Journal of Statistics and Its Application, Likelihood Vol 1 No 2, December 2025

Robustness Comparison of MLE, Clean-MLE, Bianco–Yohai, and Weighted Bianco–Yohai Estimates in Binary Logistic Regression for Heart Disease Prediction

Muhammad Zulfadhli [1]

[1] Institut Teknologi Sepuluh Nopember, Indonesia, email: muhammadzulfadhli23@gmail.com

Abstract

Logistic regression is a popular statistical method for modeling the relationship between categorical response variables and one or more explanatory variables. However, standard estimation using Maximum Likelihood Estimation (MLE) is very sensitive to the presence of outliers, which can cause bias in the model parameters. This study aims to compare the performance of classical logistic regression models (MLE and Clean-MLE) with robust logistic regression models using Bianco-Yohai (BY) and Weighted Bianco-Yohai (WBY) estimation in predicting cardiovascular disease. The data used came from the Heart Failure Prediction Dataset, which included 918 observations with response variables of heart disease status and several independent variables such as age, blood pressure, cholesterol, fasting blood sugar, and maximum heart rate. Model goodness was evaluated by comparing the smallest chi-square values through arcsin transformation. The analysis results showed the presence of outliers in the dataset used. The comparison of chi-square values showed that the Weighted Bianco-Yohai (WBY) estimation method provided the lowest value of 1531.44 compared to MLE (1540.88), Clean-MLE (1538.03), and BY (1537.30). Thus, it can be concluded that the Binary Logistic Regression method with Robust Weighted Bianco-Yohai estimation is the best model for handling data containing outliers in heart disease prediction because it produces a more accurate and robust model.

Keywords: *Robust Logistic Regression, Outlier, Bianco-Yohai, Weighted Bianco-Yohai, Cardiovascular*

Introduction

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide and continues to be a major public health challenge. The high burden of CVD highlights the need for accurate risk prediction and early detection to support clinical decision-making, preventive intervention, and public health planning. In medical research, statistical modeling is frequently used to identify the relationship between disease status and several clinical risk factors, such as age, resting blood pressure, cholesterol level, fasting blood sugar, maximum heart rate, and other physiological indicators. A reliable statistical model is expected not only to classify disease status but also to provide interpretable information about how each predictor is associated with the probability of disease occurrence.

Binary logistic regression is one of the most widely used statistical methods for modeling a dichotomous response variable, such as whether an individual has heart disease or not. This method is commonly applied in epidemiological and clinical studies because it provides a probabilistic framework and allows the effect of each predictor to be interpreted through odds ratios. Hosmer, Lemeshow, and Sturdivant (2013) emphasized that logistic regression is highly relevant in applied health research because it combines interpretability, hypothesis testing, and predictive modeling. Similarly, Agresti (2013) explained that logistic regression is a fundamental model in categorical data analysis, particularly when the response variable consists of two categories.

https://doi.org/10.30605/likelihood.1229

In classical binary logistic regression, the parameters are usually estimated using Maximum Likelihood Estimation (MLE). Under ideal data conditions, MLE has desirable statistical properties, including consistency and asymptotic efficiency. However, these properties depend on the adequacy of the model assumptions and the quality of the data. In practice, health datasets often contain outliers, leverage points, or influential observations. These abnormal observations may arise from measurement error, data entry error, biological variability, extreme clinical conditions, or sampling inconsistency. Pregibon (1981) showed that influential observations in logistic regression can strongly affect fitted probabilities and parameter estimates. Therefore, ignoring outliers may lead to biased coefficients, unstable odds ratios, misleading significance tests, and reduced predictive reliability.

The sensitivity of MLE to outliers has encouraged the development of robust logistic regression methods. Robust estimation aims to reduce the excessive influence of abnormal observations while preserving the main information structure in the data. One important robust estimator for logistic regression was introduced by Bianco and Yohai. The Bianco–Yohai estimator modifies the likelihood-based objective function so that observations with extreme residual behavior have less influence on parameter estimation. This approach can improve the stability of logistic regression when the dataset contains contamination. However, the Bianco–Yohai estimator may still be affected by leverage points in the covariate space. To address this issue, Croux and Haesbroeck proposed the Weighted Bianco–Yohai estimator, which incorporates a weighting mechanism to control the influence of high-leverage observations. Thus, the Weighted Bianco–Yohai estimator is expected to provide stronger robustness when outliers occur in both the response pattern and the predictor space.

Several empirical studies have shown that robust logistic regression can outperform classical logistic regression in the presence of outliers. Idris et al. (2024) reported that robust logistic regression methods are useful for improving parameter stability when outliers are present. Lukman et al. (2025) emphasized that robust estimation is important when logistic regression is affected by outliers and multicollinearity. Akturk et al. (2025) also showed that robust logistic-type modeling can provide more reliable results when influential observations exist in the data. These findings indicate that robust logistic regression is highly relevant for biomedical and health datasets, where extreme observations and data quality problems are common.

Despite these developments, several research gaps remain. First, many studies on heart disease prediction still rely on classical logistic regression or machine learning models without specifically evaluating the effect of outliers on parameter estimation. Second, robust logistic regression methods, particularly Bianco–Yohai and Weighted Bianco–Yohai estimation, are still less frequently applied in empirical cardiovascular disease prediction studies. Third, previous studies often emphasize predictive performance but provide limited comparison of estimator robustness, model fit, and parameter stability. Fourth, comparative studies involving MLE, Clean-MLE, Bianco–Yohai, and Weighted Bianco–Yohai on heart disease data are still limited. Therefore, further empirical evaluation is needed to determine whether robust estimation methods can provide better performance than classical MLE when outliers are present in cardiovascular disease datasets.

Based on this gap, the present study aims to compare the performance of classical and robust estimation methods in binary logistic regression for heart disease prediction. The estimation methods compared in this study are Maximum Likelihood Estimation, Clean-MLE, Bianco–Yohai, and Weighted Bianco–Yohai. The dataset used is the Heart Failure Prediction Dataset, which consists of 918 observations and several clinical predictor variables. The comparison is conducted to identify which estimation method provides the most reliable model when the data contain outliers. The novelty of this study lies in its systematic comparison of classical and robust logistic regression estimators in the context of cardiovascular disease prediction. Unlike studies that only apply standard logistic regression, this study explicitly considers the presence of outliers and evaluates robust estimation methods as alternatives to MLE. In addition, the inclusion of Clean-MLE allows the study to compare the performance of removing outlying observations with the performance of robust estimators that reduce the influence of outliers without fully discarding them. This research is expected to provide methodological insight for researchers working with binary medical outcomes and practical guidance for developing more stable and reliable heart disease prediction models..

Method

The data used in this study is secondary data from the "Heart Failure Prediction Dataset" taken from the website kaggle.com. The observation unit is 918 cardiovascular patients. The response variable (y) in this study is cardiovascular patients with 0 (heart patients) and 1 (normal). The predictor variables are presented in the following Table 1.

Table 1. *Research Predictor Variables*

Variable	Variable Name	Scale	Description
Age	Patient age	continuous	
RestingBP	Resting Blood Pressure (mmHg)	continuous	
Cholesterol	Serum Cholesterol (mm/dl)	continuous	
FastingBS	Fasting Blood Sugar (mg/dl)	Categorical	1 : if > 120 0 : others
MaxHR	Maximum Heartbeat Reached	continuous	
Oldpeak	Oldpiek	continuous	

The data analysis stage in this study is as follows.

1. Performing descriptive statistical analysis on response and predictor variables
2. Forming a logistic regression model using classical estimation
3. Forming a logistic regression model using robust estimation
4. Testing model goodness using goodness of fit
5. Interpreting the best model

In this study, binary logistic regression was applied to model the probability of heart disease status based on several clinical predictor variables. The classical model was first estimated using Maximum Likelihood Estimation (MLE), which serves as the baseline method. MLE is commonly used in logistic regression because it estimates parameters by maximizing the likelihood function. However, this method is sensitive to outliers and influential observations, which may lead to unstable parameter estimates and less reliable model interpretation.

To evaluate the effect of outliers, outlier detection was performed before model comparison. Observations identified as outliers were used as the basis for forming the Clean-MLE model, in which logistic regression was re-estimated after excluding influential observations. This approach was compared with robust estimation methods to determine whether removing outliers or reducing their influence produces better model performance.

Furthermore, robust logistic regression was estimated using the Bianco–Yohai and Weighted Bianco–Yohai methods. The Bianco–Yohai estimator reduces the effect of extreme observations by modifying the likelihood-based objective function using a bounded loss function. Meanwhile, the Weighted Bianco–Yohai estimator extends this approach by assigning smaller weights to observations with high leverage or strong influence. Therefore, this method is expected to provide more stable parameter estimates when outliers are present in the predictor space.

The performance of MLE, Clean-MLE, Bianco–Yohai, and Weighted Bianco–Yohai estimation was then compared using the goodness-of-fit value. The model with the smallest chi-square value was considered to provide the best fit to the data. This comparison was used to identify the most robust and reliable logistic regression model for heart disease prediction

Results and Discussion

Outlier detection

After conducting outlier testing, the following plot was obtained.

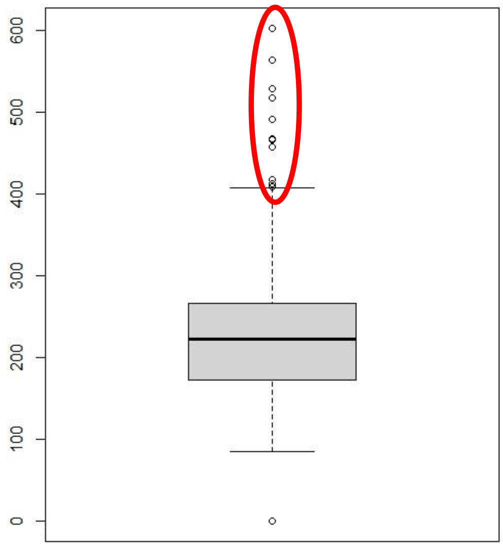

Figure 1. *Plot of Outlier Data*

Based on Figure 1, there is outlier data indicated by red circles. Therefore, it is necessary to use a method that is robust against outliers and compare the results when using classical estimation methods and robust methods.

Logistic Regression Model

The Logistic Regression Model is formed based on the coefficient values of each estimate as follows.

Table 2. *Logistic Regression Model*

Estimation Method	Logistic Regression Model
MLE	$\pi(x) = \dfrac{e^{19,80+ 0,01X_1+0,92X_2+1,39X_3-4,32X_4}}{1+ e^{19,80+ 0,01X_1+0,92X_2+1,39X_3-4,32X_4}}$
MLE-Clean	$\pi(x) = \dfrac{e^{21,05+ 0,01X_1+0,92X_2+1,41X_3-4,54X_3}}{1+ e^{21,05+ 0,01X_1+0,92X_2+1,41X_3-4,54X_3}}$
BY	$\pi(x) = \dfrac{e^{2,98-0,01X_1+0,10X_2-0,22X_3-0,55X_4}}{1+ e^{2,98-0,01X_1+0,10X_2-0,22X_3-0,55X_4}}$
WBY	$\pi(x) = \dfrac{e^{2,99- 0,01X_1+0,10X_2-0,22X_3-0,56X_4}}{1+ e^{2,99- 0,01X_1+0,10X_2-0,22X_3-0,56X_4}}$

Model Goodness Test

The goodness of the logistic regression model can be seen using the goodness-of-fit based on the smallest chi-square value. The chi-square values for each method can be seen in the following table.

Table 3. *Chi-Square Value Comparison*

Estimation Method	Chi-Square Value
MLE	1540,88
MLE-Clean	1538,03
BY	1537,30
WBY	1531,44

Based on these results, the Robust WBY method provides the smallest chi-square value. This shows that the Binary Logistic Regression method with Robust Weighted Bianco Yohai Estimation is the best method for handling outlier data. So the best model is

$$\pi(x) = \frac{e^{2{,}99-\ 0{,}01X_1+0{,}10X_2-0{,}22X_3-0{,}56X_4}}{1+\ e^{2{,}99-\ 0{,}01X_1+0{,}10X_2-0{,}22X_3-0{,}56X_4}}$$

The results indicate that the presence of outliers in the heart disease dataset affects the performance of the classical logistic regression model. This finding supports the argument that Maximum Likelihood Estimation (MLE), although widely used in logistic regression, may be sensitive to abnormal or influential observations. In medical datasets, outliers may occur due to measurement error, extreme biological variability, or unusual patient characteristics. If such observations are not properly handled, they may influence the estimated regression coefficients and reduce the stability of the model. This is consistent with Pregibon (1981), who emphasized the importance of diagnostic assessment in logistic regression because certain observations can strongly affect fitted probabilities and parameter estimates.

The comparison of goodness-of-fit values shows that the Weighted Bianco–Yohai estimator produces the smallest chi-square value compared with MLE, Clean-MLE, and Bianco–Yohai. This result indicates that the WBY method provides a better fit under the criterion used in this study. The superiority of WBY can be explained by its ability to reduce the influence of both response outliers and high-leverage observations in the predictor space. While Clean-MLE handles outliers by removing influential observations, this approach may also discard useful information from the dataset. In contrast, robust estimation methods reduce the influence of abnormal observations without completely eliminating them from the analysis.

The result is also in line with the robust logistic regression framework proposed by Bianco and Yohai, which modifies the likelihood-based estimation process to limit the effect of extreme observations. Furthermore, the Weighted Bianco–Yohai estimator extends this approach by incorporating observation weights, making it more suitable when the dataset contains leverage points. Similar findings were reported by Idris et al. (2024), who showed that robust logistic regression methods perform better than classical logistic regression when outliers are present. Lukman et al. (2025) also highlighted that robust estimation is useful in logistic regression because it can improve parameter stability under data contamination.

From an applied perspective, the use of robust logistic regression is important in heart disease prediction because clinical data often contain heterogeneous patient profiles. A model that is less affected by outliers may provide more reliable estimation and interpretation. However, the conclusion that WBY is the best method should be understood based on the goodness-of-fit criterion used in this study. To strengthen the predictive claim, future analysis should include additional performance measures such as accuracy, sensitivity, specificity, F1-score, and ROC-AUC. This is particularly important because, in medical prediction, the ability to correctly identify patients with heart disease is clinically more meaningful than relying only on a goodness-of-fit statistic.

Conclusion

Based on the results in this paper, the Robust Weighted Bianco Yohai method is the best method, as indicated by the smallest chi-square value of 1531.44. Therefore, the model used to predict food assistance acceptance in the United States is

$$\pi(x) = \frac{e^{2{,}99-\ 0{,}01X_1+0{,}10X_2-0{,}22X_3-0{,}56X_4}}{1+\ e^{2{,}99-\ 0{,}01X_1+0{,}10X_2-0{,}22X_3-0{,}56X_4}}$$

The Binary Logistic Regression Method with Bianco Yohai Robust Weighted Estimation is the best method for handling outliers. The author's suggestions are as follows. 1) Using other parameter estimation methods that are robust to outliers besides estimation using Bianco Yohai and weighted Bianco Yohai. 2) Comparing the regression models formed by adding other model goodness methods so that they can show real results.

Acknowledgment

The authors would like to express their sincere gratitude to Institut Teknologi Sepuluh Nopember (ITS) for the academic support, research facilities, and conducive research environment provided throughout the completion of this study.

References

Achmad, A. M., Zulfadhli, M., & Nurfaudzan, N. (2025). Binary logistic regression on the status of unemployment rate in West Java Province in 2023. *Likelihood: Journal of Statistics and Its Applications, 1*(1), 30–34.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Akturk, B., Beyaztas, U., & Shang, H. L. (2025). Robust functional logistic regression. *Advances in Data Analysis and Classification, 19*(1), 121–145.

Bianco, A. M., & Yohai, V. J. (1996). Robust estimation in the logistic regression model. In H. Rieder (Ed.), *Robust statistics, data analysis, and computer intensive methods* (Lecture Notes in Statistics, Vol. 109, pp. 17–34). Springer. https://doi.org/10.1007/978-1-4612-2380-1_2

Burger, D. A., van der Merwe, S., & Lesaffre, E. (2025). Addressing outliers in mixed-effects logistic regression: A more robust modeling approach. *arXiv*. https://doi.org/10.48550/arXiv.2504.13781

Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis, 44*(1–2), 273–295. https://doi.org/10.1016/S0167-9473(03)00042-2

Dey, D., Haque, M. S., Islam, M. M., Aishi, U. I., Shammy, S. S., Mayen, M. S. A., Noor, S. T. A., & Uddin, M. J. (2025). The proper application of logistic regression model in complex survey data: A systematic review. *BMC Medical Research Methodology, 25*(1), Article 15. https://doi.org/10.1186/s12874-024-02454-5

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., Hossain, R., & Hossain, K. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh. *BMC Cardiovascular Disorders*.

Idris, H. I., Mohammed, A., 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. https://doi.org/10.52589/AJMSS-YKDFCYQS

Lukman, A. F., Mohammed, S., Olaluwoye, O., & Farghali, R. A. (2025). Handling multicollinearity and outliers in logistic regression using the robust Kibria–Lukman estimator. *Axioms, 14*(1), Article 19. https://doi.org/10.3390/axioms14010019

Ma, Q. (2024). Recent applications and perspectives of logistic regression modelling in healthcare. *Theoretical and Natural Science, 36*, 185–190.

Nwohiri, A. M., Laguda, A. A., Olanite, A. A., & Olabamire, D. D. (2024). Logistic regression technique for cardiovascular disease prediction. *FUDMA Journal of Sciences, 8*(4), 266–275. https://doi.org/10.33003/fjs-2024-0804-2624

Otok, B. W., Zulfadhli, M., Pangesti, R. D., Kurniawan, M. I., Haryanto, A. E. P., & Kurniawan, I. (2026). Comparison of binary probit regression and Fourier series nonparametric logistic regression in modeling diabetes status at Hajj General Hospital Surabaya. *BAREKENG, 20*(1), 255–270.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics, 9*(4), 705–724. https://doi.org/10.1214/aos/1176345513

Saida, S., Zulfadhli, M., & Jurais, M. (2022). Analisis faktor-faktor yang mempengaruhi vaccine hesitancy (keragu-raguan vaksin) pada mahasiswa di era pandemi Covid-19. *Preventif: Jurnal Kesehatan Masyarakat, 13*(1), 144–154.

Suriaslan, A. S., Budiantara, I. N., Ratnasari, V., Zulfadhli, M., & Dhewy, R. C. (2025). An empirical study on unmet need: A statistical inference framework for truncated spline nonparametric binary logistic regression (TSNLBR). *Journal of Cultural Analysis and Social Change, 10*, 868–877. https://doi.org/10.20897/jcasc/2025.868

Tong, P., Hu, H., & Tong, L. (2025). Explainable logistic regression for heart disease risk prediction in community and clinical populations: Development and external validation study. *JMIR Cardio*. https://doi.org/10.2196/82462

Wang, H. (2024). Using logistic regression model to predict the future coronary heart disease. *Theoretical and Natural Science, 50*, 83–90. https://doi.org/10.54254/2753-8818/50/2024AU0147

Wu, H. (2024). Using logistic regression model to predict the future coronary heart disease. *Theoretical and Natural Science, 50*, 83–90. https://doi.org/10.54254/2753-8818/50/2024AU0147

Wulandari, R. (2018). Penerapan model regresi logistik robust pada status hiperkolesterolemia di RSUD Dr. Moewardi Surakarta menggunakan estimasi Bianco–Yohai. *Jurnal UNS*.

Zulfadhli, M. (2025). *Estimator regresi nonparametrik deret Fourier multivariabel untuk data kategori* [Doctoral dissertation, Institut Teknologi Sepuluh Nopember].

Zulfadhli, M., Budiantara, I. N., & Ratnasari, V. (2024). Nonparametric regression estimator of multivariable Fourier series for categorical data. *MethodsX, 13*, Article 102983. https://doi.org/10.1016/j.mex.2024.102983

Zulfadhli, M., Budiantara, I. N., Ratnasari, V., & Suriaslan, A. S. (2025a). Fourier series nonparametric regression estimator for modeling status of unmet need in East Java Province in 2023. *Communications in Mathematical Biology and Neuroscience*, Article 2025.

Zulfadhli, M., Budiantara, I. N., Ratnasari, V., & Suriaslan, A. S. (2025b). Semiparametric regression estimator of Fourier series for categorical data. *IAENG International Journal of Applied Mathematics, 55*(7), 2157–2164.

Zulfadhli, M., Budiantara, I. N., Ratnasari, V., Suriaslan, A. S., & Dhewy, R. C. (2025c). A statistical inference framework for FSNBLR: Modeling underdeveloped regional status in Eastern Indonesia. *MethodsX*, Article 103746. https://doi.org/10.1016/j.mex.2025.103746