ISRG Journal of Multidisciplinary Studies (ISRGJMS)

ISRG PUBLISHERS
Abbreviated Key Title: isrg j. multidiscip. Stud.
ISSN: 2584-0452 (Online)
Journal homepage: https://isrgpublishers.com/isrgjms/
Volume – IV, Issue - V (May) 2026
Frequency: Monthly

Robust Principal Component Regression with Wild Bootstrap for Handling Outliers, Multicollinearity, and Heteroskedasticity in Chronic Hepatitis B Data

Joseph Dedek Parhusip[1], Nusyirwan[2*], Misgiyati[3], Netty Herawati[4]

[1, 2, 3, 4] Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung, Indonesia

| **Received:** 07.05.2026 | **Accepted:** 14.05.2026 | **Published:** 19.05.2026

***Corresponding author:** Nusyirwan

Abstract

This study aims to analyze the performance of the Robust Principal Component (RPC) method combined with Wild Bootstrap in handling outliers, multicollinearity, and heteroskedasticity in chronic hepatitis B data. The data were obtained from the World Health Organization (WHO) and consist of several epidemiological indicators. The analysis methods include Principal Component Analysis (PCA), Least Trimmed Squares (LTS), and Wild Bootstrap using Wu and Liu multipliers. The results show that the dataset contains outliers, strong multicollinearity, and heteroskedasticity. The RPC-Wild Bootstrap method produces more stable parameter estimates, with RPC Boot Wu showing lower standard error and RMSE compared to RPC Boot Liu. Therefore, the RPC-Wild Bootstrap method is effective in producing more stable and reliable parameter estimates for complex real-world data.

Keywords: *Robust Principal Component, Wild Bootstrap, Outliers, Multicollinearity, Heteroskedasticity*

INTRODUCTION

Robust Principal Component (RPC) is an extension of Principal Component Analysis (PCA) designed to be resistant to the influence of outliers, providing a more stable representation of data variability in the presence of extreme observations [1]. PCA itself is widely used to transform correlated variables into orthogonal components to address multicollinearity. By integrating robustness, RPC reduces the impact of outliers while preserving essential data structure.

Wild Bootstrap is a resampling technique specifically developed to handle heteroskedasticity and non-normal error distributions, unlike classical bootstrap methods that rely on direct resampling of residuals [2]. This method generates new samples by modifying residuals using certain weighting distributions, allowing the heteroskedastic structure to be preserved [3]. As a result, the combination of RPC and Wild Bootstrap provides a more reliable framework for analyzing complex data that violate classical regression assumptions [4].

Chronic hepatitis B data exhibit complex characteristics, including variability in immunization coverage, new infection rates, the number of individuals living with chronic hepatitis, and birth dose vaccination coverage. These variations across countries are associated with differences in healthcare access, immunization

DOI: 10.5281/zenodo.20283563

policies, socioeconomic conditions, and prevention programs [5]. Such conditions often lead to the presence of outliers, multicollinearity, and heteroskedasticity.

Multicollinearity may arise due to strong linear relationships among explanatory variables [6]. In the context of hepatitis B data, such relationships may occur, for instance, between immunization coverage and infection rates. In addition, heteroskedasticity is commonly observed due to unequal variability across countries. Under these conditions, classical regression methods tend to produce inefficient and unstable parameter estimates. Therefore, robust approaches, such as RPC Boot Wu and RPC Boot Liu, are required to obtain more reliable and stable estimates [3].

Previous research [7] demonstrated that combining Robust Principal Component Regression with Wild Bootstrap yields two approaches, namely RPC Boot Wu and RPC Boot Liu, which significantly improve regression performance.

Despite these advantages, the application of RPC Boot methods to chronic hepatitis B data remains limited. Therefore, this study aims to apply RPC Boot Wu and RPC Boot Liu to obtain more accurate and stable parameter estimates.

LITERATURE REVIEW

Principal Component Analysis (PCA) is a multivariate statistical technique used to reduce data dimensionality while preserving most of the information contained in the original variables. PCA is particularly useful in overcoming multicollinearity among predictor variables, which often occurs in real-world data. The basic idea of PCA is to transform correlated variables into a set of uncorrelated principal components through linear combinations [8].

Before performing PCA, the data must be standardized to eliminate the effect of different measurement scales. This ensures that each variable contributes equally to the analysis. The standardization process is defined as:

$$X^* = \frac{x_i - \bar{x}_\iota}{\sqrt{\sum(x_i - \bar{x}_\iota)^2}}$$

After standardization, the correlation matrix is constructed to determine the structure of relationships among variables. The principal components are formed as linear combinations of the original variables, where each component represents a new uncorrelated variable derived from the original dataset.

In regression analysis, the presence of outliers can significantly affect parameter estimation and lead to unreliable results. Therefore, robust regression methods are required to produce more stable estimates. One widely used method is Least Trimmed Squares (LTS), which estimates parameters by minimizing the sum of squared residuals from a subset of observations that are least affected by outliers [9]. The LTS estimator is defined as:

$$\hat{\beta}_{LTS} = \arg\min \sum_{i=1}^{h} e_i^2$$

The LTS estimator has a high breakdown point and can produce robust parameter estimates through trimming and concentration step (C-step) procedures that reduce the influence of outliers [10]. The residuals obtained from the LTS model are used to identify influential observations. Residuals are defined as the difference between observed and predicted values:

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

To ensure robustness, residuals are standardized using a robust scale estimator, namely the Median Absolute Deviation (MAD), which is less sensitive to extreme values [9]. The MAD is defined as:

$$MAD = \frac{1}{0.6745}\, median|\varepsilon_i - median(\varepsilon_i)|$$

The standardized residual is expressed as:

$$r_i = \frac{\varepsilon_i}{MAD}$$

To further reduce the influence of extreme observations, the Tukey bisquare weighting function is applied. This function assigns smaller weights to large residuals and limits their impact on the estimation process:

$$w_i = \begin{cases} \left[1 - \left(\frac{r_i}{c}\right)^2\right]^2, |r_i| \le c \\ 0 \quad\quad\quad\quad\quad , |r_i| > c \end{cases}$$

Although LTS is robust to outliers, it does not address heteroskedasticity, which occurs when the variance of the error term is not constant. To overcome this issue, the Wild Bootstrap method is used as a resampling technique that preserves the structure of heteroskedasticity [3]. The bootstrap response is defined as:

$$y_i^{*b} = f(x_i, \hat{\beta}_{LTS}) + t_i^* \frac{\hat{\varepsilon}_i}{\sqrt{1 - h_{ii}}}$$

In the Wild Bootstrap procedure, Wu and Liu multipliers are used to generate bootstrap samples. These multipliers have zero mean and unit variance, allowing them to mimic the heteroskedastic structure and improve the accuracy of parameter estimation [7]. Bootstrap procedures in regression can also be implemented through residual bootstrap and paired bootstrap approaches, which form the basis for the development of resampling methods used in complex error structures [11].

The combination of PCA and LTS produces the Robust Principal Component (RPC) method, which is capable of handling multicollinearity and outliers simultaneously. Integrating this method with Wild Bootstrap provides a more comprehensive approach to address multicollinearity, outliers, and heteroskedasticity, resulting in more stable parameter estimates in complex data situations [7].

METHODOLOGY

The data used in this study are secondary data obtained from the World Health Organization (WHO) and consist of several epidemiological indicators of hepatitis B. The variables include the percentage of chronic hepatitis B cases in the general population (Y), the percentage of Hep B3 immunization coverage among one-year-old children (X_1), the percentage of hepatitis B birth dose vaccination coverage (X_2), the number of new hepatitis B infections (X_3), the number of people living with chronic hepatitis B (X_4), and the number of deaths due to chronic hepatitis B (X_5). These variables are used to analyze factors influencing the prevalence of chronic hepatitis B.

The analysis in this study applies the Robust Principal Component (RPC) method combined with the Wild Bootstrap approach to

DOI: 10.5281/zenodo.20283563

address multicollinearity, outliers, and heteroskedasticity. The analysis was carried out using RStudio software.

The steps of analysis are as follows:

1) Preparing the dataset obtained from the WHO database.
2) Conducting preliminary analysis, including detection of outliers, multicollinearity, and heteroskedasticity.
3) Performing data centering and scaling to standardize the variables.
4) Constructing the correlation matrix from standardized data.
5) Computing eigenvalues and eigenvectors from the correlation matrix.
6) Forming principal components by projecting standardized data into the eigenvector space.
7) Evaluating eigenvalues to determine the contribution of each principal component.
8) Estimating the regression model using the Least Trimmed Squares (LTS) method.
9) Determining initial weights based on residual values.
10) Standardizing residuals using the Median Absolute Deviation (MAD).
11) Applying Tukey bisquare weighting to obtain final weights.
12) Constructing weighted residuals (W_I^{LTS}) by multiplying the initial weight and the Tukey bisquare weight.
13) Generating bootstrap samples using the fixed-x approach and normalizing residuals using Normalized MAD (NMAD).
14) Re-estimating the LTS regression model on bootstrap samples.
15) Repeating the bootstrap process 1000 times to obtain empirical distributions of parameter estimators.
16) Evaluating model performance using bias, Root Mean Square Error (RMSE), and standard error.

RESULT AND DISCUSSION

To assess the performance of the Robust Principal Component (RPC) method combined with Wild Bootstrap in handling outliers, multicollinearity, and heteroskedasticity in chronic hepatitis B data, a descriptive analysis was first conducted. This analysis aims to provide an overview of the characteristics of the data used in the study. The results of the descriptive statistical analysis are presented in Table 1.

Table 1. Descriptive Statistics

Variable	Min	Max	Mean	Median	SD
Y	0.80	9.10	2.8675	2.20	2.128
X_1	42.0	99.00	88.85	96	14.2
X_2	1	158563	6908.32	544	25406.76
X_3	6074	17544081	1289663.7	422525	2992737
X_4	33.77	100	84.39	91.51	17.65
X_5	29	60535	4586.7	1243	10580.68

The descriptive analysis indicates substantial variability across countries in all observed variables. The prevalence of chronic hepatitis B (Y) shows notable differences, reflecting variations in endemic levels among countries. The HepB3 immunization coverage (X_1) and Hepatitis B birth dose vaccination coverage (X_4) both show relatively high average values, although variations across countries are still observed. In contrast, variables related to new infections (X_2), the number of people living with chronic hepatitis B (X_3), and mortality (X_5) exhibit wide ranges and high variability. This suggests significant disparities in disease burden, as well as the potential presence of outliers, heteroskedasticity, and strong relationships among predictor variables.

Before applying the Robust Principal Component (RPC) method with the Wild Bootstrap approach, preliminary data diagnostics were conducted to assess the presence of outliers, multicollinearity, and heteroskedasticity. The first diagnostic step involves detecting outliers. Outliers were identified using boxplot visualization and the Difference in Fits (DFFITS) method to detect influential observations.

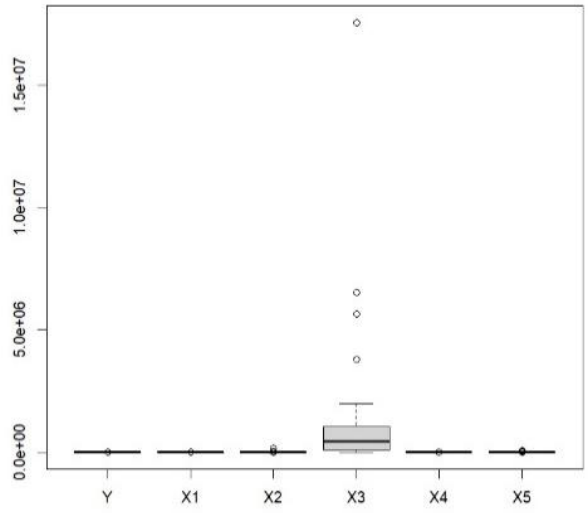

Figure 1. Boxplot outlier

Based on the boxplot in Figure 1, several observations indicate the presence of outliers across the variables, with the most prominent outliers observed in variable X_3. Further analysis using the Difference in Fits (DFFITS) method was conducted to identify influential observations. Using the calculated threshold value of 0.7746, several observations were found to exceed this limit.

Table 2. Outlier Countries

No	Country	DFFITS Value	|DFFITS|	information
8	Indonesia	-1.9104	1.9104	outlier
18	Lebanon	-1.2197	1.2197	outlier
24	Pakistan	-1.1236	1.1236	outlier
34	Timor leste	0.8703	0.8703	outlier
39	Vietnam	1.5624	1.5624	outlier

Based on Table 2, five observations have absolute DFFITS values exceeding the threshold of 0.7746, namely observations 8, 18, 24,

DOI: 10.5281/zenodo.20283563

34, and 39. These observations are considered influential and can be classified as outliers in the regression model. To further examine the data characteristics, a multicollinearity test was conducted using the Variance Inflation Factor (VIF). The results of the VIF analysis are presented in Table 3.

Table 3. VIF Results

Variable	VIF
X_1	1.589
X_2	11.974
X_3	108.321
X_4	1.958
X_5	72.042

The results indicate that variables X_2, X_3, and X_5 have VIF values greater than 10, suggesting the presence of strong multicollinearity among the predictor variables. In particular, X_3 and X_5 exhibit very high VIF values, indicating strong linear relationships with other predictors, which may lead to unstable parameter estimates under classical regression methods. In contrast, variables X_1 and X_4 show relatively low VIF values, indicating no significant multicollinearity. Furthermore, a heteroskedasticity test was conducted using the Glejser test. The results are presented in Table 4.

Table 4. Glejser Test Results

Variable	*p-value*
X_1	0.4077
X_2	0.5604
X_3	0.5052
X_4	0.0294
X_5	0.5626

The results show that variable X_4 has a p-value of 0.0294, which is less than 0.05, indicating a significant effect on the absolute residuals and confirming the presence of heteroskedasticity. Meanwhile, variables X_1, X_2, X_3, and X_5 have p-values greater than 0.05, indicating no significant effect. However, since at least one variable is significant, the data exhibit heteroskedasticity. Overall, the dataset is characterized by the presence of outliers, strong multicollinearity, and heteroskedasticity. Therefore, the application of the Robust Principal Component combined with the Wild Bootstrap method is highly appropriate, as it is capable of addressing these issues simultaneously.

After identifying the presence of outliers, multicollinearity, and heteroskedasticity, the model was estimated using the Robust Principal Component Regression combined with the Wild Bootstrap approach. First, Principal Component Analysis (PCA) was applied to transform the correlated predictor variables into a set of uncorrelated components. This transformation aims to eliminate multicollinearity while preserving most of the variability in the data. The results of the PCA are presented in Table 5, where the first principal components explain the majority of the total variance.

Table 5. Eigenvalue Results

Component	Eigenvalue	Variance Proportion	Total
PC1	3.2088	0.6418	0.6418
PC2	1.3338	0.2668	0.9085
PC3	0.3625	0.0725	0.9810
PC4	0.0895	0.0179	0.9989
PC5	0.0055	0.0011	1.0000

The eigenvalue results indicate that the first principal component contributes the largest proportion of variance, followed by the subsequent components. Although some components have relatively small eigenvalues, all principal components were retained in the analysis to preserve the information contained in the original variables, in accordance with the Robust Principal Component approach.

Subsequently, regression modeling was performed using the Least Trimmed Squares (LTS) method. This approach aims to obtain robust parameter estimates by minimizing the influence of outliers in the data. The estimated regression parameters are presented in Table 6.

Table 6. Robust Regression Parameter Estimation

Parameter	Coefficient
Intercept	2.1994
PC1	-3.6051
PC2	0.6841
PC3	0.3487
PC4	13.7491
PC5	-20.7706

Based on Table 6, the robust regression model using the LTS method with principal components as predictors is expressed as:

$$Y = 2{,}1994 - 3{,}6051(PC1) + 0{,}6841(PC2) + 0{,}3487(PC3) + 13{,}7491(PC4) - 20{,}7706(PC5)$$

After obtaining the Least Trimmed Squares (LTS) model, the next step was to calculate the initial residuals and determine the initial weights for each observation based on the inverse of the absolute values of the estimated residuals. Subsequently, the residuals were scaled and denoted as r_i, and the final weights were computed using the Tukey bisquare function.

After that, W_i^{LTS}(LTS weighted residuals) were calculated by combining the residuals with the initial weights and the Tukey bisquare weights, resulting in weighted residuals that reflect the adjusted influence of each observation. These weighted residuals were then used as the basis for the bootstrap procedure. In the next stage, bootstrap samples were generated using the fixed-x approach, and the residuals were then robustly normalized using the Normalized Median Absolute Deviation (NMAD).

Finally, the Wild Bootstrap method was applied to address heteroskedasticity using the weighted residuals obtained from the previous step. This procedure was performed through 1000 bootstrap replications to generate more reliable estimates of the model parameters.

DOI: 10.5281/zenodo.20283563

The results presented in Table 7 indicate that the Wild Bootstrap method produces stable and consistent parameter estimates. This confirms the effectiveness of the proposed approach in handling heteroskedasticity in the data.

Table 7. Result RPC wild bootstrap

Parameter	RPC Boot Wu			
	Estimate	SE	Bias	RMSE
Intercept	2.183	0.396	-0.016	0.396
PC1	-3.407	4.923	0.198	4.930
PC2	0.584	1.920	-0.100	1.924
PC3	0.328	1.856	-0.021	1.857
PC4	14.099	10.216	0.350	10.227
PC5	-18.974	41.833	1.796	41.871
Parameter	RPC Boot Liu			
	Estimate	SE	Bias	RMSE
Intercept	2.265	0.633	0.065	0.637
PC1	-5.388	8.541	-1.783	8.730
PC2	1.336	3.285	0.652	3.350
PC3	0.673	1.686	0.324	1.718
PC4	8.861	14.676	-4.888	15.476
PC5	-15.598	25.237	5.171	25.762

The results of this study indicate that the dataset is characterized by the presence of outliers, strong multicollinearity, and heteroskedasticity. These issues violate the classical assumptions of regression and may lead to unstable and inefficient parameter estimates when conventional methods are applied.

The application of Principal Component Analysis (PCA) successfully transformed correlated predictor variables into uncorrelated components, effectively addressing multicollinearity. Furthermore, the Least Trimmed Squares (LTS) method reduced the influence of outliers, resulting in more robust parameter estimates. The additional use of MAD and Tukey bisquare weighting further minimized the impact of extreme observations.

The implementation of the Wild Bootstrap method with 1000 replications proved effective in handling heteroskedasticity by producing more reliable and stable estimates. Overall, the combination of Robust Principal Component Regression and Wild Bootstrap provides a comprehensive and effective approach for analyzing complex data with multiple violations of classical assumptions.

CONCLUSIONS

This study concludes that the dataset exhibits outliers, multicollinearity, and heteroskedasticity, rendering classical regression methods unsuitable. The combination of Robust Principal Component Regression and the Wild Bootstrap approach effectively addresses these issues, producing more stable and reliable parameter estimates. Furthermore, utilizing the RPC Boot Wu scheme (Wu multiplier) provides superior performance compared to the RPC Boot Liu scheme (Liu multiplier) in terms of estimation accuracy. Therefore, the proposed method is highly suitable for analyzing complex data, particularly in epidemiological studies such as chronic hepatitis B. Future studies are recommended to apply this method to other datasets with similar characteristics and to explore alternative robust or bootstrap approaches to further enhance model performance

REFERENCES

1. M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, “ROBPCA: A new approach to robust principal component analysis,” *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.
2. B. Efron, “Bootstrap methods: Another look at the jackknife,” Ann. Stat., vol. 7, no. 1, pp. 1–26, 1979.
3. C. F. J. Wu, “Jackknife, bootstrap and other resampling methods in regression analysis,” *Ann. Stat.*, vol. 14, no. 4, pp. 1261–1295, 1986.
4. R. Davidson and E. Flachaire, “The wild bootstrap, tamed at last,” J. Econometrics, vol. 146, no. 1, pp. 162–169, 2008.
5. World Health Organization, “Hepatitis B: Key facts,” WHO, 2023.
6. D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 6th ed. New York: Wiley, 2014.
7. B. A. Rasheed, R. Adnan, S. E. Saffari, and D. M. Atiyaye, “Robust PC with wild bootstrap estimation of linear model in the presence of outliers, multicollinearity and heteroskedasticity error variance,” Int. J. Stat. Appl. Math., vol. 7, no. 3, pp. 85–93, 2022.
8. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., New York: Springer-Verlag, 2002.
9. P. J. Rousseeuw and M. Hubert, “Robust regression with continuous and binary regressors,” *J. Stat. Plann. Inference*, vol. 57, pp. 153–163, 1997.
10. C. Chen, *Robust Regression and Outlier Detection with the ROBUSTREG Procedure*. Cary, NC: SAS Institute Inc., 2002.
11. J. Shao and D. Tu, *The Jackknife and Bootstrap*. New York: Springer-Verlag, 1995.

DOI: 10.5281/zenodo.20283563