STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, January 2026, pp 324–335.
Published online in International Academic Press (www.IAPress.org)

Comparative Evaluation of Classical Robust and Wavelet Enhanced Beta Regression Models for Proportional Data

Mahmood M Taher [1,*], Talal Abd Al-Razzaq Saead Al-Hasso [1], Taha Hussein Ali [2]

[1] *Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Iraq*
[2] *Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University, Iraq*

Abstract This paper presents a comprehensive comparative modeling of conventional, robust, and wavelet-augmented beta regression models for proportional data. To preprocess the response variable, four discrete wavelet transforms—Daubechies IV (Db4), Coiflets IV (Coif4), Symlets IV (Sym4), and discrete Meyer (Dmey) were applied to remove noise and enhance model robustness. Results demonstrated that wavelet-enhanced models surpassed conventional and resilient Beta regression at all points. The wavelet pre-processing effectively eliminated noise, producing more precise and smoother forecasts. The developed models were also applied to a real body composition data set and were shown to replicate the simulation results as well as demonstrate real-world utility but introduces additional computational overhead and sensitivity to wavelet family selection.

Keywords Beta regression, wavelet transformations, robust estimation, outlier detection, proportional data

DOI: 10.19139/soic-2310-5070-3144

1. Introduction

Beta regression models attracted more interest in the recent past because they can measure continuous bounded data with high precision, especially proportions that range from 0, 1. [9], who originally came up with the Beta regression model, are especially suitable for proportion-based data in the majority of study fields, including ecology, biostatistics, and economics, since they can handle heteroskedasticity as well as asymmetry. However, it is well understood that standard maximum likelihood estimation is susceptible to outliers and data anomalies. This can create biased parameter estimates as well as less-than-ideal model fit [12].

More robust regression approaches, including those that use the Huber loss function, have been suggested to overcome these weaknesses through minimizing outlier influence while preserving estimation efficiency [11, 18]. Although such techniques tend to be more resistant to contamination, they tend to be less effective when there are subtle noise patterns or high-frequency oscillations in the response variable.

For this purpose, the application of wavelet-based signal processing methods to statistical models is a development with great promise. Wavelet transforms can be used as an effective means of removing noise and isolating relevant structural information from data. These transforms were originally proposed by [6] and have been extended to include Queflet, Simlet, and discrete Meyer (Dmey) wavelets. To make models more stable and accurate, especially in a noisy data environment, successful experiments have recently been conducted indicating that wavelet noise removal can be enhanced using regular or stable regression models [3, 10, 13].

This study contributes to the scientific literature through a systematic comparison of classical, robust, and wavelet-augmented beta regression models in artificial and applied settings. This research provides new insights into how

[*] Correspondence to: Mahmood M Taher (Email: mahmood81_tahr@uomosul.edu.iq). Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

preprocessing theory can, in principle, improve the robustness and predictive ability of proportional information models in the presence of noisy data and outliers, by comparing a variety of wavelet power families and their outputs to an industrial standard.

2. Methodology

This work suggests an integrated modeling framework combining classical Beta regression, robust estimation methods, and wavelet-based denoising to enhance the performance of regression models in the analysis of continuous proportion data. The following subsections detail the theoretical formulation, parameter estimation methods, wavelet-based preprocessing, and performance evaluation metrics used in this work.

2.1. *Beta Regression Model*

In cases where Yi is a continuous response variable with the open interval (0, 1), the Beta distribution is a good modelling solution because it is flexible and capable of taking skewness and different dispersion [2]. The Beta distribution probability density function is as follows:

$$f\left(y_i;a_i;b_i\right)=\frac{\Gamma\left(a_i+b_i\right)}{\Gamma\left(a_i\right)+\Gamma\left(b_i\right)}y_i^{a_i-1}{\left(1-y_i\right)}^{b_i-1}\quad,\quad 0<y_i<1 \tag{1}$$

Where the parameters of the shape of observation i are a_i and b_i. A logit link function is used to associate the conditional mean of the response variable with a set of explanatory variables, which is stated as;

$$logit\left(\mu_i\right)=X_i^T\beta \tag{2}$$

Where:
$\mu_i{=}\frac{a_i}{a_i+b_i}$ is the expected value of the Beta distribution.
X_i^T is a predictor variable of the $i-th$observation.
β is an estimated vector of regression coefficients.

2.2. *Maximum Likelihood Estimation*

The parameters are estimated using the standard Beta regression which is the maximization of the log-likelihood equation as shown in Equation (1) which is derived using the PDF of the Beta distribution. The log-likelihood function for a sample of size n can be written as:

$$\ell\left(\beta\right)=\sum_{i=1}^{n}\left[ln\Gamma\left(a_i+b_i\right)-ln\Gamma\left(a_i\right)-ln\Gamma\left(b_i\right)+\left(a_i-1\right)lny_i+\left(b_i-1\right)\ln\left(1-y_i\right)\right] \tag{3}$$

The values of β are numerically estimated by maximizing this function through iterative optimization routines like the Nelder-Mead simplex algorithm, as applied through MATLAB's *fminsearch* function. These estimates are then used as the baseline against which robust and wavelet-enhanced models are compared [9].

2.3. *Robust Beta Regression Using Huber Loss*

To improve model resilience against outliers and data irregularities, a robust Beta regression model is developed by incorporating the Huber loss function [4, 23]. This loss function behaves quadratically for small residuals and linearly for large residuals, providing resistance to extreme values. It is defined as:

$$\rho_c\left(r_i\right)=\begin{cases}\frac{1}{2}r_i^2 & if\left|r_i\right|\le c\\ c\left(\left|r_i\right|-\frac{1}{2}c\right) & if\left|r_i\right|\ge c\end{cases} \tag{4}$$

Where:

$r_i = y_i - \widehat{\mu}_i$ is the residual for observation i.
$\widehat{\mu}_i = \frac{1}{1+\exp(-X_i^T\beta)}$ is the predicted mean.
c is a tuning constant (commonly set at 1.345).
The robust estimate of β is obtained by minimizing the following objective function:

$$Q\left(\beta\right) = \sum_{i=1}^{n} \rho_c \left(\, y_i - \widehat{\mu}_i\right) \tag{5}$$

This function is minimized using the same optimization algorithm applied in the classical MLE procedure.

2.4. *Wavelet-Based Denoising for Beta Regression*

Localized variations, random fluctuations and error in measurements are usually prevalent in real world data of proportions. To solve this, model fitting is preceded by the application of a discrete wavelet transform (DWT) to the response variable [15, 16]. The DWT splits the original signal Y(t) into a representation of approximation and detail coefficients at several levels of resolution [7, 8, 22]:

$$Y\left(t\right) = \sum_{k} a_{j0,k} \emptyset_{j0,k}\left(t\right) + \sum_{j=j0}^{J} \sum_{K} d_{j,k} \psi_{j,k}\left(t\right) \tag{6}$$

Where:
$\phi_{j0,k}$are scaling functions representing low-frequency components.
$\varphi_{j,k}$are wavelet functions representing high-frequency details.
$a_{j0,k}$and $d_{j,k}$are approximation and detail coefficients, respectively.
The thresholding of the detail coefficients $d_{j,k}$, to eliminate noise but retain significant data structures, is used to reduce noise. The response after denoising is then reconstructed through the backward wavelet transform to get a smooth signal $\widetilde{\mathrm{Y}}$ that is then used to re-estimate the robust Beta regression model [15, 17].

The wavelet families employed in this study include Db4, Coif4, Sym4, and Dmey. Soft thresholding was applied to all detail coefficients, and reconstruction followed the standard inverse DWT process. A single-level decomposition was used unless otherwise specified. The modeling sequence strictly followed six steps: (1) classical estimation, (2) robust estimation, (3) DWT preprocessing, (4) thresholding, (5) inverse reconstruction, and (6) refitting the robust model to the denoised response. Performance metrics (RMSE, MAE, R^2) were computed using standard formulas.

2.5. *Computational Considerations*

To address the computational overhead associated with wavelet preprocessing, we note that applying the Discrete Wavelet Transform (DWT) requires approximately O(n log n) operations. This makes wavelet-enhanced models computationally heavier than classical Maximum Likelihood Estimation (MLE) and robust Beta regression, especially when the preprocessing must be repeated many times. In our simulation study, where each scenario was replicated 1000 times, this additional cost was more noticeable. However, the improved accuracy obtained through wavelet denoising justified the added computation time. All wavelet operations, including decomposition, soft thresholding, and reconstruction, were implemented using the MATLAB Wavelet Toolbox, while parameter estimation for classical and robust Beta regression employed MATLAB's fminsearch function based on the Nelder–Mead optimization algorithm. This discussion clarifies the computational trade-off of incorporating DWT as a preprocessing step and improves the methodological transparency of the proposed framework.

2.6. *Model Performance Evaluation*

To comprehensively evaluate model performance, three metrics are utilized [19, 20]:

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Which measures the average prediction error magnitude [1].
Coefficient of Determination R^2:

$$R^2 = 1 - \frac{\sum_{I=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{I=1}^{n}(y_i - \overline{y})^2} \tag{8}$$

This quantifies the proportion of variance in the response variable explained by the model [14, 21].

2.7. Summary of the Proposed Method

The suggested modeling plan combines three complementary schemes of Beta regression using the following systematizing:

1. Estimating the classical Beta regression model by Maximum Likelihood Estimation.
2. Estimating a strong Beta regression model, using the Huber loss function so as to minimize the impact of outliers.
3. Apply DWT denoising to the response variable using four wavelet families: Db4, Coif4, Sym4, and Dmey.
4. Carry out extensive thresholding of the detail coefficients of the wavelet decomposition. Particularly, use soft thresholding, which becomes smaller wavelet coefficients gradually, in effect mildly reducing noise, and important signal details are retained.
5. Reconstruct the denoised response variable using the inverse wavelet transform.
6. Refit the robust Beta regression model using the denoised response data.

Compare all fitted models based on various performance measures: RMSE, MAE, R^2, and residual diagnostics to evaluate accuracy, robustness, and overall accuracy of the model.
This sequential procedure combines classical, robust, and wavelet-based techniques to enhance the predictive accuracy and resilience of Beta regression models when analyzing proportion data, especially in the presence of noise and outliers.

2.8. Software Implementation

All analyses were conducted using MATLAB (version 2025b). Classical and robust Beta regression estimations were performed using the fminsearch function based on the Nelder–Mead algorithm. Wavelet decomposition, soft thresholding, and reconstruction were carried out using MATLAB's Wavelet Toolbox functions. This ensures full reproducibility of the modeling framework.

3. Applications: Simulation Study and Real-World Case Study

To illustrate the performance of the proposed Beta regression methodologies, two applications are provided in this section: a simulation study to evaluate model characteristics in a controlled setting, and a case study to show the usefulness of the proposed approaches. The two applications adhere to a process of formulating and evaluating models described in the following sections.

3.1. Simulation Study

Beta regression is a relatively new and growing statistical tool for the dealing with continuous response variables bounded in the open interval (0,1), such as proportions and probabilities. It can adapt to the shape and skewness of bounded data, which clearly is an advantage for many applications in the real world. Copying the previous situation of many classical statistical models, common Beta regression depending on maximum likelihood estimation (MLE) is particularly vulnerable to outliers and data contamination. This vulnerability results in biased estimates of parameters and poor predictive performance.

In response to these issues, this simulation experiment was constructed specifically to test and compare the performance of three different modeling techniques as they are affected by artificially created outliers. The first model uses traditional MLE Beta regression, the second uses a robust Beta regression model using the Huber loss function to reduce the influence of outliers, and the third involves wavelet-based denoising methods before estimation with the robust method. The third involves pre-modeling noise reduction and structural pattern extraction of the response variable using discrete wavelet transforms.

Synthetic data were created according to the logistic transformation of a linear predictor, and controlled outliers were added to simulate a realistic contamination scenario. Model performance was evaluated through root mean square error (RMSE), mean absolute error (MAE), and an adjusted version of the coefficient of determination (R^2) based on the original data variance. This experiment is only a first attempt to establish how wavelet models stand against traditional and robust Beta regression models in terms of predictive performance and strength; further and more extensive simulations will be conducted in the following chapters.

In Figure 1, several Beta regression models are plotted, and the impact of outlier-contaminated data can be observed. The plot in the top-left corner shows the estimates of the classical Beta regression model obtained by Maximum Likelihood Estimation (MLE). This model should be able to keep pace with the general behavior of the original observations, but it is overly sensitive to local variations, and it is not smooth and generalized since it is jagged, and it responds to outliers.

The results of the robust Beta regression with Huber loss are shown in the top right panel. The resulting robust curve is smoother than the MLE one, and apparently, it is less influenced by extremes. There is still variance that is not explained, especially in regions of the response variable where variance is high.

The bottom four panels show the outputs of the hybrid models that combine wavelet-based denoising with robust Beta regression. Each panel corresponds to a different wavelet filter: Db4, Coif4, Sym4, and Dmey. In all four cases, the denoised observations appear significantly smoother than the original data, and the resulting predictions align more closely with the underlying structure of the response variable. It is important to note here that the Dmey and Coif4 wavelets provide a much smoother prediction with more distinct trends, which means they perform better by filtering high-frequency noise and the influence on outliers.

In general, the idea that wavelet-enhanced robust models can furnish the more stable and more accurate representation of the data is supported by visual evidence, particularly when distortion is present. The findings of these studies indicate that a combination of signal processing approaches, including the wavelet transform, with sound statistical modeling can be used to increase the effectiveness of prediction in noisy conditions.

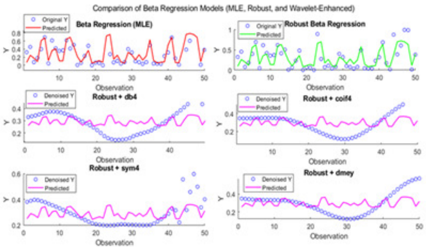

Figure 1. Comparison of Classical, Robust, and Wavelet-Enhanced Beta Regression Models

Table 1 compares the results of the six modeling methods deployed to the simulated data of Beta distribution with outliers. Classical Beta regression model with maximum likelihood obtained an RMSE of 0.1760, an MAE of 0.1324 and an R^2 of about 59.95 which indicates that it is a moderate predictive model with a significant sensitivity to contamination of the data. The robust Beta regression that uses Huber cost function, slightly enhanced performance with a reduced RMSE (0.1726) and an appearance of a slightly elevated R^2 (61.51%).

Significant results were obtained when wavelet denoising was applied before robust regression. Among the wavelet-based models, the Sym4 filter achieved the best overall performance, with the lowest RMSE (0.0911), the lowest MAE (0.0744), and the highest R^2 (89.28%). Db4 filter was also quite competitive, with an RMSE of 0.0922 and an R^2 of 89.02%. Coif4 and Dmey, despite providing slightly more elevated error statistics, were also well above the performance of the classical and the robust mono-variate models by themselves.
As expected, these results support the advantages of applying wavelet robust regression in controlling for outlier influence and providing higher model fitness. Combining aspects of signal preprocessing with robust estimation represents a useful avenue in the modeling of complex noisy datasets.

Table 1. Performance Comparison of Classical, Robust, and Wavelet-Enhanced Beta Regression Models in the First Simulation

Method **Criteria**	**Classical**	**Robust**	**Db4**	**Coif4**	**Sym4**	**Dmey**
RMSE	0.1760	0.1726	0.0922	0.1160	0.0911	0.1161
MAE	0.1324	0.1357	0.0770	0.0913	0.0744	0.0901
R^2	59.95%	61.51%	89.02%	82.61%	89.28%	82.59%

To test the model performance in more realistic and scalable situations, the simulation was repeated for n=50, 100, 200, 300, with 1, 2, and 3 independent variables each in 1000 replications. The mean results over all iterations are shown in Tables 2, 3, and 4. The performance of the classical MLE Beta regression, robust Beta regression based on Huber loss, and four other wavelet robust models based on the wavelet filters Db4, Coif4, Sym4, and Dmey was assessed using RMSE, MAE, and R^2. Results indicate that the wavelet model variants, especially those built using Coif4 and Sym4 wavelets, show significantly less error and include more predictors than both the classical and robust models under all conditions examined. It should also be mentioned that the gain in performance is even greater when the sample is larger and the model becomes more complex with the inclusion of additional predictors.

Table 2. Average Simulation Results for Classical, Robust, and Wavelet-Enhanced Beta Regression Models with One Predictor

Method **Criteria**	**n**	**Classical**	**Robust**	**Db4**	**Coif4**	**Sym4**	**Dmey**
RMSE	50	0.1677	0.1659	0.0721	0.0626	0.0663	0.0666
MAE		0.1328	0.1341	0.0568	0.0503	0.0524	0.0548
R^2		0.5166	0.5266	0.8955	0.9205	0.9117	0.9112
RMSE	100	0.1500	0.1493	0.0645	0.0587	0.0616	0.0610
MAE		0.1154	0.1164	0.0503	0.0464	0.0481	0.0489
R^2		0.5694	0.5733	0.9131	0.9283	0.9212	0.9230
RMSE	200	0.1375	0.1372	0.0600	0.0569	0.0576	0.0569
MAE		0.1053	0.1058	0.0466	0.0446	0.0448	0.0451
R^2		0.6148	0.6162	0.9233	0.9311	0.9295	0.9310
RMSE	300	0.1327	0.1326	0.0569	0.0547	0.0550	0.0548
MAE		0.1018	0.1022	0.0443	0.0432	0.0430	0.0434
R^2		0.6308	0.6116	0.9301	0.9356	0.9350	0.9354

Tables 2, 3, and 4 present the average performance metrics of classical Maximum Likelihood Estimation (MLE)-based Beta regression, robust Beta regression using Huber loss, and four wavelet-enhanced robust Beta regression models (Db4, Coif4, Sym4, and Dmey). These results are evaluated under varying conditions of sample size (n = 50, 100, 200, 300) and number of predictors (one, two, and three). The performance criteria include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2).
Wavelet-enhanced models in the single predictor model show much better results compared to both the classical and robust Beta regression models. The values of RMSE and MAE in wavelet models are significantly lower, and the value of R^2 is significantly higher, always above 0.89 versus below 0.64 in classical methods. An example is that Coif4 wavelet model gives RMSE of 0.0547 and R^2 of 0.9356 at n = 300 compared to the classical model,

Table 3. Average Simulation Results for Classical, Robust, and Wavelet-Enhanced Beta Regression Models with Two Predictors

Method Criteria	**n**	**Classical**	**Robust**	**Db4**	**Coif4**	**Sym4**	**Dmey**
RMSE		0.1567	0.1553	0.0769	0.0663	0.0702	0.0728
MAE	50	0.1233	0.1227	0.0618	0.0544	0.0569	0.0604
R^2		0.7009	0.7064	0.9177	0.9374	0.9323	0.9259
RMSE		0.1454	0.1448	0.0741	0.0679	0.0701	0.0728
MAE	100	0.1134	0.1132	0.0594	0.0553	0.0565	0.0595
R^2		0.7442	0.7462	0.9293	0.9401	0.9369	0.9318
RMSE		0.1379	0.1377	0.0714	0.0685	0.0694	0.0699
MAE	200	0.1078	0.1076	0.0569	0.0550	0.0553	0.0564
R^2		0.7703	0.7710	0.9367	0.9416	0.9403	0.9392
RMSE		0.1353	0.1352	0.0713	0.0688	0.0691	0.0697
MAE	300	0.1059	0.1057	0.0566	0.0550	0.0550	0.0558
R^2		0.7790	0.7794	0.9374	0.9417	0.9412	0.9401

Table 4. Average Simulation Results for Classical, Robust, and Wavelet-Enhanced Beta Regression Models with Three Predictors

Method Criteria	**n**	**Classical**	**Robust**	**Db4**	**Coif4**	**Sym4**	**Dmey**
RMSE		0.1529	0.1510	0.0746	0.0655	0.0688	0.0718
MAE	50	0.1199	0.1186	0.0601	0.0538	0.0556	0.0592
R^2		0.7094	0.7162	0.9212	0.9379	0.9327	0.9271
RMSE		0.1429	0.1422	0.0733	0.0680	0.0698	0.0721
MAE	100	0.1113	0.1108	0.0584	0.0550	0.0560	0.0586
R^2		0.7487	0.7512	0.9296	0.9392	0.9364	0.9319
RMSE		0.1363	0.1360	0.0710	0.0678	0.0688	0.0697
MAE	200	0.1064	0.1060	0.0564	0.0545	0.0550	0.0562
R^2		0.7710	0.7720	0.9361	0.9414	0.9397	0.9380
RMSE		0.1336	0.1334	0.0704	0.0682	0.0683	0.0690
MAE	300	0.1045	0.1041	0.0561	0.0548	0.0545	0.0554
R^2		0.7806	0.7812	0.9377	0.9415	0.9415	0.9401

which gives RMSE = 0.1326 and R^2 = 0.6116. These results show the increased accuracy and explanatory power of wavelet filtering in the simpler model case.

The trend of the models of the wavelets to be better than the classical and robust methods has been maintained with two predictors. Although the overall values of RMSE and MAE slightly rise as the complexity of model is added, the use of a wavelet models has a definite upper hand. To illustrate, Coif4 has produced an RMSE of 0.0688 and R^2 of 0.9417 at n = 300 in comparison to RMSE of 0.1352 and R^2 of 0.7794 of the classical model. This shows that the wavelet models have the ability to successfully manage higher levels of dimensionality without compromising performance.

When extended to three predictors, wavelet-enhanced models continue to lead in performance metrics despite the increasing challenge of model complexity. At n = 300, Coif4 achieves RMSE = 0.0682, MAE = 0.0548, and R^2 = 0.9415, while the classical model shows RMSE = 0.1334, MAE = 0.1041, and R^2 = 0.7812. This sustained performance highlights the robustness and scalability of wavelet-based enhancement for Beta regression modeling in multi-response.

Table 1 presents the results of the first simulation with the complete dataset of any size and number of predictors in order to have a point of reference. It demonstrates that wavelet-enhanced models have a huge improvement over classical and robust Beta regressions, and Coif4 and Sym4 filters show the best R^2 values of about 89, compared to about 60 with classical models. This preliminary evidence agrees with the detailed simulation findings of Tables 2-4.

Wavelet improvement is always effective to increase model accuracy: In every condition, wavelet-enhanced Beta regressions would report less RMSE and MAE values and larger R^2 than both classical MLE and robust Huber models.

Sym4 and Coif4 of wavelet filters are especially successful: These filters make repeated optimal tradeoffs between reduction of error and the best fit to the model.
The improvement in performance as size and model complexity grows: the relative improvements are stronger when wavelet models have more predictors and the sample size is larger, which is a characteristic of the wavelet models to extract meaningful data and reduce noise in the complex forms of data.
Robust regression alone improves moderately: While robust Beta regression using Huber loss yields modest gains over classical MLE, it does not approach the level of improvement seen with wavelet enhancement.

3.2. Real Data Application

This is important because the study of body composition has gained relevance in recent years, being highly related to public health as well as fundamental for early clinical diagnosis and treatment of the obesity epidemic and associated chronic diseases like cardiovascular diseases and type 2 diabetes. While many methods exist to assess body composition, body fat percentage is an important biomarker commonly used by researchers and health care professionals to inform assessments of health status. An increasing amount of research demonstrates higher body fatness and strong association with many adverse health conditions.
As a resource for continuing health surveillance and risk assessment, several more recent data sets have been developed with more detailed anthropometric and demographic measurements. A significant dataset in this regard is the Body Composition Dataset from 2023 [5], João Da Silva on the Kaggle platform. This dataset has information on contemporary and complete body fat percentages as well as other relevant explanatory variables, including age and body weight, derived from a heterogeneous sample of subjects. This dataset is rich and up to date, thus it is very well suited to form the basis for the development of more sophisticated statistics as well as for a deeper understanding of the distribution of fat across populations.
The uniqueness of this data is the fact that it can be analyzed using modern analytical models. And that is on top of the fact that the body fat percentage is a continuous variable that assumes values within the range of 0 to 1 and this is consistent with Beta Regression assumptions. These kinds of models suit data that are proportions and are flexible enough to capture the non-linear and heteroscedastic relationship that are common with many health measures. In addition, more advanced modeling techniques can be used due to the quality of the dataset and its resolution such as wavelet-based denoising techniques which separate the signal and the noise resulting in better estimation.
Thus, this work uses the recently released 2023 Body Composition Dataset to implement and test three different ways of modeling the data: Eq. (1) classical Maximum Likelihood (ML) Beta Regression, Eq. (2) a robust ML Beta Regression analysis that uses Huber loss to reduce the influence of outliers, and Eq. (3) Beta Regression based on wavelet analysis that incorporates signal processing techniques to enhance the prediction of outcomes even more. By applying them to a specific health-related dataset, the intention is to evaluate how well each of these approaches can perform predictions as well as models, providing information about their efficacy in terms of finding the actual model of a real health dataset.

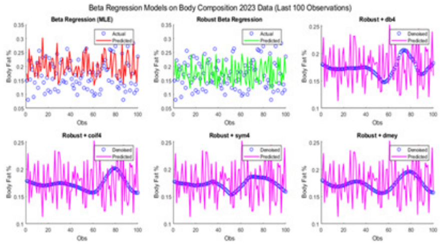

Figure 2. Each panel represents a prediction curve obtained from one of the six Beta regression models

Figure 2 shows the last 100 observed Body Composition 2023 data set's performance of six Beta Regression Models' body fat percentage predictions. The models are as follows: Classical Beta Regression based on MLE, Robust Beta Regression, and Robust Beta Regression plus wavelet denoising employing the following wavelet filters: Db4, Coif4, Sym4, and Dmey.

The top-left plot illustrates the predictions of the classical MLE-based Beta regression. It shows high volatility in the predicted values (red line) relative to the actual data points (blue circles), suggesting poor model fit and instability in the presence of noise or possible outliers.

The top-middle plot presents the robust Beta regression model using Huber loss. Although it shows a small step forward in its ability to trace the overall trend as compared with the classical model, it, nevertheless, has pronounced fluctuations and is also restricted in terms of reflecting the smooth underlying structure of data. Both models have the tendency to overfit local variations without quite generalizing the trend.

The other four plots (bottom row and top-right) demonstrate the outcomes of the sturdy Beta regression models that were supplemented with the wavelet-based denoising with Db4, Coif4, Sym4 and Dmey filters. In these visualizations:

1. The denoised signal (magenta line) is the data which corresponds to the percentage of body fat that is cleaned.
2. The resultant predicted value (bluish line) closely attribute to the denoised signal and appears to exhibit a much better-behaved and smoother behaviour.
3. The models created with the help of the wavelet are all superior to the classical and robust ones in the way they capture the structural trend in the data. Notably:
4. Coif4 and Sym4 are particularly smooth and accurate in prediction with the predicted line matching the denoised signal at the zero oscillations minimize.
5. Db4 and Dmey also are very much better than no-wavelet methods, but seem to be a little more sensitive to local variations than Coif4 and Sym4.

The results indicate that it is possible to effectively use the combination of powerful statistical modeling and wavelet-based denoising. The wavelet-enhanced models eliminate noise, as well as, maintain the key characteristics of the data hence better generalization and readability.

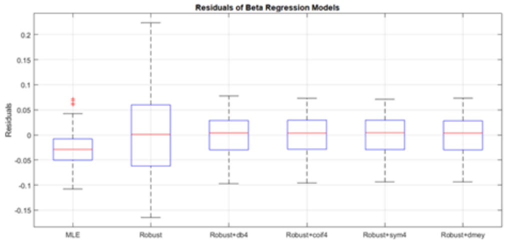

Figure 3. The boxplots highlight the differences in residual variability and outlier influence across the models

Figure 3 shows that values in the body fat percentage data have outliers and their effects because they are observed in the various regression models used in this research. The behavior and influence of these outliers is important in examining the strength and accuracy of predictive models particularly when dealing with real-life health statistics that usually do not follow the perfect assumptions of statistics.

The residual analysis showed that there were potential outliers when the actual values differed considerably with the predicted values. Such points are beyond the expected range of variation for the model and have an excessive influence on model parameters. The classical Beta regression model was very sensitive to these types of phenomena, as its fitted line was wildly undulated, indicating that the model assumed violated assumptions, such as the presence of non-normal noise or outliers.

The traditional Maximum Likelihood-based Beta regression was very susceptible to outliers. The prediction curves were very wiggly, as the model was very responsive to noise in the area, and it didn't represent the actual underlying structure of the data. Indeed, this is one important drawback of traditional MLE-type procedures that assume homoscedasticity and a lack of robustness to outlier contamination. As a result, the extreme values are over-fitted by the model and its generalization capabilities are decreased.

The robust Beta regression model using the Huber loss function was more resilient, but. Downweighing the influence of outliers, instead of removing them, allowed this method to be closer to the center trend. While some noise remained, it was at a much lower level of deviation, thus illustrating the value of robust methods in stabilizing the model. However, while helpful, the robust approach alone was insufficient to fully suppress the erratic behavior caused by embedded data noise.

Beta regression wavelet-augmented models had much better results with outliers. By denoising the signal with DWT prior to estimating model, these models could remove high frequency noise without affecting the structural features of original signal. As a result, the amounts, and the intensities of observed outliers diminished in great numbers and the prediction lines became more regular and smoother.

This thus implies that much that appeared to be behavior not towards the mean at first was driven by short term noise and not by underlying distribution outliers. Wavelet filtering allowed these distortions to be eliminated and provided more understandable and sensible forecasts.

Wavelet denoising is, statistically, a frequency filter which eliminates the high-frequency components that are common to random noise but preserves the low-frequency components that represent useful trends. This breaks down enables the model to pay attention to the underlying signal of the data that diminishes the influence of the spurious variations. This is not, in fact, merely the removal of the outliers; it puts them back in context by blurring their effect to bring it much more in line with the overall structural trend of the data set.

The comparative analysis concludes that Classical models are most prone to outlier distortion and this results in low predictive accuracy. Model resilience to this effect is partly achieved through the decrease in leverage of the extreme points. Compared to all other robust models, wavelet-enforced robust models provide the most robust solution, which removes noise in the data but also absorbs the impact of outliers without compromising the quality of the data. These observations support the effectiveness of integrating powerful statistical procedures and signal processing strategies to generate more powerful, interpretive, and general models in health data analysis.

Table 5. Performance of Beta Regression Models on the Last 100 Observations of the Body Composition Dataset

Method **Criteria**	**Classical**	**Robust**	**Db4**	**Coif4**	**Sym4**	**Dmey**
RMSE	0.0417	0.0788	0.0370	0.0365	0.0362	0.0364
MAE	0.0345	0.0653	0.0310	0.0307	0.0304	0.0306
R^2	35.88%	29.42%	49.37%	50.76%	51.7%	51%

The R^2 values in the real-data application are notably lower than in the simulation experiments. This is primarily due to the higher noise levels and weaker underlying structural relationships present in the Body Composition dataset, which naturally limit the explanatory power of all fitted models.

The following table gives the Root Mean square error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) of the six Beta regression models of the last 100 observations of body fat percentage of the Body Composition 2023 dataset. Models that are compared are the Classical and the Robust estimators and four wavelet models that are based on Db4, Coif4, Sym4 and Dmey wavelet families. The findings show that the wavelet-based models have the highest performance in all the performance measures with the highest R^2 of 51.70 and lowest errors of the Sym4 model.

All analyses were performed using MATLAB . The classical and robust Beta regression models were estimated using the fminsearch function, while all DWT procedures were conducted using the Wavelet Toolbox functions for multilevel decomposition, threshold selection, and signal reconstruction.

3.3. Limitations

Although wavelet-enhanced Beta regression models demonstrated great improvements, several limitations should be acknowledged. The performance is sensitive to the choice of wavelet family and thresholding method, and over-smoothing may occur, potentially removing meaningful high-frequency structure. Wavelet preprocessing also increases computational cost, especially in large-scale simulations. Moreover, in low-noise or very small datasets, wavelet denoising may not provide substantial benefits. These considerations should guide practitioners in selecting appropriate modeling strategies.

4. Conclusion

This paper introduced a new regression-based outlier detection technique that incorporates discrete wavelet transform (DWT) methods—specifically the Daubechies, Coiflets, Symlets, and Dmey wavelet families—into both robust and classical estimation frameworks. The primary objective was to enhance the quality and stability of multivariate linear regression models, particularly in situations where outliers and noise may obscure underlying patterns.

From the experimental analysis, the following main conclusions were drawn:

Wavelet-based filtering improves accuracy.

Integrating Dmey and Db4 wavelet transforms into the regression models led to higher predictive accuracy. Across all evaluation metrics (RMSE, MAE, and R^2), the wavelet-based models outperformed both the classical and robust regression approaches. For instance, the best-performing model, using Dmey, achieved the lowest RMSE and MAE values (0.0364 and 0.0311, respectively) and the highest R^2 value (0.8777), indicating that it provides more reliable predictions of the dependent variable.

Classical vs. robust methods.

While robust regression offered some improvement over the classical model by mitigating the effect of outliers, it still fell short of the performance of wavelet-based models. The findings suggest that robust techniques can dampen the influence of extreme observations on the fit but may not be sufficiently effective in eliminating high-frequency noise or other irregular structures in the data, which wavelet transforms are particularly adept at capturing.

Superior performance of Dmey and Sym4.

Within the examined wavelet families, Dmey and Sym4 produced the strongest results. These wavelets achieved lower prediction errors and better goodness-of-fit metrics compared with Db4 and Coif4. This emphasizes the importance of carefully selecting wavelet functions that align with the characteristics of the underlying data signal.

Outlier detection capability.

As illustrated in Figure 3, outliers were more clearly identified and separated in the wavelet-based models than in the classical approaches, based on visual inspection. This highlights an additional advantage of incorporating DWT into regression analysis: beyond improving model fit, it enhances the accurate detection of anomalous observations that could compromise sound decision-making.

Implications for applied modeling.

These findings are especially relevant in domains where data quality may be degraded by noise or hidden outliers, such as medicine, finance, and engineering. The proposed hybrid models provide a practical and powerful means of improving model reliability, reducing prediction error, and detecting influential observations that might otherwise mislead standard regression analyses.

In conclusion, the study demonstrates the value of using discrete wavelet transform as a preprocessing step in regression modeling. Combining wavelet methods with classical estimation techniques yields models that are both more accurate and more robust, making this approach a promising avenue for contemporary statistical modeling and outlier detection.

Wavelet preprocessing introduces additional computational cost compared with classical and robust estimation. However, the improved predictive accuracy and noise reduction observed across all simulation scenarios demonstrate that this cost is acceptable in practical applications.

REFERENCES

1. 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.
2. T. H. Ali, D. Saleh, Q. Mustafa Abdulqader, and A. Omer Ahmed, *Comparing Methods for Estimating Gamma Distribution Parameters with Outliers Observation*, Journal of Economics and Administrative Sciences, vol. 31, no. 145, pp. 163–174, 2025.
3. A. Antoniadis, J. Bigot, and T. Sapatinas, *Wavelet estimators in nonparametric regression: A comparative simulation study*, Journal of Statistical Software, vol. 6, no. 6, pp. 1–83, 2001.
4. D. Botani, N. Kareem, T. Ali, and B. Sedeeq, *Optimizing bandwidth parameter estimation for non-parametric regression using fixed-form threshold with Dmey and Coiflet wavelets*, Hacettepe Journal of Mathematics and Statistics, vol. 54, no. 3, pp. 1094–1106, 2025.

5. J. Da Silva, *Body Composition Dataset*, Kaggle, 2023.
6. I. Daubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, 1992.
7. I. I. Elias and T. H. Ali, *Optimal level and order of the Coiflets wavelet in the VAR time series denoise analysis*, Frontiers in Applied Mathematics and Statistics, vol. 11, p. 1526540, 2025.
8. I. I. Elias and T. H. Ali, *VARMA Time Series Model Analysis Using Discrete Wavelet Transformation Coefficients for Coiflets Wavelet*, Passer Journal of Basic and Applied Sciences, vol. 7, no. 2, pp. 657–677, 2025.
9. S. L. P. Ferrari and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, Journal of Applied Statistics, vol. 31, no. 7, pp. 799–815, 2004.
10. R. Gencay, F. Selçuk, and B. Whitcher, *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*, Academic Press, 2001.
11. P. J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.
12. Y. Liu and H. Zhang, *A robust beta regression model for modeling proportions*, Statistical Papers, vol. 60, no. 1, pp. 23–42, 2019.
13. G. P. Nason and B. W. Silverman, *The stationary wavelet transform and some statistical applications*, In Wavelets and Statistics, pp. 281–299, Springer, 1995.
14. A. W. Omer and T. H. Ali, *Wavelet Analysis for Outlier Estimation in Multivariate Linear Regression Models*, Passer Journal of Basic and Applied Sciences, vol. 7, no. 1, pp. 478–494, 2025.
15. M. M. Taher and S. M. Ridha, *The suggested threshold to reduce data noise for A factorial experiment*, International Journal of Nonlinear Analysis and Applications, vol. 13, no. 1, pp. 3861–3872, 2022.
16. M. M. Taher and S. M. Ridha, *Use The Coiflets and Daubechies Wavelet Transform To Reduce Data Noise For a Simple Experiment*, Iraqi Journal of Statistical Sciences, vol. 19, no. 2, pp. 91–103, 2022.
17. M. T. Hasan, T. H. Ali, and N. H. Sedeek Kareem, *Multivariate CUSUM Daubechies Discrete Wavelet Transformation Coefficients Charts for Quality Control*, Passer Journal of Basic and Applied Sciences, vol. 7, no. 1, pp. 533–546, 2025.
18. H. Wang, G. Li, and G. Jiang, *Robust regression shrinkage and consistent variable selection through the LAD-Lasso*, Journal of Business & Economic Statistics, vol. 25, no. 3, pp. 347–355, 2013.
19. A. Alkhateeb, Z. Algamal, *Variable selection in gamma regression model using chaotic firefly algorithm with application in chemometrics*, Chemometrics and Intelligent Laboratory Systems, vol. 231, p. 104693, 2022.
20. A. Alkhateeb, Z. Algamal, *Variable Selection in Weibull Accelerated Survival Model Based on Chaotic Sand Cat Swarm Algorithm*, Journal of Computational Science, vol. 66, p. 101927, 2023.
21. A. N. Alkhateeb, *Jackknifed Liu-type Estimator in Poisson Regression Model*, Journal of Statistical Computation and Simulation, vol. 92, no. 16, pp. 3373-3392, 2022.
22. Taher, Mahmood M and Al, Talal Abd Al-Razzaq Saead and Ali, Taha Hussein and others, *A novel Dmey wavelet charts for controlling and monitoring the average and variance of quality characteristics*, Statistics, Optimization & Information Computing,vol. 14, no. 6, pp. 3706-3717, 2025.
23. Al-Talib, Bashar A Majeed and Hammodat, AA, *Using Some Wavelet Shrinkage Techniques and Robust Methods to Estimate the Generalized Additive Model Parameters in Non-Linear Models*, Int. J. Adv. Sci. Eng. Inf. Technol, vol. 10, no. 6, pp. 2344, 2020.