Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Original Article

Dealing with the Outlier Problem in Multivariate Linear Regression Analysis Using the Hampel Filter

Amira Wali Omer [a*], **Taha Hussein Ali** [a]

[a] Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University, Erbil, Iraq.

Submitted: 11 August 2024
Revised: 12 December 2024
Accepted: 15 January 2025

*** Corresponding Author**:
amira.omer@su.edu.krd

Keywords: Multivariate Regression Model, Outliers, Hampel Filter, Window size, and Threshold parameter.

How to cite this paper: A. W. Omer, T. H. Ali, "Dealing with the Outlier Problem in Multivariate Linear Regression Analysis Using the Hampel Filter", KJAR, vol. 9, no. 2, pp. 1-17, Jun. 2024. doi: 10.24017/science.2025.1. 1

Abstract: Outliers in multivariate linear regression models can significantly distort parameter estimates, leading to biased results and reduced predictive accuracy. These outliers may occur in the dependent variable or both independent and dependent variables, resulting in large residual values that compromise model reliability. Addressing outliers is essential for improving the accuracy and robustness of regression models. In this study, proposes a Hampel filter-modified algorithm to dynamically detect and mitigate extreme values, enhancing parameter estimation and predictive performance. The algorithm optimizes window size and threshold parameters to minimize mean square errors, making it a robust approach for handling outliers in multivariate regression analysis. To assess its effectiveness, simulations and real datasets were analyzed using a MATLAB-based implementation. The algorithm was compared with the classical Hampel approach to evaluate improvements in outlier detection and suppression. The results indicate that the proposed method effectively identifies and removes extreme values, leading to improved parameter estimation accuracy, enhanced model stability, and greater predictive performance and the performance was analyzed using the Mean Squared Error (MSE). The adaptive nature of the filter minimizes the impact of outliers, ensuring a more reliable regression model. The Hampel filter-modified algorithm provides an effective and adaptive solution for handling outliers in multivariate regression models. By dynamically identifying and mitigating extreme values, it enhances model accuracy, strengthens predictive capabilities, and ensures greater resilience against data variability. This approach offers a valuable tool for researchers and practitioners working with outlier-prone datasets, significantly improving the reliability of multivariate regression analysis.

1. Introduction

In the realm of statistical analysis, multivariate linear regression serves as a cornerstone technique for understanding and predicting the complex relationships among multiple variables (dependent and independent) [1]. Multivariate linear regression aims to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable [2]. Its application spans diverse fields, from finance to healthcare, where accurate modeling can drive critical decisions. However, the presence of outliers poses a significant challenge, as these anomalous observations can skew results and undermine the integrity of the analysis. Consequently, effectively addressing outliers is crucial for ensuring the reliability of regression models [3]. Outliers are defined as data points significantly different from the rest. An outlier may be due to variability in the measurement or it may indicate experimental errors; the former are sometimes legitimate observations and the latter should be excluded from the data set. There are two types of outliers, where univariate outliers are outliers in a single variable and multivariate outliers are outliers that occur in a multivariate space and are detected based on the combination of multiple variables [4]. The significance of this study lies in its focus on the Hampel filter, a robust statistical method designed to detect and mitigate the impact of outliers within multivariate datasets. Traditional methods often rely on univariate techniques, which may overlook the interconnectedness of variables and lead to misleading conclusions. In contrast, the Hampel filter employs a more nuanced approach by utilizing robust measures of central tendency, allowing for a comprehensive assessment of the data points while minimizing the influence of outliers [5]. The steps involved in the Hampel filter are first to select a window of size (k), the number of data points to be considered on each side of the central point, and to calculate the median (M). Calculating the median of the values within the window, and determining the Median Absolute Deviation (MAD), which is a robust measure of the variability in the data, identifies outliers and compares each data point to the median within the window. If the absolute deviation of a point from the mean is greater than the chosen multiple of MAD, it is marked as an outlier. Finally, outliers can be replaced by the median value to reduce their impact but substituting outliers can lead to biased parameter estimates if the substitution is done incorrectly, i.e. replacing outliers can be useful in some cases, but it must be done cautiously to avoid introducing bias and losing valuable information. A key development of Hampel's filter centers on using median and MAD to determine outliers, providing a robust alternative to traditional methods that rely heavily on the mean and standard deviation [6].

The primary aim of the Hampel filter is to improve the robustness of the statistical analyses by reducing the influence of outliers, which can distort the results of conventional methods such as linear regression. These outliers can result from measurement errors, data entry mistakes, or the inherent variability of the data. A way to respond to this is by replacing the detected outliers with values that are more consistent with the bulk of the data. This replacement helps to reduce the impact of outliers on subsequent analyses and improves the robustness of statistical models by reducing the sensitivity to outliers. This results in more reliable and stable parameter estimates that maintain the integrity of the data's structure by only modifying outliers rather than applying broad transformations that could distort the entire dataset. In this research, focused on window size (k), and thresholding (nd), Hampel gives these parameters and the use of a method was undertaken to obtain the ideal values for these parameters to give the best estimate of the model's lowest Mean Square Error (MSE). They vary according to the data and the amount of outliers, and the parameter has recently become a powerful mathematical technique to approximate the outliers of the regression line and compare the results before and after excluding the outliers using less MSE. A model with less MSE is better and in this study, the proposed model is the best because it has the smallest MSE [5].

The layout of this paper is structured as follows: section 2 discusses related works. Section 3 outlines the methodology employed in applying the Hampel filter within the context of multivariate linear regression. Section 4 presents the results of our simulations and real-world case studies, followed by a discussion in section 5 that interprets these findings. Finally, section 6 concludes the paper and suggests avenues for future research.

2. Related Works

The challenges posed by outliers in multivariate linear regression have been the subject of extensive academic inquiry, with numerous studies highlighting their detrimental impact on model accuracy and reliability. Classical statistical methods, such as least squares estimation, are notably sensitive to extreme values which can lead to significant biases in coefficient estimation and increased prediction errors [4]. As highlighted by numerous researchers [7, 8] conventional approaches often overlook the complexity of real-world datasets, making them inadequate in scenarios where the presence of outliers is prevalent. This inadequacy constitutes the central research problem in this field—namely, the need for effective outlier detection and management strategies that enhance the robustness of multivariate linear regression analyses. Among the existing methodologies, various robust statistical techniques have been explored, including robust regression methods like Theil–Sen estimates and M-estimators, as well as robust covariance estimators [9, 10]. However, these systems still experience limitations when addressing high-dimensional data and fail to effectively isolate or mitigate the effects of outliers without sacrificing the overall model's integrity.

The primary objective of this section is to assess the efficacy of the Hampel filter as a robust outlier detection technique that specifically addresses these shortcomings in multivariate linear regression contexts. By evaluating its performance against traditional methods, this research seeks to determine the extent to which the Hampel filter can preserve data integrity while minimizing the adverse effects of outliers [11, 12] Additionally, the literature reveals a gap in comprehensive studies illustrating the integration of this filtering technique within multivariate frameworks, which this research aims to address. The significance of examining the existing literature is vast, both academically and practically. A thorough understanding of prior research provides a foundational basis for justifying the adoption of the Hampel filter in contemporary analytical practices. It also emphasizes the importance of developing advanced statistical techniques that can be reliably deployed in various fields, such as finance and healthcare, where accurate data analysis is critical [13, 14] Furthermore, as outliers can obscure true relationships in the data, the findings from this research section will contribute to fostering more accurate statistical models, thereby enhancing decision-making processes based on empirical evidence [15]. This convergence of rigorous academic inquiry and practical application underscores the importance of addressing outlier management in statistical analysis today.

3. Materials and Methods

After applying the Hampel filter to the datasets, multivariate linear regression models were constructed using the cleaned data. The models were formulated to evaluate the relationships among the independent and dependent variables, with special attention to the coefficients and their statistical significance. For comparison, traditional regression models were also developed using the original datasets, which included outliers, to demonstrate the impact of outlier presence on model performance.

3.1. Multivariate Multiple Linear Regression

It is a statistical method used to predict multiple dependent variables using more than one independent variable. It is also employed to ascertain the quantitative correlation between said variables [1]. The variable you want to predict should be continuous, and your data should meet the other assumptions such as linearity, no outliers, a similar spread across range, the normality of the residuals, and have no multicollinearity [16].

3.1.1. Multivariate Regression Model

A multivariate regression model aims to predict the dependent variables using multiple independent variables [17]. It is an extension of the multiple regression model which predicts a single dependent variable, and the multivariate regression model has specific assumptions (e.g., normality, independence, linearity) similar to the multiple regression model, Increasing the hypothesis of a linear relationship between the dependent variables [18]. Mathematically, the multivariate regression model can be written as:

$$Y = XB + E \qquad (1)$$

$$\begin{aligned} Y_1 &= \beta_{01}+\beta_{11}X_1+\cdots+\beta_{q1}X_q+e_1 \\ Y_2 &= \beta_{02}+\beta_{12}X_1+\cdots+\beta_{q2}X_q+e_2 \\ \vdots &\quad \vdots \quad \vdots \qquad\quad \vdots \qquad \vdots \\ Y_p &= \beta_{0p}+\beta_{1p}X_1+\cdots+\beta_{qp}X_q+e_p \end{aligned}$$

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix} + \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{np} \end{bmatrix} \qquad (2)$$

Where Y (n × p) is the matrix of dependent variables, X [n × (q + 1)] is the matrix of independent variables, β[(q + 1) × p] is the matrix of coefficients, and E (n × p) is the matrix of error terms. Thus, each row of (Y) contains the values of the (p) dependent variables, dependent variables must be quantitative data, while independent variables can be quantitative or qualitative. Each column of (Y) consists of the (n) observations on one of the (p) variables. Regression Coefficients (β): Each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. Here, there is no problem of multicollinearity between the independent variables, so there is no joint interpretation for the Coefficients [19].

3.1.2. Least Squares Estimation in the Multivariate Model

The goal of least square estimation (LSE) is to find the coefficient matrix (β) that minimizes the sum of squared residuals:

$$E = Y - X\beta \qquad (3)$$

The Sum of Squared Residuals (S) is:

$$S = tr(E^T E) = [(Y - X\beta)^T (Y - X\beta)] \qquad (4)$$

Where (tr) denotes the trace of a matrix [20].

Estimation of (β): To minimize (S), we take the derivative of (S) concerning (β) and set it to zero:

$$\frac{dS}{d\beta} = -2X^T(Y - X\beta) = 0, \text{ Solving for } \beta:$$

$$X^T Y = X^T X\beta$$

$$\beta = (X^T X)^{-1} X^T Y \qquad (5)$$

This equation gives the LSE for the coefficient matrix (β).

The unbiased LSE possesses the property of being an unbiased estimator, if (β) is expected [E] = 0. The expected value of (β) is equivalent to the true coefficient matrix and the least variance among all unbiased linear estimators. The estimator has a minimum variance, a property known as the Gauss-Markov theorem, and finally normality. If the Error terms (E) are normally distributed, then the LSE for (β) is also normally distributed [2, 21].

3.2. Outliers

Outliers are data points that deviate significantly from the general pattern of the data. In the context of multivariate regression, outliers can have a significant effect on the estimated coefficients and overall model fit. Identifying and handling outliers is critical to ensure robust and accurate model estimation [22].

3.2.1. Outliers in Multivariate Regression

An outlier in a multivariate dataset is an observation that lies an unusual distance from other observations. Outliers can be caused by variability in the data, measurement errors, and other anomalies [4].

3.2.2. Detection Methods

Detection techniques are essential for addressing the issue of outliers in multivariate linear regression analysis, and the Hampel filter is a widely employed tool for this purpose [23, 24]. The Mahalanobis distance, residual plots, Cook's distance, and Leverage points are among the commonly utilized detection methods in multivariate linear regression analysis [25]. The Mahalanobis distance and Hampel filter are commonly used methods for anomaly detection in multivariate linear regression analysis. Mahalanobis distance calculates the distance between a data point and the dataset's center, while the Hampel filter accurately identifies and handles outliers within the dataset [26, 27]. The formula for Mahalanobis distance is as follows:

$$D^2 = (X-\mu)^T \Sigma^{-1}(X-\mu) \qquad (6)$$

Where (X) is the observation vector, (μ) is the mean vector, and (Σ) is the covariance.

To establish an appropriate cut off value, the Mahalanobis distance can be analysed using a chi-square distribution with degrees of freedom (DF) equal to the number of variables in the data. The cutoff value is determined based on a selected significance level (α), which dictates the stringency of identifying outliers [28]. The process of detecting outliers in multivariate linear regression involves determining DF and selecting a significance level [29]. The cut off can be calculated using the chi-square cumulative distribution function. Adjusting the significance level can result in the stricter or more lenient detection of outliers [30]. Visualizing the Mahalanobis distances through a histogram or Q-Q plot can provide a helpful visual confirmation of outliers by overlaying the chi-square distribution [31]. When it comes to dealing with outliers, especially in the context of multivariate linear regression, robust methods are preferred to minimize their influence. Here are three robust methods often used alongside Mahalanobis distance:

A. Fast Minimum Covariance Determinant

The fast minimum covariance determinant (FMCD) is an effective estimator utilized in multivariate analysis to pinpoint and diminish the impact of outliers on covariance matrix computations. Common estimators such as the sample covariance matrix are vulnerable to outliers, resulting in a potential distortion of the data's structure and a substantial influence on models. It identifies a subset of data with the smallest determinant, reducing the influence of outliers on models and ensuring a more accurate covariance matrix [32]. Here are some important points to understand about FMCD, along with a detailed explanation of the key equations:

- The aim of the Minimum Covariance Determinant (MCD) is to select a subset H⊂{1,2,...,n} that consists of h observations (where h≤n) and leads to the smallest determinant of the covariance matrix computed from this subset**:**

$$\hat{\Sigma}_{MCD} = \frac{arg\ min}{\det(\Sigma_H)} \quad for\ all\ H\ of\ size\ h \qquad (7)$$

In this case, (Σ_H) is the covariance matrix that is obtained from subset H.

- The subset size, denoted as h, is often chosen to balance robustness and efficiency. The specific value of h determines the breakdown point of the estimator, where higher values of h result in increased resilience against outliers [33].
- The mean μ_H and covariance Σ_H for the chosen subset H are calculated as follows:

$$\mu_H = \frac{1}{h}\sum_{i\in H} x_i \qquad (8)$$

$$\Sigma_H = \frac{1}{h-1}\sum_{i\in H}(x_i-\mu_H)(x_i-\mu_H)^T \qquad (9)$$

- The covariance matrix determinant det (Σ_H) is iteratively minimized by refining the subset H to identify the one with the lowest determinant. This process can be described as:

$$H_{optimal} = \underset{H}{arg\min\det(\Sigma_H)} \qquad (10)$$

- Once the most effective subset $H_{optimal}$ is identified, the Mahalanobis distance $D(x_i)$ for each observation x_i is computed using μ$H_{optimal}$ and $\Sigma H_{optimal}$:

$$D(x_i) = \sqrt{(x_i-\mu_{H_{optimal}})^T \sum\nolimits_{H_{optimal}}^{-1}(x_i-\mu_{H_{optimal}})} \qquad (11)$$

This distance is utilized to pinpoint any outliers (observations that have Mahalanobis distances significantly higher than the mean of $H_{optimal}$).

- To maintain consistency, it is common to adjust the final covariance estimate by a correction factor, typically represented by c, to consider both the sample size and subset size [34]:

$$\hat{\Sigma}_{FMCD} = c * \Sigma_{H_{optimal}} \tag{12}$$

B. *Orthogonalized Gnanadesikan-Kettenring*

The Orthogonalized Gnanadesikan-Kettenring (OGK) estimator is a reliable technique for estimating the covariance matrix when dealing with outliers, especially in high-dimensional data scenarios. It overcomes the limitations of traditional estimators like sample covariance, which can generate inaccurate estimates due to outliers. The OGK estimator uses robust methods for pairwise covariance and converts the data into orthogonal components to streamline and enhance the process of estimating covariance. The OGK process can be delineated through the following set of procedures:

- Standardized Data: by employing robust univariate scale estimators like MAD to mitigate the impact of outlier values and standardize each variable effectively [21].
- Pairwise Covariance Estimation: For each pair of variables (X_i, X_j), calculate robust pairwise covariances, ($\hat{\sigma}_{ij}$) using robust bivariate statistics. The Gnanadesikan-Kettenring estimator for covariance between two variables X_i and X_j is:

$$\hat{\sigma}_{ij} = \frac{1}{4}\left(MAD\left(X_i + X_j\right)^2 - MAD\left(X_i - X_j\right)^2\right) \tag{13}$$

Where MAD represents the robust scale of the sum and the difference of the two variables.

- Initial covariance matrix: The initial covariance matrix $\hat{\Sigma}$ is constructed using these pairwise covariances ($\hat{\sigma}_{ij}$) and robust variances for each variable.
- Orthogonalization: To stabilize the initial estimate, the OGK method performs an orthogonalization (often via eigenvalue decomposition) on the initial covariance matrix $\hat{\Sigma}$:

$$\hat{\Sigma} = P\Lambda P^T \tag{14}$$

Where P is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. The data is then transformed into an orthogonal basis defined by P, with transformed data Y=XP.

- Final Covariance Matrix Estimation: After orthogonalizing, robust univariate scales (such as MAD) are applied to each component of (Y) to get a robust scale matrix. The final robust covariance matrix Σ_{OGK} is given by:

$$\Sigma_{OGK} = P \, diag\left(scale(Y_1), scale(Y_2), \dots, scale\left(Y_p\right)\right) \, P^T \tag{15}$$

where $scale(Y_i)$ is a robust measure of scale for each orthogonalized variable [31].

C. *Olive-Hawkins*

The Olive Hawkins (OH) estimator is a robust alternative for computing the covariance matrix in multivariate data analysis, particularly suited for datasets that may contain outliers. Named in honor of researchers Olive and Hawkins, this method aims to iteratively adjust the impact of outliers on the covariance structure to determine a more resilient covariance matrix. The OH estimator reduces the influence of extreme data points on the covariance matrix by assigning varying weights to the observations and iteratively adjusting these weights for a more stable estimate. It works through the following steps:

- Initial Mean and Covariance Estimate: An initial Mean Vector (μ_0) and Covariance Matrix (Σ_0) are computed using a subset of the data to provide a starting point. This initial subset can be chosen using methods like MCD or FMCD to get a preliminary robust estimate [33].
- Weight calculation based on Mahalanobis distance: The Mahalanobis distance (D_i), is calculated for each data point (x_i) based on the current mean and covariance estimates where:

$$D_i = \sqrt{D^2} \tag{16}$$

Then, a Weight (w_i) is assigned to each data point using a weight function that decreases with increasing (D_i). A common choice is Tukey's bi-weight function:

$$w_i = \begin{cases} \left(1-\left(\frac{D_i}{2}\right)^2\right)^2 & if\ D_i < c \\ 0 & otherwise \end{cases} \quad (17)$$

Where c is a cutoff parameter that determines the threshold beyond which points are considered outliers.

- Re-estimate Mean and Covariance with Weights: Using the (w_i), the Mean (μ) and Covariance (Σ) are re-estimated as follows:

$$\mu_{new} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (18)$$

$$\Sigma_{new} = \frac{\sum_{i=1}^{n} w_i (x_i - \mu_{new})(x_i - \mu_{new})^T}{\sum_{i=1}^{n} w_i} \quad (19)$$

- Iterative Refinement: The process of calculating (w_i), and updating (μ) and (Σ) is repeated until convergence, typically when the changes in (μ) and (Σ) between iterations fall below a threshold. This iterative refinement reduces the influence of points far from the central structure of the data [35].

3.3. *Hampel Filter*

A robust statistical tool that identifies and mitigates the influence of outliers without necessitating strict assumptions about data distribution as required by classical methods. It relies on the MAD and employs a rolling window for the identification of outliers [5]**.** Configuring the Hampel filter involves two parameters:

3.3.1. *Window Size*

The window size (k) parameter determines the moving window size used to evaluate each data point. It essentially defines the scope within which we look for outliers [36]. There are some references and guidelines for selecting an appropriate window size from the data characteristics, and the Hampel filtering process involves selecting the appropriate k based on the data's characteristics and the frequency of anticipated outliers. Smaller windows are better for data with many outliers, while larger windows are more effective for data with few outliers [37]. A common heuristic is to set k around 3 to 5 times the expected width of an outlier, which balances sensitivity and robustness. Fine-tuning k may be useful for data with known noise levels or varying outlier magnitudes. For periodic data, k can be chosen based on the period length to accurately capture cyclic behavior. The window size should also consider the data's sampling rate and expected outlier duration [6]**,** and the other is robustness vs. sensitivity. A larger window size enhances smoothing and reduces outlier sensitivity, while a smaller window size increases sensitivity to short-term outliers but may cause false positives due to noise [20].

3.3.2. *Threshold*

Careful threshold selection is essential to avoid triggering outlier detection for valuable data. The threshold determines how far a data point must deviate from the median to be considered an outlier. There are several steps to adjust thresholds in the Hampel filter, first calculated MAD where it is a robust measure of statistical dispersion used to set the threshold for outlier detection, calculated as the MAD from the data's median [35] where

$$MAD = median(|x_i - median(x_i)|) \quad (20)$$

where (x_i) is the data point. The other step is threshold calculation where the selection of the scaling factor for the (MAD) threshold, typically ranging from (3 to 3.5), can be influenced by the distribution of the data, levels of noise, and the objectives of the analysis. The threshold is typically set as a multiple of the (MAD) [38]. This aligns with the properties of the normal distribution. The formula for the threshold is:

$$Thres\ hold = t * MAD \tag{21}$$

When (t) is the chosen threshold factor [19]. The choice of factor for outlier detection depends on factors like Gaussian and Heavy-Tailed Probability Distributions. For normal Gaussian data, (3 to 3.5) times (MAD) is recommended but higher thresholds may be needed for heavier-tailed distributions. In high-noise environments, a factor of (3.5) or higher minimizes false positives. In low-noise and clean data, a smaller factor captures true anomalies better. Empirical tests and combining MAD with other estimators can refine the process [39]. The final step is the scaling factor for Gaussian distribution, is utilized to standardize the MAD so then it can be compared to the standard deviation of a normal distribution. This standardization is essential because the MAD is less influenced by extreme values and offers a more resilient measure of variability in comparison to the standard deviation, particularly when outliers are present [20]. The formula is as follows:

$$\text{Threshold } = t * 1.4826 * MAD \tag{22}$$

The factor of (1.4826) is derived from the statistical relationship between the MAD and the standard deviation in a normal distribution. In a normal distribution, 50% of data falls within the MAD, and 68% falls within one standard deviation. The relationship between the standard deviation and MAD is expressed as σ ≈1.4826*MAD. This relationship is based on the approximate expected value of the MAD for a normal distribution being approximately σ / sqrt {2}. The MAD is modified by the scaling factor where:

$$Adjusted\ MAD = 1.4826 * MAD \tag{23}$$

The modified MAD represents variation similar to the standard in typical situations, offering a strong substitute for the standard deviation, which is significantly impacted by outliers. This method is particularly useful in the statistical analysis of data with outliers, providing a more reliable evaluation of variability while minimizing the impact of extreme values, especially in multivariate linear regression analysis.

3.3.3. Hampel Identifier

The Hampel identifier is a variation of the three-sigma rule of statistics that is robust against outliers. Given a sequence $x_1, x_2, x_3, \dots, x_n$ and sliding window of length k, define the point-to-point median and standard-deviation estimates using [5]:

$$Local\ median = \ mi \ = median\ (x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \tag{24}$$

$$\text{Standard deviation} = \ \mathrm{\sigma i} \ = \ \mathrm{k} * \ \text{median} \ (|x_{i-k} - m_i|, \dots, |x_{i+k} - m_i|) \tag{25}$$

$$\text{where } k = \frac{1}{\sqrt{2} erf^{-1}\left(\frac{1}{2}\right)} \approx 1.4826 \text{ the quantity } \frac{\sigma_i}{k} \text{ is the MAD.}$$

If a sample x_i is such that $|x_i - m_i| > n_\sigma \sigma_i$

For a given threshold n_σ then the Hampel identifier declares x_i an outlier and replaces it with m_i. Near the sequence endpoints, the function truncates the window used to compute m_i and σ_i.

- $i < k + 1$

$$m_i = median(x_1, x_2, x_3, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \tag{26}$$

$$\sigma_i = k * median(|x_1 - m_1|, \dots, |x_{i+k} - m_i|) \tag{27}$$

- $i > n - k$

$$m_i = median(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{n-2}, x_{n-1}, x_n) \tag{28}$$

$$\sigma_i = k * median(|x_{i-k} - m_i|, \dots, |x_n - m_n|) \tag{29}$$

For expressions of erfinv (1-x), use the complementary inverse error function (erfcinv) instead. This substitution maintains accuracy. When(x) is close to 1, then (1 – x) is a small number and may be rounded down to 0 causing numerical instability in (1−x), this leads to potential rounding errors in floating-point computation. Instead, replace erfinv (1-x) with erfcinv (x)]. To ensure that the substitution is mathematically accurate and computationally stable, (x) must be specified as a positive number within the interval 0 < x ≤ 1, since the function erfcinv (x) is only defined for x > 0; for x≤0, erfcinv (x) lacks meaning or mathematical definition [35].

3.4. Mean Square Error

MSE is a common metric used to evaluate the performance of filters, including the Hampel filter. In the context of the Hampel filter, MSE can be used to quantify how well the filter reduces noise and outliers from the data, comparing the filtered data to the true or expected values. The MSE measures the average squared difference between the estimated values (filtered data) and the actual values (true data). It is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{30}$$

Where: (y_i) are the true values, $(\hat{y}_i)$ are the filtered values, and (n) the number of estimated parameters [40]. The MSE is a crucial metric for evaluating the performance of Hampel filter. It quantifies the filter's effectiveness in reducing noise and outliers [1]. Comparing the MSE of the Hampel filter with other models or methods, such as those without the filter or using different outlier detection techniques, is beneficial. A lower MSE indicates the filter's effectiveness [40].

4. Results

The proposed method involves choosing the optimal window size (k) and optimal threshold parameter (nd) values that produce the minimum MSE of the multivariate regression models (MRM) and treating outliers (in the dependent variables or both independent and dependent variables) through the following steps:

Step 1: Set k = 3 and nd = 0.1, 0.15, 0.20 …, 10. Use a Hempel filter at k and all nd values. The parameters of an MRM were estimated for the filtered data at k and all nd values. The MSE value was calculated for each estimated MRM and determined the optimal (nd) value that gave the minimum MSE for the model.

Step 2: Set optimal (nd) and k = 1, 2, …, 100. Use a Hempel filter at optimal (nd) and all k values. The parameters of an MRM were estimated for the filtered data at optimal (nd) and all k values. The MSE value was calculated for each estimated MRM and determined the optimal (k) value that gave the minimum MSE for the model.

Step 3: Set the optimal (k) and optimal (nd) values. Use a Hempel filter at optimal (k) and optimal (nd). The parameters of an MRM were estimated for the filtered data at optimal (k) and (nd) values. The MSE value was calculated for the estimated MRM for the filtered data at optimal (k) and optimal (nd) values.

4.1 Simulation Study

To compare the proposed and classical methods, data for the multivariate regression model was generated assuming three dependent variables and four independent variables, a sample size equal to (100) and a parameter matrix, chosen randomly:

$$\beta' = \begin{bmatrix} 2 & 4 & 3 & 3 & 6 \\ 4 & 3 & 6 & 5 & 4 \\ 6 & 5 & 4 & 2 & 2 \end{bmatrix}$$

A random error has a multivariate standard normal distribution with some outliers added to the dependent variables. Outliers in the dependent variables were diagnosed using Mahalanobis distances and three robust methods (FMCD, OGK, and OH) as in Figure 1.

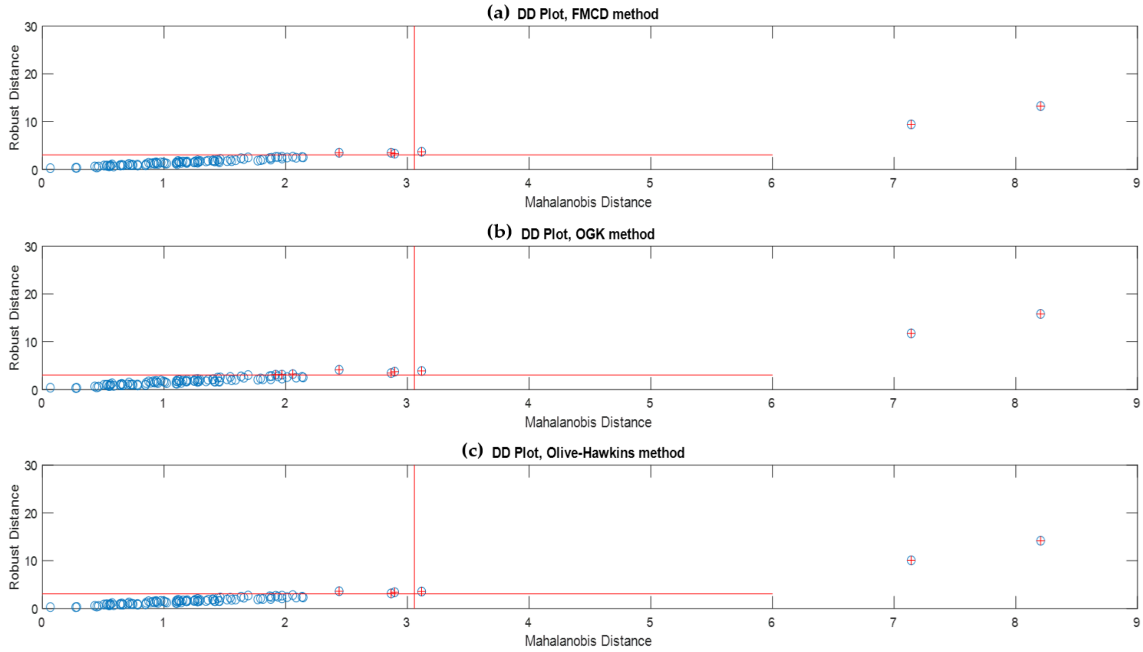

Figure 1: Identifying outliers for the dependent variables (first experiment).

Figure 1 shows how the data points are spread out based on their Mahalanobis distance on the horizontal axis and Robust Distance (FMCD, OGK, OH) on the vertical axis. Each point on the graph represents an observation in the dataset. Data points near the origin point (0,0) are considered normal, while the red lines indicate the threshold limits for the Mahalanobis and Robust distances, helping to identify potential multivariate outliers. The observations that are located beyond these limits, especially those in the top-right section, have high values for both metrics, showing that they deviate from the main data cluster. These points are considered outliers and indicate that they differ significantly from the majority of the dataset, noting the presence of several outliers in the dependent variables, which are shown in red for the first experiment. An MRM was estimated for the data from the first experiment, and then the residual values were computed as shown in figure 2.

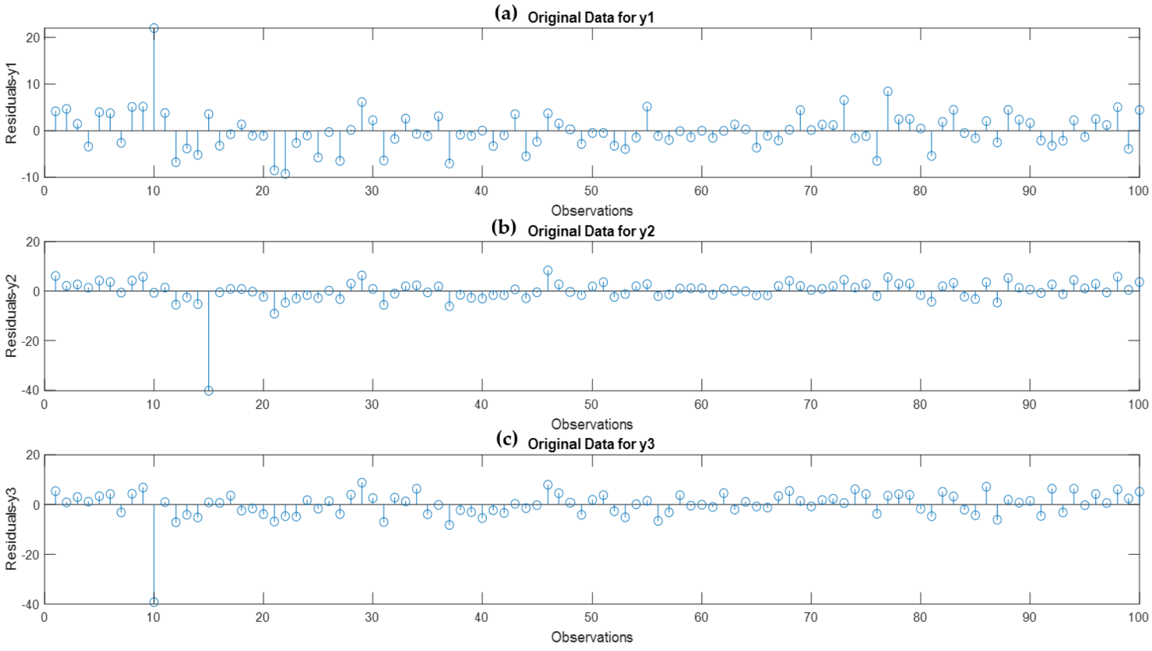

Figure 2: Residuals of MLM for the original data (first experiment).

Figure 2 shows the presence of several outliers in the dependent variables, which led to unacceptably large residual values. The method of treating outliers using the Hampel filter did not specify the window size and threshold parameter, especially to treat outliers in the analysis of an MRM, but rather relied on a default value equal to (3) in the univariate. Therefore, the Hampel filter was employed to treat outliers in an MRM using a window size parameter estimate equal to (sample size divided by 2) and a threshold parameter equal to (0.5) as a classical method (Note that the default values through the practical experiments were not effective in the treating of outliers in an MRM). The proposed method for treating outliers in the analysis of an MRM (explained in the sixth paragraph) is based on the optimal estimating of window size and threshold parameter through the three-step algorithm, which gives the minimum MSE for the model. After treating the outliers in the data of the first experiment and using the classical and proposed methods, the parameters of an MRM were estimated, and the residuals shown in figure 3 were calculated, confirming that there is a large difference in the values of the residuals in favour of the proposed method.

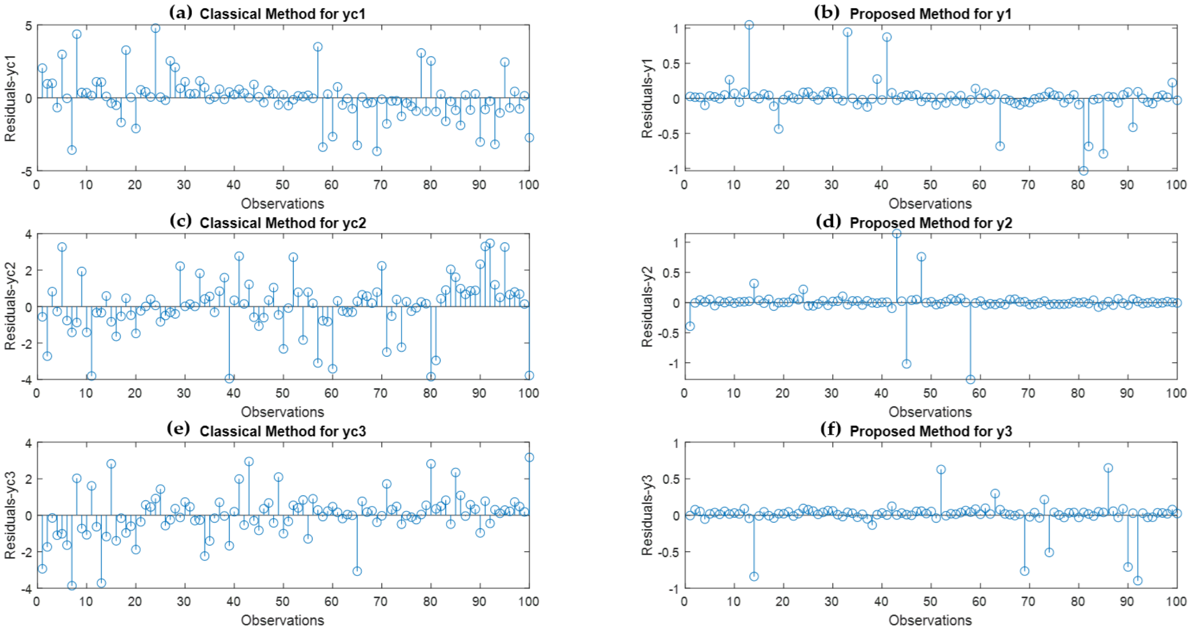

Figure 3: Residuals of MLM for the classical (a, c, f) and proposed (b, d, f) methods (first experiment).

Table 1 shows the values of the optimal window size and optimal threshold parameter with the MSE of the proposed method being lower than the classical method. The experiment was repeated 1000 times for different sample sizes (100, 150, and 200). According to the sample sizes, the classical method relied on the window size (50, 75, and 100) and threshold parameters (0.5, 0.75, and 1) respectively, and the average results with outliers in the dependent variables and outliers in the dependent and independent variables have been summarized in tables 2 and 3.

Table 1: Performance comparison of filtering methods based on MSE.

Method	k	nd	MSE
Classical	50	0.5	6.5818
Proposed	99	0.1500	0.1596
Without filter	----	----	77.3928

The results of tables 2 and 3 demonstrate the effectiveness of the classical and proposed methods in treating outliers in an MRM, and the proposed method with optimal parameters for window size and threshold was found to be more accurate than the classical method. The proposed method, with outliers in the independent and dependent variables, provided optimal parameter values with a much lower MSE than the proposed method in the presence of outliers in the dependent variables only. Different sample sizes (small, medium, large) were used to determine the effect of sample size on the proposed method by generating data with different sizes for each simulation case on this basis. It was found that increasing the sample size leads to a decrease in the effect of outliers on estimating an MRM for untreated data but the efficiency of the classical and proposed methods in treating outliers and estimating an MRM decrease. The average window size is rounded to the nearest integer, and increasing its value with the average threshold decreasing leads to higher efficiency in treating outliers compared to the classical method with outliers in the independent and dependent variables. The proposed method in the presence of outliers in the dependent variables provided larger optimal parameter values for window size and threshold compared to the classical method.

Table 2: Average MSE and optimal values with outliers in the dependent variables.

Method	Sample Size	K	nd	MSE
Classical	100	50	0.5	8.6193
Proposed		88	0.5989	5.0380
Without filter		-----	-----	77.6315
Classical	150	75	0.75	22.8768
Proposed		89	0.6933	5.9195
Without filter		-----	-----	64.7112
Classical	200	100	1	40.7089
Proposed		86	0.9013	7.7503
Without filter		-----	-----	57.4405

Table 3: Average MSE and optimal values with outliers in the dependent and independent variables.

Method	Sample Size	k	nd	MSE
Classical	100	50	0.5	8.5038
Proposed		**96**	**0.1539**	**0.2432**
Without filter		-----	-----	73.0682
Classical	150	75	0.75	24.6314
Proposed		**95**	**0.1509**	**0.3949**
Without filter		-----	-----	61.8120
Classical	200	100	1	46.4686
Proposed		**94**	**0.1507**	**0.4613**
Without filter		-----	-----	55.7157

4.2. Real Data

Real data representing diabetic patients was taken from [2], where Relative Weight (y_1) and Blood Glucose (y_2) represent the dependent variables, and the Insulin Levels $(x_1, x_2, and\ x_3)$ represent the independent variable.

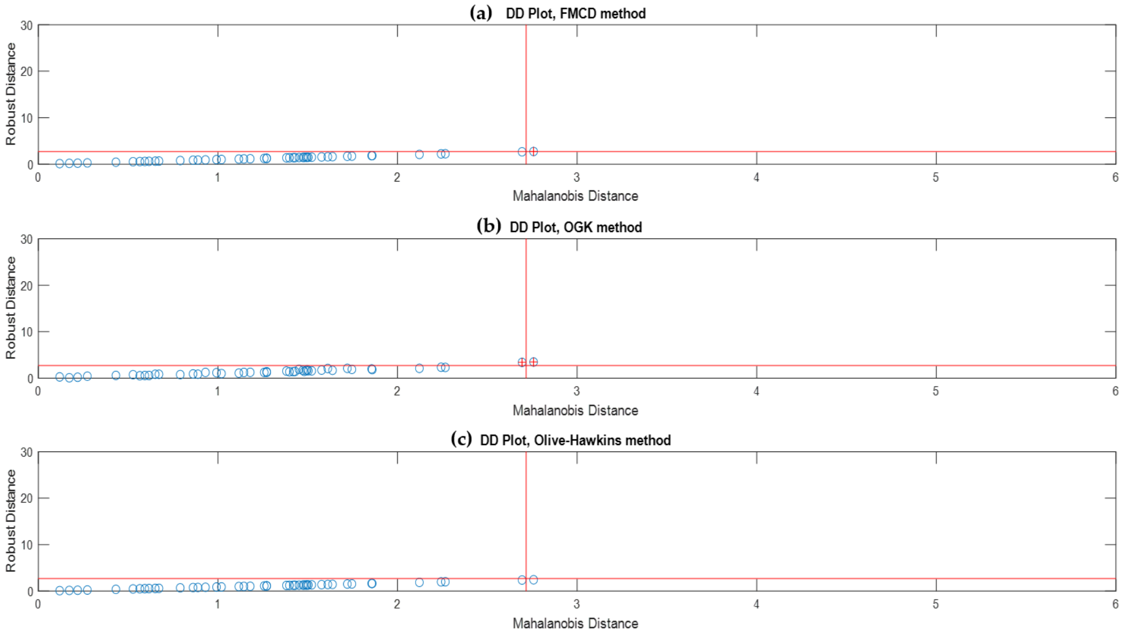

Figure 4: Identifying outliers for the dependent variables (real data).

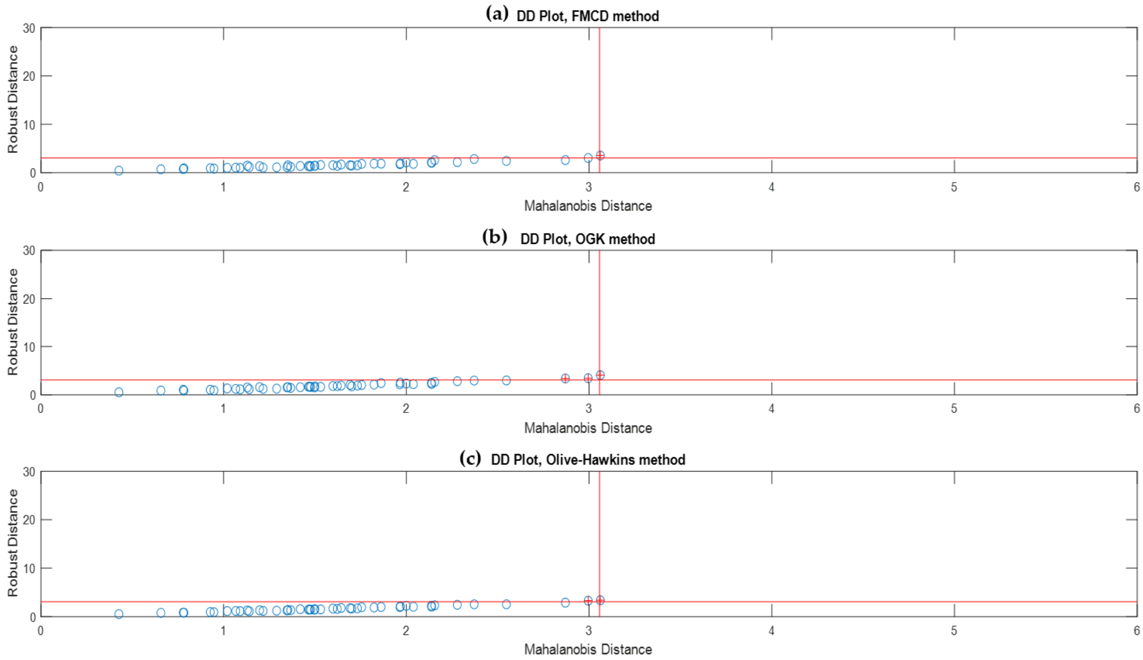

Figure 5: Identifying outliers for the independent variables (real data).

Figure 4 shows the presence of several outliers in the dependent variables (1, 2, 0) for three robust methods (FMCD, OGK, and OH), respectively, which are shown in red for the real data. Figure 5 shows the presence of several outliers in the independent variables (1, 3, 2) for the three robust methods, respectively. An MRM was estimated for the real data and then the residual values were computed, as shown in figure 6.

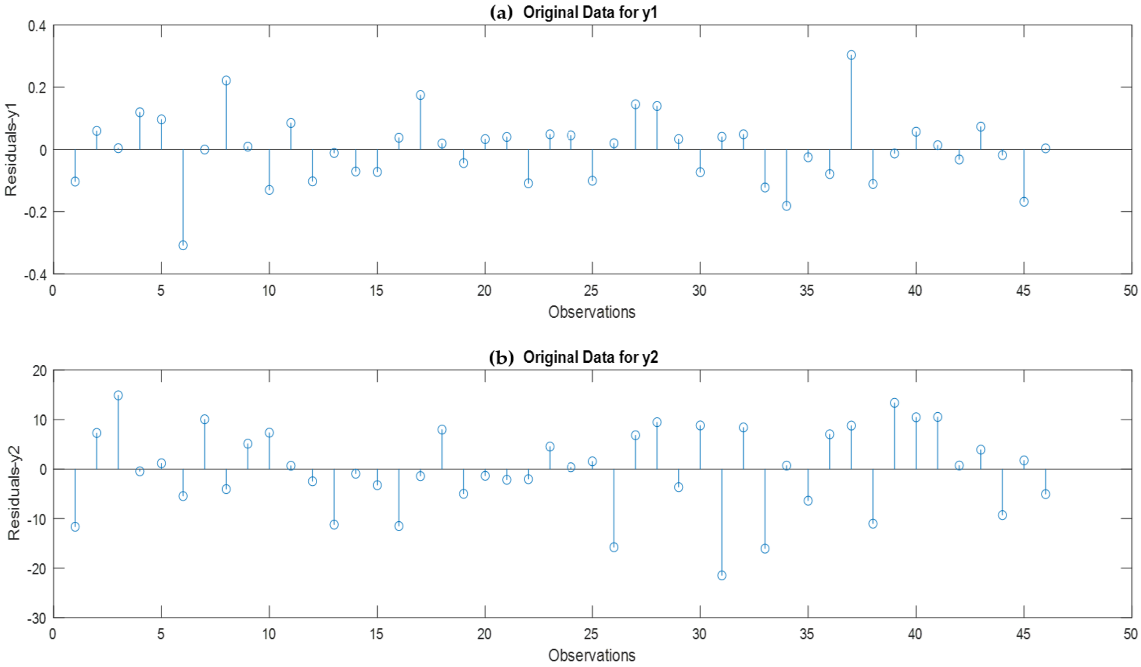

Figure 6: Residuals of MLM for real data.

Figure 6 shows the presence of several outliers in the real data, especially the second dependent variable, which led to unacceptably large residual values. After treating the outliers in the real data using the classical and proposed methods with outliers in the independent and dependent variables, the parameters of an MRM were estimated (Table 4).

Table 4: Regression coefficients of the real data.

Parameter	Classical Method		Proposed Method		Without Filter	
$\hat{\beta}_{0j}$	0.9863	98.1416	1.4851	76.6044	0.6264	83.2425
$\hat{\beta}_{1j}$	-0.0002	-0.0119	-0.0019	0.0323	0.0009	0.0287
$\hat{\beta}_{2j}$	0.0001	-0.0011	0.0003	0.0115	-0.0010	-0.0127
$\hat{\beta}_{3j}$	0.0004	-0.0420	0.0005	0.0011	0.0015	-0.0044

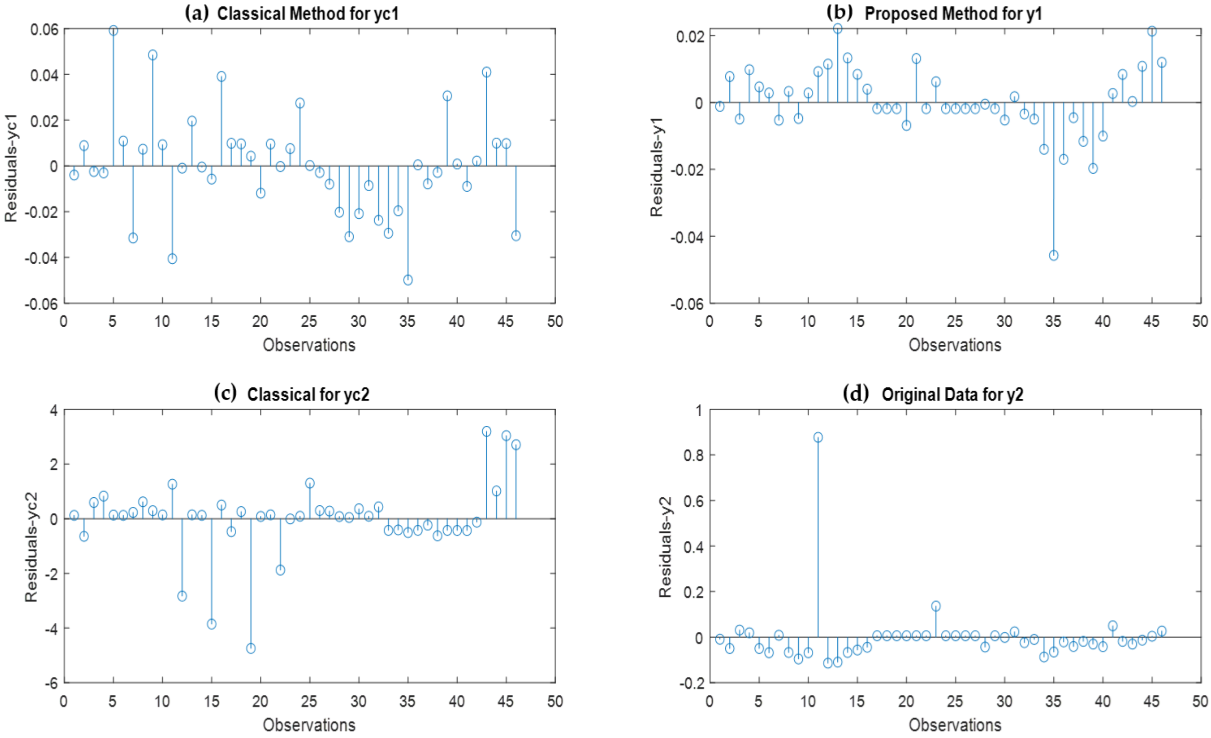

Figure 7: Residuals of MLM for both classical (a, c) and proposed (b) and original (d) methods (real data).

Figure 7 shows the residuals calculated, confirming that there is a large difference in the values of the residuals in favour of the proposed method. The results of table 5 demonstrate the effectiveness of the classical and proposed methods in treating outliers in an MRM, and the proposed method with optimal parameters for window size and threshold was found to be more accurate than the classical method. The proposed method, with outliers in the independent and dependent variables, provided optimal parameter values with a minimum MSE.

Table 5: MSE for the Real Data.

Method	k	nd	MSE
Classical	23	0.5	2.0222
Proposed	29	0.1500	0.0210
Without filter	----	----	74.3802

5. Discussion

The findings of this study demonstrate that outliers in multiple regression models can affect parameter estimates, leading to large residual values. A Hampel filter-modified algorithm is proposed to estimate optimal window size and threshold parameters. The algorithm was effective at dealing with outliers and achieving high accuracy in the estimated parameters, as assessed using simulation and real data. The diabetes data included Relative Weight y_1, representing the patient's relative body mass or a normalized weight value, possibly adjusted for height (similar to BMI). Blood Glucose y_2 was measured in mg/dL and is a key indicator in monitoring and diagnosing diabetes, and Insulin Levels are x_1, x_2, x_3. This data is quantitative. The main goal of the Hampel filter is to increase the robustness of the statistical analysis by lowering the effect of outliers, distorting the outcomes of standard techniques such as linear regression.

The Hampel filter and proposed algorithm are highly efficient and accurate when treating outliers in MRM data. The proposed method had the lowest MSE compared to simulated and real data. The method performs optimally when outliers are present in both independent and dependent variables, yielding accurate parameter estimates. Parameter adjustments for improved results increase the window size and lower the threshold parameter, enhancing the Hampel filter's ability to handle outliers in multivariate regression data. Optimal parameter values provided result in a much greater minimum MSE than the proposed method when in the presence of outliers in the dependent variables only.

These results highlight the importance of robust methods in multivariate linear regression analysis, particularly when dealing with outliers and data contamination, such as the MSE method and Hampel's MAD to address these challenges. Overall, the proposed method demonstrates superior performance, particularly in datasets with complex or multi-variable outliers, providing a robust and flexible solution with minimal MSE. In this research, the proposed method provided a lower MSE than the Hampel filter, and we obtained higher accuracy estimated parameters for the model.

6. Conclusions

In this paper, a new method is proposed for selecting optimal window size and threshold parameters to minimize the MSE in MRM. This approach specifically targets outliers in both dependent and independent variables. Although the Hampel filter was used for outlier treatment, it originally lacked guidance when it came to setting window size and the threshold parameters. The proposed method addresses this gap by optimally estimating these parameters, achieving the minimum MSE for the model. After comparing the classical and proposed methods, the residuals showed a significant difference in favor of the proposed method. The proposed method outperforms the classical method in treating outliers in an MRM, with the optimal window size and threshold parameter being lower than in the classical. Through simulations, the experiment was repeated a thousand times for each simulation case and the overall average was calculated for each case, proving the efficiency of the proposed method and its superiority over the traditional method. The results show the effectiveness of both methods.

The proposed method with optimal parameters for window size and threshold was more accurate than the classical method. However, increasing the sample size decreases the effect of outliers when estimating an MRM for untreated data. The study reveals several outliers in the dependent and independent variables for three robust methods (FMCD, OGK, and OH). The MRM was estimated for real data, and the residual values were computed. The results show there to be a large difference in residual values for the proposed method, especially in the second dependent variable. The study concludes that the proposed method outperforms classical methods. The final table shows that the classical and proposed methods effectively treat outliers in MRM, with the proposed method providing optimal parameters and minimizing MSE. MSE in the classical method is (2.0222) and MSE in the proposed method is (0.0210), while MSE Without filter is (74.3802).

Future work could focus on estimating the optimal window size and threshold parameters for outlier treatment in multivariate regression data. Comparative studies between the proposed Hampel filter and other outlier detection filters would further validate its effectiveness. Additionally, applying this approach to treat outliers in time series models presents an interesting avenue for future research.

Author contributions: Amira Wali Omer: Formal Analysis, Writing – original draftt ,Investigation, Methodology, Writing – review and editing, **Taha Hussein Ali:** Supervision, Formal Analysis, Validation.

Data availability: Data will be available upon reasonable request.

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors did not receive support from any organization for the submitted work.

References

[1] A. Omer, B. Sedeeq and T. Ali, "A proposed hybrid method for multivariate linear regression model and multivariate wavelets (simulation study)," *Polytechnic Journal of Humanities and Social Sciences*, vol. 5, no. 1, pp. 112-124, 2024.DOI: 10.25156/ptjhss.v5n1y2024.pp112-124

[2] A. Rencher and W. Christensen, "Methods of multivariate analysis", New York: Wiley series in probability and statistics, 3rd ed. Wiley, 2012. DOI: 10.1002/9781118391686

[3] D. Montgomery, E. Peck and G. Vining, introduction to linear regression analysis, New York: John Wiley & Sons, 2021. https://www.wiley.com/ -.

[4] S. Phuttisen and S. Wuttichai, "Detection of outliers method in grouped multivariate data: a method based on multiple linear regression," *Pakistan Journal of Statistics and Operation Research*, pp. 445--453, 2024. DOI: 10.18187/pjsor.v20i3.4575

[5] F. Hampel, E. Ronchetti, P. Rousseeuw and W. Stahel, Robust statistics: The approach based on influence functions, New York: Wiley, 2005. http://dx.doi.org/10.1002/9781118186435.

[6] F. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383-393, 1974. https://doi.org/10.1080/01621459.1974.10482962.

[7] V. Calabrese, G. L. Tripepi, D. Santoro, V. Cernaro, V. A. Panuccio, S. Mezzatesta, F. Mattace-Raso and C. Torino, "Impact of serum phosphate on hemoglobin level: a longitudinal analysis on a large cohort of dialysis patients," *Journal of Clinical Medicine*, vol. 13, no. 19, p. 5657, 2024. https://doi.org/10.3390/jcm13195657.

[8] A. Kochersberger, A. Coakley, L. Millheiser, J. R. Morris, C. Manneh, A. Jackson, J. L. Garrison, and E. Hariton, "The association of race, ethnicity, and socioeconomic status on the severity of menopause symptoms: a study of 68,864 women," *Menopause*, pp. 10-1097, 2023. https://doi.org/10.1097/gme.0000000000002349

[9] A. P. Rubbio, A. Sisinni, A. Moroni, M. Adamo, C. Grasso, M. Casenghi and M. B. Tusa. E.A., "Impact of extra-mitral valve cardiac involvement in patients with primary mitral regurgitation undergoing transcatheter edge-to-edge repair," *Journal of EuroPCR and the European Association of Percutaneous Cardiovascular Interventions*, vol. 19, no. 11, pp. 926-936, 2023. https://doi.org/10.4244/eij-d-23-00548.

[10] I. Deftereos, J. M. C. Yeung, J. Arslan, . V. M. Carter, E. Isenring, N. Kiss, "Assessment of nutritional status and nutrition impact symptoms in patients undergoing resection for upper gastrointestinal cancer: results from the multi-centre nourish point prevalence study," *Nutrients*, vol. 13, no. 10, p. 3349, 2021. https://doi.org/10.3390/nu13103349.

[11] A. Volfart, K. L. McMahon and G. I. de Zubicaray, "A comparison of denoising approaches for spoken word production related artefacts in continuous multiband fMRI data," *Neurobiology of Language*, vol. 5, no. 4, pp. 901-921, 2024. https://doi.org/10.1162/nol_a_00151.

[12] N. A. Ramli, Z. Zahid, S. A. S. Hussin and N. A. Ramli, "Comparison of classification models for breast cancer disease using multivariate analysis and data mining approaches," *Applied Mathematics and Computational Intelligence*, vol. 12, no. 4, pp. 1-12, 2023. http://dx.doi.org/10.58915/amci.v12i4.348.

[13] M. S. H. Talukder and R. B. Sulaiman, "Comparative analysis of epileptic seizure prediction: exploring diverse pre-processing techniques and machine learning models," *Electrical Engineering and Systems Science*, pp. 1-14, 2023. https://doi.org/10.48550/arXiv.2308.05176.

[14] G. Karthikeyan and P. Balasubramanie, "A novel attribute-based dynamic clustering with schedule-based rotation method for outlier detection," *International Journal of Business Intelligence and Data Mining*, vol. 16, no. 2, pp. 214-230, 2020. http://dx.doi.org/10.1504/IJBIDM.2020.104741.

[15] M. Verdonck, H. Carvalho, J. Berghmans, P. Forget and J. Poelaert, "Exploratory outlier detection for acceleromyographic neuromuscular monitoring: machine learning approach," *Journal of medical Internet Research*, vol. 23, no. 6, p. e25913, 2021. https://doi.org/10.2196/25913.

[16] J. Hair, W. Black, B. Babin, and R. Anderson, Multivariate data analysis: Global edition, Saddle River, New Jersey: Prentice-Hall, 2010. https://www.drnishikantjha.com/papersCollection/Multivariate%20Data%20Analysis.pdf.

[17] W. Sauerbrei, A. Perperoglou, M. Schmid, M. Abrahamowicz, H. Becher, H. Binder, D. Dunkler, F. E. Harrell Jr, P. Royston and G. Heinze, "State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues," *Diagnostic and prognostic research,* vol. 4, no. 3, pp. 1-18, 2020. https://doi.org/10.1186/s41512-020-00074-3.

[18] U. Knief and W. Forstmeier, "Violating the normality assumption may be the lesser of two evils," *Behavior Research Methods,* vol. 53, no. 6, pp. 2576-2590, 2021 https://doi.org/10.3758/s13428-021-01587-5.

[19] T. H. Ali, N. S. Albarwari and D. L. Ramadhan, "Using the hybrid proposed method for quantile regression and multivariate wavelet in estimating the linear model parameters," *Iraqi Journal of Statistical Sciences*, vol. 20, no. 1, pp. 9-24, 2023. DOI: 10.33899/IQJOSS.2023.178679.

[20] P. Huber and E. Ronchetti, Robust statistics, New York: John Wiley & Sons., 1981. https://onlinelibrary.wiley.com/doi/book/10.1002/0471725250.

[21] T. Ali, B. Sedeeq, D. Saleh and A. Rahim, "Robust multivariate quality control charts for enhanced variability monitoring," Quality and Reliability Engineering International, vol. 40, no. 3, pp. 1369-1381, 2024. https://doi.org/10.1002/qre.3472.

[22] R. Pearson, "Outliers in process modeling and identification," *IEEE Transactions on Control Systems Technology*, vol. 10, no. 1, pp. 55-63, 2002. https://doi.org/10.1109/87.974338.

[23] H. Pehlivan, "A novel outlier detection method based on bayesian change point analysis and hampel identifier for gnss coordinate time series," *EURASIP Journal on Advances in Signal Processing*, vol. 2024, no. 1, p. 44, 2024. http://dx.doi.org/10.1186/s13634-023-01097-w.

[24] J. H. Sullivan, M. Warkentin and L. Wallace, "So many ways for assessing outliers: what really works and does it matter?," *Journal of Business Research,* vol. 132, pp. 530-543, 2021. https://doi.org/10.1016/j.jbusres.2021.03.066.

[25] J. Raymaekers and P. J. Rousseeuw, "Challenges of cellwise outliers," *Econometrics and Statistics,* 2024. https://doi.org/10.1016/j.ecosta.2024.02.002.

[26] T. Li, G. Kou, Y. Peng, and P. S. Yu, "An integrated cluster detection, optimization, and interpretation approach for financial data," *IEEE Transactions on Cybernetics,* vol. 52, no. 12, pp. 13848-13861, 2021. https://doi.org/10.1109/TCYB.2021.3109066.

[27] S. Zhang, R. Yao, C. Du, E. Essah, and B. Li, "Analysis of outlier detection rules based on the ashrae global thermal comfort database," *Building and Environment,* vol. 234, p. 110155, 2023. http://dx.doi.org/10.1016/j.buildenv.2023.110155.

[28] M. Mayrhofer and P. Filzmoser, "Multivariate outlier explanations using shapley values and mahalanobis distances," *Econometrics and Statistics,* 2023. https://doi.org/10.1016/j.ecosta.2023.04.003.

[29] A. Smiti, "A critical overview of outlier detection methods," *Computer Science Review*, vol. 38, p. 100306, 2020. https://doi.org/10.1016/j.cosrev.2020.100306.

[30] S. Y. Woo and S. Kim, "Determination of Cutoff Values for Biomarkers in Clinical Studies," *Precision and Future Medicine,* vol. 4, no. 1, pp. 2-8, 2020. https://doi.org/10.23838/pfm.2019.00135.

[31] E. Cabana, R. E. Lillo and H. Laniado, "multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators," *Statistical papers,* vol. 62, pp. 1583-1609, 2021. https://doi.org/10.1007/s00362-019-01148-1.

[32] M. Mashuri, M. Ahsan, M. H. Lee, D. D. Prastyo and Wibawati, "PCA-based hotelling's T^2 chart with fast minimum covariance determinant (FMCD) estimator and kernel density estimation (KDE) for network intrusion detection," *Computers & Industrial Engineering,* vol. 158, p. 107447, 2021. https://doi.org/10.1016/j.cie.2021.107447.

[33] B. Sedeeq, Z. Muhammad, I. Ali and T. Ali, "Construction robust-chart and compare it with Hotelling's T2-chart," Zanco *Journal of Human Sciences*, vol. 28, no. 1, pp. 140-157, 2024. https://doi.org/10.21271/zjhs.28.1.11.

[34] S. Thudumu, P. Branch, J. Jin and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data,* vol. 7, pp. 1-30, 2020. https://doi.org/10.1186/s40537-020-00320-x.

[35] L. Davies and U. Gather, "The Identification of multiple outliers," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782-792, 1993. https://doi.org/10.1080/01621459.1993.10476339.

[36] V. Barnett and T. Lewis, Outliers in statistical data, New York: Wiley, 1994. ISBN: 978-0-471-93094-5.

[37] M. M. Garcez Duarte and M. Sakr, "An experimental study of existing tools for outlier detection and cleaning in trajectories," *GeoInformatica,* pp. 1-21, 2024. http://dx.doi.org/10.1007/s10707-024-00522-y.

[38] S. Jalal, D. Saleh, B. Sedeeq and T. Ali, "Construction of the Daubechies wavelet chart for quality control of the single value," *Iraqi Journal of Statistical Sciences*, vol. 21, no. 1, pp. 160-169, 2024. http://dx.doi.org/10.33899/iqjoss.2024.183257.

[39] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Outlier detection performance of a modified z-score method in time-series rss observation with hybrid scale estimators," *IEEE Access,* vol. 12, pp. 12785 - 12796, 2024. https://doi.org/10.1109/ACCESS.2024.3356731.

[40] T. H. Ali, "Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study)," *Communications in Statistics-Simulation and Computation*, vol. 51, no. 2, pp. 391-403, 2022. https://doi.org/10.1080/03610918.2019.1652319.