African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)
www.abjournals.org

EVALUATING THE PERFORMANCES OF ROBUST LOGISTIC REGRESSION MODELS IN THE PRESENCE OF OUTLIERS

Hafiza Inusa Idris[1*], Abdulmalik Mohammed[2], Umar Faruk Salisu[3], Kamalu Ibrahim Balansana[4], Danjuma Abdulazeez[5], and Nuruddeen Hassan Danrimi[6]

[1]Department of Statistics, Federal Polytechnic Nyak, Shendam, Plateau State, Nigeria.

[2]Department of Computer/Mathematics, C.O.E Dutsen Tanshi Bauchi, Bauchi State, Nigeria.

[3]Department of Statistics, Federal Polytechnic Bauchi, Bauchi State, Nigeria.

[4]Department of Statistics, Modibbo Adama University, Yola, Adamawa State, Nigeria.

[5]Department of Statistics, Federal Polytechnic Damaturu, Yobe State, Nigeria.

[6]Department of Statistics, University of Abuja, Nigeria.

[*]Corresponding Author's Email:halilu.hafiza@gmail.com

Cite this article:

Hafiza, I. I., Abdulmalik, M., Salisu, U. F., Balansana, K. I., Abdulazeez, D., Danrimi, N. H. (2024), Evaluating the Performances of Robust Logistic Regression Models in the Presence of Outliers. African Journal of Mathematics and Statistics Studies 7(4), 320-327. DOI: 10.52589/AJMSS-YKDFCYQS

Manuscript History

Received: 17 Sep 2024

Accepted: 25 Nov 2024

Published: 2 Dec 2024

ABSRACT: *Logistic regression models are widely used in the field of medical and behavioral sciences. These models are used to describe the effect of explanatory variables on a binary response variable. The maximum likelihood estimator (MLE) is commonly used to estimate the parameters of logistic regression models due to its efficiency under a parametric model. However, evidence has shown that the MLE is highly sensitive to outlying observations which might affect the parameter estimates. Robust methods are put forward to rectify this problem. This paper investigated the robustness of GM-Mallows and GM-Schweppes as an alternative to the commonly used ordinary logistic regression model in the presence of outliers. The study used a Monte Carlos Simulation, by generating a logistic regression model with Five independent normally distributed covariates. 5% of outliers was contaminated to the data on sample sizes 50, 200 and 400 respectively. The results showed that the GM-Mallows estimator perform best across all metrics having the lowest AIC, BIC, MSE and MAE except for n=50. This suggests that the robust methods, especially GM-Mallows, provide more reliable estimates in the presence of outliers. The finding suggests that if there is presence of outliers' GM-Mallows appears to be the top choice, where the GM-Schweppes offers a middle ground, providing some robustness with perhaps less extreme adjustments. The ordinary logistic regression might be preferred if simplicity and interpretability are prioritized, and there's confidence that outliers are not a significant issue in the data.*

KEYWORDS: Robust, logistic regression, GM-Mallows, GM-Schweppes, Outliers.

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)

www.abjournals.org

INTRODUCTION

Logistic regression is a proper analysis method used to model data and explain the relationship between the binary response variable and explanatory variables. Logistic regression is the most important tool for data analysis in various fields. The classical approach for estimating parameters is the maximum likelihood estimation, a disadvantage of this method is high sensitivity to outlying observations. The robust estimators for logistic regression are alternative techniques due to their robustness. Many robust estimators as an alternative to MLE have been proposed. [1] developed a diagnostic measurement of outlying observations and they showed that in the logistic regression, the MLE was very sensitively to outlying observations (see also [2]).

An outlier is an observation deviated from the other values in data and produces the large residuals. In logistic regression model, an outlier can be occurred in the response variables as well as in the predictor variables or in both. In the binary regression model, all the response variables y_i are binary, takes the numerical values 0 or 1, therefore, an outlier in the response variable can only occur as a transposition 0→1 or 1→0 discussed by [3]. An error in response variables is also well-known as a misclassification error or residual outlier. Extreme observation in explanatory variables is known as a leverage point or leverage outlier. [4] stated that the estimated parameters in logistic regression may be severely affected by outliers; hence, several robust alternatives which are much less affected by outliers are proposed in the literature (for example, [4]; [5]; [6]; [7]; [8]; [9]). A robust regression is an iterative procedure that is designed to overcome the problem of outliers and influential observations in the data and minimize their impact over the regression coefficients [10].

The main objective of robust estimation is to obtain reliable estimates/inferences for unknown parameters in the presence of outliers. [11] applied a logistic model to evaluate the risk factors for hepatitis B viral disease in Gusau local government of Zamfara State, Nigeria and recommended it as the best model for the analysis of HBV despite the fact that, the logit is not resistant to outliers which may lead to inefficient results. Therefore, this study intends to improve on their work by investigating the performances of robust logistic models namely GM Mallows and GM Schweppes robust estimators which are resistant to outliers and high leverages as an alternative to the mostly used logistic model in modelling binary response variable.

METHODOLOGY

Methods of Model estimation

A. GM Estimator

The GM estimators are known to be consistent, asymptotically normal and most efficient in the class of all estimators that do not use any extra information aside from that contained in the moment conditions.

It can be expressed as a solution of normal equations given by

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)

www.abjournals.org

$$\sum_{i=1}^{n} d_i \psi\left(\frac{\left(y_i - x_i^t \hat{\beta}\right)}{\hat{\sigma} d_i}\right) x_i = 0 \quad (1)$$

Where ψ = ρ' is an influential function and d_i = 1, 2, ..., n is the initial weight function.

The initial weight of the GM estimator is then defined as follows

$$d_i = \left[1, \left(\frac{\chi^2(0.95, \mathrm{p})}{RMD^2}\right)\right], i = 1,2, \ldots, n \quad (2)$$

The FIMGT is defined as:

$$\mathrm{FIMGT}_i = \begin{cases} \dfrac{\hat{\varepsilon}_{\mathrm{i.R}}}{\hat{\sigma}_{\mathrm{R-i}}\sqrt{1 - \mathrm{w}^*_{\mathrm{ii.R}}}} \text{ for i} \in \mathrm{R} \\ \dfrac{\hat{\varepsilon}_{\mathrm{i.R}}}{\hat{\sigma}_{\mathrm{R}}\sqrt{1 + \mathrm{w}^*_{\mathrm{ii.R}}}} \text{ for i} \notin \mathrm{R} \end{cases} \quad (3)$$

Where ($\hat{\beta}$), the parameter estimates, residuals ($\hat{\varepsilon}_{\mathrm{i.R}}$), hat values ($\mathrm{w}^*_{\mathrm{ii.R}}$), standard deviation ($\hat{\sigma}_{\mathrm{R}}$) and standard deviation with the ith case deleted ($\hat{\sigma}_{\mathrm{R-i}}$) are computed using the OLS to the remaining data, i.e R set.

Algorithm: GM Estimator

1. An arbitrary subset, H_{old} comprises of h different observations are chosen where h is smallest integer greater than or equal to$\frac{n+p+1}{2}$, p is the number of predictor variables.

2. Compute the average vector $\bar{T}_{H_{old}}$ and covariance matrix C_{Hold} of all observations that belong to H_{old}.

3. Compute the Mahalanobis Distance Squares, denoted as:$d^2_{old}(i) = (t_i - \bar{T}_{H_{old}})' C^{-1}_{H_{old}} (t_i - \bar{T}_{H_{old}})\ for i = 1,2, \ldots, n.$

4. Arrange $d^2_{old}(i)$ for I = 1,2,…,n in ascending order $d^2_{old}(\pi(1)) \leq d^2_{old}(\pi(2)) \leq \cdots \leq d^2_{old}(\pi(n))$ where π is permutation equal to { 1,2,…,n}.

5. Create $H_{new} = \{t_{\pi(1)}, t_{\pi(2)}, \ldots, t_{\pi(h)}\}$ such that its' elements comprises of the first smallest h observations acquired from step 4. Then list the new Index Set.

6. Compare $I_{new} = I_{old}$. If $I_{new} = I_{old}$, stop the process. Afterwards, equate $\bar{T}_{H_{old}}$: = $\bar{T}_{H_{new}}$, $C_{H_{old}}$: = $C_{H_{new}}$, if $I_{new} \neq I_{old}$ then recomputed $\bar{T}_{H_{new}}$,and $C_{H_{new}}$, let H_{old}: = H_{new},$\bar{T}_{H_{old}}$: = $\bar{T}_{H_{new}}$ and $C_{H_{old}}$: = $C_{H_{new}}$. Repeat Steps 3-6, until$I_{new} = I_{old}$ where at this point, $\bar{T}_{H_{new}}$ is the robust estimator of location and $C_{H_{new}}$ is the robust estimator of scatter.

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)

www.abjournals.org

B. Mallows GM-estimate

The first GM-estimate was proposed by Mallows (1975). For Mallows GM-estimate, Hat values range from 0 to 1, so weight function down weights the high leverage points. A weight of $\sqrt{1-h_i}$ ensures that observations with high leverage receive less weight than observations with small leverage (i.e if $(h_i > h_{j,} u_i < u_j$. Although this strategy seems sensible at first, it is problematic because even ‘‘good’’ leverage points that fall in line with the pattern in the bulk of the data are down-weighted, resulting in a loss of efficiency.

C. Schweppes GM-estimate

Another GM-estimate is called Schweppe GM-estimate. This method adjusted the leverage weights according to the size of the residual e_i by using v_i=w_i , where w_i is the weight function which is the same as Mallows GM-estimate and equal to $\sqrt{1-h_i}$ (see Handschin et al. 1975). However, since the weight function of this estimate only depends on x values without considering how the corresponding y values fit with the pattern of the bulk of the data, efficiency is still hindered (Krasker and Welsch 1982). Moreover, Carroll and Welsh (1988) suggested that the Schweppe estimate is not consistent when the errors are asymmetric. The breakdown points for the above two GM-estimates, although better than for regular M-estimate, are at most $1/(1+p)$, where p is the number of predictor variables (Maronna, Bustos and Yohai 1979). Thus, as dimensionality increases, their BP tends to 0.

D. Logistic regression model.

Logistic regression is a popular modeling technique used to predict binary outcomes. The model is a linear model that captures the relationship between the input variables and the output variable (binary outcomes).

The multiple binary logistics regression model is given as follows:

$$\pi(\mathrm{X}) = \frac{\exp\left(\beta 0 + \beta 1 X1 + \cdots + \beta k Xk\right)}{1 + \exp\left(\beta 0 + \beta 1 X1 + \cdots + \beta k Xk\right)} \tag{4}$$

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

$$= \frac{1}{1 + \exp\left(-X\beta\right)}$$

Where here π denotes a probability and not the irrational number 3.14…

Π is the probability that an observation is in a specified category of the binary Y variable, generally called the “success probability”. We notice that the model describes the probability of an event happening as a function of *X* variables. For instance, it may provide estimates of the probability that an older person has heart disease. With the logistic model, estimate of π from equations like the one above will always be between 0 and1 the reasons are: The numerator $exp^{(\beta_0+\beta_1+\beta_{1x1}+\cdots+\beta_{kxk})}$ must be positive, because it is power of a positive value (e). The denominator of the model is (1+numerator), so the answer will always be less than 1. With one X variable, the theoretical model for π has an elongated *“S”* shape (or sigmoidal

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)

www.abjournals.org

shape) with asymptotes at *0 and 1*, although in sample estimate we may not see this *"S"* shape if the range of *X* variable is limited.

For a sample of size n, the likelihood for a binary logistic regression is given by:

$$L(\beta; y, X) = \prod_{i=1}^{n} \pi i^{yi} \, (1 - \pi i)^{1-yi} \; = \prod_{i=1}^{n} \left(\frac{\exp{(X\beta)}}{1+\exp{(X\beta)}}\right)^{yi} \; \left(\frac{1}{1+\exp{(X\beta)}}\right)^{1-yi} \quad (5)$$

Simulation Study: Generating Data Set

A Monte Carlos Simulation study was carried out to compare the robustness of the estimators discussed above. These estimators are: Logistic model, GM-Mallows and GM-Schweppes robust logistic models

Following the simulation study similar as the one carried out by [12] and [13], a logistic regression model is generated with Five independent normally distributed covariates. The error terms ε_i are drawn from a logistic distribution defined as:

$$Y = I(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + + \beta_5 x_5 \varepsilon \geq 0.$$

A 5% outlier contaminated to the simulated data, having Five explanatory variables independently and normally distributed with zero mean and unit variance is considered. The true parameter values are β = (0, 2, 2,2,2,) with sample sizes *n* = 50, 200 and 400 representing Small, Medium and Large. The responses *y* was chosen randomly and changed from either 0 to 1 or 1 to 0.

Measure of performance and selection criteria for the best methods

1.To select the best model fitting, Akaike Information Criterion (AIC), Bayesian Information and Mean Square error (MSE) Criterion (BIC) can be used. The smaller the value of AIC or BIC, the better the model in fitting. AIC and BIC are defined as follows;

$$AIC = \; -2logp\,(L) + 2p \quad (6)$$

$$BIC = \; -2logp\,(L) + plog(n) \quad (7)$$

Where;

L is the likelihood under the fitted model,

P is the number of parameters used/ in the model,

n is the number of observations / sample size.

2. $$MSE = \left[\frac{1}{100}\sum_{i=1}^{1000}\left|\hat{\beta}_i - \beta\right|^2\right] \quad (8)$$

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)

www.abjournals.org

3. $MAE = \frac{1}{n} = \sum_{t=1}^{n} |(A_t - F_t)$ (9)|

RESULTS

Table 1: Simulated results of Logistic, GM-Mallows and GM-Schweppes with 5% outlier Contamination at sample size 50.

Sample Size		n=**50**		
Methods	**AIC**	**BIC**	**MSE**	**MAE**
Logistic	651.123	652.418	0.8647	0.9334
Mallows	641.8432	655.663	0.7021	0.4314
Schweppes	611.3513	615.4513	0.2137	0.4123

Source: *Authors' computation aided by R package v 4.1.3*

Table 2: Simulated result of Logistic, GM-Mallows and GM-Schweppes with 5% outlier Contamination at sample size 200.

Sample Size		n=**200**		
Methods	**AIC**	**BIC**	**MSE**	**MAE**
Logistic	631.3213	635.4213	0.3137	0.3123
Mallows	610.861	613.853	0.3213	0.2254
Schweppes	611.924	625.613	0.3137	0.3123

Source: *Authors' computation aided by R package v 4.1.3*

Table 3: Simulated results of Logistic, GM-Mallows and GM-Schweppes with 5% outlier Contamination at sample size 400.

Sample Size		n=**400**		
Methods	**AIC**	**BIC**	**MSE**	**MAE**
Logistic	691.12	692.31	0.1047	0.2350
Mallows	609.321	613.853	0.3213	0.2254
Schweppes	610.341	617.312	0.4137	0.2123

Source: *Authors' computation aided by R package v 4.1.3*

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 320-327)
www.abjournals.org

Table 4: Comparing the Performances of the three models simulated Results With 5% at sample sizes 50, 200 and 400 respectively.

Sample size (n)	Models	AIC	BIC	MAE	MSE
n=50	logistic	651.123	652.418	0.9334	0.8647
	GM-Mallows	641.8432	655.663	0.4314	0.7021
	GM-Schweppes	611.3513	615.4513	0.4123	0.2137
n=200	logistic	631.3213	635.4213	0.3123	0.3137
	GM-mallows	610.861	613.853	0.2254	0.3213
	GM-Schweppes	611.924	625.613	0.3123	0.3137
n=400	logistic	691.12	692.31	0.235	0.1047
	GM-Mallows	609.321	613.853	0.2254	0.3213
	GM-Schweppes	610.341	617.312	0.2123	0.4137

CONCLUSION AND RECOMMENDATION

Table 1 reports AIC, BIC, MSE and mean absolute errors (MAE) of the three models for the contaminated data on sample size 50. The result showed that GM-Schweppes gives a better result compared to the ordinary logistic and GM-Mallows having the lowest AIC (611.3513), BIC (615.4513) and MSE (0.2137) respectively. Looking at table 2, we can observe that the GM-Mallows outperformed the ordinary logistic and robust GM-Schweppes models having the lowest AIC (610.861), BIC (613.853) and MSE (0.2254) accordingly. We can also see that from the result of table 2, as the sample size increases to 200 the values of AIC, BIC, MSE and MAE of ordinary logistic and GM-Mallows have reduced drastically compared to when n is 50. In the same vain, looking at Table 3, the GM-Mallows model also found to outperform the ordinary logistic and GM-Schweppes having the lowest AIC (609.321) and BIC (613.853) values. We can also notice that the values of AIC, BIC, MSE and MAE of the GM-Schweppes reduces drastically as the sample sizes goes up to 400. Table 4 compared the performances of the three models based on sample sizes 50, 200 and 400 respectively. The result showed that the GM-Mallows gives a better result except when n=50, followed by GM-Schweppes. So, the results of the study showed that in all categories the ordinary logistic model performed less compared to the robust methods. Meaning that the ordinary logistic is highly sensitive to outliers. Therefore, this study recommended that Analyst should only use ordinary logistic regression on estimation if they are certain of no outliers in the data otherwise it would lead to unbiased results.

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS

REFERENCES

[1] Pregibon, D. (1981) Logistic Regression Diagnostics. The Annals of Statistics, 4, 705-724. https://doi.org/10.1214/aos/1176345513
[2] Kunsch, H.R., Stefonski, L.A. and Carroll, R.J. (1989) Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models. Journal of the American Statistical Association, 84, 460-466. https://doi.org/10.1080/01621459.1989.10478791
[3] Copas, J.B. (1988) Binary Regression Models for Contaminated Data. Journal of the Royal Statistical Society. Series B (Methodological), 50, 225-265. https://doi.org/10.1111/j.2517-6161.1988.tb01723.x
[4] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, *9*, 705-724.
[5] Copas, J. B. (1988). Binary regression model for contaminated data (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *50*, 225-265.
[6] Kunsch, H. R., Stefanski, L. A., & Carroll, R. J. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of American Statistical Association*, *84*, 460-466.
[7] Carroll, R. J., & Pederson, S. (1993). On robust estimation in the logistic regression model. *Journal of the Royal Statistical Society*, *Series B*, *55*, 693-706.
[8] Bianco, A. M., & Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust statistics, Data analysis and computer intensive methods*, H. Reider, Ed., 17- 34. New York: Springer Verlag.
[9] Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis Journal*, *44*, 273-295.
[10] Zaman. A, Rousseeuw.P. J, and Orhan. M, (2001). Econometric applications of high breakdown robust regression techniques, *Economics Letters*, vol.71, no.1, pp. 1–8.
[11] Olayemi T.J, Ike G.O and Bello. U. (2023). Binary Logistic Regression Analysis on Risk Factors Associated with Hepatitis B Disease in Gusau Local Government Zamfara State, Nigeria. *International Journal of Science for Global Sustainability*,9(3). https://doi.org/10.57233/ijsgs.v9i3.531 URL: https://fugus-ijsgs.com.ng.
[12] Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis Journal*, *44*, 273-295.
[13] S. Nawaz, N. Shahzad, T. R. Fraz, A. Shakil and H. R. Khuram (2022). Comparison of robust estimator in case of outliers. *Journal of Webology* Volume 19, ISSN: 1735-188X.

Article DOI: 10.52589/AJMSS_YKDFCYQS
DOI URL: https://doi.org/10.52589/AJMSS_YKDFCYQS