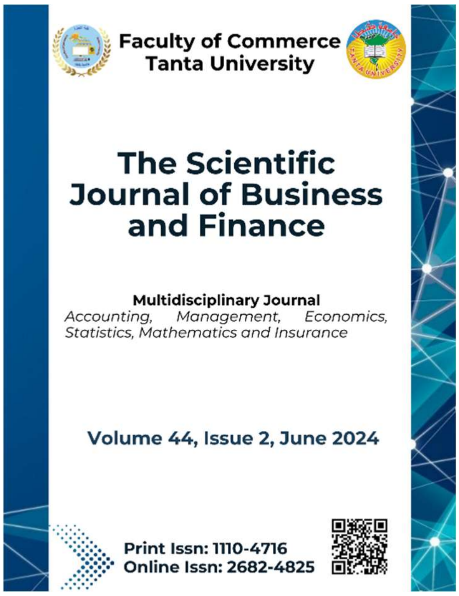


Faculty of Commerce
Tanta University

The Scientific Journal of Business and Finance

Multidisciplinary Journal
Accounting, Management, Economics, Statistics, Mathematics and Insurance

Volume 44, Issue 2, June 2024

Print Issn: 1110-4716
Online Issn: 2682-4825

The Scientific Journal of Business and Finance

https://caf.journals.ekb.eg

Efficiency of regression models in the presences of outliers

Maha Farouk[a], and Abdel Rahim Awad Bassiouni[b]

[a] Department of Statistics, Mathematics, and Insurance-Faculty of Commerce – Tanta University, Tanta-Egypt.

(2) PhD Statistics, Faculty of Commerce, Tanta University.

Published online: june 2024
To cite this article: Farouk Maha & Bassiouni Abdel Rahim Awad. **Efficiency of regression models in the presences of outliers**,The Scientific Journal of Business and Finance,
44 (2), 70-85.
DOI: 10.21608/caf.2024.360451

***Corresponding author:** **Maha.ibrahim@commerce.tanta.edu.eg,**

Efficiency of regression models in the presences of outliers

Maha Farouk

[a] Department of Statistics, Mathematics, and Insurance-Faculty of Commerce – Tanta University, Tanta-Egypt.

Abdel Rahim Awad Bassiouni

[b] PhD Statistics, Faculty of Commerce, Tanta University.

Article History

Received 10 March 2024, Accepted 20 April, Available online: june 2024

Abstract:

Outliers are those special points that differ significantly from most sample data. It can skew the data and present less accurate prediction results and detecting them is very important for obtaining more accurate predictions.The manuscript aimed to compare several regression models, including the multiple regression model using ordinary least squares (OLS), the quantile regression model (QR), and the ridge regression model (RR), to identify the model with high efficiency in the presence of extreme values in the data, both before and after treating the extreme values in the data. The comparison was applied to regression model of a sample of 62 sugary patients on kafr el- shikh university hospital, to study the effect of blood sugar level(x1), high blood pressure(x2), low blood pressure(x3), and weight (x4) on the cumulative glucose rate(y).The study was conductedasfollows:

1. Outliers were detected in the variables, namely the cumulative glucose rate variable, blood sugar level, low blood pressure, and weight, while no outliers were detected in the high blood pressure variable.This was done by relying on the Box plot.
2. The extreme values in the data were treated using the trimmed mean method.
3. The regression model was estimated in the presence of extreme values and after treating them.

It was found that the best regression models before and after treating the data were the quantile regression model, which has the lowest mean squared errors before and after treating the data.

Keywords: outliers, Quantile regression, Ridge regression, OLS, medical data.

1. Introduction:

Ordinary Least Square (OLS) method is one of the most important methods for estimating regression parameters. This method is based on a set of assumptions that may be difficult to obtain, and if they are available, the OLS method becomes the most appropriate for estimating regression parameters. If these assumptions are not available, the OLS method is invalid, and alternative methods must be sought. The violation of these assumptions used violated makes the OLS method biased and inefficient. This happens when random errors are distributed non-normally due to the presence of outliers.

Medical data usually has outliers, so using (OLS) method to estimate regression parameters, Quantile regression, and ridge regression as alternative methods are used to reach to the efficiency one of the regression models in this case.

The issue of outliers and their effect on statistical analysis has been of interest to many researchers: Carcaiso, &Grilli (2023) compared the method, which models the quantile regression coefficients using parametric functions, in place of the conventional method, which involves jittering the count variable utilized on data from university students to assess the impact of emergency remote instruction because of COVID-19 on the number of credits the students obtained. The result showed that the selection of the parametric functions is still guided by the jittering technique. (Li & Leeuwen, (2023) developed connections between dependency-based traditional anomaly detection methods for outliers detection and contextual anomaly detection methods using Quantile Regression Forests. (Muspratt , &Mammadov,(2023)) applied a modified version of anomaly detection algorithm by enacting refined targeting capability based on the identification of sub-extreme anomalies. The result showed that final candidate volumes are controlled with greater accuracy and sensitivity.

Also, (Fernández, et al., (2022)) proposed a new supervised outlier estimator by pipelining an outlier detector with a supervised model to reach the targets of the later supervise how all the hyperparameters involved in the outlier detector are optimally selected. Furthermore, (Xiyujiao,and & Felixpretis (2022) proposed two sets of tests on the presence of outliers in regression models applying to a panel difference-in-differences model of transport CO2 emissions in response to the introduction of North America’s first major carbon tax.. (Abu El-Ela, (2020)) addressed the problems of multicollinearity and outliers using the method of combining the active regression model and ridge regression, by applying it to the pure water stations of the Water Holding Company in Egypt. (Abonazel, (2020) is designed Robust estimation method in case of outliers in the data in linear regression model to the influence of outliers using OLS. AL Rezami (2020), regression analysis was applied to the relationship between Semester average and Cumulative average using algorithm was presented based on the simple and multiple of determination coefficient, and the sum of averages to estimate multiple outliers when outliers are real. Vignotto, &Engelke (2020) proposed two algorithms for anomaly detection relying on approximations from extreme value theory that are more robust and showed the effectiveness of them on real data sets and in simulations. (Affindi, et al., (2019), compared the performance between ridge MM and ridge LTS estimators depending on The Root Mean Square Error (RMSE) and Bias. Ridge regression estimator was suggested to handle severe multicollinearity, Least Trimmed Squares (LTS) estimator and MM estimator are recommended in tackling the outlier issues and when the two problems occur simultaneously.

It is clearly seen that all previous studies agreed with the current study in the necessity of discovering outliers and their impact on the regression models, and they differed with the current paper on treatment methods. The remainder of the paper is divided into the second section, research methodology, the third section, Results and Discussion, and the fourth section, results, and recommendations.

2. Methodology:

The primary goal of this research represents a comparison between the estimate of the multiple regression model using three methods (OLS), (QR), and the (RR) to reach the optimal model, which is the least sensitive in the presence of outliers.

The presence of extreme values can be encountered in three types: in the dependent or independent variables or both. Also, the regression model affects the estimates of the model's parameters and the statistics associated with it (Cousineau & Chartier (2010)). Detecting extreme values is one of the objectives of statistical analysis, and there are several methods for detecting extreme values, in this paper we will depend on Box – Plot graph. which is used to detect outliers in data.

2.1. Treatment Methods:

After detecting the extreme values, they are treated using one of the following methods:

2.1.1. Deletion method:

After identifying outliers in the data, they are excluded or deleted from the data. Rahman & Al Amri (2011) emphasize that outliers are often removed to improve the accuracy of estimation parameters.

2.1.2. Trimmed mean method:

In the Trimmed mean method, the data is sorted in ascending order and the median value is calculated (Júnior, et al (2019)). Then, the outliers are estimated based on their size relative to the rest of the data as follows:

- If the outliers are smaller than the median, the largest value in the data and the outlier to be estimated are deleted, and the mean of the remaining values is calculated, which is an estimate of the outliers.
- If the outliers are larger than the median, the smallest value in the data and the outlier are deleted, and the mean of the remaining values is estimated, which is the estimate of the outliers, and so on.

2.2. Estimation methods:

To estimate the multiple regression model of medical data with outliers, this paper is using three methods (OLS), (QR), and the (RR) to reach the optimal method.

a) Ordinary least squares (OLS)

If the linear regression model is as follows: (Li & Zhu; 2008)

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

where,

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 $\in \approx N(0, \sigma^2)$

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

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

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

$$SSE = (y - xB)^{'}\ (y - xB) \qquad (3)$$

It is well known that the OLS estimator is unstable and more sensitive in the presence of outliers, i.e. in the case where the random error does not follow a normal distribution. In this case, its estimates cannot be relied upon the prediction process.

b) Quantile regression model (QR):

Quantitative regression is one of regression methods analysis and an extension of linear regression in case the conditions of linear regression are not met. Therefore, quantile regression is the best alternative (li & Zhu 2008). It does not assume that the random errors follow a normal distribution, unlike linear models, and it is also not affected by outliers because there are regression lines that pass near these values. Thus, it can be said that it provides a statistical model that is more comprehensive than classical models (OLS). Therefore, quantile regression is one of the robust models.

$$y_i = \grave{X}_\iota\ \mathrm{B}_p + \in_i \qquad i = 1,2, \ldots\ldots\ldots\ldots.n \qquad (4)$$

where;

y_i: dependent variable.

x_i: vector of independent variables.

B_p: The vector of parameters at the quantity (P) where 0<P<1.

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

$$\int_{-\infty}^{0} F_P(\in_i)\ d\in_i = P \quad \text{Or } F_e^{-1}(p) = 0 \qquad (5)$$

and assuming that the distribution of the random error is: (Koenker & Bassett; 1978)

$$F_p(\in_i) = p\ (1 - p)\ \ exp^{\{-\rho_p(\in_i)\}} \qquad (6)$$

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

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

We can show graphically both the OLS, Quantile regression as illustrated in figure (1), and figure (2).

Figure 1 quantile Regression method

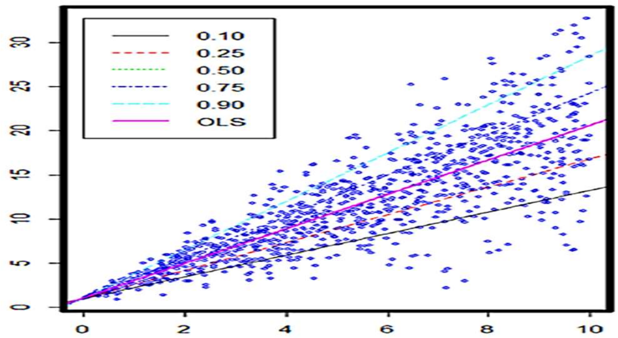
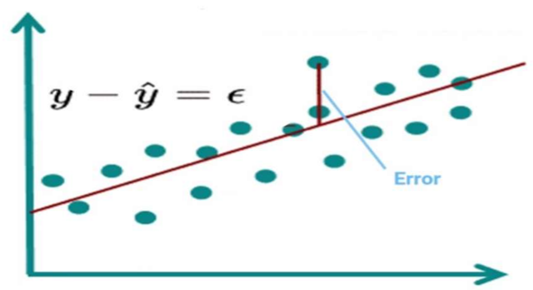


Figure 2 Ordinary Least Square method



From figure (1), we note that the estimated regression line uses the ordinary least squares (OLS) method. However, sometimes cannot provide us with complete information about the relationship between the dependent variable and the independent variables due to one of the reasons that have been mentioned before. While quantile regression gives us a clear picture of the relationship between the dependent variable and the independent variables because many regression lines are estimated at different quantiles as shown in the figure (2) where there are four regression lines estimated by quantile regression (QR) at four different levels.

c) Ridge Regression model (RR):

Ridge regression is one of the specialized methods in multiple regression analysis when it suffers from the problem of multicollinearity. It has shown effective results in eliminating this problem, as the presence of this problem leads to inflation of the variance of the parameters when using the least squares method in estimation. This is the same symptom as the presence of extreme values, and therefore it is expected to do the same task (Lukman, et al. (2014)). The idea of ridge regression is to find the value of the constant (K), which is called the bias parameter, which is a positive quantity added to the main diagonal elements in the matrix (X'X) that leads to reducing the variance of the estimated parameters. Where adding the constant (K) with small values causes a rapid change in the values of the estimated parameters. With the increase of the value of (K), the values of the parameters begin to stabilize gradually until they reach a limit where the change is very small. The ridge regression model estimates the parameters by minimizing the sum of squared errors (SSE) as follows:

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

The shrinkage equation is as follows:

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

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:

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

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

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

Where;

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

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

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

$$\hat{\mathcal{B}}_{RR} = Z_{RR}\, \hat{\mathcal{B}}_{ols} = \left[I_p + k(\grave{x}x)^{-1}\right]^{-1} \quad \hat{\mathcal{B}}_{ols} \qquad (13)$$

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

$$MSE_R = vanance\, (\hat{B}_R) + (bias\ in\ \hat{B}_R)^2 \qquad (14)$$

K: bias parameter.

The larger the value of the parameter K, the greater the amount of bias and the smaller the variance. Therefore, the value of K must be chosen so that the decrease in the value of the variance is greater than the increase in the square of the bias. At that time, the mean squared error of ridge regression is less than the variance of least squares estimates. Also, increasing the value of K reduces the value of the (R^2). Therefore, ridge regression estimates are not necessarily the best fitting model for the data, but they are looking for the best equation with fixed coefficients (unbiased with increasing K) (Alkhamisi; 2007).

3. Results and Discussion:

To compare the performance of three regression models: multiple regression using Ordinary Least Squares (OLS), Quantile Regression (QR), and Ridge Regression (RR) in the presence of extreme values in the data. The data was processed using the Trimmed Mean method to identify the optimal model that works efficiently in the presence of extreme values, i.e., the model that is least sensitive to the effect of extreme values. The comparison was based on several statistical criteria, including the (R^2) and the mean squared error (MSE). The models were applied to a sample of 62 sugary patients were,

the dependent variable (y) represents the average cumulative sugar level, while the independent variables are:

(X_1) blood sugar level,

(X_2) high blood pressure,

(X_3) low blood pressure, and (X_4) weight.

The analysis was conducted using several statistical packages, including Stata 15, Stat graphic 18, and EViews 12.

3.1. Outlier Detection

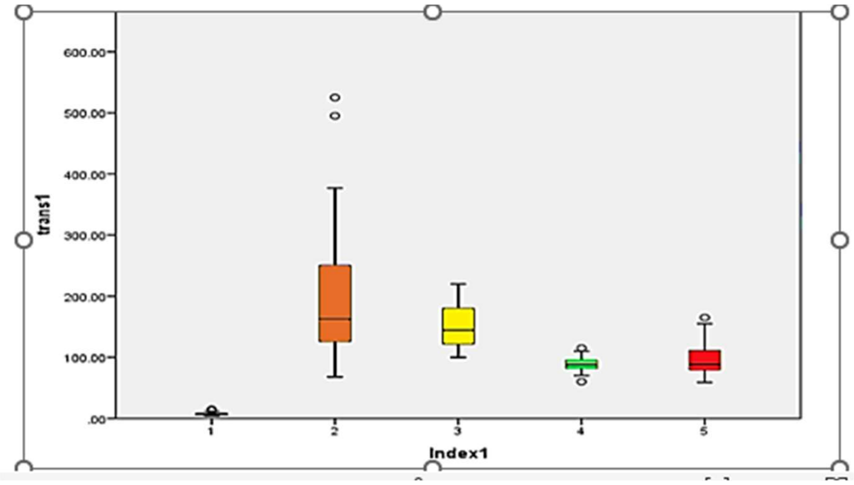
Detecting extreme values in data is one of the first steps in statistical analysis, and this is done using a Boxplot, as illustrated in the following figures.

Through figure (3), and table (1) we observe that both the dependent variable (Y), and the independent variables (X_1, X_3, X_4) have extreme values

Table 1, extreme values in the data variables

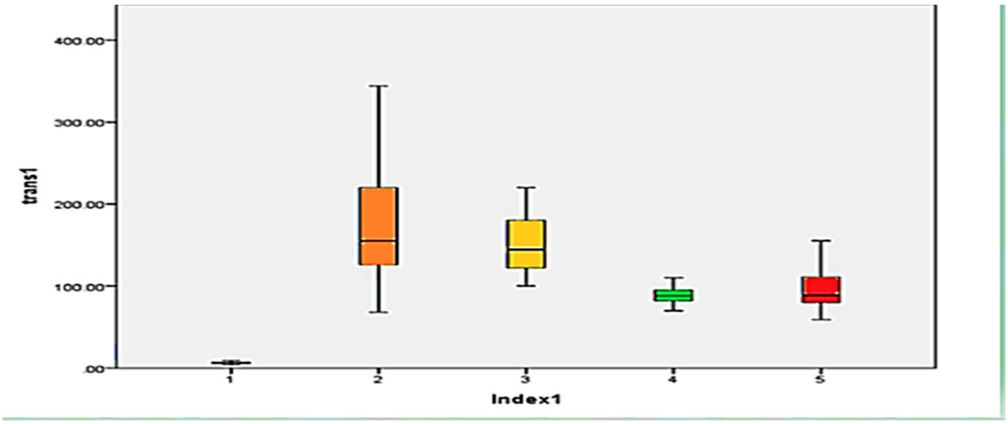
Variables	Outliers
Y(the average cumulative sugar level)	(12 ,13, 39, 40, 41)
X_1 (blood sugar level)	(40, 13)
X_2 (high blood pressure)	No outliers
X_3 (low blood pressure)	(10, 48).
X_4 (weight)	(41)

Figure 3 Data base with outliers



After processing the extreme values using the trimmed mean method, all variables were free of extreme values, as shown in figure (4).

Figure 4: Data base After processing outliers



To confirm the effect of extreme values on the independent and dependent variables, descriptive statistics for the variables were examined before and after processing the extreme values, as shown in table (2).

Table 2: Descriptive statistics

Variables	Mean		Median		S. Deviation	
	with outliers	after processing	with outliers	after processing	with outliers	after processing
Y	7.226	6.8	6.8	6.87	2.49	2.32
X_1	191.967	162.5	162.5	163	96.34	94.33
X_2	152.887	144.5	144.5	144.5	33.32	33.32
X_3	89.081	88	88	88.5	11.41	11.012
X_4	97.323	88.5	88.5	89	24.35	21.12

From table (2), we noticed some changes that occurred in the descriptive statistics of the variables that had extreme values, either by a decrease in the standard deviation or a slight increase in the mean and median.

Estimation Methods of 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	(OLS) with outliers		
	Value	St. error	P- value
β_0	1.285	2.265	0.9195
β_1	1.983	1.698	000
β_2	0.6818	0.04554	0.1397
β_3	0.549	0.1182	0.6440
β_4	1.8029	1.6445	0.007
R^2	0.8355		
MSE	1.094		
F	72.351		0.000
Normality Test (Jb)	0.8447		0.9586
MAE	2.135		
Variables	OLS after processing outliers		
	Value	St. error	P- value
β_0	1.315	2.153	0.067
β_1	2.02	1.376	0.000
β_2	0.8178	.0213	0.091
β_3	0.7176	.0971	0.537
β_4	1.976	1.496	0.000
R^2	0.8517		
MSE	1.073		
F	81.83		000
Normality Test (Jb)	1.725		0.0416
MAE	1.821		

Table (3) shows the Ordinary Least Squares (OLS) estimates before and after processing the variables **with extreme values**, we observed that:

1. in the presence of extreme values, the F-statistic was significant at a level of 0.05 or less, indicating that the null hypothesis was rejected, and the alternative hypothesis was accepted, i.e., the significant regression model.
2. $only, \beta_1, \beta_4$ are significant.

After processing the extreme values,

3. the significance of the model accepted, i.e., the significant regression model, as indicated by the F-value.
4. all parameters are significant except β_3
5. that processing the extreme values led to an increase in the interpretive power of the model from 83.55% to 85.17%, and the mean squared error decreased from 1.094 to 1.073.

6. After testing the normality assumption of the data using the Jarque Bera test before processing the variables with extreme values, the P-value was 0.9586, which was greater than the 5% significance level, indicating that the null hypothesis of non-normality was accepted.
7. the Jarque Bera test values changed to 0.0416, which was less than 0.05, indicating that the null hypothesis was rejected, and the alternative hypothesis was accepted, i.e., the variables followed a normal distribution, as shown in the following figures:

Figure 5 Normality with outlier

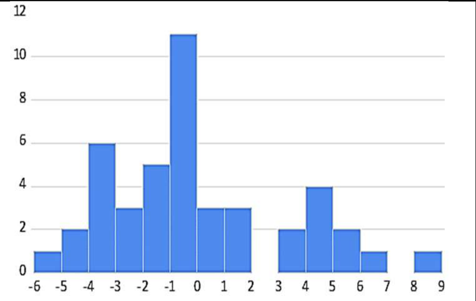
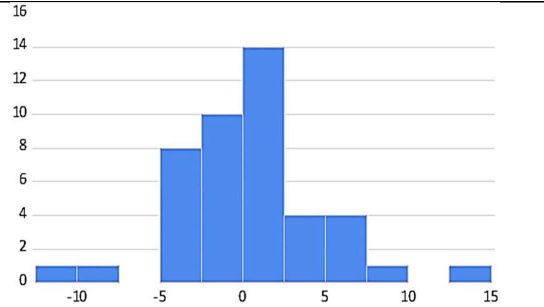


Figure 6 Normality after processing outlier



3.2.2. Ridge regression (RR)

Based on the results obtained using the Ordinary Least Squares (OLS) method, which is misleading in the presence of extreme values, another method called Ridge Regression was used to estimate the model parameters using the Ridge Trace method. This was done to determine the optimal value of the Ridge parameter (K), which ranges from 0 to 1, with an increment of 0.005, that leads to the most stable model with the lowest mean squared error. The analysis was conducted using the statistical package Stat graphic 18, and the optimal value of K was found to be 0.005. The estimation results are as follows:

Table 4: The result of ridge regression:

	(RR)with outliers		
Variables	Value	St. error	P- value
β_0	1.302	2.351	0.2753
β_1	1.972	1.596	0.0012
β_2	0.6819	0.0254	0.076
β_3	0.546	0.1171	0.615
β_4	1.814	1.653	0.001
R^2	83.23%		
MSE	1.045		
MAE	0.7851		
(RR)after processing outliers			
Variables	Value	St. error	P- value
β_0	1.42	2.112	0.231
β_1	2.135	1.302	0.000
β_2	0.765	0.0115	0.0231
β_3	0.549	0.0762	0.520
β_4	2.013	1.3510	0.000
R^2	86.1%		
MSE	1.023		
MAE	0.6135		

From table (4), we observed that:

1. $only, \beta_0, \beta_3$ are not significant in both regression models **with** outliers and after processing outliers.
2. the interpretive power of the Ridge Regression model increased from 83.23% to 86.1% after processing the extreme values, with a decrease in the mean squared error from 1.045 to 1.023.

This indicates the robustness of the Ridge Regression model to extreme values. Additionally, some variables showed high significance in the Ridge Regression model, while they were insignificant in the Ordinary Least Squares method.

3.2.3. Quantile Regression (QR):

Using data with extreme values can lead to significant errors in data analysis when using traditional methods such as Ordinary Least Squares (OLS). Therefore, Quantile Regression (QR) was used to analyze this relationship by testing quantiles equal to p=.50 and p=.95. The reason for this test is that the data is always centered around the median, i.e., when p=.50, while the other value, which is the extreme quantile value, is when p=.95. We chose this value because if the regression line is good at this value, it is also good at quantile values close to it, i.e., the regression line at any other quantile, whether p=.25 or p=.75, is a good line for prediction and estimation.

Table 5 :The results of quantile Regression

Variables	**QR** with outliers		
	Value	St. error	P- value
β_0	0.6625	1.644	0.607
β_1	2.135	0.025	0.000
β_2	0.6397	0.0672	0.43
β_3	0.0291	0.1613	0.085
β_4	1.853	0.1113	0.0109
R^2	0.7915		
MSE	1.023		
MAE	1.012		
Normality Test	0.4108		0.8144
Variables	**(QR)**after processing outliers		
	Value	St. error	P- value
β_0	1.231	1.35	0.532
β_1	2.416	0.0179	0.000
β_2	0.842	0.0524	0.2215
β_3	0.3512	0.1529	0.0462
β_4	2.0146	0.1013	0.0215
R^2	0.8214		
MSE	0.9875		
MAE	0.9012		
Normality Test	12.15		0.002

From table (5), we observed that:

1. The Quantile Regression model achieved an interpretive power of 79.15% in the presence of extreme values, and after processing the extreme values, the interpretive power of the model increased to 82.14%.
2. the mean squared error decreased from 1.023 to 0.9857.
3. $only, \beta_1, \beta_4$ are significant in regression models **with** outliers and after processing outliers$\beta_1, \beta_4, \beta_3$ are significant. It means that some variables showed high significance after processing the extreme values.
4. The normality assumption of the data was tested using the Jarque Bera test in the presence of extreme values, where the P-value was 0.8144, which was greater than the 5% significance level, indicating that the null hypothesis of non-normality was accepted.
5. After processing the extreme values, the P-value was 0.002, which was less than the 5% significance level, indicating that the null hypothesis was rejected and the alternative hypothesis was accepted, i.e., the data followed a normal distribution.

Therefore, the model became more stable after processing the data from the extreme values.

3.2.4. Comparison of Models Estimations

To identify the optimal and least sensitive model that works efficiently in the presence of the problem of extreme values, the following comparison was made:

Table 6: Comparison of Estimation Methods

	Data base with outliers		
	OLS	RR	QR
R^2	0.8355	0.8323	0.7915
MSE	1.094	1.045	1.023
MAE	1.085	1.07851	1.012
significant Variables	X1, x4	X1, x4	X1, x4
Variables	after processing outliers		
	OLS	RR	QR
R^2	0.8517 (1.94%)	0.8618 (3.54%)	0.8214 (3.77%)
MSE	1.073 (-1.92%)	1.023 (-2.1%)	0.9875 (-3.47%)
MAE	1.072 (-1.19%)	1.06735 (-1.035%)	0.9072 (-10.35%)
significant Variables	X1, x4, x3	X1, x4, x2	X1, x4, x3

Based on the output of Table (6), we found that:

1. In the presence of outliers, Although, the QR regression has the smallest R^2, it also has the smallest MSE and MAE. So, it is the best method estimation in this case.
2. After treating outliers, the superiority of the QR regression model was evident, as it was found to have the largest increase in the R^2 after treating the extreme values, and the highest percentage decrease in the mean squared errors, followed by the ridge regression model. It means that is the best model for outliers' data.

4. Conclusion and Recommendations:

The research aimed to compare several regression models, including the multiple regression model using ordinary least squares (OLS), the quantile regression model (QR), and the ridge regression model (RR), to identify the model with high efficiency in the presence of extreme values in the data, both before and after treating the extreme values in the data. The study was conducted as follows:

1. Outliers were detected in the variables, namely the cumulative glucose rate variable, blood sugar level, low blood pressure, and weight, while no outliers were detected in the high blood pressure variable. This was done by relying on the Box plot.
2. The extreme values in the data were treated using the trimmed mean method.
3. The regression model was estimated in the presence of extreme values and after treating them. It was found that the best regression models before and after treating the data were the quantile regression model, which has the lowest mean squared errors before and after treating the data.

It was also the model with the highest percentage increase in the coefficient of determination and the highest percentage decrease in the mean squared errors, followed by the ridge regression model.

Therefore, the researchers recommend:

1. The quantile regression model should be used as one of the models that are preferred to be used in the presence of the problem of extreme values especially on the medical data.
2. Assumptions of regression analysis should be verified before using regression to obtain more accurate results.
3. It is necessary to pay attention to the process of estimating extreme values in future research studies without trying to delete those values, due to their importance and impact on the accuracy of the results.

References:

Abo El-Ala, H. H., (2020). “A statistical model for predicting the quantities of production and consumption of pure water in water stations affiliated with the Holding Company in Egypt using the integration method between the active regression and ridge regression models”. Scientific Journal of Economics and Trade, Vol Vol. 50(2), pp. 225-266.

Abonazel M. R., (2020).” Handling Outliers and Missing Data in Regression Models Using R: Simulation Examples”. Academic Journal of Applied Mathematical Sciences Vol. 6, pp. 187-203. https://arpgweb.com/journal/journal/17

Affindi A. N., Ahmad S. & Mohamad M., (2019). “A Comparative Study between Ridge MM and Ridge Least Trimmed Squares Estimators in Handling Multicollinearity and Outliers”. Journal of Physics: Conference Series Vol.1366.

doi:10.1088/1742-6596/1366/1/012113

AL Rezami A. Y., (2020) “Effect of outliers on the coefficient of determination in multiple regression analysis with the application on the GPA for student”. International Journal of Advanced and Applied Sciences. Vol .7(10), Pp. 30-37.

Alkhamisi, M. A., & Shukur, G., (2007). “A Monte Carlo study of recent ridge parameters”. Communications in Statistics-Simulation and Computation, Vol (36(3)), pp (535-547). https://doi.org/10.1080/03610910701208619

Carcaiso V. , & Grilli L.,(2023).” Quantile regression for count data: jittering versus regression coefficients modelling in the analysis of credits earned by university students after remote teaching”, Statistical Methods & Applications), Vol. 32, pp .1061–1082

Cousineau D. & Chartier S., (2010). “Outliers detection and treatment: a review”. International Journal of Psychological Research, Vol. (3-l), pp (59-68).

Fernández Á., &Bella J., Dorronsoro J.R., (2022). “Supervised outlier detection for classification and regression”. Neurocomputing, vol .486, pp.77-92. https://doi.org/10.1016/j.neucom.2022.02.047

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E; Tatham, R.L. (20*10*). Multivariate data analysis. (*7*end edition). Pearson Education, New Jersey.

Hideo, K., & Genya .K.,(2011) .“Gibbs sampling methods for Bayesian quantile regression”. Journal of Statistical Computation and Simulation, Vol. 81 (11), pp. 1565-1578. https://doi.org/10.1080/00949655.2010.496117

Júnior, A.C. A., Souza, P. D. D., Assis, A. L. D., Cabacinha, C. D., Leite, H. G., Soares, C. P. B., ... & Castro, R. V. O. (2019). “Artificial neural networks, quantile regression, and linear regression for site index prediction in the presence of outliers”. Pesquisa Agropecuária Brasileira, Vol .54.

doi: 10.1590/S1678-3921.pab2019.v54.00078.

Koenker, R., & Bassett, G. (1978). “Regression quantiles”. Econometrica, Vol. 46(1), pp.33-50.

Li, Y., & Zhu, J. (2008). “L 1-norm quantile regression”. Journal of Computational and Graphical Statistics, Vol 17(1), pp (163-185).

Li,Z. & Leeuwen M.V., (2023).” Explainable contextual anomaly detection using quantile regression forests .’Data Mining and Knowledge Discovery, Vol .37, pp.2517–2563. https://doi.org/10.1007/s10618-023-00967-z

Lukman, A., Arowolo, O., & Ayinde, K. (2014). “Some robust ridge regression for handling multicollinearity and outlier”. International Journal of Sciences: Basic and Applied Research, Vol.16(2), pp.192-202.

Muspratt R., & Mammadov, M., (2023).” Anomaly Detection with Sub-Extreme Values: Health Provider Billing”. Data Science and Engineering. https://doi.org/10.1007/s41019-023-00234-7

Rahman, M. S., & Amri, K. A. (2011). Effect of outlier on coefficient of determination. International Journal of Education Research, Vol. 6(1), pp. 9-20.

Vignotto E., &Engelke S., (2020).” Extreme value theory for anomaly detection – the GPD classifier “. Extremes, Vol (23), pp. 501–520.https://doi.org/10.1007/s10687-020-00393-0

Xiyujiao, & Felixpretis,(2022).” Testing the Presence of Outliers in Regression Models” oxford bulletin of economics and statistics, Vol .84(6). doi: 10.1111/obes.12511.