1 - h_{ii} - \frac{r_{p_i}^2}{\sum r_{p_i}^2}$

Covariance Ratio (CR) measure the influence of ith observation on the variances of the estimates[29]. The observation is suspected to be influential when $\left(1+3\frac{p}{n}<CR<1-3\frac{p}{n}\right)$ (Amin, 2016). For GLM the CR is defined as $CVR_i=\frac{\left[n-p-s^2_{rp_i}\right]^{(p+1)}}{1-h_{ii}}$

Welsch's Distance (WD) is the modified form of DFFITS[30] and the observation are suspected to be influential if the value of WD is greater than $3\sqrt{p}$. The WD for GLM is defined as $WD_i=(n-1)t_i^2\frac{h_{ii}}{(1-h_{ii})^2}$

1.2 Stepwise Logistic Regression

Logistic regression is a commonly used model to predict the occurrence of binary outcomes[31]. The logit function of GLM defined in Eq. (1) is as

$$\pi_i=g^{-1}(X\beta) \tag{2}$$

Where X is known as $n\times p$ design matrix and β is $p\times 1$ parameter vector. The binary logistic model for the multiple predictors[32] can be written as

$$P(y/X_1,X_2,...,X_p)=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}} \tag{3}$$

The stepwise logistic regression (SLR) is then evaluated by adopting backward selection approach. The stopping criteria was chosen on the base of minimum AIC and BIC.

1.3 Stepwise Logistic Regression After Deletion

The observations commonly diagnose by all the outlier detection methods are considered as outliers. In this approach logistic regression model is applied after the group deletion of outliers. Thus, the Eq. (1) for logistic regression after deletion is as

$$\pi_i=g^{-1}\left(X_{(n-d)}\beta\right) \tag{4}$$

where d is the number of outliers and $X_{(n-d)}$ is the matrix of observation does not contain bad outliers. Finally, Stepwise Logistic regression after deletion (SLRAD) model was used by following backward selection approach.

1.4 Artificial Neural Network

The Artificial Neural Network (ANN) model is generally considered as the machine learning approach and is considered as the ideal for the prediction of tisease occurrence in individuals. The model consists of hidden layers which lead the input variables to the output. The training of the model is done by using backpropagation. There is no need to distributional assumption in input variables while using the ANN model[33]. Additionally, this model is not much sensitive of outliers in data.

The weights of the input variables show their importance in the model. The impact of a variable on output is depends on the size of its weight. To avoid the over selection of the variables we have proposed a way to selected important variables. In this procedure the weights are normalized by the average absolute weights and variables having a weight more the 0.12 are selected as the important variables and hence used in the final model.