Impact of Outliers on Regression Models Performance: A Comparative Analysis of Diabetes Data

By

Dr. Abdelreheem Awad Bassuny

Lecturer at the Higher Institute of Management in EL Mahalla El-Kubra

dr-AbdelreheemBassuny@outlook.com

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.6, No.2, Part 1., July 2025

APA Citation

Bassuny, A. A. (2025). Impact of Outliers on Regression Models Performance: A Comparative Analysis of Diabetes Data, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 6(2)1, 241-269.

Website: https://cfdj.journals.ekb.eg/

Impact of Outliers on Regression Models Performance: A Comparative Analysis of Diabetes Data

Dr. Abdelreheem Awad Bassuny

Abstract:

This study used a dataset of 150 diabetic patients from Kafr El-Sheikh, Egypt, collected between 2000 and 2024, to examine the impact of outliers on the performance of different regression models: OLS, RR, QR, and SVR. Outliers were addressed using the Trimmed Mean method, and performance was evaluated using R^2, MSE, and MAE. The results showed that the OLS model was the most sensitive to outliers, while the QR model was relatively robust. For example, the MSE for the SVR model decreased by 50.59% upon removing outliers, whereas the changes in RR and QR were less significant. Without outliers, the RR model achieved the highest R^2 value (0.8618), and the QR model had the lowest MSE (0.9875) and MAE (0.9072). These findings highlight the critical need to carefully select regression techniques and outlier handling methods, even with seemingly robust models like QR, to ensure valid and reliable statistical inferences. Future research should explore alternative outlier handling methods, investigate the causes of outliers, and develop data and model-specific outlier treatment strategies.

Keywords: outliers, Quintile Regression, Ridge Regression, Support vector regression.

1. Introduction:

Outliers, or extreme values, are data points that deviate significantly from the rest of the dataset. If not addressed appropriately, these values can negatively impact the validity and reliability of analyses and results. One common approach to handling outliers is their removal. While outlier removal can improve the accuracy and quality of analyses, it must be implemented carefully to avoid losing crucial information or distorting the underlying data structure. Random removal can lead to a distorted view of the data, whereas systematic removal strategies can enhance the accuracy and quality of findings.

The challenge of dealing with outliers has been a subject of extensive research, leading to the development of diverse methods for their detection and management. The following are summaries of recent studies that address this issue:

Alves et al. (2024): Conducted a systematic literature review to evaluate methods for removing both outliers and inliers in agricultural data, specifically in the context of precision agriculture. By analyzing 102 studies, they identified techniques, including Chebyshev's inequality, boxplots, principal component analysis, and the local Moran's index, along with 13 software packages commonly used for data cleaning. The study underscored the importance of addressing outliers and inliers to ensure the accuracy of agricultural analyses and the validity of thematic maps.

Taghikhah et al. (2024): Introduced QuantOD, a novel unsupervised outlier detection method that utilizes quantile-based maximum likelihood training, a technique previously unexplored in this area. This approach employs a normalizing flow to estimate the density of feature representations derived from a pre-trained image classifier, to enhance discrimination between inliers and outliers. Experimental results demonstrate QuantOD's superiority over state-of-the-art methods for outlier detection, making it a promising solution for reducing reliance on negative data sampling.

Atif et al. (2024): Explored the impact of outliers on cluster evolution in temporal datasets, using both synthetic and real-world data (Melbourne house prices). Their research assessed the effect of outliers on both external and internal cluster transitions, using a survival ratio metric and history costs. Their findings revealed that outliers significantly affect internal cluster transitions, highlighting the need for appropriate outlier handling techniques to achieve reliable clustering results.

Fröhlich (2024): Proposed an alternative outlier identification procedure for time series data, based on a nonlinear model estimated with support vector regressions (SVR). This method identifies additive outliers and applies to short time series (less than 3 years of observations). The researcher noted that while traditional methods like RegArima (used in X13-Arima or Tramo/Seats) can identify different types of outliers, SVR provides an effective alternative, especially for short time series.

Thériault et al. (2024): Presented a practical introduction to identifying statistical outliers using the R language, focusing on the easystats package. The study covers univariate, multivariate, and model-based outlier detection methods, discussing recommended detection thresholds, standard output, and plotting techniques. The authors review different theoretical types of outliers, discussing whether they should be excluded or winsorized while emphasizing the importance of transparency in outlier handling.

Nijhuis and van Lelyveld (2023): Explored "Outlier Detection with Reinforcement Learning for Costly to Verify Data," proposing the integration of reinforcement learning with statistical outlier detection methods. This approach uses an ensemble of outlier detection methods with reinforcement learning to dynamically adjust the weights of these methods based on each additional piece of information. The results demonstrated that this approach enabled outlier identification and improved results by optimizing the model's weights.

Geetha Mary and Sangeetha (2023): Explored outlier detection within a single universal set, proposing a methodology based on an intuitionistic fuzzy proximity relation combined with a rough set approach using complement entropy and a weighted density method. The results of their study indicated that the proposed methodology accurately identified outliers.

Lei, Chen, and Li (2023): Focused on functional outlier detection for density-valued data, and its application in robustifying distribution-to-distribution regression, proposing a tree-structured transformation system for feature extraction and shape outlier detection, as well as a multiple detection strategy to manage uncertainty.

Boiar et al. (2022): Developed an enhanced method for outlier detection in noisy data utilizing a leave-out strategy in conjunction with Support Vector Machine (SVM) models.

Benslimane et al. (2022): Aimed to divide outlier detection methods in regression models into three methods, including statistical methods based on normal distribution, geometric methods such as dispersion and deviation, and machine learning methods such as unsupervised learning.

Lorenzo and Saracco (2021): Addressed the problem of outlier detection in sliced inverse regression (SIR) models, proposing three novel computational approaches based on in-bag (IB) or out-of-bag (OOB) prediction errors from resampling techniques.

Hendricks et al. (2019): Proposed a method to improve deep anomaly detection in machine learning by training anomaly detectors against an auxiliary dataset of outliers, which they called "Outlier Exposure.

Shi et al. (2019) addressed the challenge of outlier detection in wireless sensor networks by introducing ID-SVDD, an enhancement of the traditional SVDD method. Their approach incorporates data density, utilizing density estimations and a refined distance calculation, to improve the accuracy of identifying anomalous data points. The resulting ID-SVDD method shows promise for applications like real-time water quality monitoring due to its more precise outlier identification.

Dhhan, Midi, & Alameer (2017) investigate the issue of outliers in Support Vector Regression (SVR) models. They introduce Double Support Vector Regression (DSVR), a robust method based on Fixed Parameter SVR (FP-SVR). DSVR seeks to lessen the influence of outliers and leverage points on both linear and nonlinear functions. The paper evaluates DSVR using real and simulated datasets and compares it to standard SVR, indicating DSVR's superiority by demonstrating smaller Mean Squared Errors (MSE).

These studies collectively highlight the importance of addressing outliers across various domains and demonstrate the diverse range of techniques being developed to detect and handle them. The overarching goal remains to improve the accuracy and reliability of analyses and results while emphasizing the need for careful consideration and transparency when dealing with outliers.

This study is distinguished by its comprehensive comparison of the impact of outliers on diverse regression models, from classic ordinary least squares to the modern Support Vector Machine (SVM). The specific focus on SVM as an AI model provides novel insights into its robustness against anomalous data, alongside a comparative analysis of different models' performance in the presence of outliers. This approach, combined with an innovative experimental methodology and practical applications, offers added value beyond previous studies, contributing to a better understanding of how to effectively handle outliers in regression models.

2. Methodology:

Outliers, or extreme values, present a significant challenge in data analysis. These data points deviate substantially from the typical pattern, potentially stemming from measurement errors or reflecting genuine, noteworthy phenomena. Handling outliers is a subject of ongoing debate, particularly concerning the optimal methodology. Removing outliers can significantly influence the final results, potentially introducing bias or leading to the loss of crucial information. Consequently, a comprehensive understanding of the data's characteristics and the analysis's objectives is essential before deciding whether to remove or modify these values. Several methods exist for detecting extreme values, the most prominent of which include:

2.1: Box – Plot Method:

The box plot serves as a visual tool for outlier detection in data. It comprises a rectangular box representing the interquartile range (IQR), which is the difference between the first quartile (Q1) and the third quartile (Q3). This box encompasses 50% of the data observations, specifically those falling between Q1 and Q3. A line

indicating the median (MD) is drawn within the box. Values considered significantly distant from the box, signaling potential outliers, are depicted as individual points lying outside the box's boundaries. Once outliers are identified using this method, they are typically addressed through one of several treatment approaches, which include:

2.2: Deletion method:

Identified outliers are commonly excluded from the data. Rahman & Al Amri (2010) emphasize that outlier removal improves the accuracy of parameter estimation.

2.3: Trimmed mean method:

The data is first sorted in ascending order, and the median is calculated. Outliers are then estimated based on their relative magnitude compared to the rest of the data, according to the following procedure:

- If the outlier is smaller than the median: The largest value in the dataset and the outlier are removed. The mean of the remaining values is then calculated, which serves as an estimate for the outlier's true value.
- If the outlier is larger than the median: The smallest value in the dataset and the outlier are removed. The mean of the remaining values is then calculated, providing an estimate of the outlier's true value. This process is repeated for subsequent outliers.

2.4: Estimation Regression Models:

This study investigates the impact of outliers on the performance of several regression models, encompassing both traditional and contemporary approaches. Specifically, we examine the sensitivity of ordinary least squares (OLS) regression, ridge regression, quantile regression, and support vector Regression (SVR) – a modern machine learning technique – to the presence of extreme values. By comparing these models, we aim to provide a nuanced understanding of how outliers affect parameter estimation and predictive accuracy across diverse regression methodologies, ultimately informing best practices for outlier management in statistical modeling**.**

2.4.1: Quantile Regression model (QR):

Quantile regression is a method within regression analysis that extends beyond the limitations of linear regression when its assumptions are not satisfied (Li & Zhu, 2008).

$$y=X\,B+\in \qquad (1)$$

- **whereas:**

y: response variable vector (n x1)

X: a matrix of degree (n x p)

$\in$: random error vector of degree (n x1) And it was $\in \rightarrow N(0,\sigma^2)$

Using the ordinary least squares (OLS) method, if the conditions are met, the estimates are:

$$\widehat{\mathcal{B}}_{ols}=(x'x)^{-1}(x`y) \qquad (2)$$

The estimates are obtained by minimizing the sum of squared residuals as follows:

$$SSE=(y-xB)'\,(y-xB)$$

Ordinary Least Squares (OLS) estimation is known to be susceptible to instability and highly sensitive to the presence of outliers, particularly when the assumption of normally distributed random errors is violated. Consequently, OLS-based predictions may become unreliable in such circumstances. Quantile regression offers a robust alternative (Koenker, 1978) due to its distribution-free nature. Unlike linear models, it does not rely on the assumption of normally distributed errors and exhibits resilience to the influence of outliers as its regression lines can accommodate these extreme values. This makes quantile regression a more comprehensive statistical modeling tool compared to traditional OLS approaches, and positions it as a robust alternative.

$$y_i=\grave{X}_\iota\,\mathrm{B}_p+\in_i \qquad i=1,2,\dots\dots\dots\dots.n \qquad (3)$$

- **whereas:**

y_i: dependent variable.

x_i: vector of independent variables.

B_p: The vector of parameters at the quantity (P) where 0<P<1 and assuming that the distribution of the random error is: (Koenker & Bassett; 1978)

$$F_p(\in_i)=p\,(1-p)\;\; exp^{\{-\rho_p(\in_i)\}} \qquad (4)$$

- **whereas:**

$\in_i$: represents the random error that has a constrained distribution (Hideo & Kobayashi; 2011)

$$\int_{-\infty}^{0} F_P(\in_i)\ d\ \in_i = P \qquad (5)$$

Or $F_e^{-1}(p)=0$ (6)

The estimation of the parameters of the quantile regression model is done by minimizing the following loss function:

$$\min \sum_{i=1}^{n} \rho_p\ \left(y_i - \grave{x_\iota}\ \mathcal{B}_p\right) \qquad (7)$$

It is in the following form:

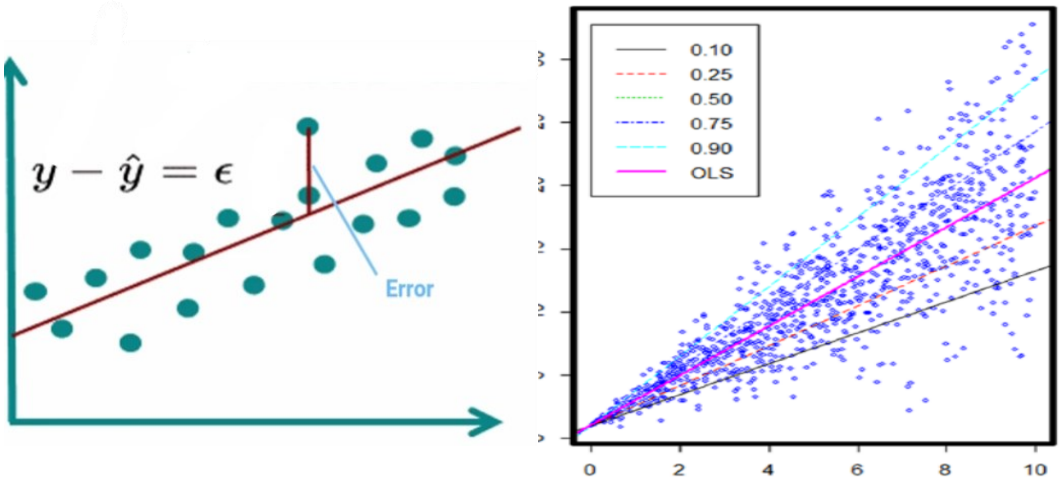


Fig (2) Ordinary Least Square method **Fig(1) Quantile regression method**

Figure 2 illustrates the estimated regression line derived from the ordinary least squares (OLS) method. However, as previously discussed, this model may not always provide a complete understanding of the relationship between the dependent and independent variables. In contrast, quantile regression offers a more detailed representation of this relationship. As depicted in the figure on the right, quantile regression (QR) estimates multiple regression lines across different quantiles, providing a more nuanced perspective. In this example, four regression lines, representing four distinct quantile levels, are shown.

2.4.2: Ridge Regression model (RR):

Ridge regression is a specialized technique within multiple regression analysis designed to address the issue of multicollinearity. It effectively reduces this problem, as multicollinearity can lead to inflated variance in parameter estimates when using the ordinary least squares method.

This inflated variance mirrors the symptoms observed when extreme values are present, suggesting ridge regression may also be beneficial in handling outliers (Lukman, 2014). The core concept of ridge regression involves identifying an optimal value for a constant, denoted as (K), also known as the bias parameter. This positive value is added to the diagonal elements of the (X'X) matrix, effectively reducing the variance of the estimated parameters. Adding a constant (K) with small values typically induces rapid changes in the estimated parameters. However, as the value of (K) increases, the parameter values begin to stabilize progressively, eventually reaching a point where further change is minimal. Ridge regression estimates the parameters by minimizing the sum of squared errors (SSE), according to the following formula:

$$\widehat{\mathcal{B}} = \min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=1}^{p} \mathrm{B}_j x_{ij})^2 \qquad (8)$$

The shrinkage equation is as follows:

$$\widehat{\mathcal{B}} = \min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=1}^{p} \mathrm{B}_j x_{ij})^2 + k \sum_{j=1}^{p} {\mathrm{B}_j}^2) \quad (9)$$

"The equation consists of two parts: the first is the sum of squared errors (SSE), and the second is called the penalty function, as shown below:

$$\widehat{\mathcal{B}} = SSE + k \sum_{j=1}^{p} {\mathrm{B}_j}^2 \qquad (10)$$

"By minimizing the sum of squared errors, we obtain the following ridge regression estimates:

$$\widehat{\mathcal{B}}_{RR} = \left[\dot{x}x + k\, I_p\right]^{-1} (\dot{x}y) \qquad (11)$$

- **Whereas:**

$\widehat{\mathcal{B}}_{RR}$: The vector of estimated parameters in ridge regression is given by the following equation:

K: bias parameter.

$$\widehat{\mathcal{B}}_{RR} = [(\dot{x}x + k\,(\dot{x}x)(\dot{x}x)^{-1}]^{-1}(\dot{x}y) \qquad (12)$$

$$\widehat{\mathcal{B}}_{RR} = \left[I_p + k(\dot{x}x)^{-1}\right](x\dot{x})^{-1}\,\dot{x}y$$

$$\widehat{\mathcal{B}}_{RR} = Z_{RR}\,\widehat{\mathcal{B}}_{ols} \qquad (13)$$

- **whereas:**

$$Z_{RR} = \left[I_p + k(\dot{x}x)^{-1}\right]^{-1}$$

"From the above, ridge regression estimates are a linear transformation of least squares estimates, where:"

$$MSE_R = vanance\left(\hat{B}_R\right) + (bias\ in\ \hat{B}_R)^2$$

In ridge regression, increasing the value of the biasing parameter, K, leads to a larger bias but reduces the variance of the parameter estimates. Thus, the optimal K value is one where the reduction in variance outweighs the increase in the squared bias. Under such conditions, the mean squared error (MSE) of the ridge regression estimates can be lower than the variance of the least squares estimates. However, a larger K also results in a decrease in the coefficient of determination (R^2). Consequently, ridge regression may not necessarily produce the best-fitting model in terms of maximizing R^2, but rather aims to identify an equation with coefficients that are stable and relatively unbiased as K increases (ALKhamisi, 2007).

2.4.3: Support vector Regression:

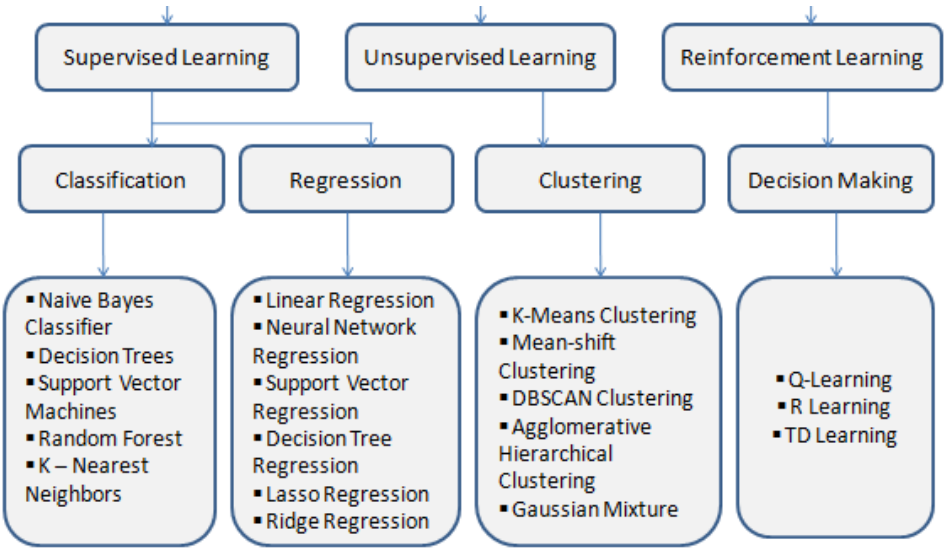


Fig (3) Machine Learning

Support Vector Regression (SVR) is a robust and versatile technique employed for data analysis and forecasting. Notably, it excels at handling non-linear data, making it well-suited for time series prediction, particularly in fields like finance and economics where non-linear patterns are common. SVR operates by classifying input data and identifying an optimal hyperplane to separate the data points even in cases where a clear separation isn't readily apparent. The general formulation is as follows (Cao and Tay; 2003):

$$f(x,w) = w^t \varphi(x_i) + b \qquad (14)$$

Where:

w: is the regression coefficients vector.

$\varphi(x_i)$: plots in a high-dimensional space.

b: is the bias term, and the prediction error is made using the loss function f(x). (Rosenbaum et al; 2013)

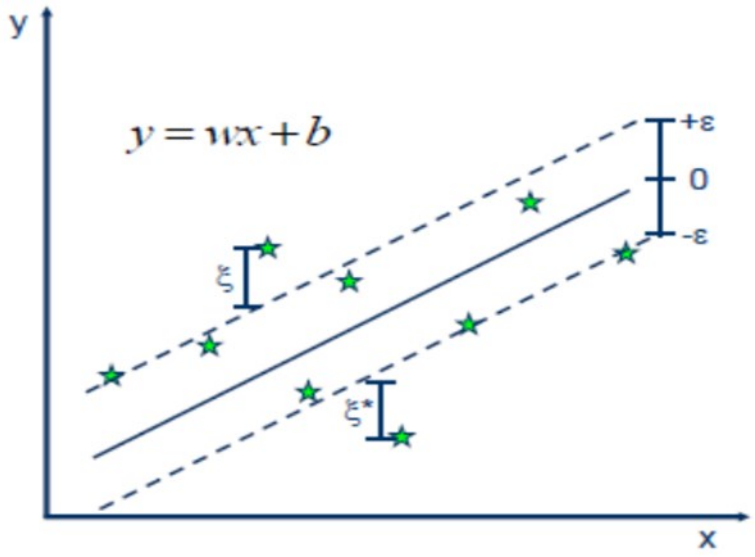


Figure (4) algorithm of the Support Vector Regression (SVR)

A key goal of Support Vector Regression (SVR) is to achieve accurate prediction and classification. The underlying principle involves separating different groups of data using more than one line, aiming for the optimal boundary represented by a hyperplane. The optimal hyperplane is selected to maximize the margin—the distance between it and the closest data points from each class. These critical data points are called Support Vectors. Increasing the margin leads to better generalization; the wider the separation between the closest points, the more effectively the model can classify new, unseen data. The loss function can be represented by the following equation:

$$L_{\varepsilon}\left(y, f\left(x,w\right)\right)=\begin{cases}0 & \left|y-f\left(x,w\right)\right|\leq\varepsilon;\\ \left|y-f\left(x,w\right)\right|-\varepsilon & otherwise,\end{cases} \quad (15)$$

The objective of using SVR is to find the function f (w, x) that agrees with the deviation ϵ for all training data in the prediction model. The problem is formulated as follows:

$$\text{Minimize: } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\delta_i + \delta_i^*)$$

$$\text{Subject to } \begin{cases} y_i - f(x_i,w) - b \leq \epsilon + \delta_i^* \\ f(x_i,w) + b - y_i \leq \epsilon + \delta_i \end{cases}$$

$$\delta_i^*, \delta_i \geq 0$$

Sometimes it's not possible to predict all training data within the deviation ϵ. Therefore, slack variables are introduced. Consequently, the problem is reformulated as follows:

$$\text{Minimize: } \frac{1}{2}\ \|w\|^2$$

$$\text{Subject to } \begin{cases} y_i - f(x_i,w) - b \leq \epsilon + \delta_i \\ f\ (x_i,w) + b - y_i \leq \epsilon + \delta_i \end{cases}$$

Whereas:

C: is the regularization parameter, where C > 0, also known as the regularization term.

δ_i*, δ_i: are variables that measure deviations larger than ϵ.

The Lagrange equation is used to solve the problem as follows:

$$f(x) = \sum_{i}^{n} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad (16)$$

Whereas:

$k(x_i,x)$: his function is known as Kernel, and it is used when linear separation of data is challenging. The data is transformed from the original space to a high-dimensional space, allowing for better data separation. (Cosgun et al.; 2011).

3. Applied Study:

This research aims to compare the performance of various regression models – including traditional models represented by Ordinary Least Squares (OLS) multiple regression, Quantile Regression (QR), and Ridge Regression (RR) – alongside a machine learning model, Support Vector Regression (SVR). This comparison is conducted both with and without outlier treatment using the Trimmed Mean method. The objective is to identify the most efficient and least sensitive model to the influence of outliers. This approach avoids the common practice of deleting outliers, which can lead to misleading results, loss of potentially valuable information, and a departure from real-world conditions. The evaluation relies on several statistical criteria: the coefficient of determination (R^2) and the Mean Squared Error (MSE). The analysis is applied to a dataset of 150 diabetic patients from the records of the General and University Hospitals in Kafr El-Sheikh, covering the period from 2000 to 2024. The dependent variable (y) represents the Glycated Hemoglobin (HbA1c) level. The independent variables are: (X_1) Blood Glucose Level, (X_2) Systolic Blood Pressure, (X_3) Diastolic Blood Pressure, and (X_4) Weight. The analysis is performed using statistical software packages including Stata 15, Stat Graphics 19, and EViews 12.

3.1: Outlier Detection in Data:

Outliers, those observations distant from the general data pattern, pose a genuine challenge in statistical analysis. They directly impact descriptive measures such as the mean and standard deviation, distorting our understanding of data distribution and central tendency. They also weaken the power of statistical tests and increase the likelihood of reaching erroneous conclusions about hypotheses. In regression models, outliers can alter the slope of the line and distort parameter estimates, affecting the accuracy of predictions. Therefore, a thorough understanding of the

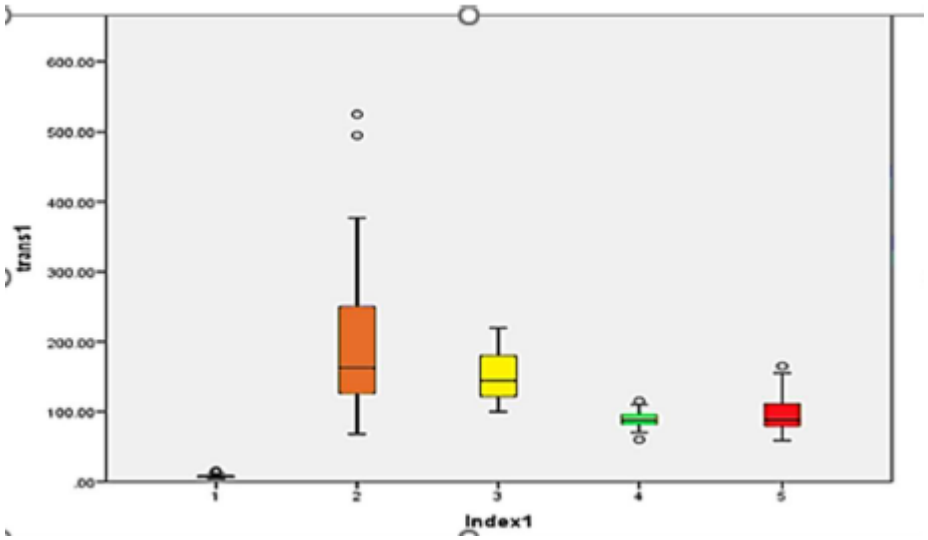
nature of outliers and handling them with care is crucial to ensuring the credibility of statistical analysis and the validity of the derived results.

Detecting extreme values in data is one of the first steps in statistical analysis, and this is done using a box plot, as illustrated in the following figures.

Table(1) extreme values in the data variables

Variables	Outliers
Y(the average cumulative sugar level)	(12,13, 39, 40, 41)
X_1 (blood sugar level)	(40, 13)
X_2 (high blood pressure)	No outliers
X_3 (low blood pressure)	(10, 48).
X_4 (weight)	(41)

The table(1) summarizes the extreme values (outliers) detected in each variable of the dataset. Notably, Y (the average cumulative sugar level), X_1 (blood sugar level), X_3 (low blood pressure), and X_4 (weight) have identified outliers, while X_2 (high blood pressure) shows none. This information is crucial for outlier treatment and robust model selection in the research.



Fig(5) Database with outliers

Figure 5 displays boxplots for five different variables revealing varying degrees of data spread and outlier presence. Variable 2 exhibits a wider interquartile range and notable outliers, suggesting higher variability and potential extreme values. Variables 1, 4, and 5 show relatively compact distributions with fewer outliers compared to Variable 2.

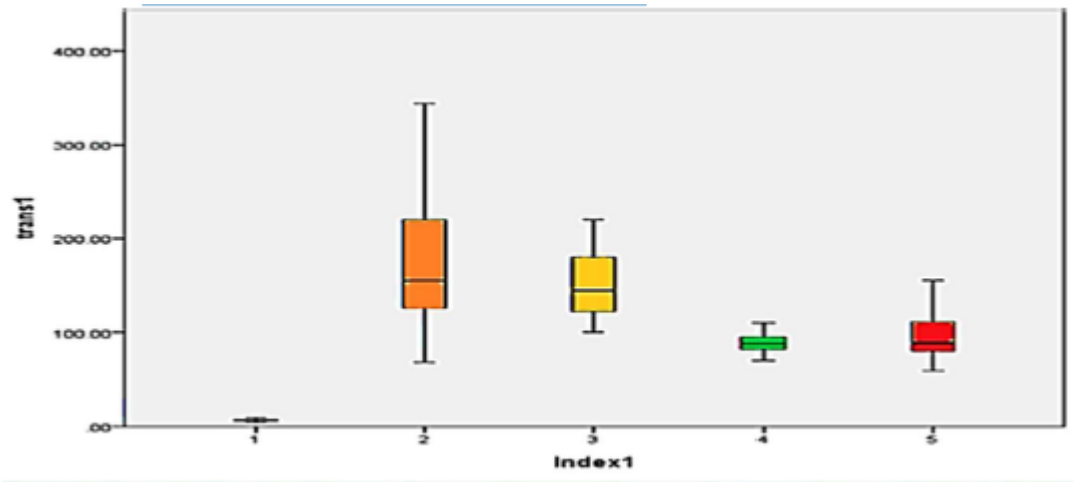


Fig (6) Database After processing outliers

Figure 6 presents boxplots after outlier processing, showing changes in data distributions compared to Figure 3. The removal or adjustment of outliers has likely reduced the range and skewness of some variables, particularly noticeable in variable 2. Overall, the data appears more condensed and potentially better suited for subsequent statistical analyses after outlier handling.

Table (2) Descriptive statistics

Variables	**Database with outliers**			**Database without outliers**		
	Mean	**Median**	**S. deviation**	**Mean**	**Median**	**S. deviation**
Y	7.22	6.80	2.47	7.40	6.82	2.30
X_1	191.97	162.51	96.32	192.76	163	94.30
X_2	152.887	144.5	33.32	152.887	144.5	33.32
X_3	89.081	88	11.41	90.12	88.5	11.01
X_4	97.32	88.5	24.35	98.02	89	21.12

Table (2) compares descriptive statistics (mean, median, standard deviation) for five variables, with and without outliers. Removing outliers noticeably changes the mean and standard deviation for variables like X1 and X4, implying extreme values skewed the average and increased data variability. X2, however, shows little

change, suggesting minimal outlier influence. The median remains fairly consistent across variables, highlighting its robustness to extreme values. Therefore, outliers impact specific variables' central tendency and spread, while the median offers a more stable representation of the typical value. The results show, the outliers greatly influence means and standard deviations.

3.2: Estimation Regression Models

Outliers can substantially distort regression model results by unduly influencing coefficient estimates and inflating error terms. While tempting to remove these extreme data points, outright deletion poses risks, potentially leading to biased samples and the loss of valuable information. Instead, robust statistical methods, such as robust regression or data transformations, should be considered. A thorough investigation and careful consideration of the data's context are critical to making informed decisions about the strategy for managing outliers effectively.

3.2.1: Multiple Regression Model:

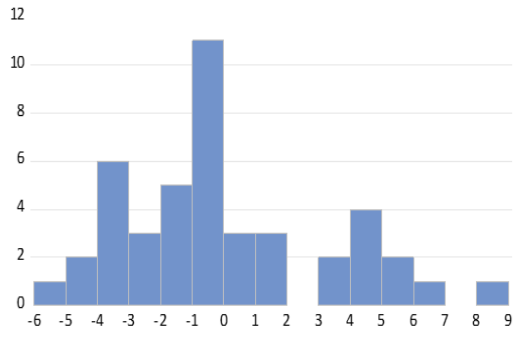
A multiple regression model was estimated using ordinary least squares (OLS) and the results were as follows:

Table (3) The results of ordinary Least squares:

Variables	Database with outliers			Database without outliers		
	Value	St. error	P- value	Value	St. error	P- value
β_0	1.285	2.265	0.9195	1.315	2.153	0.067
β_1	1.983	1.698	000	2.02	1.376	0.000
β_2	0.6818	0.04554	0.1397	0.8178	.0213	0.091
β_3	0.549	0.1182	0.6440	0.7176	.0971	0.537
β_4	1.8029	1.6445	0.007	1.976	1.496	0.000
R^2	0.8355			0.8517		
MSE	1.094			1.073		
F	72.351		0.000	81.83		000
Normality Test (Jb)	0.8447		0.9586	1.725		0.0416
MAE	2.135			1.821		

The table highlights the impact of outliers on an Ordinary Least Squares (OLS) regression model. After removing outliers, the model fit improves, with R-squared increasing from 0.8355 to 0.8517 and MSE decreasing from 1.094 to 1.073. The F-statistics also rise, indicating a more significant overall model. Notably, the Jarque-Bera statistic suggests better adherence to the normality assumption of residuals post-outlier removal. While coefficients of significant variables like β1 (changing

from 1.983 to 2.02) and β4 (changing from 1.8029 to 1.976) remain relatively consistent, removing outliers affects non-significant coefficients like β2 (changing from 0.6818 to 0.8178) and β3 (changing from 0.549 to 0.7176). Overall, the model's predictive accuracy is enhanced with a reduction in Mean Absolute Error (MAE). However, improvements do not always signify a better model, and careful interpretation is crucial before removal to assess if outliers reflect data errors or genuine phenomena.



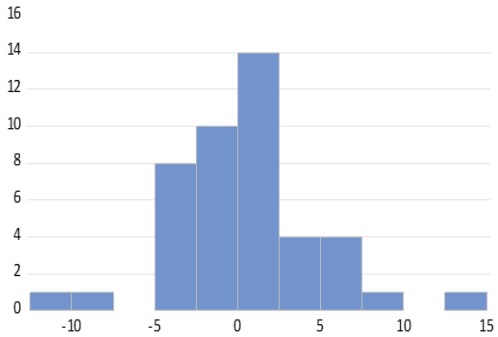


Fig (7) Normality with outliers **Fig (8) Normality without outlier**

3.2.2: Ridge regression Method (RR):

Recognizing that Ordinary Least Squares (OLS) regression can be significantly influenced by outliers, Ridge Regression was utilized as a robust alternative to estimate model parameters. To determine the optimal Ridge parameter (K), minimizing the impact of extreme values and stabilizing the model, the Ridge Trace method was employed, evaluating K values ranging from 0 to 1 in increments of 0.005. Using Stat Graphics 19, the analysis identified K=0.005 as the value yielding the lowest Mean Squared Error (MSE) and thus the most stable model, with the corresponding estimation results presented below. This approach is particularly beneficial when dealing with data prone to outliers or exhibiting multicollinearity.

Table (4) The result of ridge regression:

Variables	Database with outliers			Database without outliers		
	Value	St. error	P- value	Value	St. error	P- value
β_0	1.302	2.351	0.2753	1.42	2.112	0.231
β_1	1.972	1.596	0.0012	2.135	1.302	0.000
β_2	0.6819	0.0254	0.076	0.765	0.0115	0.0231
β_3	0.546	0.1171	0.615	0.549	0.0762	0.520
β_4	1.814	1.653	0.001	2.013	1.3510	0.000
R^2	83.23%			86.1%		
MSE	1.045			1.023		
MAE	0.7851			0.6135		

Table (4) presents the Ridge Regression results comparing a dataset with outliers to one without. Overall, the Ridge Regression appears to have mitigated the impact of outliers, showing relatively smaller changes compared to OLS results (if previously available).

- **Coefficient Stability**: While the coefficients change, Ridge Regression tends to maintain coefficients of already significant variables relatively close. For example, β1 (1.972 vs 2.135) and β4 (1.814 vs 2.013) remain significant ($p < 0.05$) even with the removal of outliers, with only a shift in their estimated values. β2 shows an interesting change: while previously insignificant with outliers ($p=0.076$), it becomes significant without outliers ($p=0.0231$).
- **Model Fit Improvement**: R-squared increases from 83.23% (with outliers) to 86.1% (without outliers), indicating an improved model fit by explaining more variance in the dependent variable when outliers are removed. This shows that Ridge Regression, though less sensitive, still benefits from cleaner data.
- **Error Reduction**: The Mean Squared Error (MSE) decreases from 1.045 to 1.023 after removing outliers, which suggests that the model's predictive accuracy is enhanced, even if only slightly, by removing the outliers. The Mean Absolute Error (MAE) also decreases from 0.7851 to 0.6135, indicating more accurate predictions on average.
- **Impact on Significance**: Note that the significance of variables isn't drastically altered by removing outliers. The variables that significant initially remain so (except for β2 which shows a reversal), reinforcing Ridge Regression's ability to handle data with extreme values.

The Ridge Regression results show a degree of robustness against the influence of outliers, with relatively consistent coefficient estimates and significant levels even despite having extreme values. Removing outliers results in enhancements in model fit (R-squared), and predictive accuracy (MSE and MAE). Ridge Regression seems to be effective in balancing bias reduction and variance control, mitigating the outlier influence more effectively than OLS. However, even with Ridge Regression, it's clear that removing outliers produces a model with improved goodness-of-fit and predictive power. This underscores the importance of identifying and appropriately addressing outliers in regression modeling, regardless of the chosen method.

3.2.3: Quantile Regression Model (QR):

Quantile Regression is a statistical method used to estimate the relationship between independent variables and specific quantiles of the dependent variable, offering a more complete picture than OLS regression, which focuses solely on the mean. A key advantage is its relative robustness to outliers, as it models different parts of the distribution. While less susceptible than OLS, Quantile Regression is not entirely immune; extreme outliers can still influence the estimation of quantiles, particularly at the distribution's tails. Therefore, data examination and outlier assessment remain important even within the Quantile Regression framework**.**

To mitigate the significant errors that extreme values can introduce in Ordinary Least Squares (OLS) analysis, Quantile Regression (QR) was employed to examine the relationships within the data. The analysis focused on quantiles $p = 0.50$ and $p = 0.95$. The median ($p = 0.50$) was chosen because data are inherently centered around it. Simultaneously, the extreme quantile $p = 0.95$ was analyzed with the logic that a well-fitted regression line at this extreme indicates robustness across quantiles. Consequently, if the regression line performs well at $p = 0.95$, it is inferred to provide reliable prediction and estimation across all quantiles, such as $p = 0.25$ or $p = 0.75$.

Table (5) The results of quantile Regression

Variables	Database with outliers			Database without outliers		
	Value	St. error	P- value	Value	St. error	P- value
β_0	0.6625	1.644	0.607	1.231	1.35	0.532
β_1	2.135	0.025	0.000	2.416	0.0179	0.000
β_2	0.6397	0.0672	0.43	0.842	0.0524	0.2215
β_3	0.0291	0.1613	0.085	0.3512	0.1529	0.0462
β_4	1.853	0.1113	0.0109	2.0146	0.1013	0.0215
R^2	0.7915			0.8214		
MSE	1.053			0.9875		
MAE	1.012			0.9012		
Normality Test	0.4108		0.8144	12.15		0.002

Table (5) presents Quantile Regression results comparing a dataset with and without outliers. The overall trend indicates that while Quantile Regression is generally more robust than OLS, outliers still influence the model.

- **Coefficient Changes**: While Quantile Regression is intended to be robust, it's evident that coefficients shift with the removal of outliers. For instance, the coefficient for β_0 nearly doubles (0.6625 to 1.231), though it remains statistically insignificant.
- **Significance Shifts**: Some variables experience changes in statistical significance. Specifically, β_3 transitions from being statistically insignificant with outliers ($p = 0.085$) to becoming significant without them ($p = 0.0462$).
- **Model Fit Improvement**: Removing outliers results in a slight increase in R-squared (from 0.7915 to 0.8214), suggesting an improved model fit in terms of variance explained, even if the improvement is modest.
- **Error Reduction**: There's a reduction in both Mean Squared Error (MSE) and Mean Absolute Error (MAE) upon removing outliers. MSE decreases from 1.023 to 0.9875, and MAE decreases from 1.012 to 0.9012, both indicating enhanced predictive accuracy when outliers are addressed.
- The "**Normality Test**" result is crucial. Without outliers, the p-value is 0.002, which is statistically significant. This indicates that the residuals (errors) are NOT normally distributed at a significant level. On the other hand, with outliers included the p-value is 0.8144. This implies a failure to reject the null hypothesis of normality which means we can assume that residuals are normally distributed.

Impact of Outliers on Quantile Regression:

1-**Coefficient Estimation**: Outliers cause coefficients to be biased, though to a lesser extent than in OLS.

2- **Significance of Variables**: Removal of outliers can alter whether a variable is deemed statistically significant or not, impacting conclusions about the relationship between predictors and the dependent variable.

3- **Model Fit and Accuracy**: Even for a robust method like Quantile Regression, addressing outliers still improves overall model fit (R-squared) and predictive accuracy (MSE and MAE), though the degree of improvement may be smaller than what would be observed in OLS.

- While Quantile Regression has some built-in robustness, it still appears that outliers had some influence.
- The improvement in model fit and accuracy with outlier removal points to the importance of outlier detection and handling, even when using techniques designed to be less sensitive to them.
- The Normality test demonstrates that the residuals are not normally distributed when outliers are removed and this can change the interpretation of the model.

3.2.4: Support vector Regression (SVR):

Support Vector Regression (SVR) offers a robust approach to regression modeling, exhibiting a degree of resilience to outliers due to its epsilon-insensitive loss function. Instead of strictly minimizing the error between predicted and actual values, SVR aims to fit a model within a defined margin of error (ε). Data points falling within this margin have no impact on the model. However, SVR is not entirely immune to outliers. A significant presence of extreme values can lead to an inflated epsilon margin, potentially compromising the model's accuracy and generalization ability. Furthermore, outliers that become support vectors can unduly influence the model, causing it to overfit these anomalies and perform poorly on unseen data. Therefore, proper data preprocessing and outlier management techniques are essential when deploying SVR.

Table (6) The results of Support vector Regression

Variables	Database with outliers		Database without outliers	
	Value	P- value	Value	P- value
R^2	0.769		0.840	
MSE	2.085		1.025	
MAE	1.241		0.963	
Normality Test	3.698	0.2587	22.78	0.0356

Impact of Outliers on the Explanatory Power of the Model (R^2):

The coefficient of determination (R^2) measures the proportion of variance in the dependent variable that is explained by the model. With outliers present, the value of R^2 is 0.759. After removing the outliers, the value of R^2 increased to 0.840. This means the model explains a greater proportion of the variance in the data after removing outliers, indicating an improvement in the model's explanatory power by 8.1%.

Impact of Outliers on Prediction Accuracy (MSE and MAE):

Mean Squared Error (MSE) and Mean Absolute Error (MAE) measure the prediction accuracy of the model. With outliers present, the value of MSE is 2.045, and the value of MAE is 1.241. After removing the outliers, the value of MSE decreased to 1.025 (a decrease of 50%), and the value of MAE decreased to 0.963 (a decrease of 22.4%). This indicates the model became more accurate in predicting values after removing outliers.

Impact of Outliers on the Distribution of Residuals (Normality Test):

Outliers influence the distribution of residuals and consequently the value of the normality test. With outliers present, the normality test value is 3.698. After removing the outliers, the test value increased to 22.78. This large change indicates that the outliers were influencing the distribution of the residuals, causing it to deviate from a normal distribution.

These numbers clearly demonstrate how outliers negatively affect the regression model, and how their removal significantly improves its performance.

3.3: Comparison of Regression Models:

Table (4) presents a comprehensive comparison between four different regression models (OLS, RR, QR, SVR) to evaluate their performance in the presence and absence of outliers. The analysis aims to determine the impact of outliers on the accuracy of the different models and to highlight the model that is most resistant to these values. The models are evaluated based on three main metrics: the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE):

Table (7) Comparison of Regression Models

	Database with outliers				Database without outliers			
	OLS	RR	QR	SVR	OLS	RR	QR	SVR
R^2	0.8355	0.8323	0.791	0.769	0.851	0.8618	0.8214	0.840
MSE	1.094	1.045	1.023	2.085	1.073	1.023	0.9875	1.025
MAE	1.085	1.07851	1.012	1.241	1.072	1.06735	0.9072	0.963

Table 7 compares the performance of four regression models (OLS, RR, QR, SVR) on datasets both with and without outliers, using R^2, MSE, and MAE as key metrics. Without outliers, RR exhibits the highest R^2 (0.8618), while QR demonstrates the lowest MSE (0.9875) and MAE (0.9072), indicating the most efficient model and with outliers having an MAE of 1.012 and MSE of 1.023. When assessing sensitivity to outliers, SVR displays the most significant change in performance, revealing high sensitivity. Conversely, QR appears the least sensitive, maintaining comparatively stable MSE and MAE values regardless of outlier presence, indicating robustness against extreme data points.

4. Discussion:

This research examined the impact of outliers on regression model performance (OLS, RR, QR, and SVR) using a dataset of 150 diabetic patients from the General and University Hospitals in Kafr El-Sheikh, Egypt, between 2000 and 2024. Key findings reveal varying model sensitivities to outliers. For example, SVR's MSE decreased substantially from 2.085 to 1.025 upon outlier removal, signifying its responsiveness. Quantile Regression (QR), while exhibiting relative stability, still showed improved MAE, decreasing from 1.012 to 0.9012 after outlier treatment, reinforcing the need for outlier handling despite robustness. Ridge Regression (RR) also demonstrated resilience compared to OLS, showing smaller changes in model results. These results highlight the importance of model selection considerations in regression modeling, relating to dataset selection and preprocessing.

5. Limitations:

This study is limited by its scope. The dataset is restricted to 150 diabetic patients from Kafr El-Sheikh, Egypt, between 2000 and 2024, potentially limiting generalizability. Outlier treatment used the Trimmed Mean method; alternative techniques might yield different outcomes. The study also didn't extensively explore the causes of outliers, which could represent genuine phenomena. While R^2, MSE, and MAE were used to assess model sensitivity, additional diagnostic metrics could offer a more nuanced insight.

6. Conclusion:

This study comprehensively examined the impact of outliers on regression model performance, utilizing a real-world dataset of diabetic patients in Kafr El-Sheikh, Egypt (2000-2024). The findings confirm that outliers significantly influence regression outcomes, affecting model fit, coefficient estimates, and predictive accuracy. Specifically, OLS regression demonstrated the highest sensitivity to outliers, whereas Quantile Regression (QR) emerged as a robust alternative, maintaining relative stability and improving accuracy, even with outlier presence. This is further supported by the percentage improvement in predictive accuracy (as measured by MSE and MAE) shown in the table below. These findings support the application of robust techniques like QR when dealing with data potentially contaminated by extreme values. While outlier detection and handling methods improved overall model fit for all approaches studied, understanding the unique characteristics of data combined with carefully selected outlier detection methods will likely allow the data scientist to enhance overall model fitness for a variety of modeling approaches. Furthermore, these methods should be investigated, particularly when the outliers are more likely to represent legitimate but unusual data patterns. Ultimately, this study reinforces the importance of thorough data exploration, careful outlier assessment, and the selection of appropriate regression techniques to ensure the validity and reliability of statistical inferences, especially in the presence of extreme data values.

Table (8): Percentage Improvement After Outlier Removal

Model	R^2 Improvement (%)	MSE Reduction (%)	MAE Reduction (%)
OLS	1.94	1.92	15.14
RR	3.45	2.10	21.86
QR	3.78	6.23	11.01
SVR	9.23	50.59	22.40

The table displays the percentage improvement in R^2 and the percentage reduction in MSE and MAE following the removal of outliers for each regression model. These percentages reflect the extent to which each model is affected by outliers and how its performance changes after outlier treatment.

- **R^2 Improvement:** All models show an improvement in R^2 after removing outliers, indicating that the quality of the model's fi to the data has increased. However, the percentage improvement varies between the models, with the SVR model recording the highest percentage improvement (9.23%) and the OLS model recording the lowest (1.94%).

- **MSE Reduction:** All models show a reduction in MSE after removing outliers, indicating an improvement in the model's accuracy in predicting values. The largest MSE reduction was observed for the SVR model (50.59%), while the smallest reduction was observed for the OLS model (1.92%).
- **MAE Reduction:** All models show a reduction in MAE after removing outliers, confirming the improvement in the model's accuracy in predicting values. The largest MAE reduction was achieved by the RR model (21.86%), while the QR model recorded the smallest MAE reduction (11.01%).

Ranking of Models by Efficiency in the Presence of Outliers:

Based on the performance of the models in the presence of outliers and their ability to maintain relatively good performance:

1. Quantile Regression (QR): QR demonstrates good efficiency in the presence of outliers, as it maintained relatively low MSE and MAE values (1.053 and 1.012, respectively) even before removing outliers. Followed by a reduction in MAE (11.01%).

Ranking of Models by Sensitivity to Outliers (Most to Least):

Based on the extent of the change in model performance after removing outliers, they can be ranked as follows:

1. Support Vector Regression (SVR): SVR shows the largest change in performance after removing outliers, with MSE decreasing significantly (50.59%), indicating that it is the most sensitive to outliers.

2. Ridge Regression (RR): RR appears as the second-best decrease in MAE, suggesting that it has a significant impact on outliers.

3. Ordinary Least Squares (OLS): OLS appears to be the least sensitive to outliers.

In Summary and Based on the Preceding Analysis:

- The QR model demonstrates notable efficiency in the face of outliers, rendering it a fitting selection for datasets where these values are known or suspected to exist.
- The SVR model exhibits marked sensitivity to the presence of outliers. Thus, in its application, meticulous consideration of techniques to properly account for outliers becomes vital to optimize the performance of regression analysis.
- The RR model finds a middle ground between exhibiting efficient performance and sensitivity to outliers.

- The quantitative findings will help inform the decision-making process when choosing a regression model best suited for the characteristics of a given dataset, especially when values are believed to fall outside the normal distribution. While considering this summary, it is of particular interest for the data scientist to factor in the aims of each study.

Recommendations for Future Research:

1. Enhance Outlier Handling: Develop and evaluate novel methods for handling outliers, focusing on selecting the most appropriate approach for each data type and regression model, with the aim of minimizing their negative impact on model accuracy.

2. Understand the Nature and Impact of Outliers: Identify the causes of outlier presence in different datasets (measurement errors vs. genuine rare cases), and study their effect on various regression models and across diverse fields.

3. Analyze Model Sensitivity and Develop Custom Solutions: Investigate how outliers affect model coefficients, evaluate the performance of different models (including neural networks and decision trees) under varying outlier proportions, with a focus on developing tailored handling methods for specific datasets and models.

References:

1. Abdelwahab, M., Elhoseny, M., & Benslimane, M. A. (2022). A novel anomaly detection approach for regression models. arXiv preprint arXiv:2202.01353.
2. Alkhamisi, M. A., & Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. Communications in Statistics—Simulation and Computation®, 36(3), 535-547.
3. Alves, F., de Souza, E. G., Sobjak, R., Bazzi, C. L., Hachisuca, A. M. M., & Mercante, E. (2024). Data processing to remove outliers and inliers: A systematic literature study. Revista Brasileira de Engenharia Agrícola e Ambiental, 28(9), e278672. https://doi.org/10.1590/1807-1929/agriambi.v28n9e278672
4. Atif, M., Farooq, M., Shafiq, M., Alballa, T., Alhabeeb, S. A., & Khalifa, H. A. E.-W. (2024). Uncovering the impact of outliers on clusters' evolution in temporal data-sets: an empirical analysis. Scientific Reports, 14(1), 30674. https://doi.org/10.1038/s41598-024-75928-7

5. Benslimane, M. A., Elhoseny, M., Abdelwahab, M., & Elhoseny, M. (2020). Anomaly detection in regression models: A survey. arXiv preprint arXiv:2001.01018.
6. Boiar, D., Liebig, T., & Schubert, E. (2022). LOSDD: Leave-Out Support Vector Data Description for Outlier Detection. arXiv preprint arXiv:2212.13626.
7. Cao, L. J., & Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. IEEE Transactions on Neural Networks, 14(6), 1506-1518.
8. Cosgun, E., Limdi, N. A., & Duarte, C. W. (2011). High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. Bioinformatics, 27(10), 1384-1389.
9. Dhhan, W., Midi, H., & Alameer, T. (2017). Robust support vector regression model in the presence of outliers and leverage points. Modern Applied Science, 11(8), 92-97. https://doi.org/10.5539/mas.v11n8p92
10. Fröhlich, M. (2024). Outlier identification and adjustment for time series. Statistical Journal of the IAOS, 40(2), 389-402. https://doi.org/10.3233/SJI-230109
11. Geetha Mary, A., & Sangeetha, T. (2023). Outlier Detection in a Single Universal Set using Intuitionistic Fuzzy Proximity Relation based on A Rough Entropy-Based Weighted Density Method. International Research Journal on Advanced Science Hub, 5(05S), 501-506. https://doi.org/10.47392/irjash.2023.S067
12. Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606.
13. Koenker, R., & Bassett, G. (1978). Regression quantiles. Econometrica: J. Econometric Society, 46(1), 33.
14. Kozumi, H., & Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. Journal of Statistical Computation and Simulation, 81(11), 1565-1578.
15. Lei, X., Chen, Z., & Li, H. (2023). Functional outlier detection for density-valued data with application to robustify distribution to distribution regression. arXiv preprint arXiv:2212.11270.
16. Li, Y., & Zhu, J. (2008). L 1-norm quantile regression. Journal of Computational and Graphical Statistics, 17(1), 163-185.

17. Lorenzo, H., & Saracco, J. (2021). Computational outlier detection methods in sliced inverse regression. In Advances in Contemporary Statistics and Econometrics (pp. 101-122). Springer, Cham.
18. Lukman, A., Arowolo, O., & Ayinde, K. (2014). Some robust ridge regression for handling multicollinearity and outlier. International Journal of Sciences: Basic and Applied Research, 16(2), 192-202.
19. Mohammadi, A., Khakpour, S. A., & Hosseini, M. S. (2021). Anomaly detection in regression models with missing data. arXiv preprint arXiv:2103.01260.
20. Mohammadi, A., Khakpour, S. A., & Hosseini, M. S. (2023). Anomaly detection in regression models with extreme values. arXiv preprint arXiv:2301.05979.
21. Nijhuis, M., & van Lelyveld, I. (2023). Outlier Detection with Reinforcement Learning for Costly to Verify Data. Entropy, 25(6), 842. https://doi.org/10.3390/e25060842
22. Rahman, M. S., & Amri, K. A. (2011). Effect of an outlier on the coefficient of determination. International Journal of Education Research, 6(1), 9-20.
23. Shi, P., Li, G., Yuan, Y., & Kuang, L. (2019). Outlier detection using improved support vector data description in wireless sensor networks. Sensors, 19(21), 4712.
24. Taghikhah, M., Kumar, N., Šegvić, S., Eslami, A., & Gumhold, S. (2024). Quantile-Based Maximum Likelihood Training for Outlier Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 3, pp. 21610-21618).
25. Thériault, R., Ben-Shachar, M. S., Patil, I., Lüdecke, D., Wiernik, B. M., & Makowski, D. (2024). An introduction to identifying statistical outliers in R with easy stats! Check your outliers. Behavior Research Methods, 56, 4162-4172. https://doi.org/10.3758/s13428-024-02406-5

تأثير القيم الشاذة على أداء نماذج الانحدار: تحليل مقارن ل بيانات مرضى السكري

الملخص:

استخدمت هذه الدراسة مجموعة بيانات تضم ١٥٠ مريضًا بمرض السكري من مدينة كفر الشيخ في مصر، تم جمعها بين عامي ٢٠٢٠ و٢٠٢٤، لدراسة تأثير القيم الشاذة على أداء نماذج الانحدار المختلفة: الانحدار الخطي العادي (OLS)، الانحدار الريدج (RR)، الانحدار الكمي (QR)، وانحدار دعم المتجهات (SVR). تم التعامل مع القيم الشاذة باستخدام طريقة المتوسط المقتطع (Trimmed Mean)، وتم تقييم الأداء باستخدام مقاييس مثل معامل التحديد (R^2)، متوسط مربعات الخطأ (MSE)، ومتوسط الخطأ المطلق (MAE). أظهرت النتائج أن نموذج الانحدار الخطي العادي (OLS) كان الأكثر حساسية للقيم الشاذة، بينما كان نموذج الانحدار الكمي(QR) أكثر مقاومة نسبيًا. على سبيل المثال، انخفض متوسط مربعات الخطأ (MSE) لنموذج SVR بنسبة ٥٠,٥٩٪ بعد إزالة القيم الشاذة، بينما كانت التغيرات في نموذجي RR وQR أقل وضوحًا. وبعد إزالة القيم الشاذة، حقق نموذج الانحدار الريدج (RR) أعلى قيمة لمعامل التحديد (R^2) وهي ٠,٨٦١٨، بينما سجل نموذج الانحدار الكمي (QR) أقل قيم لمتوسط مربعات الخطأ (٠,٩٨٧٥) ومتوسط الخطأ المطلق (٠,٩٠٧٢). تُبرز هذه النتائج الحاجة الملحة لاختيار تقنيات الانحدار وطرق التعامل مع القيم الشاذة بعناية، حتى مع النماذج التي تبدو قوية مثل الانحدار الكمي (QR)، لضمان استنتاجات إحصائية دقيقة وموثوقة. يُوصى في الأبحاث المستقبلية باستكشاف طرق أخرى للتعامل مع القيم الشاذة، ودراسة أسباب ظهورها، وتطوير استراتيجيات مخصصة للتعامل معها بناءً على البيانات والنموذج المستخدم.

الكلمات المفتاحية: القيم الشاذة، الانحدار الكمي، الانحدار الريدج، انحدار دعم المتجهات.