Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Comparison of Wavelet Shrinkage and Hampel Filter in the Analysis of Multivariate Linear Regression Models

Amira Wali Omer[1] Taha Hussein Ali[2]

[1,2]Statistics and Informatics Department- College of Administration & Economics - Salahaddin University – Erbil, Iraq.

Article information

Article history:
Received: September 10,2024
Revised: December 28,2024
Accepted: April 10,2025
Available: December 1,2025

Keywords:
Multivariate Linear Regression Models, Outliers, Hampel Filter, Wavelets Shrinkage, and Threshold.

Correspondence:
Amira Wali Omer
amira.omer@su.edu.krd

Abstract

The presence of outliers in the data of a multivariate regression model affects the accuracy of the estimated model parameters and leads to unacceptably large residual values. Therefore, some filters, including the Hempel filter, are usually used to handle outliers (or use some robust method). This paper proposes to employ wavelet shrinkage to address the problem of outliers in multivariate regression model data by using wavelets (Coiflets, Daubechies, and Demy) with a universal threshold method and soft rule. To illustrate the efficiency of the proposed method (Wavelet Shrinkage filter) was compared with the traditional method (Hampel filter) based on the mean square error criterion through simulation and real data. A program has been designed in MATLAB to do this. The results proved that the Wavelet shrinkage filter method was more efficient than the traditional method in dealing with the outlier problem and obtaining more accurate multivariate model parameters than the Hampel filter method.

DOI 10.33899/iqjoss.v22i2.54068,

1. Introduction

In the realm of statistical analysis, multivariate linear regression is a fundamental tool used to model the relationship between a dependent variable and multiple independent variables. However, the presence of noise and outliers can significantly distort the results, leading to inaccurate predictions and interpretations. This paper explores two advanced techniques, wavelet shrinkage and the Hampel filter to mitigate these issues. The primary objective is to compare their performance in enhancing the robustness of multivariate linear regression models. Multivariate linear regression has been extensively studied, with numerous applications across various fields such as economics, biology, and engineering (Montgomery et al., 2012). Despite its widespread use, the method is sensitive to noise and outliers, which can skew results (Kutner et al., 2004).

An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves unusually, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities that impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights (Aggarwal, 2017).

Wavelet shrinkage is a popular technique for handling outliers and noise in the analysis of multivariate linear regression models. The method involves the use of wavelet transforms to decompose the data into different frequency bands, and then

selectively shrinking or thresholding the wavelet coefficients to remove unwanted components. The key advantage of wavelet shrinkage is its ability to preserve important localized features in the data, while effectively removing noise and outliers. (Li et al., 2015)

A threshold might refer to a cut-off value above or below which data is considered an outlier or irrelevant. For example, in outlier detection, any data point that lies beyond a certain threshold value may be considered an outlier. A threshold might be applied to residuals to determine outliers. Points with residuals beyond a certain threshold might be considered problematic or outliers. (Barnett & Lewis, 1994). The concept you're referring to, where a threshold is used as a cut-off value in outlier detection, is a standard approach in statistical analysis and data preprocessing. This approach is often used in techniques like z-score, interquartile range (IQR), and robust statistics such as the Hampel filter. (Aggarwal, 2013).

The Hampel filter is another approach for dealing with outliers in multivariate linear regression models. The Hampel filter operates by replacing each data point with a robust estimate of the central tendency in a local neighbourhood, effectively down-weighting or removing outliers. Unlike wavelet shrinkage, the Hampel filter does not require any data decomposition and can be applied directly to the original data. (Hampel et al., 1986). Now, the Hampel filter is more like a no-nonsense bouncer at a fancy club - it identifies and deals with outliers in your data. It's named after John R. Hampel, who had a keen eye for spotting troublemakers in your dataset. The Hampel filter helps clean up your data by replacing outlier values with more sensible ones, ensuring your regression model doesn't get thrown off by rowdy data points. (Hampel, 1974).

Wavelet shrinkage and Hampel filter are two fancy-sounding techniques that can help make sense of complex data in multivariate linear regression models. They're like the cool cousins at the family reunion of statistical analysis methods. Both wavelet shrinkage and the Hampel filter have been successfully applied to the analysis of multivariate linear regression models (Tibshirani, 1996; Unser, 2002; Li et al., 2022; Li et al., 2015). While wavelet shrinkage can effectively remove noise and preserve important localized features, the Hampel filter offers a more straightforward and computationally efficient approach to outlier removal. (Li et al., 2022) Empirical studies have suggested that the performance of these two methods may depend on the specific characteristics of the data and the underlying regression model. (Najafi & Hakim, 1992).

The primary aim of this research is to address the problem of outliers in multivariate regression model data, which can severely impact the accuracy of estimated model parameters and lead to unacceptably large residuals. While traditional methods like the Hampel filter are often employed to mitigate the influence of outliers, this paper explores a novel approach: using wavelet shrinkage as a robust alternative for handling outliers. Specifically, the study investigates the performance of various wavelet families (Coiflets, Daubechies, and Demy) combined with a universal threshold method and soft rule to effectively reduce the impact of outliers. The research aims to answer the following key questions:

- Can wavelet shrinkage improve the accuracy of estimated parameters in multivariate regression models with outliers compared to traditional methods such as the Hampel filter?
- How do different wavelet families (Coiflets, Daubechies, and Demy) perform in terms of their ability to handle outliers and reduce the mean square error (MSE)?
- Is the proposed wavelet shrinkage filter method more efficient than the traditional Hampel filter in both simulated and real-world datasets? The hypotheses of this research are:

H1: The Wavelet Shrinkage filter provides more accurate parameter estimation in multivariate regression models with outliers compared to the traditional Hampel filter.

H2: The Wavelet Shrinkage filter results in a lower Mean Squared Error (MSE) in the presence of outliers than the Hampel filter.

H3: The performance of the Wavelet Shrinkage filter varies depending on the choice of wavelet family (Coiflets, Daubechies, and Demy), but all outperform the Hampel filter.

Through comprehensive simulation and real data experiments, the paper demonstrated that the wavelet shrinkage filter offers superior performance over the Hampel filter, leading to more accurate multivariate model parameter estimation and better handling of outliers.

2. Methodology

2.1. Multivariate Linear Regression Models

Multivariate regression is a statistical technique used to model the relationship between a dependent variable and multiple independent variables (Omer et al., 2024). Unlike simple linear regression, which only considers a single predictor, multivariate regression allows for a more comprehensive analysis encompassing multiple factors. This methodology is crucial in numerous fields, including economics, medicine, and social sciences, where the outcome is often influenced by several variables simultaneously (Tabachnick & Fidell, 2019). Despite its widespread applicability, multivariate regression

is fraught with challenges. Multicollinearity, where independent variables are highly correlated, can distort the results. Overfitting, where the model becomes too complex and tailored to the training data, reduces its predictive power on new data. Additionally, the presence of outliers can significantly influence the regression coefficients, leading to biased results (Montgomery et al., 2012). Before diving into the specifics of wavelet shrinkage and the Hampel filter, let's quickly touch base on multivariate linear regression models. These models are like the bread and butter of regression analysis, allowing us to understand the relationship between multiple input variables and an outcome of interest (Cook, 1977). They're the workhorses of predictive analytics, helping us make sense of complex data relationships (Greene, 2018). The multivariate regression model can be expressed mathematically as follows:

$$Y = XB + E \tag{1}$$

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

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 measured on a subject. 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 (Hair, et al. 2010).

2.2. Least Squares Estimation in the Multivariate Model

Least Squares Estimation (LSE) is a technique used to determine the coefficient matrix β that minimises the sum of squared residuals, equivalent to $E = Y - X\beta$. The sum of squared residuals is $S = tr(E^T E) = [(Y - X\beta)^T (Y - X\beta)]$ where tr denotes the trace of a matrix (Huber, 1981; Omer et al., 2020). To minimise S, the derivative of S with respect to β is set to zero. The least squares estimator for β is determined by solving for:

$$X^T Y = X^T X\beta \text{ and } \beta = (X^T X)^{-1} X^T Y \tag{3}$$

The properties of the least squares estimator include unbiasedness, minimum variance, and normality. The Gauss-Markov theorem states that the least squares estimator has the minimum variance among all linearly unbiased estimators (Rencher & Christensen, 2012).

2.3. Outlier Problem

An outlier problem in statistical analysis, particularly in multivariate regression models, is the existence of data points that exhibit substantial deviation from the statistical pattern of the bulk of the data. Outliers can have a substantial impact on the results of an analysis, leading to biased estimates, distorted model fits, and reduced predictive accuracy (Aggarwal, 2017). The outlier problem's impact on regression analysis is one of its key points. Regression coefficients can be disproportionately impacted by outliers, which might provide models that are inaccurate representations of the underlying data patterns. This is especially troublesome for multivariate regression, which evaluates several variables' associations at once (Hampel et al. 1986). Identification and Care, there are several ways to find outliers, statistical tests, robust estimating approaches, and graphical tools like boxplots and scatterplots. Outliers can be dealt with after they are identified by techniques like transformation, elimination, or the application of strong statistical techniques that lessen their influence (Breunig, et al., 2000). Problems: It might be difficult to determine whether a data point is an outlier and how to manage it. Outliers can reflect legitimate but exceptional events that are also indicative of flaws in data collection (Rousseeuw & Leroy, 2003).

2.4. Handling Outliers

Robust regression techniques, such as M-estimators, least trimmed squares (LTS), and S-estimators, are less sensitive to outliers than ordinary least squares (OLS) (Huber, 1981). Variable transformations, such as logarithmic, square root, and Box-Cox transformations, can reduce the impact of outliers. Data cleaning is appropriate for outliers due to data entry or measurement errors (Rousseeuw & Leroy, 2003). Winsor zing transforms outliers to the nearest value within a certain percentile, reducing their impact without removing them. Imputation techniques replace outliers with estimated values based

on remaining data (Montgomery et al. 2012). Robust standard errors provide more reliable standard errors for hypothesis testing and confidence intervals. Bayesian methods incorporate prior information about data distribution and outliers, providing a probabilistic framework for handling outliers (Kutner et al., 2004).

2.5. Hampel Filter

A statistical method for locating and lessening the effect of outliers in data is the Hampel filter. To find any outliers, it compares each data point with its neighbours to see whether there is a substantial deviation (Hampel, 1974). This filter is helpful for real-time signal processing and noise reduction because it works especially well on datasets where outliers have the potential to skew the study. The main advantage of the Hampel filter is that it uses standard deviation and median rather than the more conventional mean-based techniques, making it resistant to outliers. Regression models constructed with reprocessed data are more reliable since this guarantees that the fundamental structure of the data is preserved. However, the filter needs an adjustable window size, can be computationally intensive, particularly for large datasets, and may not be as successful in datasets with non-symmetrical noise distribution. (Hampel et al., 1986). There are two parameters to configure the Hampel filter:

2.5.1. Window Size (k)

The size of the moving window that is utilised to assess each data point is determined by this parameter. It establishes the parameters that we use to search for outliers. The following sources and recommendations might help you choose the right window size (Hampel et al., 1986). The window size k should be chosen based on the nature and the frequency of expected outliers. A common heuristic is to set k around 3 to 5 times the expected width of an outlier. For periodic data, k should be chosen based on the period length to accurately capture cyclic behaviour. The window size should account for the data's sampling rate and periodic patterns (Pearson, 2002). If outliers are expected to last a specific duration, the window size should be large enough to capture them effectively. Robustness vs. sensitivity can be achieved by using larger window sizes for smoothing and reducing sensitivity to outliers, while smaller window sizes increase sensitivity to short-term outliers but may lead to false positives due to noise. Adaptive window size can be beneficial in some cases, requiring more sophisticated algorithms (Rousseeuw & Leroy, 2003).

2.5.2. Threshold

Selecting thresholds carefully is necessary to prevent useful data from being detected as outliers. The threshold establishes the amount of deviation a data point must have from the median to qualify as an outlier (Rousseeuw & Leroy, 2003). Below is a thorough breakdown of how to create thresholds in the Hampel filter, along with pertinent citations (Pearson, 2002).

2.5.3. Rules for Determining Thresholds

The median absolute deviation (MAD) is a statistical measure that measures the median of absolute deviations from the data's median

$$MAD = median(|x_i - median(x)|) \tag{4}$$

It is used in the Hampel filter to set a threshold for outlier detection (Ali et al., 2021). The threshold is typically set as a multiple of the MAD, with a factor of 3 to 3.5 times the MAD to align with the properties of the normal distribution

$$Threshold = t * MAD \tag{5}$$

where t is the chosen threshold factor. A scaling factor, often 1.48261, is applied to the MAD to make it comparable to the standard deviation for normally distributed data.

$$Threshold = t * 1.4826 * MAD \tag{6}$$

A threshold factor (t) of 3 is commonly used, with higher sensitivity (lowering) and reduced sensitivity (raising). The threshold factor can be adjusted based on the data's nature and application (Huber, 1981).

2.5.3.1. Experimental Data:

- In controlled experimental settings, the data are often structured and may follow assumptions close to the normal distribution. Hence, the threshold factor is typically set to 3 (i.e., t=3) to detect significant outliers while preserving most of the core data.
- The scaling factor of 1.4826 is used to make MAD comparable to the standard deviation under the assumption of normality, yielding the threshold as:

Threshold=3*1.4826*MAD

This method is chosen because the assumption of normality is valid, and the threshold aims to balance sensitivity with the risk of marking valid data as outliers.

2.5.3.2. Real-World Data:

- Real-world data is often noisier and less likely to adhere to strict normality. The presence of non-Gaussian distributions or heavy tails may require adjusting the threshold to increase or decrease sensitivity.
- A threshold factor t of 3 is still common, but it may be adjusted depending on the nature of the data:

Higher sensitivity (lowering the threshold) might involve setting t=2.5, capturing more potential outliers. Lower sensitivity (raising the threshold) could mean increasing t to 3.5 or more to focus only on extreme outliers.

The method of threshold determination can depend on exploratory analysis, expert knowledge, or iterative fine-tuning based on performance in outlier detection. In some cases, cross-validation or empirical testing may help define the optimal t for specific real-world applications.

2.5.4. 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 point-to-point median and standard-deviation estimates using (Hampel et al., 1986):

$$Local\ median = m_i = median(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \quad (7)$$

$$Standard\ deviation = \sigma i = \ k, median(|x_{i-k} - m_i|, \dots, |x_{i+k} - m_i|) \quad (8)$$

Where $k = \frac{1}{\sqrt{2}erf^{-1}\left(\frac{1}{2}\right)} \approx 1.4826$ the quantity $\frac{\sigma_i}{k}$ is known as the median absolute deviation (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.

$$m_i = median(x_1, x_2, x_3, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \ \ if\ i < k + 1 \quad (9)$$

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

$$m_i = median(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{n-2}, x_{n-1}, x_n)\ if\ i > n - k \quad (10)$$

And $\sigma_i = k.median(|x_{i-k} - m_i|, \dots, |x_n - m_n|)$

For expressions of the form 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. Instead, replace erfinv(1-x) with erfcinv(x) (Pearson, 2002).

2.6. Wavelet

Wavelets refer to short-lived oscillations characterised by amplitudes that begin at zero, increase or decrease, and then return to zero. Scientists have developed a taxonomy of wavelets based on their quantity and polarity.

2.6.1. Coiflets Wavelets

Coiflets wavelets are a family of wavelets designed by Ingrid Daubechies and Ronald Coifman, known for their balance between smoothness and compact support. They are orthogonal wavelets, with vanishing moments in both the scaling function and wavelet function, making them suitable for approximating smooth signals. For Coiflets of order N, both wavelet and scaling functions have N vanishing moments. They are nearly symmetric, minimising phase distortion in signal processing (Nielson, 2001). Coiflets have compact support, making them efficient for computational purposes. They are widely used in signal processing, image processing, feature extraction in machine learning, and pattern recognition. They are used for denoising and compressing signals while preserving important features, image compression algorithms like JPEG 2000, and feature extraction in machine learning and pattern recognition. Overall, Coiflets wavelets are highly useful in signal processing and other applications due to their excellent time-frequency localisation and smoothness (Strang & Nguyen, 1996). Coiflets wavelets are defined by a scaling function ϕ(t) and a wavelet function ψ(t), which are related through the refinement equation:

$$\phi(t) = \sum_k h_k \phi(2t-k) \tag{11}$$

$$\psi(t) = \sum_k g_k \phi(2t-k) \tag{12}$$

Where h_k and g_k are the filter coefficients related to the scaling and wavelet functions, respectively (Mallat, 1989).

2.6.2. Daubechies Wavelet (Db):

Normal orthogonal wavelets, named after Ingrid Daubechies, originated in 1988 and enabled discrete wavelet analysis. They are named after her. The wavelet function's vanishing or ephemeral moments are represented by 4D and (Db), while (N) is the candidate's length and (L_1) is the number of ephemeral moments. The second-ranked person in this family is L_1, corresponding to N, $L_1 = N/2$ is a family of small waves of order n, with an anchor on the period [0, 2n-1]. Each wave has n ephemeral moments, and analytes increase with rank.

$$\left|\frac{d^j}{dX^j}\psi(X)\right| < \infty \quad \Rightarrow \quad \int X^j \psi(X) dX = 0\,, \;\; 1 \le j \le n \tag{13}$$

The family has (rn) continuous derivatives, with a rank of about 0.2. The small wave is a member of this family.

2.6.3. Meyer Wavelet (Demy)

The Meyer wavelet is widely used in signal processing and data analysis, particularly in image processing and compression. It is known for its smoothness and good localisation properties in both time and frequency domains. The wavelet is constructed using a smooth function defined in the frequency domain, offering high smoothness for continuity and differentiability. It is often used in conjunction with wavelets that do not have compact support. The Meyer wavelet also allows multi-resolution analysis, a key feature of wavelet transforms. It is commonly used in areas like image denoising and feature extraction (Guo et al., 2022). The Meyer wavelet's ability to balance time and frequency localisation makes it a valuable tool in various analytical contexts. Many bivariate wavelet filters are used for the stationary two datasets, such as the Meyer wavelet (demy) (Mustafa & Ali, 2013). The Meyer wavelet is defined using a function ν(ω) in the frequency domain, where ω denotes the angular frequency. Meyer Wavelet Function ψ(ω) in the Frequency Domain. The Meyer wavelet is defined in the frequency domain by its Fourier transform $\hat{\psi}(\omega)$:

$$\hat{\psi}(\omega) = (2\pi)^{\frac{-1}{2}} e^{\frac{i\omega}{2}} sin\left(\frac{\pi}{2} v\left(\frac{3}{2\pi}|\omega| - 1\right)\right) \; if \; \frac{2\pi}{3} \le |\omega| \frac{4\pi}{3} \tag{14}$$

$$\hat{\psi}(\omega) = (2\pi)^{\frac{-1}{2}} e^{\frac{i\omega}{2}} cos\left(\frac{\pi}{2} v\left(\frac{3}{4\pi}|\omega| - 1\right)\right) \; if \; \frac{4\pi}{3} \le \frac{8\pi}{3} \tag{15}$$

$$\hat{\psi}(\omega) = 0 \; if \; \notin \left[\frac{2\pi}{3}, \frac{8\pi}{3}\right] \tag{16}$$

where ν(ω) is a smooth function defined as: $\nu(\omega) = 0 \;\; for \;\; \omega \le 0$

$$\nu(\omega) = exp\left(\frac{-1}{\omega^2} exp\left(\frac{-1}{\omega^2 - 1}\right)\right) \; for \; 0 < \omega < 1 \; and \; \nu(\omega) = 1 \;\; for \;\; \omega \ge 1 \tag{17}$$

This function ν(ω) ensures that the wavelet is smooth and has the desired properties in the Fourier domain. Meyer Wavelet in the Time Domain: The inverse Fourier transform of $\hat{\psi}(\omega)$ gives the Meyer wavelet ψ(t) in the time domain. However, the time-domain expression is generally not simple and is usually not provided in closed form due to the complexity of the Fourier transform of the function defined above (Arts & van den Broek, 2022).

2.7. Shrinkage

Shrinkage, a sometimes-misinterpreted phrase, is the process of lowering sample sizes to minimise the impact of sampling error. It is frequently employed in statistical methods, such as the Risk Metrics volatility estimator, a frequently used instrument in this domain.

2.7.1. Universal Method

presented the formula (18) for the universal threshold approach:

$$\delta^U = \hat{\sigma}_{MAD}\sqrt{2\ log(n)} \tag{18}$$

The wavelet coefficients of interest have a median absolute deviation of 0.6745, which indicates the standard error of the estimate for them (Donoho & Johnstone, 1994).

2.7.2. Soft Rule

A soft threshold is a more lenient criterion for identifying outliers. It allows for a broader range of data points to be considered as "potential outliers." Soft thresholds are often used when the aim is to detect outliers that might still be part of the natural variability of the data, particularly in exploratory data analysis (Tukey, 1977). The Median Absolute Deviation (MAD) is a robust measure of statistical dispersion. A soft threshold can be established by considering data points that deviate from the median by a multiple of MAD. It is calculated as follows:

$$MAD = median(|\ Xi - median(X)\ |) \tag{19}$$

$$Threshold\ Boundaries = median(X) \pm k * MAD \tag{20}$$

Where k is a soft multiplier, often chosen as 1.5 or 2 for softer thresholds. Points outside this range are considered outliers. The choice of k controls the softness of the threshold; smaller values of k allow more points to be classified as outliers (Hampel, 1974).

2.8. Mean Squared Error (MSE)

Mean Squared Error (MSE) is a metric used to assess the performance of filters, such as the Hampel filter and Wavelet shrinkage. It quantifies the filter's ability to reduce noise and outliers in data by comparing the filtered data to the true or expected values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{21}$$

where: y_i are the true values, $\hat{y}_i$ are the filtered values, and n is the number of data points (Hair et al., 2010). Mean Squared Error (MSE) is a commonly used statistic to compare the efficacy of wavelet shrinkage and the Hampel filter approaches in decreasing noise or outliers while maintaining the underlying structure of the data (Hampel et al., 1986).

2.9. Proposed Filter

Wavelet shrinkage, which addresses data noise, was employed to address an outlier problem in the multivariate regression model data by:

Step 1. Select the appropriate wavelet for the dependent variables data, such as (Coiflets, Daubechies, and Demy).

Step 2. The data of the dependent variables and the selected wavelets the maximal overlap discrete wavelet transformation coefficients are calculated.
Step 3. Choose the appropriate level to minimise the MSE of the estimated models.
Step 4. Choose the appropriate thresholding method to minimise the MSE of the estimated models such as the universal threshold, and estimate the threshold parameter.
Step 5. Apply the soft threshold rule using the estimated universal threshold on the maximal overlap discrete wavelet transformation coefficients by kill or keep.
Step 6. Calculate the inverse of the modified maximal overlap discrete wavelet transformation coefficients to obtain denoised data and handle outliers.
Step 7. Estimation of parameters of a multivariate linear regression model based on filtered data.

3. Result

3.1. Simulation Study

For tackling outliers in the multivariate linear regression model data, in the context of comparing Wavelet Shrinkage and the Hampel Filter, the objective could be to assess which method is more effective in handling outliers in multivariate linear regression models. the random error of the model was generated with a multivariate normal distribution function, a zero-mean vector, and a variance-covariance matrix E shown in Table 1. Different numbers of predictor (p=2 and 3) and response variables (q=1, 2, 3, and 4) were used with different sample sizes (100, 150, and 200), the Var-Covariance matrix are equals to [1 2 3;2 1 2;3 2 1] and the regression coefficients (β) is equal to [2 4 6;4 3 5;3 6 4;3 5 2;6 4 2] if a setup might specify (p=3 and q=4), and if (p=2, q=1) the Var-Covariance and the regression coefficients β matrix's is given by [1 2; 2 1]; [2 4; 4 3] respectively, if (p=2, q=2) that the regression coefficients (β)=[2 4; 4 3; 3 6], Var-Covariance = [1 2; 2 1], also if (p=2, q=3,4) the β=[2 4; 4 3; 3 6; 3 5]; [2 4 6;4 3 5] respectively, The generated data and applied to a multivariate linear model to get the dependent variables. An estimation of the regression coefficients for the multivariate linear models was conducted on the unfiltered data, followed by using the Hampel filter, and ultimately the wavelet filter (Coif5, Db20, Dmey). it is also clear that the average of mean square estimation for simulation data for all methods as shown in Table 1, and compared results turns out that Demy is better than them because it has the lowest variance in all possibilities after repeating the process (1000) times. Utilizing the real data, the multivariate regression model was estimated employing five distinct methodologies: Unfiltered, Hampel filter, Coif5, Db20, and Dmey wavelets filter. Each model was evaluated using the Mean Squared Error (MSE), and Coif5 was determined to be the most appropriate model because of its lowest contrast. A summary of the results is provided in Table 2.

Table (1) MSE Average for Simulation

p	q	Sample Siz	Without Filt	Hampel Filte	Proposed Filter		
					Coif5	Db20	Dmey
3	4	100	77.6315	8.6193	3.9546	3.8576	**1.7986**
		150	64.7112	9.4944	4.1733	3.9243	**2.5106**
		200	57.4405	9.8111	4.2685	3.7678	**3.0363**
2	1	100	32.7052	10.5883	2.6059	1.6823	**1.4390**
		150	25.1766	10.5329	1.9709	1.3124	**1.2051**
		200	21.4125	10.5597	1.5711	1.0845	**1.0347**
2	2	100	32.8532	12.6988	2.0725	1.5688	**0.9840**
		150	25.8104	12.6335	1.8926	1.4410	**1.1152**
		200	21.4333	12.5082	1.7384	1.3446	**1.1870**
2	3	100	33.7146	15.0454	2.0677	1.8446	**0.9291**
		150	25.7816	14.4695	2.1734	1.9365	**1.2993**
		200	21.6743	14.3117	2.1327	1.8241	**1.4848**
2	4	100	34.1757	18.8092	2.2949	2.5301	**1.0387**
		150	26.2801	17.6298	2.6838	2.6767	**1.6847**
		200	21.9232	17.1362	2.6349	2.4709	**1.9053**
3	1	100	76.3679	37.3840	4.1852	2.5760	**2.1220**

		150	63.1671	37.3939	3.3297	2.1667	**1.9587**
		200	56.5530	37.3239	2.8330	1.8678	**1.8463**
		100	77.0316	40.3226	3.5111	2.6583	**1.5928**
3	2	150	63.6847	40.0397	3.3908	2.6994	**1.9875**
		200	56.8063	39.8525	3.1639	2.4973	**2.1666**
		100	77.2126	42.3117	3.6533	3.1400	**1.6496**
3	3	150	64.0612	42.2358	3.7357	3.2229	**2.2489**
		200	56.7162	41.6745	3.5031	2.9545	**2.4282**
		100	77.6315	47.3913	3.9546	3.8576	**1.7986**
3	4	150	64.7112	46.0795	4.1733	3.9243	**2.5106**
		200	57.4405	9.8111	4.2685	3.7678	**3.0363**

Table 1 presents the effectiveness of different regression filtering methods across various settings, comparing the use of no filter, Hampel Filter, and the three proposed filters Coif5, Db20, and Dmey. The data is organized by different combinations of the number of predictors (p), the number of responses (q), and sample sizes (100, 150, and 200). For each configuration, the table reports the Mean Squared Error (MSE) of the regression models. Generally, the "Without Filter" method shows the highest MSE values across all configurations, indicating the least accuracy; Coif5 significantly enhances model performance.

Figure 1 presents a comparison of filter performance by plotting the mean squared error (MSE) against the sample number. It seems to have four different lines representing the performance of various filters. Here are the key details concerning a sample size of 100. The x-axis represents the sample number, which ranges from 0 to 1000. If it were a sample size of 100, we should focus only on the first 100 samples (from 0 to 100), The y-axis represents the Mean Squared Error, which is a common metric used to evaluate the difference between predicted and true values in filtering or estimation processes. In this plot, it ranges from 0 to 120. Each sample could be a data point where the filter's performance is being evaluated. The y-axis represents the MSE values; higher values indicate worse performance, while lower values indicate better accuracy. MSEH (The MSE of a Hampel filter (indicated by black circular markers) values are significantly higher and more variable than the others, fluctuating widely between 40 and 100, suggesting this filter has higher error rates. MSEw1 (Red line) This represents another filter or method's MSE values. It fluctuates a lot but remains relatively low compared to MSEH. MSEw2 (Blue line) Another filter's performance, showing even lower values than MSEw1. Finally, the MSEw3 (Green line) The green line represents a filter with the lowest MSE values across the samples, indicating the best performance among the shown methods. From the previous explanation for a sample size of 100, you would focus on the first segment of the plot, which would likely show similar trends but over a smaller subset. The black circles (MSEH) would still fluctuate more than the other MSE series, with more pronounced outliers compared to the smoother performance of MSEw1, MSEw2, and MSEw3 among the three filters (MSEw1, MSEw2, MSEw3). MSEw3 is the best.

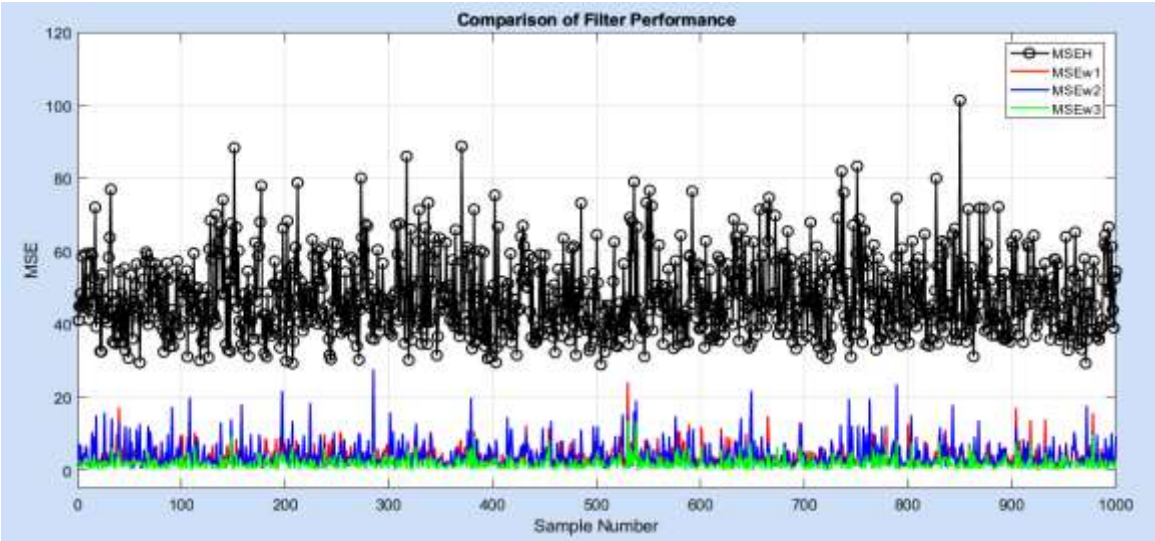

Figure 1. Comparison of Filter Performance by using MSE for a sample size of 100

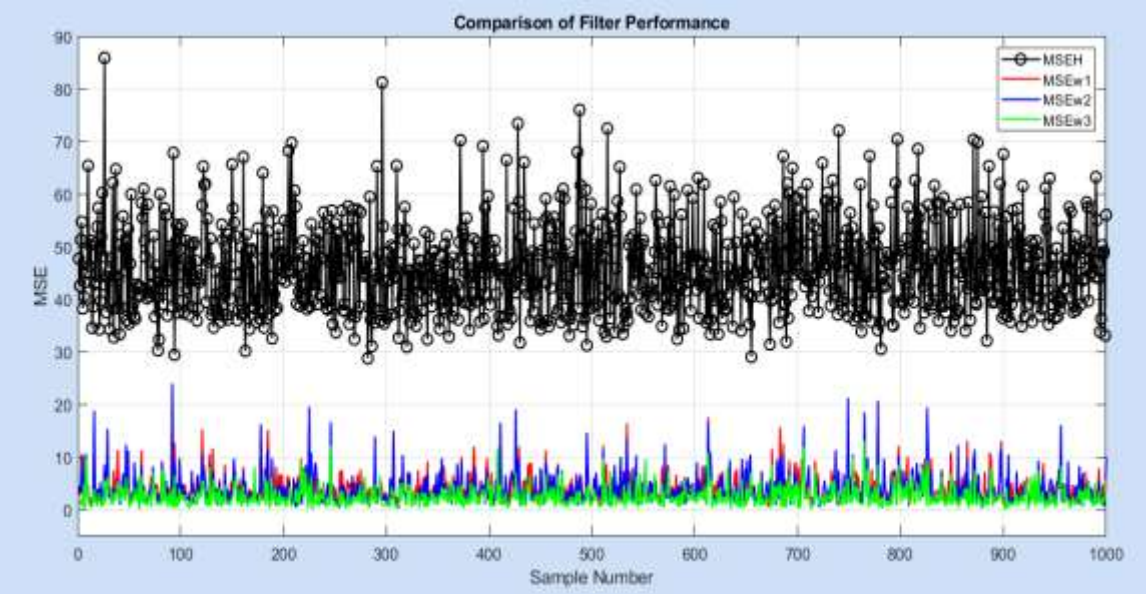

Figure 2. Comparison of Filter Performance by using MSE for a sample size of 150

Figure 2 shows the efficiency of the proposed method and its superiority over the traditional method at a sample size of 150.

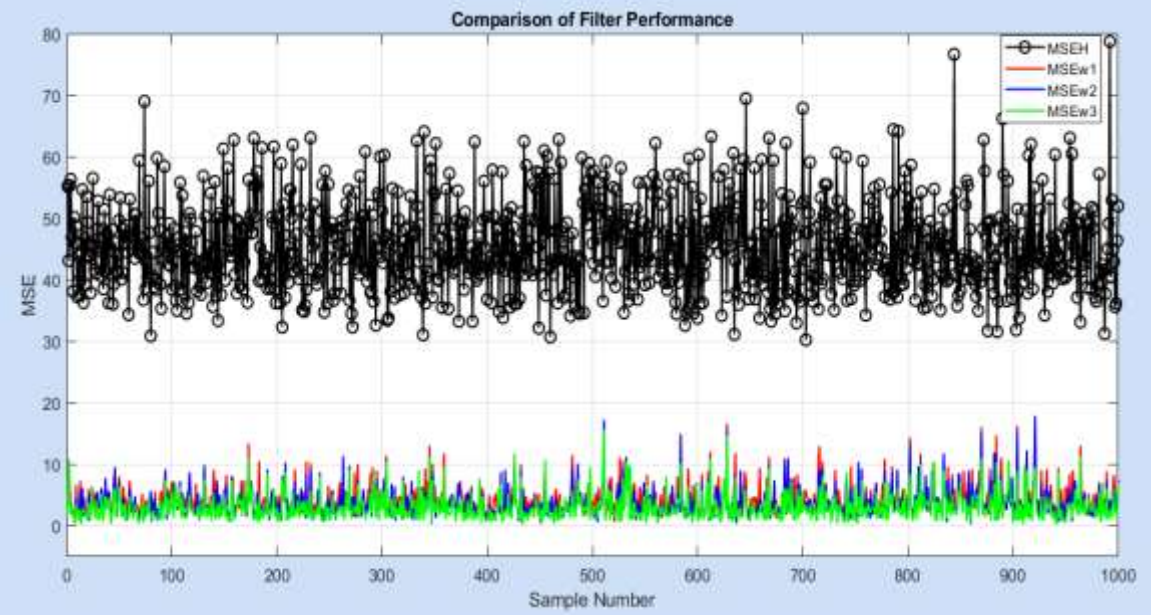

Figure 3. Comparison of Filter Performance by using MSE for a sample size of 200

Figure 3 shows the efficiency of the proposed method and its superiority over the traditional method at a sample size of 200.

3.2. Real Data

The real data from (Rencher, 2012) represent blood glucose measurements on three occasions (fasting). The multivariate regression model was estimated by the five methods (Without the filter, Hampel filter, and Coif5, Db20, and Dmey wavelets filter) with the MSE calculated for each model and the results are summarized in Table 2.

The results evaluate the effectiveness of various regression filtering methods by comparing their coefficients and Mean Squared Error (MSE). The "Without Filter" method yields the highest MSE of 1676.7000, indicating a poor model fit and high prediction error. Its regression coefficients show considerable variability, with large differences in the impact of predictors. In contrast, the Hampel Filter, designed to mitigate the effect of outliers, achieves a lower MSE of 208.0334 and offers more stable coefficient estimates, but it still does not match the performance of the proposed filters. Among the proposed filters, the Coif5 filter stands out with the lowest MSE of 7.9213, reflecting the best model accuracy and precision. This method produces the most consistent regression coefficients, demonstrating the highest effectiveness in reducing error. The Db20 filter also performs well, with a notable MSE of 46.4108, but it is less effective than Coif5. The Dmey filter, with an MSE of 13.2069, provides a balance between error reduction and coefficient stability. Overall, the proposed filters, especially Coif5, significantly outperform both the Hampel Filter and the "Without Filter" method, highlighting their superior capability in enhancing regression model performance.

Table (2) Regression Coefficients and MSE for Real Data

Methods	Regression Coefficients			MSE
Without Filter	47.2961	25.0896	32.5241	1676.7000
	0.5773	0.0583	0.5938	
	-0.1858	0.7636	-0.3520	
	0.4854	0.2554	0.8311	
Hampel Filter	90.0219	66.1235	80.2017	208.0334
	-0.0500	0.2773	0.1462	
	0.1816	0.2237	-0.0369	
	0.1761	-0.0363	0.2810	
Proposed Filter Coif5	105.3209	105.0360	118.4777	**7.9213**
	0.0525	-0.0279	-0.0204	
	0.0178	0.0269	-0.0717	
	0.0056	-0.0051	-0.0116	
Proposed Filter Db20	104.4113	104.8545	131.5511	46.4108
	0.0066	0.0233	0.0698	
	0.0673	-0.0202	-0.2788	
	0.0080	-0.0170	-0.0742	
Proposed Filter Dmey	108.6301	102.5845	120.3460	13.2069
	-0.0129	-0.0166	0.0426	
	0.0244	0.0296	-0.1224	
	0.0140	0.0130	-0.0465	

From Figure 4, summing up the figure which provides the residuals of the multivariate regression model for each of the responses (y_1, y_2, and y_3) without using any filter, the range of the residuals for the three models lies between ±50. This range is notably wide compared to the residuals obtained using other filtering methods. The residuals for y_1 exhibit the largest errors and most significant outliers, while y_2 and y_3 display relatively smaller residuals with fewer outliers. y_3 shows the most balanced residuals, though there are still some isolated deviations. These patterns suggest that the regression model for y_1 may need the most refinement or outlier handling. Or Residuals for y_1 (Top Plot), several large residuals, with some extreme values reaching close to ±50. This suggests significant model errors or potential outliers in the data. Residuals for y_2 (Middle Plot) are moderate residuals compared to y_1, with some points still deviating significantly from zero. Fewer extreme values, but still, some outliers are present. Residuals for y_3 (Bottom Plot) are relatively more balanced residuals with smaller deviations compared to the other plots. A few observations still show significant errors, especially around observations 20 and 40.

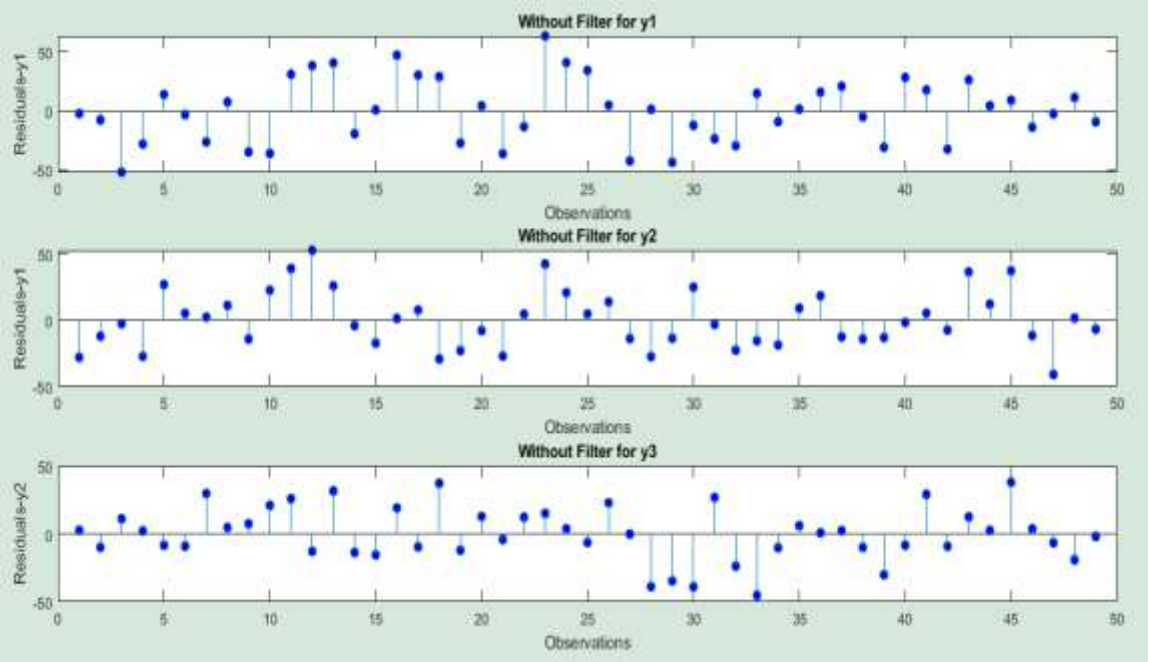

Figure 4. Residuals of the Multivariate Regression Model without filter data

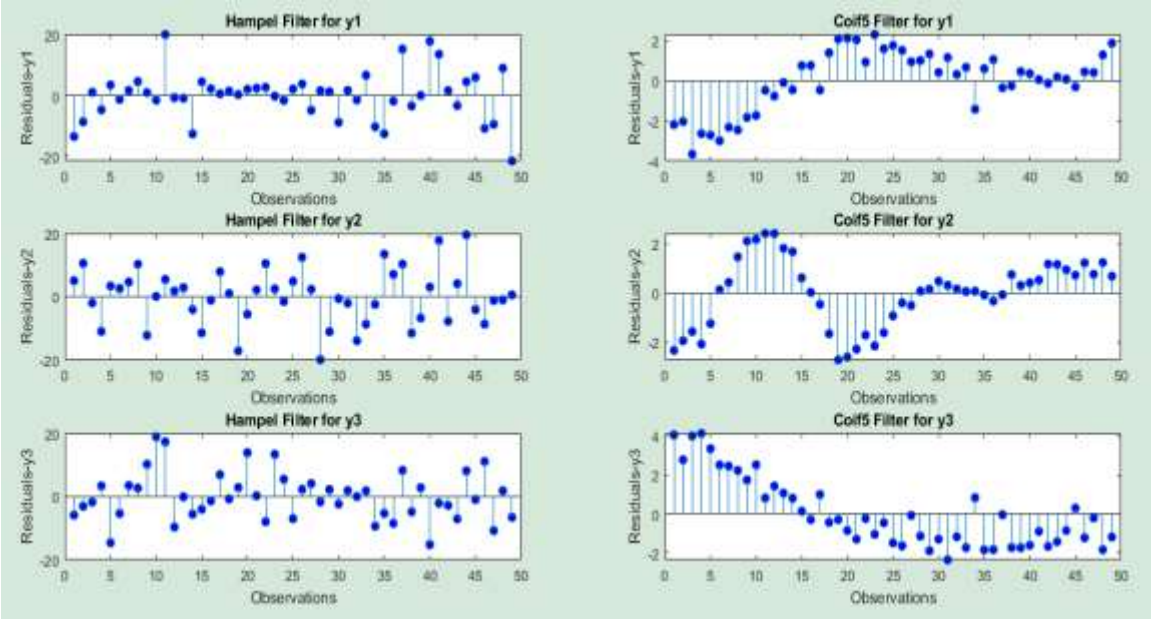

Figure (5) Residuals of Multivariate Regression Model for Hampel and Coif5 filter data

From Figure 5, when combining the residuals from the multivariate regression model for each response variable (y_1, y_2, and y_3) with the Hampel and Coif5 filters, the results show that for the Hampel filter, the residuals range from ±20 for all models. Similarly, the Coif5 filter's residuals range from (2 to -4) for y_1, ±2 for y_2, and finally (4 to -2) for y_3. This range is considerably less than the residuals obtained by using no filter.

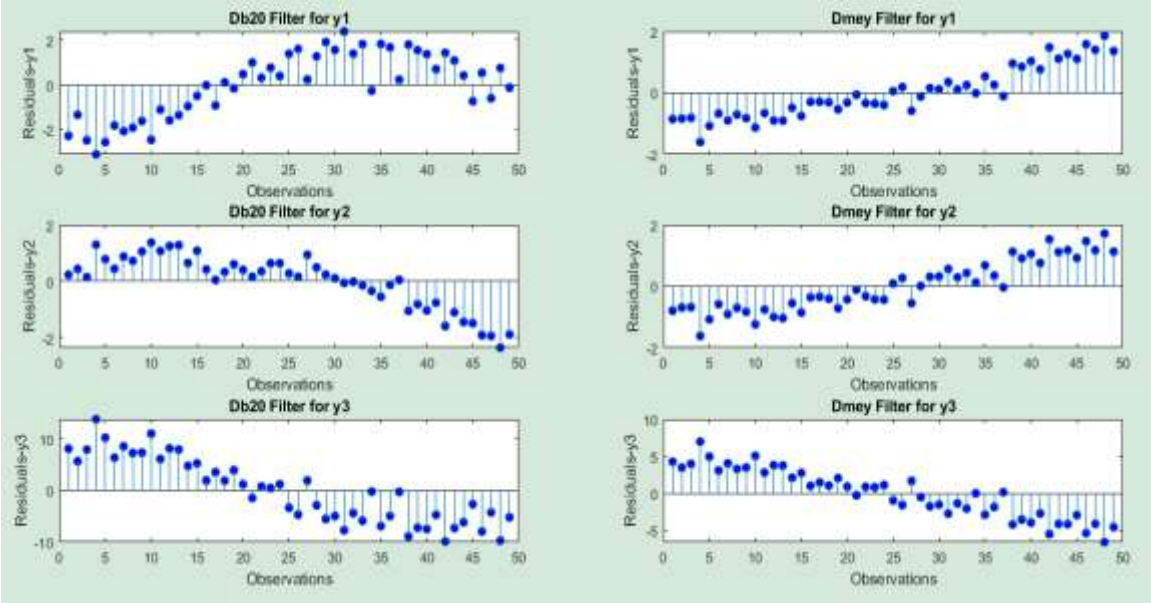

Figure 6. Residuals of Multivariate Regression Model for Db20 and Dmey filter data

From Figure 6, when aggregating the residuals from the multivariate regression model for each response variable (y_1, y_2, and y_3) with the Db20 and Dmey filters applied, for the Db20, the residuals fall within a range of ±2 for (y_1 and y_2) and ±10 for (y_3). And for the Dmey filter, also the residuals fall within a range of ±2 for (y_1 and y_2) and (-5 to 10) for (y_3). This range is notably wide compared to the residuals obtained using other filtering methods.

4. Conclusions

In conclusion, the evaluation of various regression filtering methods reveals significant differences in their effectiveness. Using no filter demonstrates the highest Mean Squared Error (MSE), indicating a poor fit and high prediction errors, with considerable variability in regression coefficients. The Hampel Filter improves performance by reducing MSE and stabilizing coefficients, but falls short compared to the proposed filters. Among these, the Coif5 filter proves to be the most effective, achieving the lowest MSE and most consistent coefficients, thereby offering superior accuracy and precision. The Db20 filter also shows strong performance with a relatively low MSE, though it is less effective than Coif5. The Dmey filter strikes a balance between reducing error and maintaining coefficient stability. Overall, the proposed filters, particularly Coif5, significantly enhance regression model performance, surpassing both the Hampel Filter and the no filter.

5. Recommendations

Based on the evaluation of regression filtering methods, it is recommended to utilize the Coif5 filter for its outstanding performance, as it achieves the lowest Mean Squared Error (MSE) and the most consistent regression coefficients, thereby offering the highest accuracy and precision. When a slightly less optimal but still effective alternative is needed, the Db20 filter is a suitable choice due to its relatively low MSE. The Dmey filter is also a good option for balancing error reduction with coefficient stability. It is advisable to avoid the "Without Filter" method due to its high MSE and variability in coefficients, and while the Hampel Filter provides some improvement over no filtering, it does not match the effectiveness of the proposed filters. Implementing these recommendations will enhance the accuracy and reliability of regression models.

References

1. Aggarwal, C. C. (2017). Outlier Analysis (2nd ed.). Springer. https://link.springer.com/book/10.1007/978-3-319-47578-3
2. Arts, L. P. A. & van den Broek, E. L. (2022). The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time-frequency analysis. Nature Computational Science. https://www.nature.com/articles/s43588-021-00183-z

3. Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data (3rd ed.). New York: Wiley. https://www.amazon.com/Outliers-Statistical-Data-V-Barnett/dp/0471930946
4. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93-104). LOF: identifying density-based local outliers
5. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18.. https://doi.org/10.1080/00401706.1977.10489493
6. Donoho, D. L., & Johnstone, I. M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. Biometrika, 81(3), 425-455. https://doi.org/10.1093/biomet/81.3.425
7. Greene, W. H. (2018). Econometric Analysis (8th ed.). Pearson. Econometric Analysis - VitalSource
8. Guo, T., Zhang, T., Lim, E., Lopez-Benitez, M., Ma, F., & Yu, L. (2022). A review of wavelet analysis and its applications: Challenges and opportunities. IEEe Access, 10, 58869-58903. https://ieeexplore.ieee.org/abstract/document/9785993/
9. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis: Global edition. Multivariate Data Analysis(Internet Archive)
10. Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association, 69(346), 383-393. https://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962
11. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). Robust Statistics: The Approach Based on Influence Functions. New York: Wiley. http://dx.doi.org/10.1002/9781118186435
12. Huber, P. J., & Ronchetti, E. M. (1981). Robust statistics john wiley & sons. *New York*, *1*(1). https://onlinelibrary.wiley.com/doi/book/10.1002/0471725250
13. Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models. McGraw-Hill. https://thuvienso.hoasen.edu.vn/handle/123456789/9564
14. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2015). Feature Selection: A Data Perspective. ACM Computing Surveys, 50(6), Article 94. https://doi.org/10.1145/3136625
15. Li, J., Liu, M., Li, X., & Wang, S. (2022). A Comprehensive Review on Outlier Detection with Data Imputation Techniques. Information Sciences, 610, 222-245. https://doi.org/10.1145/3645108
16. Mallat, S. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7), 674-693. https://ieeexplore.ieee.org/abstract/document/192463/
17. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis 5th ed. Introduction to Linear Regression Analysis
18. Mustafa, Q., & Ali, T. H. (2013). COMPARING THE BOX-JENKINS MODELS BEFORE AND AFTER THE WAVELET FILTERING IN TERMS OF REDUCING THE ORDERS WITH APPLICATION. *Journal of Concrete & Applicable Mathematics*, *11*(2). (DPU Online)(University of Zakho)
19. Najafi, M., & Hakim, A. (1992). Robust Estimation and Outlier Detection in Multivariate Data. Communications in Statistics - Theory and Methods, 21(5), 1495-1509.
20. Nielsen, M. (2001). On the construction and frequency localization of finite orthogonal quadrature filters. *Journal of Approximation Theory*, *108*(1), 36-52. https://doi.org/10.1006/jath.2000.3514
21. Omar, C., & Ali, T. H., Hassn, K. (2020). Using Bayes weights to remedy the heterogeneity problem of random error variance in linear models. *Iraqi Journal of Statistical Sciences*, *17*(2), 58-67. DOI: 10.33899/iqjoss.2020.167391
22. Omer, A. W., Sedeeq, B. S., & Ali, T. H. (2024). A proposed hybrid method for Multivariate Linear Regression Model and Multivariate Wavelets (Simulation study). *Polytechnic Journal of Humanities and Social Sciences*, *5*(1), 112-124. https://journals.epu.edu.iq/index.php/ptjhss/article/view/1452
23. Pearson, R. K. (2002). Outliers in Process Modeling and Identification. IEEE Transactions on Control Systems Technology, 10(1), 55-63. https://doi.org/10.1109/87.974338
24. Rencher, A.C. and Christensen, W.F. (2012). Wiley Series in Probability and Statistics. In Methods of Multivariate Analysis (eds A.C. Rencher and W.F. Christensen). https://doi.org/10.1002/9781118391686.scard
25. Rousseeuw, P. J., & Leroy, A. M. (2003). Robust Regression and Outlier Detection. Wiley. Robust regression and outlier detection
26. Strang, G., & Nguyen, T. (1996). Wavelets and Filter Banks. Wellesley-Cambridge Press. https://books.google.iq/books?id=Z76N_Ab5pp8C&source=gbs_navlinks_s
27. Tabachnick, B. G., & Fidell, L. S. (2019). Using Multivariate Statistics (7th ed.). Pearson. Pearson Higher Ed (PearsonHigherEd)
28. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

29. Tukey, J. W. (1977). Exploratory data analysis. *Reading/Addison-Wesley*. https://link.springer.com/content/pdf/10.1007/978-3-031-20719-8_2?pdf=chapter%20toc
30. Unser, M. (2002). Sampling—50 Years After Shannon. Proceedings of the IEEE, 90(5), 742-765. https://doi.org/10.1109/TAC.2002.1000281

Appendix

```
clc
clear all
n=200;q=4; p =3;beta=[2 4 6;4 3 5;3 6 4;3 5 2;6 4 2]; randn('seed',1234);
for j=1:1000
   x=randn(n,q); E=randn(n,p)*[1 2 3;2 1 2;3 2 1]; X=[ones(n,1) x];
yc= X*beta+E;yc(10,1)=35;yc(15,2)=-30;yc(10,3)=-35;
% without filter
[betah sigma Eh c logl]=mvregress(X,yc); Ehc=yc-X*betah; EhpEhc=Ehc'*Ehc;
MSE(j)=trace(EhpEhc)/(n-q-1);
% Classical Hampel Filter
[yCC,i,xmedian,xsigma] = hampel(yc,50,3); [betah sigma Eh c logl]=mvregress(X,yCC);
EhCC=yCC-X*betah; EhpEhCC=EhCC'*EhCC; MSEH(j)=trace(EhpEhCC)/(n-q-1);
% Wavelet Coif5 Filter
yw1= wdenoise(yc,6,'Wavelet','coif5', 'DenoisingMethod','universal','ThresholdRule','soft');
[betaw1 sigma Eh c logl]=mvregress(X,yw1); Ehw1=yw1-X*betaw1; Ew1=Ehw1'*Ehw1;
MSEw1(j)=trace(Ew1)/(n-q-1);
% Wavelet Db2 Filter
yw2= wdenoise(yc,6,'Wavelet','Db20', 'DenoisingMethod','universal','ThresholdRule','soft');
[betaw2 sigma Eh c logl]=mvregress(X,yw2); Ehw2=yw2-X*betaw2; Ew2=Ehw2'*Ehw2;
MSEw2(j)=trace(Ew2)/(n-q-1);
% Wavelet Dmey Filter
yw3= wdenoise(yc,6,'Wavelet','dmey', 'DenoisingMethod','universal','ThresholdRule','soft');
[betaw3 sigma Eh c logl]=mvregress(X,yw3); Ehw3=yw3-X*betaw3; Ew3=Ehw3'*Ehw3;
MSEw3(j)=trace(Ew3)/(n-q-1);
end
MMSE=mean(MSE), MMSEH=mean(MSEH), MMSEw1=mean(MSEw1)
MMSEw2=mean(MSEw2), MMSEw3=mean(MSEw3)
```

المقارنة بين مرشح التقليص المويجي وهامبيل في تحليل نماذج الانحدار الخطي متعدد المتغيرات

[1]أميره ولي عمر، [2]طه حسين علي

قسم الإحصاء والمعلوماتية، كلية الإدارة والاقتصاد، جامعة صلاح الدين ، اربيل، العراق.

الخلاصة: إن وجود القيم الشاذة في بيانات نموذج الانحدار الخطي متعدد المتغيرات يؤثر على دقة المعلمات المقدرة للنموذج ويؤدي إلى قيم بواقي كبيرة غير مقبولة. لذلك، نستخدم بعض المرشحات، بما في ذلك مرشح هامبل لمعالجة القيم الشاذة (أو استخدام بعض الطرائق الحصينة). يقترح هذا البحث توظيف مرشح التقليص المويجي لمعالجة مشكلة القيم الشاذة في بيانات نموذج الانحدار الخطي متعدد المتغيرات باستخدام الموجات (Coiflets, Daubechies, Demy) مع طريقة العتبة الشاملة والقاعدة الناعمة. لتوضيح كفاءة الطريقة المقترحة (مرشح التقليص المويجي) تمت مقارنتها بالطريقة التقليدية (مرشح هامبل) بناءً على معيار متوسط الخطأ التربيعي من خلال المحاكاة والبيانات الحقيقية. تم استخدام برنامج MATLAB للقيام بذلك. أثبتت النتائج أن طريقة مرشح التقليص المويجي كانت أكثر كفاءة من الطريقة التقليدية في معالجة مشكلة القيم الشاذة والحصول على معلمات نموذج متعدد المتغيرات أكثر دقة من طريقة مرشح هامبل.

الكلمات المفتاحية: نماذج الانحدار الخطي متعدد المتغيرات، القيم الشاذة، مرشح هامبل، التقليص المويجي، وقطع العتبة.