Int. J. Nonlinear Anal. Appl. 13 (2022) 2, 1325–1333
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6255

New estimation method to reduce the high leverage points effect in quantile regression

Mohammad Abdul Kareem*, Taha Alshaybawee

Department of Statistics, Faculty of Administration and Economics, University of Al-Qadisyiah, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

Quantile regression is a powerful statistical method for modeling and analyzing the impact of explanatory and response variables at different points in the conditional distribution of the response variable. Many research papers have indicated that Quantile Regression (QR) estimator is only resistant to vertical outliers. Quantile regression like other regression M-estimators and Least Absolute Deviation LAD can be very sensitive to outliers in explanatory variables (Leverage Points). To overcome this drawback, at first, we have to use a robust, effective and efficient method to identify high leverage points if there is masking and swamping problems. In literature, the usage of Generalized M-estimator (GM-estimator) is proposed to estimate the unknown parameters against high leverage points. In this paper, we proposed weighted method's the generalized- M for quantile regression namely (GMQu), and improve the algorithm of this method by adapting the Improved Diagnostic Robust Generalized Potential (IDRGP) method. So that the calculation of the initial weights in this algorithm depends on (IDRGP), we're going to symbolize that method by (GMQuID). Simulation study and real data are considered to verify the performance of our proposed methods compared to other methods.

Keywords: weighted quantile regression, high leverage points, GMQu, GMQu (IDRGP).
2020 MSC: 62G08

1 Introduction

Quantile regression (QR) model has been introduced by [22] as an extension from the notion of ordinary quantiles in a location model to a more general class of linear models in which the conditional quantiles have a linear form. Quantile regression model have been successfully used in a wide range of scientific applications, for instance: Economics, Biology, Ecology and Finance. It can reveal relationships between model variables that are difficult to be captured by a traditional mean regression. One of the most virtues of quantile regression is that it allows us to make inference on the entire conditional distribution of response by estimating a number of different quantiles. Also, these estimators can resist the damaging effect of outlier's observations in y- direction. In addition, QR does not impose any distributional assumption on the error except the requirement about the zero conditional quantiles [21]. Considers the following

*Corresponding author
Email addresses: stat.post06@qu.edu.iq (Mohammad Abdul Kareem), sirtaha12@qu.edu.iq (Taha Alshaybawee)

Received: November 2021 *Accepted:* January 2022

regression model:

$$y_i = x_i^t \beta + \zeta_i \qquad i \ = \ 1, 2, ..., \ n \tag{1.1}$$

$$\widehat{\beta}_\tau = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^t \beta \right) \tag{1.2}$$

where $\rho_\tau(u) = u\left(\tau - I_{\{u<0\}}\right), I_{\{A\}}$ is the indicator function of the event A, $\rho_\tau(u)$ is a quantile loss function it can be defined as

$$\rho_\tau(u) = \begin{cases} (\tau - 1)u & \text{if } u < 0 \\ \tau u & \text{if } u \geq 0 \end{cases} \tag{1.3}$$

Quantile regression estimators can be highly sensitive to outliers with high leverage [18]. Some studies dealing with relevant on this issue for example see [15, 7, 2, 31, 28].

The crucial point here is that the leverage points in multiple regression data is very hard to identify when the number of covariates exceeds tow. The classical Mahalanobis Distance (MD) measure was most the familiar choice to identify the leverage points. The MD is suffering from masking problem [35], so it is a non robust measure. Hat matrix which is traditionally used as a measure of leverage points in regression analysis. However, [4] pointed out that the hat matrix may fail to identify the HLPs due to the effect of HLPs in leverage structure. So he introduced potential measure. Still this method was unable to detect all of the HLPs. [33] introduced another diagnostic tool which is called Generalized Potential (GP).Recently, [8], developed Diagnostic Robust Generalized Potential (DRGP) to determine outlying points in multivariate data sets. The mine weakness of (DRGP) is that it has small rate of masking and swamping effects, especially for small percentage of high leverage points between (5% and 10%). [30] improve the performance of DRGP (MVE) called (IDRGP) by adding new step pertaining two cases to algorithm of (DRGP).

The rest of this paper is organized as follows, section 2, Generalized M- Estimator, Section 3, proposed methods and algorithms, In Section 4, the simulation study has been done to assess the performance of our proposed method compared to other methods, Section 5 real data and some results, Section 6 conclusion of this research.

2 Generalized M- Estimator

[10] showed that the M-estimator does not have Bounded Influence Function (BIF), because it fails to account for leverages, and also implicitly assumes that the model matrix (X) is measured without errors, to overcome these drawbacks of M-estimator. Generalized M- Estimator (GM-estimator) which is usually called bounded influence estimation had proposed by Schweppe [13, 3]. In general, GM-estimator ofβ in regression model is to produce weights that consider both Y and X-direction. The general form is given by

$$\sum_{i=1}^{n} \pi_i \psi \left(\frac{y_i - {x_i}^t \widehat{\beta}}{\pi_i} s \right) x_i = 0, \tag{2.1}$$

where, $\pi_i, \ i = 1, \ldots, n$ is initial weight controls weights that given for leverages. Equation (2.1) can be solved by using IRLS technique, and then the GM convergence can be written as

$$\widehat{\beta}_{\mathrm{GM}} = \left(x^t \mathrm{wx} \right)^{-1} x^t wy \tag{2.2}$$

$$\omega_i = \frac{\psi \left(\frac{y_i - {x_i}^t \widehat{\beta}_{\mathrm{GM}}}{\pi_i s} \right)}{\left(\frac{y_i - {x_i}^t \widehat{\beta}_{\mathrm{GM}}}{\pi_i s} \right)} \tag{2.3}$$

3 Improved Diagnostic Robust Generalized Potential (IDRGP)

[30] He improved the performance of DRGP (MVE). He notices that there is an impact of swamping and masking cases when the percentage of high leverage points is between 5% and 10%. Therefore [30] suggested to adding a new diagnostic step to second algorithm step of DRGP .The algorithm of IDRGP can be summarized as follows

Step1: computing the location and scale estimators by using MVE.

Step2: Based on MVE calculate Robust Mahalanobis Distance RMD. Any i-th exceed Median $(\mathrm{RMD}_i)+c\times MAD\,(\mathrm{RMD}_i)$ $i=1,2,\ldots,n$ that means the i-th row having the suspected observation as HLP.

Step3: the row i-th which diagnosed from the previous step will be deleted from design matrix X and putting in a new sub-matrix denoted as X_D. The remaining rows that are diagnosed as clean will be putting in sub- matrix denoted as X_R .

Step4: constructing weights matrix as follows

$$w=\begin{bmatrix} U_R & V \\ V^t & U_D \end{bmatrix}$$

When D rows are omitted, the ${w_{ii}}^{(D)}$ is the i-th diagonal elements of

$${x_i}^t\left({x^t}_R x_R\right)^{-1}x_i\ ,i=1,2,\ldots,n$$

and $R=(N-D)\times p$. Deletion the i-th diagonal elements from x_R makes $R=(N-1)\times p$, in this case ${w_{ii}}^{(-i)}$will be a single diagnostic equivalent to potential measure ${w_{ii}}^{(-i)}={x_i}^t\left({x^t}_{(i)}x_{(i)}\right)^{-1}x_i$.

The group deletion measure can be written as follows

$$p_{ii}=\begin{cases} {w_{ii}}^{(D)} & for\ i\in D \\ \frac{{w_{ii}}^{(D)}}{1-{w_{ii}}^{(D)}} & for\ i\in R \end{cases}$$

When$p_{ii}>median\ (p_{ii})+c\ \times mad(p_{ii})$ is confirmed i-th row having HLP, to improve DRGP [30] adding a further step to algorithm through the diagnostic of LPs by using hat matrix and then compared with the first diagnosis.

1. If the observations diagnosed as HLPs are the same as $'D_2'$ and $'D'$ thus the algorithm will the announcement of this diagnosis and then stop.
2. If the number of HLPs in $'D_2'$ are more than those in 'D', then the algorithm works on move those observations that are not matched with 'D' to the matrix R one by one according to P_{ii} value. If the value of P_{ii} for certain observation exceeds the cutoff point $(P_{ii}+3P_{ii})$, stay in $'D_2'$ matrix, otherwise it move to R matrix.
3. If the number of the HLP in D_2 is less than the number of what is diagnosed in D which means new observations that have not been diagnosed before, that the algorithm works on merging between D and D_2. Re-checking the suspected observations by using the generalized potential measure P_{ii} is crucial to confirm whether these observations are HLP or clean. So, clean observation should be moved to R matrix.

4 Proposed methods

GM-estimator is in the definition of π_i –weight that depended on hat matrix which could fail to identify high leverage point due to the effect of these points in leverage structure. This problem can be alleviated by assigning weights to each term of Eq.(1.2) that is decreasing functions of their leverage [5]. [6] pointed out that down-weighting the leverage points can improve the conditional breakdown point of LAD regression (special case of QR when $\tau=0.5$). [7] proposed a weighted LAD estimator (special case of QR when $\tau=0.5$)

$$\widehat{\beta}_{\mathrm{LAD}}=\arg\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}\pi_i\left|y_i-x_i^t\beta\right|, \tag{4.1}$$

From (4.1), the weighted quantile regression estimators can be obtained by minimizing the problem

$$\widehat{\beta}_{\tau}^{\pi}=\arg\min_{\beta\in\mathbb{R}^p}\left[\tau\sum_{y_i\geq x_i^t\beta}\pi_i\left|y_i-x_i^t\beta\right|+(1-\tau)\sum_{y_i\leq x_i^t\beta}\pi_i\left|y_i-x_i^t\beta\right|\right] \tag{4.2}$$

The minimization of Equation (4.2) can be quite complex because it is non-differentiable. [21] suggested using computational algorithm based on linear programming techniques. In order to minimize Equation (4.2) via linear programming, the model in (1.1) it can be written as

$$y_i \ = x_i^t\beta_\tau + (u_{i\tau} - v_{i\tau}\ )\big|_{\zeta_{i\tau}=u_{i\tau}-v_{i\tau}} \tag{4.3}$$

where $u_{i\tau} = |\zeta_{i\tau}|\, I(\zeta_{i\tau} > o)$ and $v_{i\tau} = |\zeta_{i\tau}|\, I(\zeta_{i\tau} < o)$. Therefore the linear programming can be show

$$\widehat{\beta}_\tau = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \pi_i \rho_\tau \left(y_i - x_i^t\beta\right) \iff \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \pi_i \left(\tau u_{i\tau} + (1-\tau)\, v_{i\tau}\right) \tag{4.4}$$

$\widehat{\beta}_\tau$ free, $u > 0$, $v > 0$. π_i is the weigh and set to be $0 < \pi_i \leq 1$. From (4.3), we can see that the difference of two non-negative residuals$(u_{i\tau} - v_{i\tau}\ ) = y_i - x_i^t\widehat{\beta}_\tau$, then the weight of the loss function will be equal to

$$\sum_{i=1}^{n} \pi_i \left(\tau u_{i\tau} + (1-\tau) v_{i\tau}\right) = \tau \sum_{(i/y_i \geq x_i^t\widehat{\beta}_\tau)} \pi_i \left(y_i - x_i^t\widehat{\beta}_\tau\right) + (1-\tau) \sum_{(i/y_i < x_i^t\widehat{\beta}_\tau)} \pi_i \left(y_i - x_i^t\widehat{\beta}_\tau\right). \tag{4.5}$$

In This study, we will propose two new methods based one GM-estimator method. The first proposed method GM-estimator method will be modified for the quantile regression. The algorithm of GMQu is summarized as follows

Step 1: Calculate the initial weight π_i by applying the form $\pi_i = \sqrt{1 - h_{ii}}$,

Step 2: Compute quantile regression estimate $\widehat{\beta}_\tau = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^\top \beta\right)$ to use as initial estimates.

Step 3: Compute the residuals r_i and scale from the form $S = 1.4826$ [median largest $(n-p)$ of $|r_i|$] based on step 1.and then compute standardized residuals (t_i) . where $t_i = \frac{r-i}{\pi_i \times s}$.

Step 4: Using standardized residuals t_i in first iteration (weighted quantile regression), then ω_i are chosen based on Huber function.

Step 5: Update $\widehat{\beta}_\tau$ using weighted least squares with the weightsω_i.

Step 6: Steps (3-5) are repeated until convergence.

In the second proposed method the precision of GM-estimator can be improved by utilizing more effective diagnostic method. This motivates us to consider the IDRGP which is proposed by [30] for identifying the HLPs. The weights will be calculated based on IDRGP from the following form:

$$\pi_i = min\left\{1, \frac{\text{median}\,(p_{ii}) + 3MAD(p_{ii})}{p_{ii}}\right\}.$$

5 Simulation study

Monte Carlo simulation example is considered in this section to evaluation and compare the performance the proposed methods Generalized- M for quantile regression (GMQu) and generalized M based on IDRGP to the existing methods classical quantile regression QR [22] and least trimmed quantile regression LTQR [31]. The following model was considered to generated data

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

where, $\beta_0 = 2,\ \beta_1 = 1,\ \beta_2 = 0,\ \beta_3 = 1.7,$ three predictor's variables are considered $p = 3$ the covariates $x_{\text{ij}},\ j = 3$ are generated from uniform distribution $U\,(-1, 1)$ and the error term generated from normal distribution $\varepsilon_i \sim N\,(0, 3)$. Two sample sizes are used (50 and 100). We have contaminated the generated data by replacing the first k observations, where k compute from $n \times \epsilon$, where n is a sample size and ϵ is the contamination rate, (0.10, 0.20 and 0.30) is the contamination level used in this simulation study. The first k observations of by using uniform distribution $U\,(-30, 30)$ .We have used five levels of quantiles $\tau \in (0.10,\ 0.30,\ 0.50,\ 0.70,\ 0.90)$ to estimate the parameters vector β_τ.

To compare the performance of all used methods existing and proposed, three measurements are considered bias, mean squares error MSE and mean absolute error MAE. These measures are computed as follows:

$$Bias = \overline{\beta}_j - \beta_j^{\text{true}}, \quad where\ \overline{\beta}_j = \frac{1}{m}\sum_{r=1}^{m}\widehat{\beta}_j^{(r)}$$

$$MSE = \frac{1}{n}\sum_{r=1}^{n}(y_i - \widehat{y}_i)^2$$

$$MAE = \frac{1}{n}\sum_{r=1}^{n}|y_i - \widehat{y}_i|$$

Where, the replication m is used 100 and β_j^{true} is the true parameter. To compute QR, LTQR, GMQu and GMQuID R code is used.

In the following tables we will summarize the results of the simulation study. In Table (1) and Table (2), we reported the bias for the existing and proposed methods when the sample sizes 50 and 100. In these two tables, we see the bias at three different contamination rates (10%, 20% and 30%) and the five levels of quantiles (0.10, 0.30, 0.50, 0.70, and 0.90).

Table 1: Bias values for QR, LTQR, GMQR and GMQuID methods at five quantiles (0.10,0.30,0.50,0.70,0.90) and different contamination levels 10%,20% and 30% when the sample size $n = 50$.

		Contamination (10%)				*Contamination (20%)*				*Contamination (30%)*			
Quantile	*Method*	*Bias* β_0	*Bias* β_1	*Bias* β_2	*Bias* β_3	*Bias* β_0	*Bias* β_1	*Bias* β_2	*Bias* β_3	*Bias* β_0	*Bias* β_1	*Bias* β_2	*Bias* β_3
0.10	*Qu*	*2.007*	*0.272*	*0.674*	*2.555*	*2.391*	*0.176*	*0.257*	*2.562*	*2.298*	*0.166*	*0.173*	*2.573*
	LTQRe	*2.236*	*0.248*	*0.149*	*2.523*	*2.962*	*0.288*	*0.414*	*2.455*	*2.843*	*0.129*	*0.461*	*2.634*
	GM-Qu	*1.375*	*0.259*	*0.759*	*2.553*	*1.542*	*0.122*	*0.072*	*2.619*	*1.445*	*0.206*	*0.023*	*2.542*
	GMQuID	*0.808*	*0.005*	*0.003*	*0.033*	*0.717*	*0.023*	*0.019*	*0.002*	*0.728*	*0.025*	*0.134*	*0.020*
0.30	*Qu*	*0.907*	*0.273*	*0.456*	*2.501*	*1.053*	*0.061*	*0.408*	*2.662*	*0.849*	*0.272*	*0.371*	*2.464*
	LTQRe	*1.342*	*0.297*	*0.621*	*2.371*	*1.441*	*0.073*	*0.359*	*2.657*	*1.211*	*0.273*	*0.400*	*2.466*
	GM-Qu	*0.584*	*0.246*	*0.503*	*2.521*	*0.653*	*0.088*	*0.011*	*2.645*	*0.503*	*0.237*	*0.376*	*2.499*
	GMQuID	*0.336*	*0.002*	*0.043*	*0.035*	*0.307*	*0.001*	*0.037*	*0.042*	*0.285*	*0.001*	*0.041*	*0.046*
0.50	*Qu*	*0.086*	*0.249*	*0.325*	*2.484*	*0.086*	*0.031*	*0.003*	*2.686*	*0.191*	*0.247*	*0.559*	*2.487*
	LTQRe	*0.053*	*0.211*	*0.139*	*2.508*	*0.092*	*0.078*	*0.063*	*2.638*	*0.320*	*0.250*	*0.685*	*2.483*
	GM-Qu	*0.027*	*0.204*	*0.284*	*2.532*	*0.027*	*0.061*	*0.031*	*2.664*	*0.126*	*0.219*	*0.503*	*2.512*
	GMQuID	*0.003*	*0.016*	*0.015*	*0.044*	*0.028*	*0.004*	*0.024*	*0.006*	*0.011*	*0.001*	*0.009*	*0.021*
0.70	*Qu*	*0.909*	*0.207*	*0.048*	*2.514*	*1.031*	*0.021*	*0.274*	*2.692*	*1.279*	*0.167*	*0.620*	*2.563*
	LTQRe	*1.165*	*0.168*	*0.116*	*2.495*	*1.492*	*0.008*	*0.266*	*2.711*	*1.682*	*0.198*	*0.614*	*2.530*
	GM-Qu	*0.663*	*0.171*	*0.087*	*2.548*	*0.624*	*0.044*	*0.162*	*2.675*	*0.834*	*0.194*	*0.544*	*2.537*
	GMQuID	*0.298*	*0.013*	*0.050*	*0.051*	*0.309*	*0.016*	*0.065*	*0.063*	*0.356*	*0.012*	*0.005*	*0.086*
0.90	*Qu*	*2.174*	*0.113*	*0.030*	*2.574*	*2.240*	*0.080*	*0.678*	*2.630*	*2.400*	*0.097*	*0.636*	*2.621*
	LTQRe	*2.497*	*0.127*	*0.036*	*2.594*	*2.687*	*0.096*	*0.761*	*2.621*	*2.596*	*0.143*	*0.635*	*2.561*
	GM-Qu	*1.465*	*0.137*	*0.006*	*2.556*	*1.508*	*0.066*	*0.444*	*2.653*	*1.628*	*0.154*	*0.531*	*2.567*
	GMQuID	*0.747*	*0.020*	*0.029*	*0.125*	*0.814*	*0.003*	*0.029*	*0.052*	*0.799*	*0.005*	*0.080*	*0.041*

From Table 1 and Table 2, we can see that the at different levels of contamination and at all quantiles was the highest for QR and LTQR methods. That means QR method was so affected by the high leverage point observations. Whereas LTQR is a robust method but it cannot treat the effect of the high leverage points. The proposed methods GMQu and GMQuID get the smallest values of . Therefore, these two methods overcome the effects of the high leverage point observations. In addition, GMQuID has a smaller than GMQu method, its remedy the influence of high leverage point better than GMQu method.

In Figure 1 and Figure 2, we summarize the MSE and MAE values for the proposed and existing methods. It is clear that the LTQR method is getting the largest values of MSE and MAE especially in the extreme quantiles levels. MSE and MAE for the QR method is also large but it's smaller than LTQR method. The proposed method GMQuID gets the smallest values of MSE and MAE at all the different quantile levels and the contamination rates. The GMQu method was better than the existing methods, where the values of MSE and MAE for this method were smaller than that for the existing methods.

Table 2: *values for QR, LTQR, GMQR and GMQuID at five quantiles (0.10,0.30,0.50,0.70,0.90) and different contamination levels 10%,20% and 30% when the sample size* $n = 100$.

Quantile	*Method*	*Contamination (10%)*				*Contamination (20%)*				*Contamination (30%)*			
		Bias β_0	*Bias* β_1	*Bias* β_2	*Bias* β_3	*Bias* β_0	*Bias* β_1	*Bias* β_2	*Bias* β_3	*Bias* β_0	*Bias* β_1	*Bias* β_2	*Bias* β_3
0.10	*Qu*	*2.338*	*0.172*	*0.247*	*2.547*	*2.617*	*0.235*	*0.262*	*2.500*	*2.184*	*0.169*	*1.036*	*2.567*
	LTQRe	*2.836*	*0.070*	*0.005*	*2.674*	*3.202*	*0.177*	*0.280*	*2.585*	*2.648*	*0.146*	*1.042*	*2.611*
	GM-Qu	*1.538*	*0.183*	*0.327*	*2.549*	*1.801*	*0.251*	*0.212*	*2.486*	*1.392*	*0.205*	*0.860*	*2.532*
	GMQuID	*0.857*	*0.012*	*0.073*	*0.029*	*0.841*	*0.006*	*0.054*	*0.013*	*0.818*	*0.017*	*0.002*	*0.050*
0.30	*Qu*	*0.957*	*0.252*	*0.169*	*2.469*	*1.241*	*0.321*	*0.089*	*2.383*	*0.773*	*0.241*	*0.914*	*2.485*
	LTQRe	*1.537*	*0.265*	*0.158*	*2.447*	*1.853*	*0.369*	*0.242*	*2.327*	*1.353*	*0.268*	*1.278*	*2.458*
	GM-Qu	*0.574*	*0.216*	*0.228*	*2.503*	*0.846*	*0.277*	*0.163*	*2.438*	*0.459*	*0.214*	*0.694*	*2.513*
	GMQuID	*0.329*	*0.004*	*0.018*	*0.035*	*0.316*	*0.002*	*0.003*	*0.011*	*0.288*	*0.013*	*0.020*	*0.004*
0.50	*Qu*	*0.094*	*0.277*	*0.341*	*2.430*	*0.186*	*0.317*	*0.232*	*2.383*	*0.252*	*0.268*	*0.471*	*2.451*
	LTQRe	*0.111*	*0.310*	*0.503*	*2.401*	*0.093*	*0.318*	*0.264*	*2.381*	*0.339*	*0.305*	*0.516*	*2.414*
	GM-Qu	*0.107*	*0.223*	*0.247*	*2.489*	*0.194*	*0.281*	*0.196*	*2.426*	*0.184*	*0.238*	*0.543*	*2.483*
	GMQuID	*0.017*	*0.007*	*0.024*	*0.064*	*0.004*	*0.007*	*0.017*	*0.003*	*0.021*	*0.006*	*0.002*	*0.019*
0.70	*Qu*	*1.141*	*0.224*	*0.165*	*2.462*	*0.897*	*0.266*	*0.317*	*2.435*	*1.150*	*0.226*	*0.350*	*2.490*
	LTQRe	*1.902*	*0.157*	*0.056*	*2.538*	*1.540*	*0.247*	*0.307*	*2.457*	*1.735*	*0.195*	*0.262*	*2.523*
	GM-Qu	*0.762*	*0.219*	*0.203*	*2.477*	*0.468*	*0.264*	*0.294*	*2.439*	*0.824*	*0.218*	*0.407*	*2.500*
	GMQuID	*0.331*	*0.004*	*0.029*	*0.005*	*0.326*	*0.003*	*0.032*	*0.017*	*0.344*	*0.003*	*0.001*	*0.001*
0.90	*Qu*	*2.381*	*0.154*	*0.016*	*2.525*	*2.152*	*0.198*	*0.341*	*2.506*	*2.325*	*0.183*	*0.252*	*2.529*
	LTQRe	*2.921*	*0.085*	*0.141*	*2.608*	*2.711*	*0.141*	*0.357*	*2.566*	*2.799*	*0.136*	*0.178*	*2.575*
	GM-Qu	*1.638*	*0.184*	*0.067*	*2.506*	*1.376*	*0.233*	*0.270*	*2.473*	*1.650*	*0.199*	*0.290*	*2.517*
	GMQuID	*0.802*	*0.007*	*0.051*	*0.007*	*0.810*	*0.007*	*0.009*	*0.030*	*0.822*	*0.009*	*0.030*	*0.027*

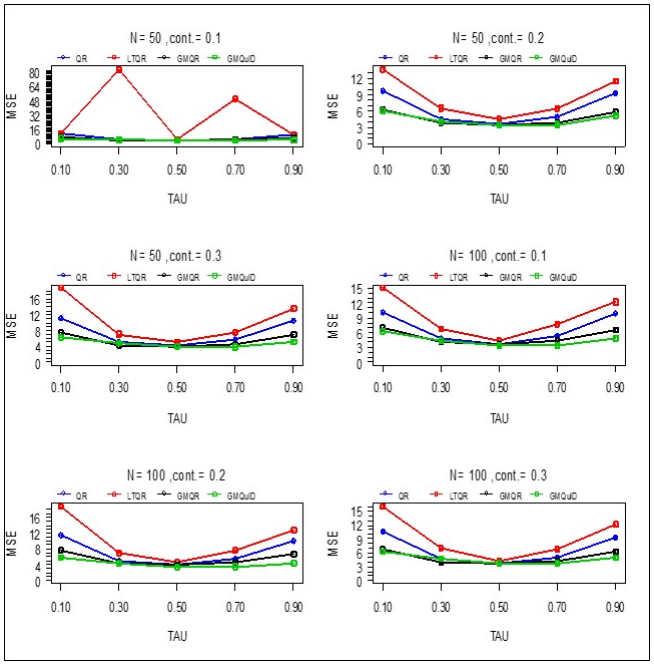

Figure 1: show the MSE for QR, LTQR, GMQu and GMQuID methods at five quantiles (0.10,0.30,0.50,0.70, 0.90) and three contamination levels, when $N = (50, 100)$.

6 Real data Example

To verify the performance of the existing and proposed methods, Star Cluster CYG OB1 dataset is used. This dataset was presented by [24]. It is consists of 47 observations and two variables, the dependent variables (the logarithm of light intensity) whereas the explanatory variable shows the logarithm of the effective temperature at the surface of stars. The scatterplot of these data was showing two groups of observations. The first group was the majority of data and it's consisted of 43 observations, while the second group consisted of four observations 11, 20, 30 and 34, these data are classified as leverage points.

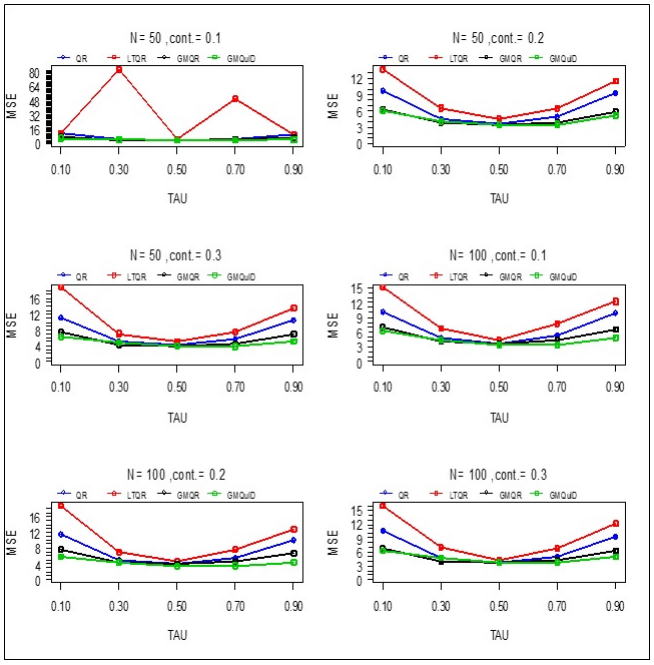

Figure 2: show the MAE for QR, LTQR, GMQu and GMQuID methods at five quantiles (0.10,0.30,0.50,0.70, 0.90) and three contamination levels, when $N = (50, 100)$.

As same as simulation study five quantiles are considered (0.10, 0.30, 0.50, 0.70, and 0.90)

Table 3: Show the MSE and MAE values for QR, LTQR, GMQu and GMQuID, at three quantiles (0.10, 0.30, 0.50, 0.70, 0.90) for Star cluster CYG OB1 dataset.

	Methods	*0.10*	*0.30*	*0.50*	*0.70*	*0.90*
MSE	***Qu***	*0.3114114*	*0.09528817*	*0.1077508*	*0.1207724*	*0.1449797*
	LTQRe	*0.4391339*	*0.10198449*	*0.1196301*	*0.1281637*	*0.1493265*
	GM-Qu	*0.1823753*	*0.08582650*	*0.1026927*	*0.1140047*	*0.1266969*
	GMQuID	*0.1644511*	*0.08154381*	*0.0975173*	*0.0834345*	*0.1078837*
MAE	***Qu***	*0.4810517*	*0.1762131*	*0.1642249*	*0.1753095*	*0.2099951*
	LTQRe	*0.5883964*	*0.1751335*	*0.1674482*	*0.1981649*	*0.2260934*
	GM-Qu	*0.3690579*	*0.1863123*	*0.1656058*	*0.1693746*	*0.1856701*
	GMQuID	*0.2618719*	*0.1382675*	*0.1209018*	*0.1215717*	*0.1400892*

In Table 3, we have reported the MSE and MAE values for the existing methods QR and LTQR, and the proposed methods GMQu and GMQuID for all quantile levels. The values of MSE and MAE for LTQR method is the largest among the other methods, which means this method has failed in diagnosing the high leverage points and trims them. The proposed method gets the smallest values of MSE and MAE at all quantiles; this indicates this proposed methods success in diagnosing the high leverage point and down weights them. The classical method is better than LTQR whereas it gets smaller values of MAE and MAE in most cases. The proposed method GMQu gets MSE and MAE smaller than the existing methods.

7 Conclusion

In this paper, to overcome the effect of high leverage point observation we have adapted generalized – M and also use generalized M based on IDRGP where the weight computed depend on IDRGP. To check the performance of our proposed new methods compare to two existing methods QR and LTQR simulation study and real data Star cluster CYG OB1 dataset are considered. MSE and MAE criteria are using to evolution the performance of the methods in this study.

The result that reported in Table 1 indicates the proposed methods GMQu and GMQuID are getting the smallest at all the different of contamination rates and quantile levels. In addition, the proposed methods have got the smallest

MSE and MAE compare to the existing in both simulation study and real dataset. So that, we can conclude that the new proposed method GMQuID is better all method in this study also GMQu do better than the other existing method. Therefore, we can see that the new method can reduce the effect of high leverage point observations.

References

[1] M. Abosaooda, W.J. Majid, E.A.Hussein, A.T. Jalil, M.M. Kadhim, M.M. Abdullah and H.A. Almashhadani, *Role of vitamin C in the protection of the gum and implants in the human body: theoretical and experimental studies*, Int. J. Corros. Scale Inhib. **10** (2021), no. 3, 1213–1229.

[2] J. Adrover, R. A. Maronna and V. J. Yohai, *Robust regression quantiles*, J. Statist. Plann. Infer. **122** (2004), no. 1-2, 187–202.

[3] R. Andersen, *Modern methods for robust regression*, Sage, 2008.

[4] S.H. Dilfy, M.J. Hanawi, A.W. Al-bideri and A.T. Jalil, *Determination of chemical composition of cultivated mushrooms in iraq with spectrophotometrically and high performance liquid chromatographic*, J. Green Engin. **10** (2020), 6200–6216.

[5] D.L. Donoho and P.J. Huber, *The notion of breakdown point*, A Festschrift for Erich L. Lehmann, 1983.

[6] S.P. Ellis and S. Morgenthaler, *Leverage and breakdown in L 1 regression*, J. Amer. Statist. Assoc. **87** (1992), no. 417, 143–148.

[7] A. Giloni, J.S. Simonoff and B. Sengupta, *Robust weighted LAD regression*, Comput. Statist. Data Anal. **50** (2006), no. 11, 3124–3140.

[8] M.M.R. Habshah Norazan and A.H.M. Rahmatullah Imon, *The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression*, J. Appl. Statist. **36** (2009), no. 5, 507–520.

[9] A.S. Hadi, *A new measure of overall potential influence in linear regression*, Comput. Statist. Data Anal. **14** (1992), 1–27.

[10] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, *The approach based on influence functions*, Robust Statistics, Wiley, 1986.

[11] Z.K. Hanan, M.B. Saleh, E.H. Mezal and A.T. Jalil, *Detection of human genetic variation in VAC14 gene by ARMA-PCR technique and relation with typhoid fever infection in patients with gallbladder diseases in Thi-Qar province/Iraq*, Materials Today, Proceedings, 2021.

[12] X. He, J. Jurečková, R. Koenker and S. Portnoy, *Tail behavior of regression estimators and their breakdown points*, Econometrica **1990** (1990) 1195–1214.

[13] R.W. Hill, *Robust regression when there are outliers in the carriers*, Doctoral Dissertation, Harvard University, 1977.

[14] D.C. Hoaglin and R.E. Welsch, *The hat matrix in regression and ANOVA*, Amer. Statist. **32** (1978), 17–22.

[15] M. Hubert and P.J. Rousseeuw, *The catline for deep regression*, J. Multivariate Anal. **66** (1998), no. 2, 270–296.

[16] A.T. Jalil, A.H. D. Al-Khafaji, A. Karevskiy, S.H. Dilfy and Z.K. Hanan, *Polymerase chain reaction technique for molecular detection of HPV16 infections among women with cervical cancer in Dhi-Qar Province*, Materials Today: Proceedings, 2021.

[17] A. T. Jalil, W.R. Kadhum, M.U. Faryad Khan, *Cancer stages and demographical study of HPV16 in gene L2 isolated from cervical cancer in Dhi-Qar province, Iraq*, Appl. Nanosci, 2021.

[18] A. T. Jalil, S. H. Dilfy, A. Karevskiy and N. Najah, *Viral Hepatitis in Dhi-Qar Province: Demographics and Hematological Characteristics of Patients*, Int. J. Pharmac. Res. **12** (2020), no. 1.

[19] A.T. Jalil, M.T. Shanshool, S.H. Dilfy, M.M. Saleh and A.A. Suleiman, *Hematological and serological parameters for detection of COVID-19*, J. Microbio. Biotechnol. Food Sci. **11** (2022), no. 4, 4229–4229

[20] J. Jumintono, S. Alkubaisy, D. Yánez Silva, K. Singh, A. Turki Jalil, S. Mutia Syarifah, Y. Fakri Mustafa, I.

Mikolaychik, L. Morozova and M. Derkho, *The effect of cystamine on sperm and antioxidant parameters of ram semen stored at 4 ˚c for 50 hours*, Arch. Razi Instit. **76** (2021), no. 4, 115.

[21] R. Koenker, *Quantile regression*, Cambridge University Press, Cambridge, UK., 2005.

[22] R. Koenker, and G. Bassett Jr, *Regression quantiles*, Econometrica J. Econ. Soc. **1978** (1978), 33–50.

[23] W.S. Krasker and R.E. Welsch, *Efficient bounded-influence regression estimation*, J. Amer. Statist. Assoc. **77** (1982), no. 379, 595–604.

[24] A.M. Leroy and P.J. Rousseeuw, *Robust regression and outlier detection*, Wiley series in probability and mathematical statistics, 1987.

[25] F. Marofi, O. Abdul-Rasheed, H. Sulaiman Rahman, H. Setia Budi, A.T. Jalil, Y. Valerievich and M. Jarahian, *CAR-NK cell in cancer immunotherapy: A promising frontier*, Cancer Sci. **112** (2021), no. 9, 3427.

[26] F. Marofi, H.S. Rahman, Z.M.J. Al-Obaidi, A.T. Jalil, W.K. Abdelbasset, W. Suksatan and M. Jarahian, *Novel CART therapy is a ray of hope in the treatment of seriously ill AML patients*, Stem Cell Res. Therapy, 12 (2021), no. 1, 1–23.

[27] R.A. Maronna and V.J. Yohai, *Asymptotic behavior of general M-estimates for regression and scale with random carriers*, Z. Wahrscheinl. verwandte Gebiete **58** (1981), no. 1, 7-20.

[28] H. Midi, T. Alshaybawee and M. Alguraibawi, *Modified least trimmed quantile regression to overcome effects of leverage points*, Math. Probl. Engin. **2020** (2020).

[29] S. Moghadasi, M. Elveny and H.S. Rahman, *A paradigm shift in cell-free approach: the emerging role of MSCs-derived exosomes in regenerative medicine*, J. Transl. Med. **19** (2021), no. 302.

[30] A.M. Mohammad, *Rbust estimation methods and robust multicollinearity diagnostics for multiple regression model in the presence of high leverage collinearity -influential observation*, Thesis submitted to the School of Graduate Studies, UPM., 2015.

[31] N.M. Neykov, P. Čížek, P. Filzmoser and P.M. Neytchev, *The least trimmed quantile regression*, Comput. Statist. Data Anal. **56** (2012), no. 6, 1757–1770.

[32] N. Ngafwan, H. Rasyid, E.S. Abood, W.K. Abdelbasset, S.G. Al-Shawi, D. Bokov and A.T. Jalil, *Study on novel fluorescent carbon nanomaterials in food analysis*, Food Sci. Technol. **42** (2022), 37821.

[33] A.H.M. Rahmatullah Imon, *Identifying multiple influential observations in linear regression*, J. Appl. Statist. **32** (2005), no. 9, 929–946.

[34] A.B. Roomi, G. Widjaja, D. Savitri, A. Turki Jalil, Y. Fakri Mustafa, L. Thangavelu and S. Aravindhan, *SnO2: Au/Carbon quantum dots nanocomposites: synthesis, characterization, and antibacterial activity*, J. Nanostruct. **11** (2021), no. 3, 514–523.

[35] P. Rousseeuw and B. Van Zomeren, *Unmasking multivariate outliers and leverage points*, J. Amer. Statist. Assoc. **85** (1990), 633–639.

[36] M.M. Saleh, A.T. Jalil, R.A. Abdulkereem and A.A. Suleiman, *Evaluation of immunoglobulins, CD4/CD8 T lymphocyte ratio and interleukin-6 in COVID-19 patients*, Turk. J. Immunol. **8** (2020), no. 3, 129–134.

[37] I. Sarjito, M. Elveny, A.T. Jalil, A. Davarpanah, M. Alfakeer, A.A.A. Bahajjaj and M. Ouladsmane, *CFD-based simulation to reduce greenhouse gas emissions from industrial plants*, Int. J. Chem. Reactor Engin. **19** (2021), no. 11, 1179–1186.

[38] A. Turki Jalil, S. Hussain Dilfy, S. Oudah Meza, S. M. Aravindhan, M. Kadhim and M. Aljeboree, *CuO/ZrO2 nanocomposites: facile synthesis, characterization and photocatalytic degradation of tetracycline antibiotic*, J. Nanostruct. **11** (2021), no. 2, 333–346.

[39] S. Vakili-Samiani, A.T. Jalil, W.K. Abdelbasset, A.V. Yumashev, V. Karpisheh, P. Jalali and F. Jadidi-Niaragh, *Targeting Wee1 kinase as a therapeutic approach in Hematological Malignancies*, DNA Repair **107** (2021), 103203.

[40] G. Widjaja, A.T. Jalil, H.S. Rahman, W.K. Abdelbasset, D.O. Bokov, W. Suksatan and M. Ahmadi, *Humoral Immune mechanisms involved in protective and pathological immunity during COVID-19*, Human Immunol. **82** (2021), no. 10, 733–745.