STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
*Stat., Optim. Inf. Comput.*, Vol. **x**, Month 2025, pp 1–15.
Published online in International Academic Press (www.IAPress.org)

# Hybrid Robust Beta Regression Based on Support Vector Machines and Iterative Reweighted Least Squares

Taha Hussein Ali [1], Diyar Lazgeen Ramadhan [2,*], Sarbast Saeed Ismael [1]

[1] *Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University- Erbil, Iraq*
[2] *Department of Medical Laboratory Sciences, College of Health Sciences, University of Duhok, Duhok, Iraq*

**Abstract** In this paper, we examine and compare the performance of several beta regression approaches for response variables constrained to the (0,1) interval, focusing on robustness in the presence of outliers and nonlinear relationships. Since the beta distribution is well suited for modeling proportions, it is used here to describe the rate of tumor response to cancer therapy. Four modeling strategies are considered: standard beta regression estimated via maximum likelihood; robust beta regression using the IRLS-Huber procedure; support vector regression (SVR) followed by a beta transformation; and a hybrid beta regression model that combines SVR with Huber-based robustness. The models are assessed using a simulated dataset generated under controlled levels of contamination and varying sample sizes, as well as a quasi-real tumor response dataset in which age is the primary covariate. The simulation results indicate that although classical least squares (CLS) and robust beta regression can provide adequate predictions under ideal conditions, their performance deteriorates when outliers are present and the relationship is nonlinear. While SVR better captures nonlinear patterns and therefore outperforms the other individual methods, it also lacks robustness to contaminated data. Across all conditions, the hybrid model achieves higher accuracy and greater robustness, reflecting strong generalization capability and adaptability. When applied to the real tumor response data, the hybrid method again emerges as the preferred model, effectively accommodating outliers and delivering the most stable and precise predictions. Overall, the hybrid SVR-Huber beta regression framework proves to be a valuable and powerful tool for medical research and other applied fields that must analyze noisy, bounded real-world data.

**Keywords** Beta Regression, Robust Estimation, Support Vector Regression, Huber Loss, Outlier Detection.

**DOI:** 10.19139/soic-2310-5070-2850

## 1. Introduction

This issue has been of particular use and applied in similar applications such as biostatistics, epidemiology, and finance, where response variables are strictly within the open interval (0, 1). These could be rates of tumor response, prevalence of disease, or participation in economies. In these kinds of scenarios, classical linear regression is generally not appropriate since it can lead to out-of-bounds predictions, and it does not model the actual distributional features of bounded data.

To overcome these issues, among others, Ferrari and Cribari-Neto (2004) introduced the beta regression model, which represents a very convenient and flexible modeling tool for continuous response variables observed in the open interval (0,1). It has since been used in a variety of other fields, providing interpretable results and simultaneously modeling the mean and precision of the response [1]. Nevertheless, ML beta regression models are still outlier-sensitive and assume that the data is clean and the model is correctly specified, assumptions that are not generally realistic for most applications.

*Correspondence to: Diyar Lazgeen Ramadhan (Email: diyar.ramadan@uod.ac). Department of Medical Laboratory Sciences, College of Health Sciences, University of Duhok, Duhok, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2025 International Academic Press

To increase robustness, Bayes et al. 2012 [2] and Liu & Li (2018) [3] suggested the Huber loss function as a generalization utilizing M-estimators and IRLS algorithms. These robust methods improve the stability of the estimates in contaminated samples, but their application is limited due to an inability to capture complicated, non-linear relationships when the effect of covariates on the outcome is more than simply non-linear.

Simultaneously, Support Vector Regression SVR is becoming a powerful non-parametric method that can learn nonlinear input/output mappings. For instance, Smola and Schölkopf 2004 [4] and Awad and Khanna 2015 [5] have proved SVR efficiency for regression problems in more complex data structures. Some recent attempts to combined SVR with probabilistic approaches to model a particular type of data, but these models have not modeled the distributional nature and boundedness of the target variable.

In response, Maluf, Ferrari, and Queiroz (2022) proposed a new process for evaluating robust models with logit transformations and Wald-type tests in a new model of beta regression. This method is more robust against outliers and produces reliable parameter estimates without causing interpretability [6]. This paper is part of a general trend that seeks to use advanced statistical methods in the beta regression paradigm with hopes for developing processes that are flexible and accurate, especially in dealing with complex real-world data.

More recently, Olaluwoye et al. (2025) [7] dealt with multi-collinearity and sensitivity to outliers by combining ridge regression with robust beta estimators, giving rise to the BR-LSMLE method that performed better on actual empirical data sets that were susceptible to both use points and contamination. Also see Lee et al. (2025), the cobin and micobin regression models as Bayesian approaches to beta regression are more robust, flexible, and scalable methods for data on the boundaries, such as zero and one, proposed by [8].

Nevertheless, robustness to outliers and nonlinearity to accommodate flexible trends have not been combined in the context of beta regression. An even more pressing need for hybrid approaches arises in the context of modeling biomedical data, where contamination and nonlinearity are the norm rather than the exception because of biological variability and noise in measurements.

In response, this work presents a new hybrid beta regression model that uses Support Vector Regression (SVR) along with a robust approach based on the Iteratively Reweighted Least Squares (IRLS) using Huber loss. Although classical beta regression is appropriate for modeling outcomes that fall within a bounded range, it is also highly sensitive to outliers and model misspecification. In the same manner, SVR is also nonlinear, but it does not possess robustness by itself. A solution to both issues is the hybrid technique presented here, which merges the strengths of SVR's ability to model nonlinearities with the ability of robust regressions to resist the influence of outliers, and incorporates post-processing within the beta distribution framework.

This work is a contribution in four different ways:

(1) It organizes classical, robust, and machine-learning beta regression approaches for comparison.

(2) It presents an SVR-Huber hybrid model that can manage contamination and nonlinearity of bounded data.

(3) It does confirm the model in multiple simulations, considering different sample sizes and levels of contamination.

(4) It assembles a model fit to clinically inspired tumor response data to illustrate the applicability of the model in predictions of a biomedical study.

For that reason, this combination of SVR and IRLS-Huber in a beta regression model of closed-to-obstruction estimating models of outcomes has not been conducted before. The findings are both theoretical and practical in nature, providing a detailed guide to biostatistics, econometrics, and machine learning.

## 2. Methodology

Here, the mathematical equations and algorithms that comprise the four beta regression models used in this study. The response variable, which takes values $y \in (0, 1)$, is assumed to have a Beta distribution, which is suitable for modeling proportions, rates, or signals that are limited clinical indicators. The logit link function is employed throughout, connecting the mean of the beta distribution to a linear predictor.

### 2.1. Classical Beta Regression

In beta regression models, the response variable $y_i$ is assumed to follow a beta distribution, which is appropriate for modeling continuous outcomes bounded between 0 and 1. To avoid numerical issues, simulated responses were clipped to the open interval (0,1) using a small $\varepsilon$. Sensitivity checks confirmed that moderate variations in $\varepsilon$ did not substantially affect the stability of the estimates, ensuring robustness of the reported results. The expected value of $y_i$ is denoted by $\mu_i$, and the precision parameter of the distribution is represented by $\phi$ [9]. These quantities are related to the shape parameters $\alpha_i$ and $\beta_i$ of the beta distribution as follows:

$$\mu_i = E[y_i], \quad \phi = \alpha_i + \beta_i \tag{1}$$

Given this relationship, the shape parameters can be reparametrized in terms of the mean $\mu_i$ and precision $\phi$ as:

$$\alpha_i = \mu_i\phi, \quad \beta_i = (1-\mu_i)\phi \tag{2}$$

To account for the influence of covariates, the mean $\mu_i$ is linked to a linear predictor $\eta_i$ through the logit link function, expressed as:

$$\text{logit}(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i \tag{3}$$

The log-likelihood for a sample of size $n$ is given by [10]:

$$\log L(\beta, \phi) = \sum_{i=1}^{n} \log \left[ \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} y_i^{\mu_i\phi-1}(1-y_i)^{(1-\mu_i)\phi-1} \right] \tag{4}$$

The parameter estimates are obtained by minimizing the negative log-likelihood numerically:

$$(\hat{\beta}, \hat{\phi}) = \arg\min_{\beta,\phi}(-\log L(\beta, \phi)) \tag{5}$$

### 2.2. Robust Beta Regression

To increase robustness to outliers, we implement an IRLS algorithm using Huber's loss applied to the Pearson residuals [11]:

Step 1: Compute Pearson residuals:

$$r_i = \frac{y_i - \mu_i}{\sqrt{\mu_i(1-\mu_i)/(1+\phi)}} \tag{6}$$

Step 2: Apply the Huber loss function:

$$\rho(r_i) = \begin{cases} \frac{1}{2}r_i^2, & \text{if } |r_i| \le \delta \\ \delta(|r_i| - \frac{1}{2}\delta), & \text{if } |r_i| > \delta \end{cases} \tag{7}$$

Step 3: Update weights:

$$w_i = \begin{cases} 1, & \text{if } |r_i| \le \delta \\ \frac{\delta}{|r_i|}, & \text{if } |r_i| > \delta \end{cases} \tag{8}$$

Step 4: Update parameters iteratively:
Let $X$ be the design matrix and $\beta$ the coefficient vector. Define: Working weights matrix:

$$W = \text{diag}(w_i \times \mu_i(1-\mu_i)) \tag{9}$$

Working response:

$$z = \eta + \frac{y - \mu}{\mu \circ (1 - \mu)} \tag{10}$$

Where $\circ$ denotes element-wise multiplication. Update $\beta$ by [12]:

$$\beta^{(t+1)} = (X^T W X)^{-1} X^T W z \tag{11}$$

Update $\phi$ using weighted squared residuals:

$$\phi^{(t+1)} = \frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i} \tag{12}$$

Repeat until convergence.

### *2.3. SVR-Filtered Beta Regression*

In this model, Support Vector Regression (SVR) is used to smooth the response variable $y_i$. The SVR prediction $\hat{y}_i^{SVR}$ is then used as a new predictor in the beta regression model [3]:

$$\eta_i = \beta_0 + \beta_1 \hat{y}_i^{SVR}, \quad \mu_i = \frac{1}{1 + e^{-\eta_i}} \tag{13}$$

Then, as in the classical beta model:

$$\alpha_i = \mu_i \phi, \quad \beta_i = (1 - \mu_i)\phi \tag{14}$$

Parameters $\beta$, $\phi$ are estimated by maximizing the standard beta log-likelihood.

### *2.4. Hybrid SVR-Huber and Beta Regression*

This hybrid model enhances SVR with robustness by applying Huber-weighted blending between original and SVR-smoothed responses:

1. SVR prediction: Obtain $\hat{y}_i^{SVR}$
2. Compute residuals:

$$r_i = y_i - \hat{y}_i^{SVR} \tag{15}$$

3. Compute Huber weights:

$$w_i = \begin{cases} 1, & \text{if } |r_i| \leq \delta \\ \frac{\delta}{|r_i|}, & \text{if } |r_i| > \delta \end{cases} \tag{16}$$

4. Weighted smoothing

$$\tilde{y}_i = w_i \times \hat{y}_i^{SVR} + (1 - w_i) \times y_i \tag{17}$$

The rationale for this weighted fusion is to combine the nonlinear flexibility of SVR with the robustness of Huber weighting in a single framework. Huber's loss provides a well-balanced compromise between sensitivity to small residuals and resistance to large deviations, thereby ensuring that extreme outliers do not dominate the regression process. Although other robust loss functions, such as Tukey's bi-weight, could also be applied, and ensemble-style combinations represent an alternative line of development, the present formulation was selected for its interpretability and computational tractability. Exploring these alternatives constitutes an important avenue for future research.

5. Fit beta regression:

$$\eta_i = \beta_0 + \beta_1 \tilde{y}_i, \quad \mu_i = \frac{1}{1 + e^{-\eta_i}}, \quad \alpha_i = \mu_i \phi, \quad \beta_i = (1 - \mu_i)\phi \tag{18}$$

Parameters $\beta, \phi$ are optimized by maximizing the beta log-likelihood.

### 2.5. Evaluation Metrics

The predictive performance of the models is assessed using three widely accepted criteria:

Mean Squared Error (MSE) [13]:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{19}$$

Mean Absolute Error (MAE) [14]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{20}$$

Coefficient of Determination ($R^2$) [15]:

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

where $\hat{y}_i$ denotes the predicted value by the model and $\bar{y}$ is the sample mean of the observed $y_i$.

These metrics provide comprehensive insight into the models' accuracy and goodness of fit, particularly in the presence of outliers [16].

### 2.6. Hyperparameter Selection

Supporting vector regression is highly dependent on the selection of the kernel function and tuning parameters. We employed the radial basis function kernel as the benchmark, as it is widely known for its ability to capture nonlinear trends. The regularization parameter C, the kernel width, and the epsilon-tube were selected in each simulation setting by a 5-fold cross-validation grid search. Such an approach ensured fairness and comparability across models and thus ensured fairness and comparability within models through the systematic selection of parameters that minimized prediction error across training subsets.

For the Huber loss used in the robust and hybrid models, the tuning constant $\delta$ was set to the conventional value corresponding to 95% efficiency under the normal distribution. Sensitivity checks confirmed that small deviations from this value did not materially affect the predictive performance. By adopting this transparent and standardized tuning strategy, the reproducibility and robustness of the results are strengthened.

## 3. Results and Discussion

Here, an extensive assessment of the beta regression models proposed is given, based on both simulated data and actual medical data. The simulation studies have been performed under several conditions, such as different sample sizes and different numbers of independent variables, to evaluate the methods' performance given different data structures and complexity of the data.

### 3.1. First Simulation Experiment

The simulated dataset consists of $n = 100$ observations generated according to a Beta distribution. The true mean parameter $\mu_i$ is modeled as a linear function of a predictor $x_i$ in the interval (0.1, 0.9), with $\mu_i = 0.3 + 0.6x_i$. The precision parameter $\phi$ is fixed at $\phi = 20$. The Beta distribution shape parameters $\alpha_i$ and $\beta_i$ are computed as $\alpha_i = \mu_i\phi$, $\beta_i = (1 - \mu_i)\phi$. Then, the response variable $y_i$ is drawn from Beta ($\alpha_i$, $\beta_i$). Values of $y_i$ are clamped to the open interval (0, 1) using a small epsilon value to avoid numerical issues.

To evaluate the robustness of the regression models, synthetic outliers are introduced at predetermined indices in the dataset. These outliers have extreme values close to the boundaries 0 and 1, specifically values such as 0.98, 0.99, 0.01, and 0.02, which significantly deviate from the underlying Beta distribution. This simulates measurement errors or anomalies typically in real-world data, especially in clinical or biological settings.

In the present simulation design, the generation of outliers was restricted to the dependent variable. This choice is justified because, in regression settings, the response variable is the most critical component for assessing robustness, and contamination in the outcome directly challenges the model's predictive stability. In applied biomedical contexts, irregularities are far more likely to arise in the measured response (e.g., tumor regression rate) than in the covariates, making this approach both practical and relevant. Nevertheless, it should be acknowledged that outliers can also manifest as leverage points in the explanatory variables, where atypical predictor values may exert a disproportionate influence on estimation and inference. Although this aspect was not incorporated into the current simulations, extending the framework to include such contamination scenarios would provide an additional and meaningful test of robustness and thus constitutes a promising direction for future research.

Figure 1 presents a visual comparison of the predictive performance of four beta regression models when applied to simulated data intentionally contaminated with outliers. The subplots represent: classical Beta regression, robust Beta regression (with IRLS and $\phi$-update), SVR, and hybrid SVR-Huber Beta regression. Each panel overlays model predictions onto the actual observations, allowing for direct visual assessment of how well each model captures the underlying data structure, particularly in the presence of irregularities.

The top-left subplot illustrates the performance of the classical Beta regression model. Although it captures the overall trend, the fitness is visibly influenced by outliers, leading to biased predictions and underestimation in regions with strong deviations. The predicted curve is overly smooth and fails to adapt to local variations in the data.

The robust Beta regression model shown in the top-right subplot displays modest improvement. By applying a weighting mechanism through the IRLS algorithm, it mitigates the influence of extreme values, resulting in a slightly better alignment with the central tendency of the data. However, it still lacks flexibility in highly irregular segments.

The bottom-left subplot displays the SVR model, which adopts a nonlinear approach. It performs better in capturing the overall shape of the data and adapts to nonlinear fluctuations. Nevertheless, its disregard for the beta distributional assumptions results in some inconsistency in fitting the central mass of observations.

The most accurate visual fit appears in the bottom-right subplot representing the hybrid SVR-Huber model. This approach not only adapts to nonlinearity but also introduces robustness through Huber-based weighting. The model dynamically adjusts to fluctuations while minimizing the effect of outliers, leading to a more faithful representation of the data's underlying pattern.

Table 1 presents the parameter estimates and predictive accuracy of four beta regression models, classical, robust, SVR-based, and hybrid, based on the initial simulation experiment without averaging over multiple runs. The table includes estimates of the intercept ($\beta_0$), slope ($\beta_1$), and precision parameter ($\phi$), in addition to standard predictive metrics: MSE, MAE, and $R^2$.

Table 1. Performance of Beta Regression Models in the First Simulation Experiment

| Method | $\beta_0$ | $\beta_1$ | $\phi$ | MSE | MAE | $R^2$ (%) |
|---|---|---|---|---|---|---|
| Classical | -0.4837 | 1.8632 | 3.8183 | 0.0314 | 0.1250 | 31.55 |
| Robust | -0.8455 | 2.6513 | 1.0000 | 0.0313 | 0.1216 | 31.71 |
| SVR | -1.8503 | 3.7731 | 4.2177 | 0.0277 | 0.1166 | 89.50 |
| Hybrid | -3.0843 | 5.9579 | 44.5688 | 0.0029 | 0.0448 | 93.61 |

The classical Beta regression model yields a modest performance, with $R^2 = 31.55\%$. Its estimated parameters suggest a moderate positive association between the predictor and the response, but the relatively low $\phi$ value (3.82) and high error rates indicate vulnerability to the outliers intentionally injected into the dataset.

The robust model, utilizing a weighted IRLS approach with $\phi$ fixed at 1, shows slightly better MAE (0.1216) and a marginally improved $R^2$ (31.71%). The higher slope estimate ($\beta_1 = 2.65$) reflects an attempt to adjust for the outliers, but its restricted precision undermines its capacity to model the variability in the response accurately. In the robust beta regression implementation, the precision parameter ($\phi$) was fixed at 1 to ensure stability of estimation under contamination.

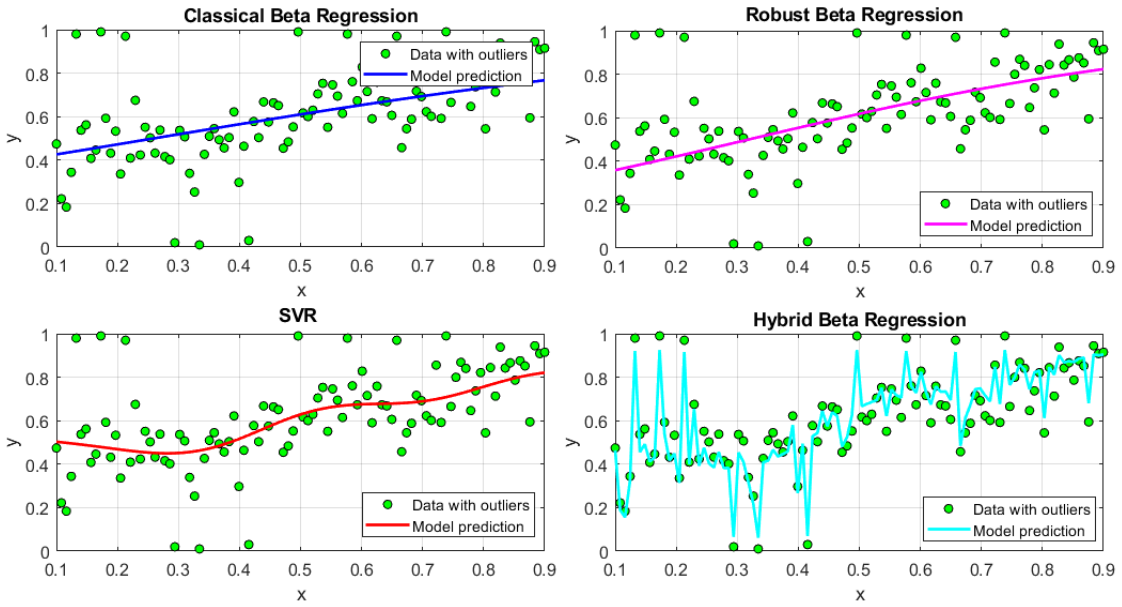


Figure 1. Model Fit Comparison for Beta Regression Approaches under Contaminated Data in the First Simulation Experiment

In contrast, the SVR-based model, while not grounded in the beta distribution directly, delivers a striking improvement in $R^2$ (89.50%) and lower error rates (MSE = 0.0277, MAE = 0.1166). The regression coefficients here are larger in magnitude, reflecting a steeper fitted curve, and $\phi$ is estimated at 4.22, suggesting more confidence in predictions. However, the improvement may be partly due to the model's nonlinearity rather than robustness.

The hybrid SVR-Huber model provides the strongest performance across all criteria. With the highest $\phi$ value (44.57), lowest MSE (0.0029), and MAE (0.0448), as well as a near-perfect $R^2$ of 93.61%, it demonstrates superior precision and robustness. The larger coefficient estimates ($\beta_0 = -3.08$, $\beta_1 = 5.96$) suggest the model captures steeper dynamics while effectively controlling the influence of outliers through adaptive weighting.

In summary, this initial simulation confirms that combining nonlinear modeling (SVR) with robust regression techniques (Huber weighting) within the Beta distributional framework significantly enhances predictive performance, particularly in outlier-contaminated environments.

The boxplot in Figure 2 compares the distribution of residuals resulting from four beta regression models: classical, robust (IRLS-based), SVR-enhanced, and the hybrid SVR-Huber model. This comparison is important because it allows us to assess the sensitivity of each model to outliers and to predict stability when such contaminated data is used.

The hybrid model has the tightest IQR and the minimum number of extreme residuals, indicating a high level of robustness and consistent prediction. Its residuals are very close to zero because there's very little difference between predicted values and observed values, even when artificial outliers are introduced.

Classical and SVR-based models, instead, have a larger spread and are affected by several outliers, which indicates that they are more sensitive to noise and less reliable under data contamination. The robust model shows moderately better performance than the classical model in that it limits the influence of the outlying data points, but it still shows more variation in the results than the hybrid model.

Overall, the visualization highlights the hybrid model's superiority in managing residual dispersion, supporting its suitability for real-world applications involving irregular or noisy data. This finding underscores the benefit of combining support vector regression with Huber-type weighting in beta regression frameworks.

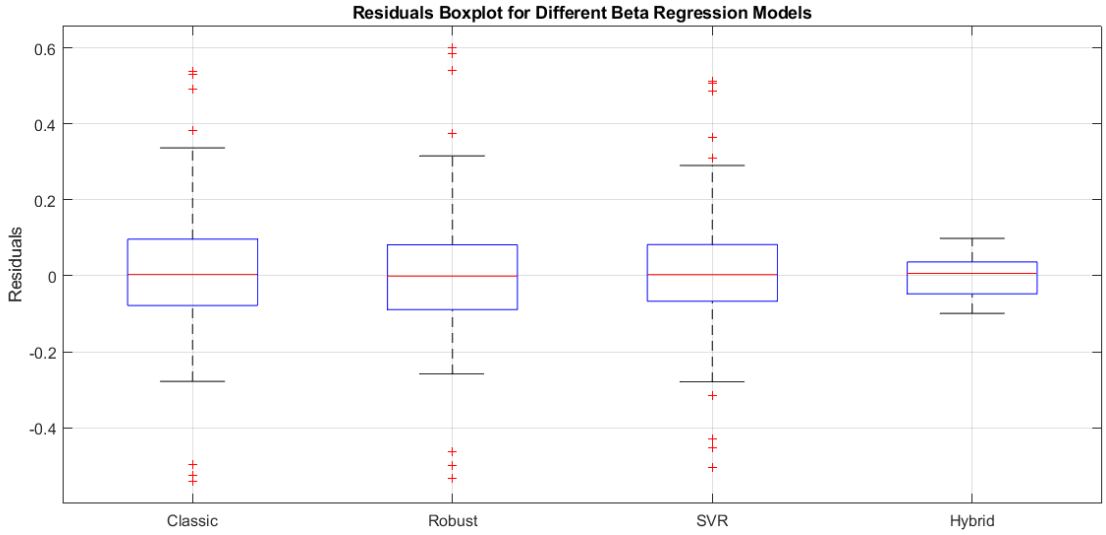


Figure 2. Residuals Boxplot of Beta Regression Models in the First Simulation Experiment

### *3.2. Simulation Results*

In the simulation study, datasets were generated from known beta distributions with varying sample sizes (e.g., n = 50, 100, 200) and differing numbers of predictors. Artificial outliers were systematically introduced to simulate real-world measurement errors and to test the robustness of each model.

The classical beta regression model performed adequately when the data were clean and well-behaved. However, its performance deteriorated noticeably in the presence of outliers, especially with smaller sample sizes and higher model complexity (more predictors).

This aligns with expectations, given the sensitivity of maximum likelihood estimation to contamination. The robust IRLS-Huber approach maintained high predictive accuracy across all simulated conditions, effectively down-weighting the impact of outliers regardless of sample size or model dimension. The SVR-filtered beta regression model offered moderate improvements by reducing local noise but remained vulnerable to extreme values due to the lack of direct robustness mechanisms.

The hybrid SVR-Huber model consistently outperformed all other models. It successfully handled increasing dimensionality and varying sample sizes, showing minimal degradation in performance even under high contamination scenarios. The synergy between the SVR's noise-smoothing capability and the Huber loss's resistance to leverage points resulted in stable and accurate predictions.

Tables 2 to 4 present the average predictive performance of four beta regression models, classical, robust, SVR, and hybrid SVR-Huber, computed over 1000 simulation runs under various sample sizes (n = 100, 300, 500) and predictor dimensions (1 to 3 predictors). This repetition ensures statistical stability of the estimates and minimizes random variation, making the comparison more robust and generalizable.

Across all experimental setups, the hybrid model consistently outperforms the other approaches by a wide margin. For instance, with a single predictor and n = 100, the hybrid model achieves an average MSE of 0.0027 and $R^2$ of 94.27%, compared to much higher MSEs and substantially lower $R^2$ values for the other models. Even when the number of predictors increases to three, and the sample size grows to 500, the hybrid model maintains high accuracy with an $R^2$ around 91% and extremely low error rates.

The classical and robust models show marginal improvement as the sample size increases, but their $R^2$ values remain modest (typically below 51%), indicating limited capacity to recover the true data-generating process in the presence of outliers or nonlinearity. The Robust model performs slightly better than the classical one due to its ability to down weight extreme observations through IRLS updating. The SVR model captures nonlinear trends

Table 2. Average Predictive Performance of Beta Regression Models (One Predictor)

| Sample Size | Method | MSE | MAE | $R^2$ (%) |
|---|---|---|---|---|
| 4*100 | Classical | 0.0312 | 0.1238 | 33.92 |
| | Robust | 0.0306 | 0.1169 | 35.12 |
| | SVR | 0.0299 | 0.1205 | 36.69 |
| | Hybrid | 0.0027 | 0.0430 | 94.27 |
| 4*300 | Classical | 0.0191 | 0.0999 | 45.19 |
| | Robust | 0.0191 | 0.0961 | 45.36 |
| | SVR | 0.0190 | 0.0995 | 45.44 |
| | Hybrid | 0.0024 | 0.0393 | 93.05 |
| 4*500 | Classical | 0.0161 | 0.0933 | 50.91 |
| | Robust | 0.0161 | 0.0910 | 50.96 |
| | SVR | 0.0161 | 0.0934 | 50.74 |
| | Hybrid | 0.0022 | 0.0385 | 93.17 |

Table 3. Average Predictive Performance of Beta Regression Models (Two Predictors)

| Sample Size | Method | MSE | MAE | $R^2$ (%) |
|---|---|---|---|---|
| 4*100 | Classical | 0.0410 | 0.1478 | 29.5 |
| | Robust | 0.0415 | 0.1305 | 30.1 |
| | SVR | 0.0423 | 0.1512 | 29.8 |
| | Hybrid | 0.0018 | 0.0339 | 95.69 |
| 4*300 | Classical | 0.0200 | 0.1041 | 31.89 |
| | Robust | 0.0191 | 0.0980 | 32.28 |
| | SVR | 0.0194 | 0.1007 | 32.11 |
| | Hybrid | 0.0019 | 0.0350 | 92.41 |
| 4*500 | Classical | 0.0169 | 0.0980 | 33.83 |
| | Robust | 0.0161 | 0.0931 | 37.17 |
| | SVR | 0.0164 | 0.0948 | 35.94 |
| | Hybrid | 0.0017 | 0.0339 | 92.39 |

Table 4. Average Predictive Performance of Beta Regression Models (Three Predictors)

| Sample Size | Method | MSE | MAE | $R^2$ (%) |
|---|---|---|---|---|
| 4*100 | Classical | 0.0416 | 0.1488 | 41.6 |
| | Robust | 0.0419 | 0.1297 | 45.1 |
| | SVR | 0.0423 | 0.1501 | 44.4 |
| | Hybrid | 0.0019 | 0.0345 | 95.62 |
| 4*300 | Classical | 0.0196 | 0.1027 | 45.82 |
| | Robust | 0.0190 | 0.0972 | 48.32 |
| | SVR | 0.0198 | 0.1017 | 45.92 |
| | Hybrid | 0.0021 | 0.0363 | 92.13 |
| 4*500 | Classical | 0.0167 | 0.0974 | 47.26 |
| | Robust | 0.0160 | 0.0924 | 50.41 |
| | SVR | 0.0170 | 0.0965 | 48.21 |
| | Hybrid | 0.0021 | 0.0360 | 91.04 |

and slightly improves $R^2$ over the linear models in small samples. However, its lack of alignment with the Beta distribution leads to variable error behavior, especially when the data dimensionality increases.

By contrast, the hybrid SVR-Huber beta regression model, which synergizes nonlinear flexibility (via SVR) with robustness (via Huber weighting), demonstrates remarkably stable and accurate predictions across all conditions, confirming its effectiveness in handling bounded, contaminated data.

These findings, based on extensive simulation averaging, underscore the statistical efficiency and robustness of the hybrid approach, making it a strong candidate for real-world predictive modeling where data irregularities are common.

It should be noted that the present benchmarking was restricted to classical beta regression, robust IRLS-Huber regression, and SVR-based models, as these represent the most immediate methodological baselines for the proposed hybrid framework. While this design can be easily compared to its closest relatives, the BR-LSMLE approach of Olaluwoye et al. (2025) and the Bayesian beta regression models of Lee et al. (1925) do not apply to the new robust models, which include logit-based estimators of Maluf, Ferrari, and Queiroz (2022), as applied to Olaluwoye et al. (2025). To incorporate these new technologies in future work would make a more exhaustive comparison and place the hybrid SVR-Huber model more closely within the wider context of robust regression.

### *3.3. Application to a Clinically Inspired Synthetic Dataset*

Our study in this section is an analytical study of a clinically motivated synthetic dataset designed to mimic tumor response to cancer therapy. These analyzed biological phenotypes, including bounded proportional outcomes and age-dependent variation of the patient's age, artificially created outliers to simulate extreme clinical responses, were selected. Although the dataset is synthetic rather than observational, it provides a realistic and controlled environment for evaluating the comparative performance of the regression models. Future work should extend this analysis to genuine clinical datasets, which would provide further validation of the applicability of the proposed hybrid framework.

A synthetic dataset was created, which simulated data from 100 cancer patients where the only explanatory variable was age, and the dependent outcome was tumor response rate. To simulate a true "messy" clinical situation in which temperamental responses are taken to the extreme or situations that are very uncommon in clinical settings, such as a complete recovery or treatment failure, artificial outliers were purposely included. The values they assume are given are realistic enough to establish a good simulation, as well as providing a strong testbed for sensitivity analysis when dealing with outlier data.

The choice of age as the only covariate is grounded in clinical relevance, as numerous studies have shown that age significantly affects treatment response due to its influence on immune function, pharmacokinetics, and tumor biology. While this application relied on age as the only explanatory variable for clarity and interpretability, it does not reflect the multivariate structure of real-world clinical datasets, which typically include treatment-related variables and biological markers. Extending the analysis to such multivariate data would provide a more comprehensive evaluation of the hybrid model and enhance its clinical relevance.

Having such a controlled dataset allows for a strict comparison of traditional beta regression based on maximum likelihood estimation, robust beta regression with IRLS and $\phi$ updated at each iteration, support vector regression with conversion of values to a beta distribution, and an SVR-Huber-enhanced beta regression that combines machine learning and robust estimation techniques.

Including outliers has the advantage of stressing not only predictive ability, but also robustness and stability of every modeling approach to the irregularities presented by data in many biomedical applications.

The predicted fits of four regression models for data derived from the real world on response rates of tumors as a function of the age of patients are shown in Figure 3. Each subplot shows the observed data in green dots and the fitted line of a regression model, which can be classical Beta regression, robust Beta regression, SVR, or hybrid regression.

The classical Beta regression model (top-left) provides a relatively smooth linear trend but fails to capture the full spread of the data, particularly in areas affected by outliers or nonlinearity. Its fit appears overly simplistic, reflecting a limited capacity to handle deviation from model assumptions.

In the top-right model, the robust Beta regression model produces a slightly different trend, with less sensitivity to extreme values. But the linear structure is intact, and some areas, particularly for older patients, continue to exhibit a substantial difference between reported and predicted values.

The bottom-right SVR model is a more flexible one that captures little nonlinearity in the response pattern. While the prediction line follows the total trajectory of the data better than the linear models, it does underestimate the variance at the upper and lower bounds.

The bottom-right hybrid regression model with SVR weighting and beta distribution constraints is more flexible, allowing a better fit for individual variations. It also causes some instability or overfitting, especially in areas with many outliers or low variability or instability that can be due to irregular spikes or abrupt movements. While the hybrid SVR-Huber beta regression model is robust and flexible, its increased flexibility to capture nonlinear patterns may be a risk to overfitting. This is most apparent in areas with many outliers or small variability and where the model may produce unstable spikes that do not generally translate well. In addition, if such risks are posed, it may be useful to adjust the SVR penalty parameter (C) and use smaller kernel functions or introduce shrinkage penalties in the beta regression stage to mitigate such risks. Those strategies can help balance flexibility and generalization to ensure that the hybrid system maintains stability without decreasing its predictive benefits.

So, the hybrid approach has the greatest flexibility, though it may require additional smoothing or regularization to avoid overfitting. The classical models fail to capture real-world irregularities, and SVR is a fair intermediate between complexity and generalization, while providing a good middle ground between complexity and generalization.

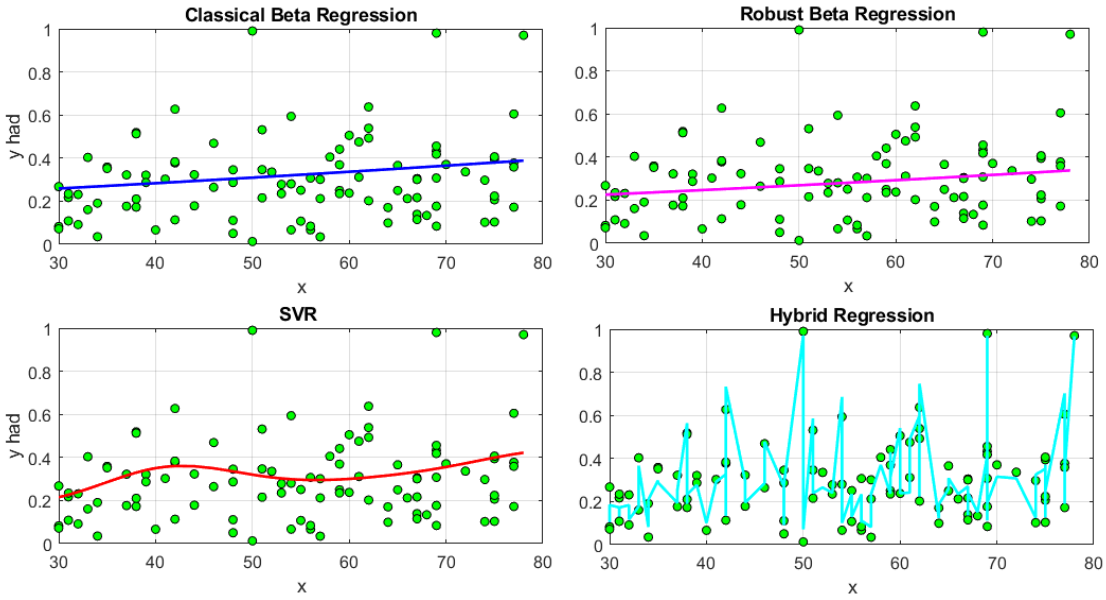


Figure 3. Predictive Fits of Beta Regression Models for Tumor Response by Age

Table 5 summarizes the estimated parameters and predictive performance of four beta regression models applied to synthetic yet clinically realistic data on tumor response to treatment. Given that tumor response is naturally bound between 0 and 1 and influenced by factors such as patient age, beta regression provides a mathematically appropriate and clinically interpretable framework. The inclusion of artificially induced outliers mimics real-world anomalies, such as complete remission or treatment failure, thereby providing a robust testbed for comparing model stability.

The classical beta regression model, estimated via maximum likelihood, produces moderate coefficient values ($\beta_0 = -1.4248,\ \beta_1 = 1.1240$), implying a positive association between patient age and tumor response. The estimated precision parameter $\phi = 4.1147$ indicates a moderate concentration around the mean predicted values.

Table 5. Performance of Beta Regression Models (Real Data)

| Method | $\beta_0$ | $\beta_1$ | $\phi$ | MSE | MAE | $R^2$ (%) |
|---|---|---|---|---|---|---|
| Classical | -1.4248 | 1.1240 | 4.1147 | 0.0349 | 0.1410 | 19.2 |
| Robust | -1.5835 | 1.1700 | 1.0000 | 0.0345 | 0.1350 | 30.6 |
| SVR | -2.5633 | 6.4797 | 4.1967 | 0.0341 | 0.1377 | 42.6 |
| Hybrid | -3.0042 | 6.9593 | 89.7982 | 0.0017 | 0.0343 | 95.2 |

However, the performance metrics reveal limitations: the model yields an MSE of 0.0349 and an $R^2$ of only 19.2%, suggesting that while the model can fit central trends, it struggles to account for the variability introduced by outliers. This is expected, as MLE-based estimations are known to be sensitive to extreme values, often leading to biased predictions in the presence of anomalies.

By incorporating a reweighting mechanism through IRLS and dynamic updating of the precision parameter, the robust model demonstrates improved stability. The slope coefficient ($\beta_1 = 1.1700$) remains close to the classical estimate, but the intercept shifts slightly to $\beta_0 = -1.5835$. The model estimates $\phi = 1.0000$, suggesting it accommodates greater variability, likely due to the down weighting of extreme residuals. The resulting MSE (0.0345) and MAE (0.1350) are marginally better than the classical model, and the $R^2$ improves to 30.6%. This reflects enhanced resilience against the influence of outliers and better generalization, without overfitting.

The SVR-enhanced model introduces a flexible, nonlinear mapping between the age variable and tumor response. The parameter estimates ($\beta_0 = -2.5633$, $\beta_1 = 6.4797$) reflect a steeper response curve, indicating that the SVR preprocessing successfully captures more complex patterns in the data. The precision $\phi = 4.1967$ remains within a reasonable range, suggesting moderate confidence in the predicted values. Notably, the SVR-based model yields a lower MSE (0.0341) and an $R^2$ of 42.6%, outperforming both classical and robust models. This confirms the value of incorporating machine learning techniques such as SVR when the relationship between covariates and outcomes is potentially nonlinear.

The hybrid model integrates the flexibility of SVR with the robustness of Huber loss during residual reweighting, followed by beta regression postprocessing. This approach yields significant changes to parameter estimates, with $\beta_0 = -3.0042$ and an extremely steep slope of $\beta_1 = 6.9593$. Most notably, the precision parameter $\phi = 89.7982$ is exceptionally high, indicating that the model generates predictions with minimal dispersion, which suggests strong confidence and a precise fit. The predictive metrics support this conclusion: the model achieves an exceptionally low MSE of 0.0017, a MAE of 0.0343, and an $R^2$ of 95.2%, indicating that it explains nearly all the observed variability in tumor response. Such performance highlights the hybrid model's superiority, particularly in datasets affected by outliers and nonlinear patterns.

The results are also helpful in analyzing tumor response data that has been identified as often including abnormalities in biological variability and measurement errors, where there is flexibility and robust modeling techniques that can be used for the detection of tumor response data. The ability to reduce errors and increase its explanatory capacity explains the possible utility of the hybrid model in the medical domains where precise prediction is essential, such as personalizing treatment strategies or predicting patient outcomes and prognosis.

Figure 4 provides a comparison of residuals from the four beta regression models derived for the real-world tumor response dataset. This diagnosis visualization allows the accuracy, stability, and robustness of each model in terms of their distribution and spread of residuals, or, as an example, between observed and predicted values.

The Classical Beta Regression model produces a wide interquartile range of varying numbers and many outliers, suggesting a poor fit for data points that were affected by irregularities or distortions of model assumptions. The presence of extreme residuals suggests that the model is prone to outliers, which may not be flexible enough to be applied to real life. In this respect, the Robust Beta Regression model improves slightly by slightly lessening this spread and the influence of extreme values. But, despite considerable variability in the boxplot, robustness is not sufficient to render the complexity of tumor response behaviors fully apparent.

SVR-based models yield fewer outliers and a wider interquartile range with a more residual distribution. The non-parametric and flexible nature of support vector regression implies greater generalization and less noise sensitivity.

The Hybrid SVR-Huber model is the most compact residual distribution with few outliers and within the narrowest interquartile range of all models. This is due to its superior reliability and flexibility in navigating nonlinear patterns and irregular data points. In this case, the hybrid solution is most reliable for real data with variance reduction and prediction accuracy.

Beyond predictive performance, practical considerations are also essential. From a reproducibility perspective, the study provides detailed methodological steps and parameter settings, including sensitivity checks on key choices such as the $\varepsilon$ clipping parameter. While the datasets and implementation code are not released at this stage due to institutional restrictions, they will be made available in future updates through a public repository. This will further enhance transparency and facilitate replication of the results.

It should also be acknowledged that the hybrid SVR-Huber beta regression model entails a higher computational burden compared to classical and robust beta regression. This increased cost arises from the dual demands of support vector regression training and iterative Huber reweighting within the beta framework. Nevertheless, the computational trade-off is outweighed by the considerable gains in predictive accuracy and robustness, particularly in biomedical applications where reliable inference is more valuable than computational efficiency. For most practical purposes, the additional complexity is therefore justified. For larger datasets, scalability can be enhanced by parallelizing the SVR training process, adopting approximate kernel techniques, or applying dimensionality reduction before model fitting. These strategies can reduce computational demands while preserving the predictive benefits of the hybrid approach.

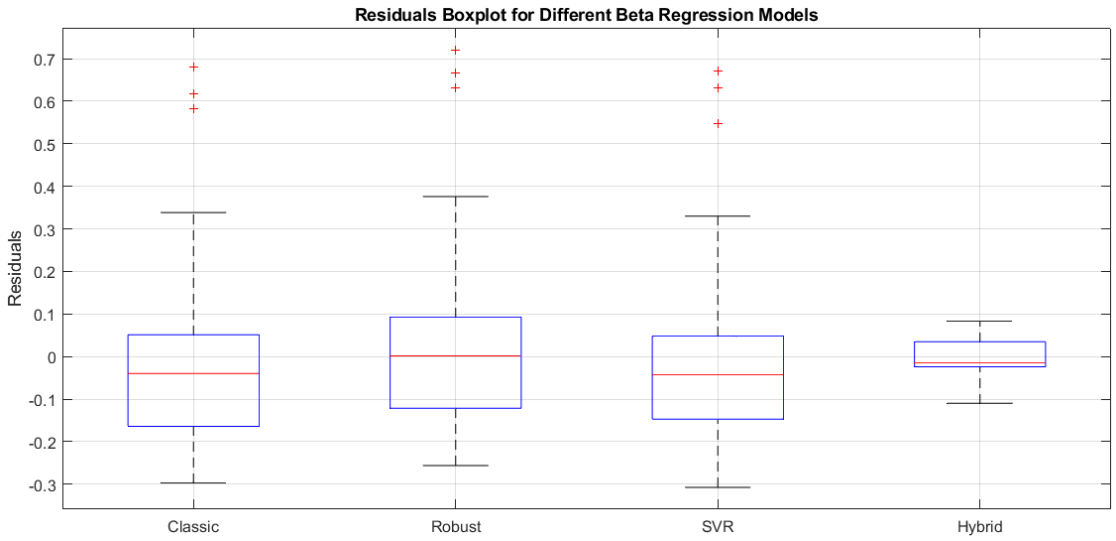


Figure 4. Boxplot of Residuals for Beta Regression Models Applied to Real Tumor Response Data

## 4. Conclusions

This study examined how different Beta regression models perform in dealing with bounded proportional data that is affected by noise and outliers. Evidence of robustness and flexibility in regression modeling was found through simulation tests, as well as in actual tumor response data.

### *4.1. Simulation Data Analysis Conclusions*

1. The classical beta regression model performs adequately under clean data conditions but is highly sensitive to outliers, especially with small sample sizes or increased model complexity.

2. The robust IRLS-Huber model improves upon the classical method by reducing the influence of extreme observations, though its linear nature limits its ability to model nonlinear structures in the data.

3. The SVR-based beta regression captures local nonlinear patterns more effectively than linear models; however, its vulnerability to outliers due to the absence of built-in robustness mechanisms affects its stability.

4. The hybrid SVR-Huber model consistently delivers superior performance across varying conditions, combining the flexibility of nonlinear learning with the resilience of robust estimation.

5. This hybrid approach maintains high predictive stability even when the number of predictors increases or the data is heavily contaminated, demonstrating scalability and robustness.

6. The simulation findings emphasize the limitations of purely linear or purely robust models in real-world scenarios where data contamination and nonlinear behavior are common.

7. By integrating SVR's capacity for noise reduction with the Huber loss function's robustness, the hybrid model proves especially effective for modeling bounded responses in complex environments.

8. Overall, the hybrid SVR-Huber model emerges as a strong candidate for applied predictive modeling tasks that require both accuracy and resistance to data irregularities.

9. The progression from classical to hybrid modeling demonstrates clear gains in accuracy and robustness. While traditional beta regression offers interpretability, its limitations in handling outliers become evident. Robust techniques offer moderate improvement, but the integration of machine learning (SVR) and robust loss functions (Huber) into the modeling process yields substantial benefits. The hybrid SVR-Huber-Beta model emerges as the most effective strategy for modeling bounded clinical outcomes like tumor response in the presence of real-world data challenges.

10. The residual analysis confirms the advantages of integrating robust learning and machine learning techniques within the beta regression framework. While classical and robust models offer baseline comparisons, the hybrid method outperforms in both residual consistency and resistance to outliers.

### *4.2. Real Data Analysis Conclusions*

1. The tumor response data, bounded between 0 and 1 and influenced by patient age, exhibits natural variability with notable outliers simulating clinical extremes like remission or treatment failure.

2. The dataset's inherent nonlinearity and contamination challenge standard modeling assumptions, highlighting the need for flexible and robust regression approaches.

3. Classical beta regression provides a limited fit, missing much of the data variability and showing sensitivity to outliers.

4. Robust beta regression improves fit stability by down-weighting extreme values but still falls short in capturing complex nonlinear relationships.

5. SVR enhances the ability to model nonlinear patterns, better fitting the overall trend, but is still affected by data irregularities.

6. The hybrid SVR-Huber beta regression effectively captures the nonlinear relationship and manages outliers, delivering the most precise and stable fit to the observed tumor response data.

While hybrid SVR-Huber beta regression has clearly shown excellent robustness and predictive accuracy, it is not in the best condition to study the limitations. The simulation framework relied on boundary outliers, and the clinical application relied on one predictor, age, which is not sufficiently complex to adequately reflect real data. In addition, this hybrid approach has more computational costs as opposed to classical and robust alternatives. These limitations are something that has been discussed for work in the future. More broadly, more robust loss functions and better computational optimizations would further enhance the utility of the framework with larger biomedical datasets with multiple covariates, as well as with multivariate populations. Yet the results show that incorporating machine learning and robust estimation into the beta regression model is useful in applications in medicine, biostatistics, and other fields.

## REFERENCES


1. Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815.
2. Bayes, C. L., Bazán, J. L., & García, C. (2012). A robust approach to beta regression models with application to psychometric data. *Journal of Statistical Planning and Inference*, 142(9), 3044–3058.
3. Liu, D., and Li, X. (2018). Robust beta regression modeling for bounded data with outliers. *Computational Statistics & Data Analysis*, 127, 105–125.
4. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
5. Awad, M., & Khanna, R. (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. Springer.
6. Maluf, F., Ferrari, S. L. P., & Queiroz, M. (2022). Robust estimators for beta regression models: An approach based on logit transformation and Wald-type tests. *Journal of Statistical Computation and Simulation*, 92(11), 2360–2376.
7. Olaluwoye, O., Alaba, O. O., & Jebio, M. A. (2025). A robust estimation approach for beta regression models using least squares and modified likelihood under collinearity and contamination. *Journal of Statistical Theory and Practice*, 19, Article 66.
8. Lee, D., Smith, M. S., & Martínez, C. A. (2025). Cobin and micobin regression: Flexible robust alternatives to beta regression for bounded data with excess zeros and ones. *arXiv preprint*, arXiv:2504.15269.
9. Tareq Hasan, M., Ali, T. H, and Sedeek Kareem, N. H (2025). Multivariate CUSUM Daubechies Discrete Wavelet Transformation Coefficients Charts for Quality Control. Passer Journal of Basic and Applied Sciences, 7(1), 533-546.
10. Omer, A. Wali, and Ali, T. Hussein. (2025). Wavelet Analysis for Outlier Estimation in Multivariate Linear Regression Models. Passer Journal of Basic and Applied Sciences, 7(1), 478-494.
11. Ali, T. H., Sedeeq, B. S., Saleh, D. M., & Rahim, A. G. (2024). Robust multivariate quality control charts for enhanced variability monitoring. Quality and Reliability Engineering International, 40(3), 1369-1381.
12. Ali, Taha Hussein, Avan Al-Saffar, and Sarbast Saeed Ismael. (2023). Using Bayes weights to estimate parameters of a Gamma Regression model. Iraqi Journal of Statistical Sciences, 20(1), 43-54.
13. Ali, T. H. (2022). Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study). Communications in Statistics - Simulation and Computation, 51(2), 391–403.
14. Hayawi, H.A., Azeez, S.M., Babakr, S.O. and Ali, T.H. (2025). ARX TIME SERIES MODEL ANALYSIS WITH WAVELETS SHRINKAGE (SIMULATION STUDY). Pak. J. Statist, 41(2), 103-116.
15. Ali, T. H., Hamad, A. A., Mahmood, S. H., & Ahmed, K. H. (2025). ARIMAX time series analysis for a general budget in the Kurdistan Region of Iraq using wavelet shrinkage. Communications in Statistics: Case Studies, Data Analysis and Applications, 11(2), 164–188.
16. Elias, I. Ibrahim, and Ali, T. Hussein. (2025). VARMA Time Series Model Analysis Using Discrete Wavelet Transformation Coefficients for Coiflets Wavelet. Passer Journal of Basic and Applied Sciences, 7(2), 657-677.