*International Journal of Innovative Mathematics, Statistics & Energy Policies*
*12(4):14-19, Oct-Dec., 2024*

 **www.seahipublications.org** **ISSN: 2467-852X**

**DOI:10.5281/zenodo.13998812**

# A Robust Logistics Regression Model In The Presence Of Outliers, Multicollinearity And High Leverage Points On Hepatitis B Virus

**Idris, H.I., Rasheed, B.A., Bawa, M.U., Idi H. and. Abdulkadir. A.**

**Department of Statistics, Federal Polytechnic Nyak Shendam, Plateau State, Nigeria**
**Department of Mathematical Sciences, Gombe State University Gombe, Nigeria**
**Department of Statistics, Abubakar Tafawa Balewa University Bauchi, Nigeria**
**Department of Mathematics and Statistics, Federal Polytechnic, Bauchi, Nigeria.**
**Department of Statistics, Abubakar Tafawa Balewa University Bauchi, Nigeria**
**Corresponding author: Hafiza Inusa Idris Halilu.hafiza@gmail.com**

**ABSTRACT**
This study has investigated the performance of Robust logistic regression as an alternative to ordinary logistic regression model in the presence of outliers, Multicollinearity and High leverage points on modelling Hepatitis B Virus data for the period of Six months i.e. from 1st July 2023 to December 2023. The study used Hepatitis B as the response variable and Gender (GD), Marital Status (MS), History of Blood Transfusion (HBT), Multiple Sexual Partner (MSP), Alcohol consumption (AC) and Infected Family Member (IFM) are the independent variables. Respectively. The HBV data was subjected to boxplot visualization so as to detect the presence of outliers in the data and the result revealed that out of the seven variables in the HBV model only two found to be outliers free. All the variables under consideration were tested for multillinearity and the result shows that the VIF values for **Gender** (1.011606), **IFM** (1.061440), **MSP** (1.047772), **HBT** (1.087136), **MS** (1.060032) and **AC** (1.024398) are close to 1, indicating very low multicollinearity among the predictors. This is good for the model's stability. The result of the study showed that the GM-Mallows estimator perform best across all metrics having the lowest AIC, BIC, MAE, MSE, RMSE and highest MPA.
**Keywords:** Robust logistic regression, Multicollinearity, GM- estimators and **Schweppes GM-estimate**

## INTRODUCTION

Hepatitis B Virus was discovered in 1965 by Dr. Baruch Blumberg who won the Nobel Prize for his discovery. Originally, the virus was called the "Australia Anti-gen" because it was named for an Australian aborigine's blood sample that reacted with an American haemophilia patient. He developed the blood test that is used to detect the virus and invented the first hepatitis B vaccine in 1969. Hepatitis is an inflammation of the liver that is caused by a variety of infectious viruses and non-infectious agents leading to a range of health problems, some of which can be fatal liver damage. Hepatitis B is a vaccine preventable disease caused by the hepatitis B virus (HBV) that can induce potentially fatal liver damage.
The prevalence of chronic HBV infection continues to be highly variable, ranging over 10% in some Asian and Western Pacific Countries to under 0.5% in the United States and northern European countries. The current global estimate of the number of HBV infected individuals is 350 million. There are approximately 50 million chronic carriers of Hepatitis B Virus (HBV) in Africa, with a 25% mortality

risk. In Sub-Saharan Africa, carrier rates range from 9-20%. Many studies have suggested that HBV transmission in Africa occurs predominantly in childhood, by the horizontal rather than the prenatal route. More than 90 million people are living the hepatitis in the Region, accounting for 26% of the global total. Nigeria has a prevalence rate of 8.1 and 1.1 % for HBV and HCV among adult aged 15-64 years respectively according to the Nigeria HIV-AIDS indicator and impact survey, 2018 (NAIIS 2018).

The accurate forecasting of hepatitis B can be obtained by analysing the sufficient historical data. However, in China and perhaps some other developing countries, the current public health surveillance system does not collect detailed essential epidemiological information as they are often difficult to obtain. The forecasted of hepatitis B will be inaccurate only by the limited data. Therefore, it is significant to make the limited data-processing.

However most of the regression estimation method that are used to estimate the hepatitis B. Virus model are best on ordinary least square method and the data usually does not completely satisfy the assumptions often made by researchers which result in a dramatic effect on the quality of statistical analysis. The ordinary least square regression will produce wrong estimate when Outliers, multicollinearity and high leverage exists in the matrix of explanatory variables**.** The existing of robust regression is not widely known, and the challenge of having outliers multicollinearity and high leverage in the data set is the focus that this thesis is going to emphasize**.** Robust regression is a regression method that is used when the distribution of residual is not normal or there are some outliers that affect the model (Yuliana et al., 2014). This method is an important tool for analyzing the data which is affected by outliers so that the resulting models are stout against outliers (Drapper & Smith, 1998). 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 (Zaman *et al*., 2001). The main objective of robust estimation is to obtain reliable estimates/inferences for unknown parameters in the presence of outliers. The robust procedure replaces the sum of squared residuals of the OLS with some other function that is being less influenced by the unusual observations. These procedures first fit a regression to the data and then identify the outliers as those observations having large residuals Increases. For example, OLS has a breakdown point of 0% which represents that even a single outlier is sufficient to distort the OLS estimators. The robust techniques have 50% of breakdown point which is considered as the highest breakdown point. The property of bounded influence measures the resistance of the estimator against bad observations. The study concluded that, robust logistic regression model such as GM Mallows estimator and GM Schweppes can offer a significant estimate for hepatitis incidences and could assist in estimating outlier, multicollinearity, and high leverage point and would aid decision makers to control the disease. However, to the best of our knowledge, the use of this model on HBV detection studies is still limited.

The aim of this paper is to estimate a Robust Logistic Regression in the presence of outliers, Multicollinearity and high leverage points On Hepatitis B Virus

## METHOD AND MATERIALS

### 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 model uses a sigmoid function to convert the linear output into a binary outcome. Here are the steps for the development of a logistic regression model:

I. Data Preparation: We need to prepare the data by removing missing values, scaling the features, and handling categorical variables
II. Feature Selection: We can use techniques like univariate analysis, correlation matrix, and feature importance to select the most relevant features for our model
III. . . Model Selection: We need to split the data into training and validation sets and use techniques like cross-validation to tune the hyper parameters of the model.

The multiple binary logistics regression model is as follows:

$$\pi(X) = \frac{\exp(\beta 0 + \beta 1X1 + \cdots + \beta kXk)}{1 + \exp(\beta 0 + \beta 1X1 + \cdots + \beta kXk)}$$
$$= \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$
$$= \frac{1}{1 + \exp(-X\beta)}$$

Where here π denotes a probability and not the irrational number.

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 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:

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

**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.

**Schweppes GM-estimate**

Another GM-estimate is called Schweppes 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.

**RESULTS AND DISCUSSIONS**

The results of data analysis and discussions on Hepatitis B virus and its determinant factors are presented accordingly in tables. The results include the following: detection of outliers and high leverage points on Hepatitis B data using Box plot diagrams, checking for the presence of Multicollinearity in the Hepatitis B data and its determinants parameters using Variance inflation factor (VIF). Then, the investigation of the performances of Robust Logistic Regression (GM-mallows) and Robust Logistic (GM-Schweppes) as an alternative to the ordinary logistic regression for handling outliers in modelling the Hepatitis B virus.

**Test of Multicollinearity for HBV and its determinant factors**

**Table 1: s Result of Variance inflation factor(VIF) of HBV data**

| Gender | IFM | MSP | HBT | MS | AC |
|---|---|---|---|---|---|
| 1.011606 | 1.061440 | 1.047772 | 1.087136 | 1.060032 | 1.024398 |

**Source:** Authors' computation aided by R package v4.1.3.

The Table above presented the VIF (Variance Inflation Factor) values for each variable of the HBV data. The rule of thumb for the VIF is that VIF values close to 1 indicate little or no multicollinearity. All VIF values are well below 5, suggesting that multicollinearity is not a concern in this model. Where VIF greater than or equal to 5 suggests multicollinearity. So, in our own case it can be seen the VIF values for all the variables: Gender (1.011606), IFM (1.061440), **MSP** (1.047772), **HBT** (1.087136), **MS** (1.060032) and AC (1.024398) are close to 1, indicating very low multicollinearity among the predictors. This is good for the model's stability.

**Table 2: Ordinary logistic regression results of HBV data.**

| Variables | Coefficient | S.E | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -2.2033 | 0.2449 | -8.996 | 2e-16 *** |
| GENDER | 0.5897 | 0.2203 | 2.676 | 0.00744 ** |
| IFM | 1.7041 | 0.2432 | 7.006 | 2.46e-12 *** |
| MSP | 0.9669 | 0.2228 | 4.340 | 1.43e-05 *** |
| HBT | 0.1058 | 0.2926 | 0.362 | 0.17825 |
| MS | 0.9024 | 0.2555 | 3.532 | 0.00041 *** |
| AC | 0.3417 | 0.5124 | 0.667 | 0.71771 |
| AIC | 538.1435 | | | |
| BIC | 566.908 | | | |

**Source:** Authors' computation aided by R package v4.1.1. Note: (**) and (***) denote significance at 5% and 1% respectively.

Table 2. Presents the results of a logistic regression model predicting a binary outcome (HPB) based on several predictor variables. The table shows the estimated coefficients, standard errors, z-values, and p-values for each predictor. Gender, Infected Family member(IFM), MSP (Multiple Sexual Partner), and MS (Marital Status) are statistically significant predictors of Hepatitis B ($p < 0.01$), while History of Blood Infusion (HBT) and Alcohol Consumption (AC) are not significant. The model's performance indicated by AIC of value (538.1435), BIC (566.908), MAE (0.2866667), MSE (0.286667), RMSE (0.5354126), and MPA (0.7133333). The null deviance (607.3) and residual deviance (497.1) indicate that the model explains some of the variation in the data, but there's still unexplained variance.

**Table 3: Robust Logistic Regression (GM-mallows) result of HBV data.**

| Variables | Coefficient | S.E | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -2.5314 | 0.3118 | -8.119 | 4.69e-16 *** |
| GENDER | 0.9215 | 0.2325 | 3.963 | 7.41e-05 *** |
| IFM | 1.1856 | 0.2451 | 4.837 | 1.32e-06 *** |
| MSP | 0.9473 | 0.2366 | 4.004 | 6.23e-05 *** |
| HBT | 0.4381 | 0.3254 | 1.346 | 0.17825 |
| MS | 0.9024 | 0.2555 | 3.532 | 0.00041 *** |
| AC | 0.1591 | 0.5328 | 0.298 | 0.76537 |
| AIC | 520.7689 | | | |
| BIC | 549.53336 | | | |

**Source:** Authors' computation aided by R package v4.1.1. Note: (**) and (***) denote significance at 5% and 1% respectively.

Robustness weights w.r * w.x: 418 weights are ~= 1. The remaining 32 ones are summarized as

| Min | 1st Quarter | Median | Mean | 3rd quarter | maximum |
|---|---|---|---|---|---|
| 0.08912 | 0.69235 | 0.69980 | 0.71523 | 0.83745 | 0.94120 |

Number of observations: 450

Table: 3 Presents the results of a robust logistic regression model using the GM-Mallows method to predict a binary Hepatitis B (HPB) based on the predictor variables. Where it shows the estimated coefficients, standard errors, z-values, and p-values for each predictor. Gender, IFM (Infected Family Member), Multiple Sex Partner (MSP), and Marital Status (MS) are statistically significant predictors of Hepatitis b+ ($p < 0.001$), while History of Blood transfusion (HBT) and Alcohol Consumption are not significant. Compared to the ordinary logistic regression, the coefficients and their significance levels have changed slightly, which could indicate the presence of influential observations in the dataset. The model's performance shows that AIC has a value of (520.7689), BIC (549.53336), MAE (0.253333), MSE (0.25333), RMSE (0.503331), and MPA (0.745667).

The robustness weights provide information about how the model treats different observations. Out of 450 observations, 418 have weights close to 1, indicating they are treated similarly to how they would be in ordinary logistic regression. The remaining 32 observations have lower weights, ranging from 0.08912 to 0.94120, suggesting these points may be outliers or influential observations. This weighting helps the model produce more reliable estimates in the presence of such data points. The model was fitted in 6 iterations, with specific algorithmic parameters provided for transparency.

**Table 4: Robust Logistic (GM-Schweppes) result of HBV data**.

| Variables | Coefficient | S.E | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -2.4255 | 0.3021 | -8.0255 | 1.101.69e-15 *** |
| GENDER | 0.9002 | 0.2377 | 3.991 | 6.58e-05 *** |
| IFM | 1.1564 | 0.2298 | 4.8661 | 1.14e-06 *** |
| MSP | 0.9248 | 0.2393 | 4.032 | 5.55e-05 *** |
| HBT | 0.4276 | 0.3156 | 1.355 | 0.17532 |
| MS | 0.8809 | 0.2478 | 3.556 | 0.00038 *** |
| AC | 0.1553 | 0.5168 | 0.300 | 0.76392 |
| AIC | 529.4563 | | | |
| BIC | 558.2218 | | | |

**Source:** Authors' computation aided by R package v4.1.1. Note: (**) and (***) denote significance at 5% and 1% respectively.

Table 4 presented the result of GM-Schweppes robust logistic model which is less fit than GM-Mallows but better than the ordinary logistic regression. The coefficients of the GM-Schweppes are slightly different from both the ordinary logistic regression and the GM-Mallows model, positioned between the two in terms of values. The significance levels of the variables are somehow similar to both models, with Gender, IFM, MSP, and MS being highly significant ($p < 0.001$). The robustness weights summary shows that 414 weights are close to 1, while 36 have lower weights. This is between the ordinary logistic regression (which would have all weights at 1) and the GM-Mallows model (which had 418 weights close to 1). The model converged in 5 iterations, the same as the ordinary logistic regression but one less than GM-Mallows, suggesting it might be computationally less intensive than GM-Mallows. The algorithmic parameters are the same as in the previous outputs, maintaining consistency. The GM-Schweppes robust logistic regression compares to both the ordinary logistic regression and the GM-Mallows model demonstrates a middle ground between the two in terms of robustness and that it's less fit than GM-Mallows but still an improvement over ordinary logistic regression.

**Table 5: Model Comparison based on coefficients of OL, GM-M and GM-S on HBV data**

| Variables | Ordinary Logistic | GM-Mallows | GM-Schweppes |
|---|---|---|---|
| (Intercept) | -2.2033 | -2.5314 | -2.4255 |
| GENDER | 0.5897 | 0.9215 | 0.9001 |
| IFM | 1.7041 | 1.1856 | 1.1564 |
| MSP | 0.9669 | 0.9473 | 0.9248 |
| HBT | 0.1058 | 0.4381 | 0.4277 |
| MS | 0.9024 | 0.9024 | 0.8809 |
| AC | 0.3417 | 0.1591 | 0.1553 |

From Table 5 above, the robust methods generally have slightly larger absolute coefficient values indicating that there might be less affected by potential outliers. The GM-Mallows tends to have the largest absolute coefficients, suggesting it might be the most aggressive in handling outliers. The direction (sign) of the effects remains consistent across all models. The statistical Significance shows that all the three models show similar patterns of statistical significance. Where Gender, IFM, MSP, and MS are statistically significant predictors of HPB. ($P < 0.001$) in all models. The variable IFM has the strongest effect (largest coefficient and highest significance). HBT and AC remain non-significant across all models. This consistency suggests that the overall interpretation of which variables are important doesn't change much between models. The positive coefficients indicate that increases in these variables are associated with higher odds of HPB.

## CONCLUSION

The result of the study shows that, all the three models under consideration, that's Logistic regression model, GM-Mallows robust logistic regression model and GM-Schweppes robust logistic regression model approaches can be used to fit and predicted Hepatitis B virus data. But the GM-Mallows robust model gives a better result than the logistic model and GM-Scheweppes robust model having the lowest AIC and BIC values. And this means that its more resistant to effect of outliers over the other two models.Therefore, the study concluded that the best model to be used for the analysis of HBV in the presence of outliers and multicollinearity is the GM-Mallows robust logistic regression, because of its more resistibility against outliers.

## REFERENCES

Ahmed, IdrissAbdelmajid, Cheng, Weihu,(2020) , " The Performance of Robust Methods in Logistic Regression Model", Open Journal of Statistics, , 10, 127-138 https://www.scirp.org/journal/ojs ISSN Online: 2161-7198 ISSN Print: 2161-718X, DOI: 10.4236/ojs.2020.10.

Anderson, Cynthia, and Randall E. Schumacker. (2003). "A Comparison of Five Robust Regression Methods with Ordinary Least Squares Regression: Relative Efficiency, Bias, and Test of the Null Hypothesis." Understanding Statistics 2 (2): 79–103. https://doi.org/10.1207/s15328031us0202_01.

Andrews, D F. (1974). "A Robust Method for Multiple Linear Regression." Technometrics 16 (4): 523–31. https://doi.org/10.1080/00401706.1974.10489233.

Ajuwon *et al*. (2021). Hepatitis B virus infection in Nigeria: a systematic review and meta-analysis of data published between 2010 and 2019. *BMC Infectious Diseases (2021) 21:1120* https://doi.org/10.1186/s12879-021-06800-6

Anteneh Z.A, Wondaye E, Mengesha E.W (2021). Hepatitis B virus infection and itsdeterminants among HIV positive pregnant women: Multicenter unmatched case-control study. *PLOS ONE* 16(4): e0251084. https://doi.org/ 10.1371/journal.pone.0251084

Almongy, Hisham Mohamed, and Ehab Mohamed Almetwaly. n.d. (2021) "Comparison between Methods of Robust Estimation to Reduce the Effect of Outliers": Researchgate.Net:https://www.researchgate.net/profile/Ehab_Almetwally/publication/