https://doi.org/10.26682/sjuod.2023.26.1.2
Journal of University of Duhok., Vol. 26, No.1 (Pure and Engineering Sciences), Pp 12-21, 2023

A CLASSIFICATION OF OUTLIERS IN TRANSFORMED VARIABLES

ROJEEN **T**AHA **A**HMAD and **S**HELAN **S**AIED **I**SMAEEL
Dept. of Mathematics, Faculty of Science**,** University of Zakho**,** Kurdistan Region-Iraq

(Received: July 24, 2022; Accepted for Publication: January 22, 2023)

ABSTRACT

The diagnostic of outliers is very essential since of their responsibility for producing large interpretative problems in linear regression analysis and nonlinear regression analysis. There has been a lot of work accomplished in identifying outliers in linear but not in nonlinear regression. In practice, it is often the case that the assumption of linear regression is violated, such as when highly influential outliers exist in the dataset, which will adversely impact the validity of the statistical analysis. Finding outliers is important because they are responsible for invalid inferences and inaccurate predictions as they have a larger impact on the computed values of various estimations. The outliers must be divided into vertical outliers (VO), good leverage points (GLP), and bad leverage points (BLP) since only the vertical outliers and bad leverage have an undue effect on parameter estimations. We compare several outlier detection techniques using a robust diagnostic plot to correctly classify good and bad leverage points and vertical outliers, by decreasing both masking and swamping effects for both the untransformed variables and transformed variables. The main idea is to detect of outliers before transformation (original data) and after transformation. The results of generation study and numerical indicate that modified generalized DIFFITS (different of fit) against the Diagnostic Robust Generalized Potential (MGDFF-DRGP) successfully detect outliers in the data.

KEYWORDS: Robust estimations, outliers, MM-estimate, multiple high leverage points.

I. INTRODUCTION

The assumption of linear regression is used by the majority of statistical techniques for the analysis of multivariate data. In the majority of the cases, the results obtained from the standard application of these techniques are not robust to departures in this assumption. These departures can be either systematic (caused by model misspecification) or isolated (caused by the presence of the outliers). A severe problem is that a small number of outliers may disguise systemic departures and hide one another due to masking. This article introduces data transformations of variables and diagnostic analysis leading to the detection of outliers.

To make the linear regression model, data transformations of the variables (predictor variable (X), response variable (Y), or both (X) and (Y)) are sufficient. The R program's trafo package, which offers a simple, user-friendly framework for selecting a suitable transformation for data, was used to select the response variable transformations.

You may estimate, choose and compare different transformation families using the trafo package. The transformation families in the trafo package are listed below: Gpower, Log, Log-shift opt, Manly, Modulus, Neglog [1], Glog [2], Bickel-Doksum [3], Dual, Reciprocal and Yeo-Johnson, Box-Cox [4]. The package makes it simpler to compare linear models with transformed and untransformed dependent variables, and also linear models with different transformations applied to the dependent variable.

There are several types of outliers in regression problems, including residual outliers, high leverage points (HLPs), and vertical outliers. A residual outlier is any observation that has a large residual. The observations that are severe or outlying in the y-coordinate are known as y-outliers or vertical outliers (VO). However, HLPs are observations that are extreme or outlying in the x-coordinate. Good leverage points (GLPs) and bad leverage points (BLPs) are two types of HLPs. Outliers in the explanatory variables are known to be GLPs when they follow the pattern of the majority of the data, whereas BLPs do the opposite. BLPs have a larger effect on the computed values of different estimations. On the other hand, GLPs contribute to the efficiency of an estimate (see [5],[6],[7],[8],[7]). As a result, in the computation of the weighting function in any robust technique, only BLPs should be weighted

down whereas GLPs should not. However, it is now evident that most robust methods attempt to lessen the impact of outliers by weighting the outliers down, irrespective of whether they are GLPs or BLPs [28]. There are several good studies on the detection of HLPs in the literature (see [9],[10],[11],[12]). Nevertheless, those detection techniques are mainly focused only on the identification of HLPs without taking into consideration their classification as bad and good. Making the classification is important since only the BLPs are responsible for a misleading conclusion about the regression model's fit. Using a graphical method, it is difficult to capture the existence of several outlier versions in multiple regression analysis [10]. When just one independent variable is taken into account, the four kinds of outliers can simply be observed from a scatter plot of y against the x variables. But, it's really difficult to detect these outliers from a scatter plot when there are several predictor variables. Rousseeuw and Van Zomeren [6] have suggested a robust diagnostic plot, also known as an outlier map, which is more effective than a non-robust plot in classifying observations into 4 kinds of data points: regular or good observations (VO), (BLPs) and (GLPs). We believe that this figure fails to detect HLPs and multiple outliers. To detect outliers and empirical influences as well as to find the influence of observations, Pison and Van Aelst [13]proposed a new plot using robust distance obtained using robust location and scale estimators. They create a graphical tool for multivariate models that is similar to the Rousseeuw and Van Zomeren plot. Even though Pison and Van Aelst plot may be very useful when developing a model to evaluate the quality of a fit based on number of the outliers , it does not focus on classifying unusual data into GLP, BLP and VO (see [13]). To draw the outliers map or diagnostic plot, on the horizontal axis, plot the robust score distance of each observation and on the vertical axis.

The transformation methods and Identification of outliers are discussed in the methodology which is in Section 2. Section 3 gives a generation study and two numerical illustrations. Finally, the concluding remark is given in Section 4.

II. METHODOLOGY

1- Transformations method:

The equation describing and summarizing the relationship between several discrete or continuous variables x and a continuous dependent variable y is defined by $y_i = \beta x_i^T + e_i$, with $i = 1, \dots, n$. This is also known as the linear regression model, and it consists of deterministic and random components. These components are based on several assumptions, including linearity, normality, and homoscedasticity. There are several ways that may be used to satisfy the model assumptions. We focus on the detection of outliers before and after transformation. the trafo package's transformations for the dependent variable and the independent variable (Square Root, Inverse, Logarithmic) [14] are used. However, the majority of them correct other assumptions at the same time.

In the package trafo, the values are by default shifted by (a) deterministic shift (a) if the response variable includes negative values, so that $y + a > 0$. A square root transformation with deterministic shift is given in one instance [15]. We demonstrated how the trafo package makes possible for users to easily determine which transformations are suitable for satisfying the model's assumptions. such as linearity, normality, and homoscedasticity. The trafo is the only R package that we know of that supports this decision process.

Table (1): Diagnostic Checks Provided In The Package Trafo

Assumption	Diagnostic check
1) Normality	Skewness and kurtosis Shapiro-Wilk test Quantile-quantile plot Histograms
2) Homoscedasticity	Breusch-Pagan test Residuals vs. fitted plot Scale-location
3) Linearity	Scatter plots between y and x Observed vs. fitted plot

The method of estimating with the maximum likelihood finds the values of the transformation parameters that maximize the likelihood function of the dataset under the given transformation[16]. Several of the R packages listed above [17], [18] employ this standard approach. Because the only way for the user to know if the transformation is beneficial is to check the above assumptions, the trafo package comes with a wide range of diagnostic tests [19], [20]. A smaller selection is used for the quick check, which determines whether a transformation is beneficial. The diagnostic checks that have been implemented for the untransformed and transformed models, as well as two different transformed models, are shown in Table 1 including the diagnostics that are conducted during the quick check. In order to help in the detection of outliers, graphs such as the Cook's distance plot by the residuals vs leverage plot are also provided. Below are some of the transformed methods that used it.

1.1 **Dual Transformation**

The function transforms the dependent variable of a linear model using dual transformation [21].

$$y_i^*(\lambda) = \begin{cases} \dfrac{(y_i{}^{\lambda} - y_i{}^{-\lambda})}{2\lambda} & \text{if } \lambda > 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases} \quad , \quad y_i > 0$$

1.2 **Manly Transformation**

The function transforms the dependent variable of a linear model using the Manly transformation [22].

$$y_i^*(\lambda) = \begin{cases} \dfrac{e^{\lambda y_i} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ y_i & \text{if } \lambda = 0 \end{cases} \qquad y_i \epsilon R$$

2- Identification of outliers

Statistical literature has suggested various identification processes for outliers. Outlier identification has been proposed using residual plots based on different residual types [23]. In regression, outliers are often identified as data that correlate to residuals that exhibit an unusual pattern.

2.1 Identification of vertical outlier

The term "vertical outliers" describes observations with large residuals. To find these outliers in a data set, several diagnostic methods have been proposed in the statistical literature, such as residual plots, which are based on residuals [23]. Several analytical methods may be used to detect these outliers, and have their basis on the estimation of the residuals' scale, such as **standardized residuals (SR)** denoted by $SR_{OLS} = \frac{e_i}{\hat{\sigma}},\ i = 1,2,\dots,n$, where e_i refers to the ordinary least square residual for the i^{th} case and the $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$. The i^{th}observation is considered as an outlier if $|SR_{OLS}| > 2.5\ or\ 3$ (see [24)].

2.2 Identification of high leverage points

In regression analysis, it might be vital to determine whether a certain set of X-values is influencing the regression model's fitting too much. An influential set of X-values is referred to as HLPs. A single case deletion measure named a Potentials matrix was proposed by Hadi [25]. The potential matrix's diagonal elements, represented by "p_{ii}" are provided by (see [5],[25],[26])

$$p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i\ , \qquad i = 1,2,\dots,n,$$

In this case, $X_{(i)}$ is the matrix X excluding the i^{th} row. Additionally, [25] suggested a cut off point for p_{ii} based on the manner below.

$$Median(p_{ii}) + cMAD(p_{ii})$$

Where $MAD(p_{ii}) = \text{Median}\{|p_{ii} - \text{Median}(p_{ii})|\}/0.6745$, where c represents an appropriate constant, like two or three.

Mahalanobis Distance (**MD**), which measures the distance between the observation x_i and the middle of the bulk of data was developed by Rousseeuw and Leroy [27]. The arithmetic mean

(T_x), and covariance matrix (C_x), can be defined as

$T_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $C_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - T_x)^t(x_i - T_x)$

Consequently, the classical MD used to the i^{th} a case can be defined as follows:

$$MD_i = \sqrt{(x_i - T_x)^t C_x^{-1}(x_i - T_x)} \qquad i = 1,2,\dots,n$$

To compare it to the cut-off $\sqrt{x^2_{p+1,0.95}}$, the MD_i can be calculated for each $i = 1,2,\dots,n$. When an observation exceeds the cutoff point, it is considered to be an HLP.

Based on hat values, the MD_i can also be given as follows:

$$MD_i^2 = (n-1)\left[h_{ii} - \frac{1}{n}\right]$$

Robust Mahalanobis Distance (**RMD**) based on MVE was proposed by Rousseeuw [28] as

$$RMD_i(MVE) = \sqrt{(x_i - T(x))^t C(x)^{-1}(x_i - T(x))} \ , i = 1,2,\dots,n$$

Where $T(x)$ denotes the MVE's robust locations and $C(x)$ denotes the MVE's scatter.

The robust alternative diagnostic methods like RMD_{MVE} are capable of correctly detecting the HLPs.

To improve the detection rate of HLPs, [29] developed the **Diagnostic Robust Generalized Potential (DRGP)** using the MVE as a basis. There are two steps in the DRGP. The robust technique was used in the first step to identify the suspected HLPs. The suspicion is confirmed using a generalized potential diagnostic approach in the second step. The goal is to enhance the rate of correct HLPs detection while minimizing the impacts of swamping and masking. Furthermore, it is established that the DRGP performs better than other commonly used HLPs detection methods. Nevertheless, DRGP suffers with its tendency to swamp cases with low leverage for 5% and 10% HLPs.

MGti-DRGP was first developed by [30], and it has exhibited to be quite successful in classifying observations into regular observations, BLPs, GLPs, and VOs. Nevertheless, it also still suffers from some of the impacts of swamping and masking.

3- Modified Generalized DFFITS (MGDFF)

MGDFF was created by Habshah et al. (2015b) to identify multiple influential observations. following is a definition of the MGDFF:

$$MGDFF_i = \begin{cases} \sqrt{\dfrac{w_{ii}(R)}{1 - w_{ii}(R)}}\, MGt_i & \text{for } i \in R \\ \sqrt{\dfrac{w_{ii}(R)}{1 + w_{ii}(R)}}\, MGt_i & \text{for } i \notin R \end{cases}$$

The conventional cutoff value is $CP_{MGDFF} = Median(MGDFF_i) + c * MAD(MGDFF_i)$

Furthermore, the Modified Generalized Studentized Residuals (MGt_i) are formulated based on the RLS and the DRGP as initial estimators. MGt_i for the whole data set is created as follows: [12]

$$MGt_i = \begin{cases} \dfrac{\hat{e}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1 - h_{ii(R)}}} , & \text{for } i \in R, \\ \dfrac{\hat{e}_{i(R)}}{\hat{\sigma}_{R}\sqrt{1 + h_{ii(R)}}} , & \text{for } i \notin R, \end{cases}$$

$\hat{\sigma}_R$ is the standard deviation as well as $\hat{\sigma}_{R-i}$ is the standard deviation of R groups excluding the i^{th} case.

4- Diagnostic Plots for Classifying Observations into Four Categories

For classifying observations into regular observations, VOs, GLPs, and BLPs, [6] suggested a robust diagnostic plot that is more effective than the nonrobust plot. The standardized_ MM _residual against robust Mahalanobis distance (RMD) based on minimum volume ellipsoid (MVE); this plot is denoted by the (MM-RMD). The nonrobust plot draws the standardized _OLS _residuals against the Mahalanobis distance (MD), and we called this plot an (OLS-MD) plot. We believe that since the robust MM-RMD diagnostic plot is based on the robust Mahalanobis distance, which suffers from swamping effects, it is not very effective at classifying the observations into respective categories. Additionally, this plot employs standardized residual, which performs badly in identifying multiple outliers. [29]shown that the DRGP was very successful in detecting HLPs. Additionally, this plot, known as the (MGDFF-DRGP) plot, against the MGDFF is able to detect multiple outliers, as seen in Figure 1's classification. The observations from [31] Habshah et al. (2021) are classified as follows:

i. An Observation is described as a "RO" if

ii.
$$|MGDFF_i| \leq CP_{MGDFF} \quad and \quad p_{ii} \leq Median(p_{ii}) + cMad(p_{ii})$$

iii. An Observation is described as a "VO" if

$$|MGDFF_i| > CP_{MGDFF} \quad and \quad p_{ii} \leq Median(p_{ii}) + cMad(p_{ii})$$

iv. An Observation is described as a "GLP" if

$$|MGDFF_i| \leq CP_{MGDFF} \quad and \quad p_{ii} > Median(p_{ii}) + cMad(p_{ii})$$

An Observation is described as a "BLP" if

$$|MGDFF_i| > CP_{MGDFF} \quad \text{and} \quad p_{ii} > Median(p_{ii}) + cMad(p_{ii})$$

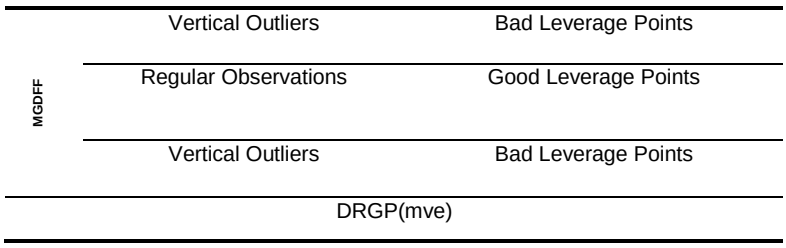

Fig. (1): The DRGP (mve) against MGDFF

III. APPLICATION

This section includes a generation study and two real data are designed to evaluate how well the (MGDFF-DRGP) plot method performs in classifying observations into VO, bad, and good HLPs. The OLS-MD and MM-RMD plots are compared to the MGDFF-DRGP plot in this case. Based on the rate of correct outlier detection, these plots' performances are evaluated. To obtain the transformation function and outlier detection methods for these variables, the data were analyzed using the R program.

1-Detecting Outliers in Real data

Our first example is the dataset which is taken from [32]. Age is the only independent variable in the dataset, while SBP is the dependent variable (Systolic Blood Pressure). There are 45 observations in this dataset.

The data scattered around the ideal curve is assumed to follow nonlinearity and non-normality using the least squares to fit a polynomial regression model with the influence of a single predictor (x).

$$\text{SBP} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Age}^2$$

A **manly** function in the trafo package for the response variable (SBP) was used to transform y for the best results. The assumptions of linear regression, such as linearity, normality, and homogeneity in the data, are approximately satisfied by the result of this transformation, whereas they were not met before the transformation.

Figures 2 show the classification of data into VO and BLPs. Figures 2(a), 2(b), and 2(c) show

that one VO was found using the nonrobust plot (OLS-MD). One VO and one BLPs were found by the (MGDFF-DRGP) plot and one VO by the (LMS-RMD) plot before the transformation. After response variable transformation, this result was changed. Figures 2(d), 2(e), and 2(f) show that the MGDFF-DRGP dropped, the MM-RMD rose, and the OLS-MD remained steady.

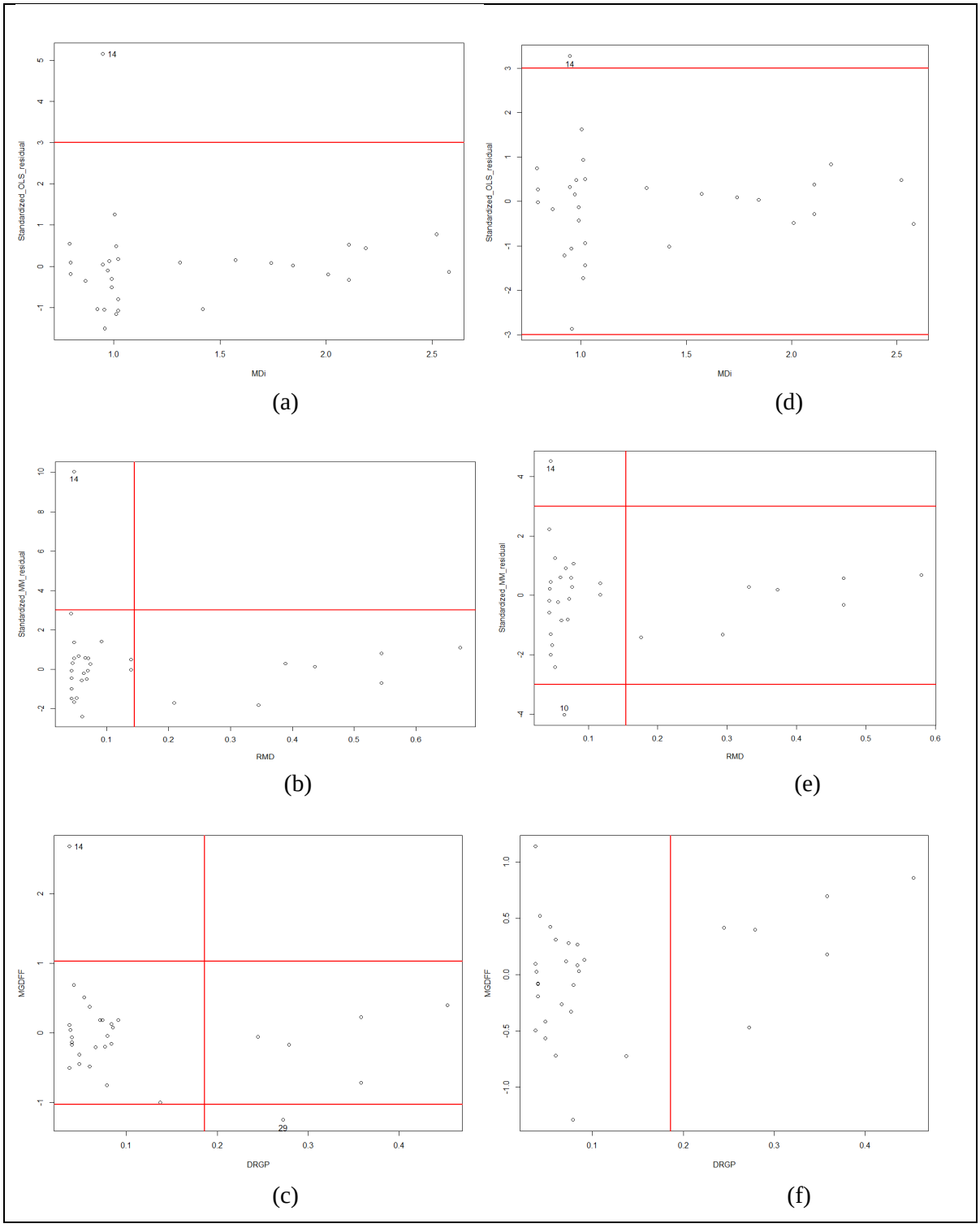

Fig. (2): The OLS-MD, MM-RMD, and MGDFF-DRGP (a, b, and c) plots for the original data and the plots d, e, and f of transformation data.

The second example is the Duhok data set about COVID-19 data. There are 105 observations for multiple regressions. For the daily new cases, active cases, and new recovered from October 1, 2021, to January 13, 2022, it can be found at [33]. The dependent variable y is represented by active cases and the independent variables x_1 and x_2 are represented by new recovered cases and new cases respectively. In order to get the best results, both x and y were transformed using a dual function for the dependent variables (y) and a logarithmic function for the independent variable (x_1) [34]. The result of this transformation is approximate to the assumption of linear regression in the data, while the linear regression assumptions were not met before the transformation.

The three plots (OLS-MD, MM-RMD, and MGDFF-DRGP) were then applied to both original and transformed data. According to Figure 3, which shows the results of the diagnostic methods and the preceding plots, the OLS-MD plot correctly identified two VO in the original data. However, the LMS-RMD plot identified five BLP with 2 VO in the original data, whereas the MGDFF-DRGP plot can classify the observations into 3 VO with (12) BLP in the original. However, one VO with (3) BLP in the transformed data, as shown in Figures 3(a), 3(b), and 3(c) for the original data, it is evident that all classical and robust diagnostic plots can correctly detect the BLP and VO. But, for transformed data, it is interesting to observe that only the MGDFF-DRGP plot is able to detect and classify the 4 outlying into VO and BLPs correctly see Figure 3(f). Although the MM-RMS plot can detect one VO as shown in Figure 3(e). The classical OLS-MD can only detect one VO as shown in Figure 3(d).

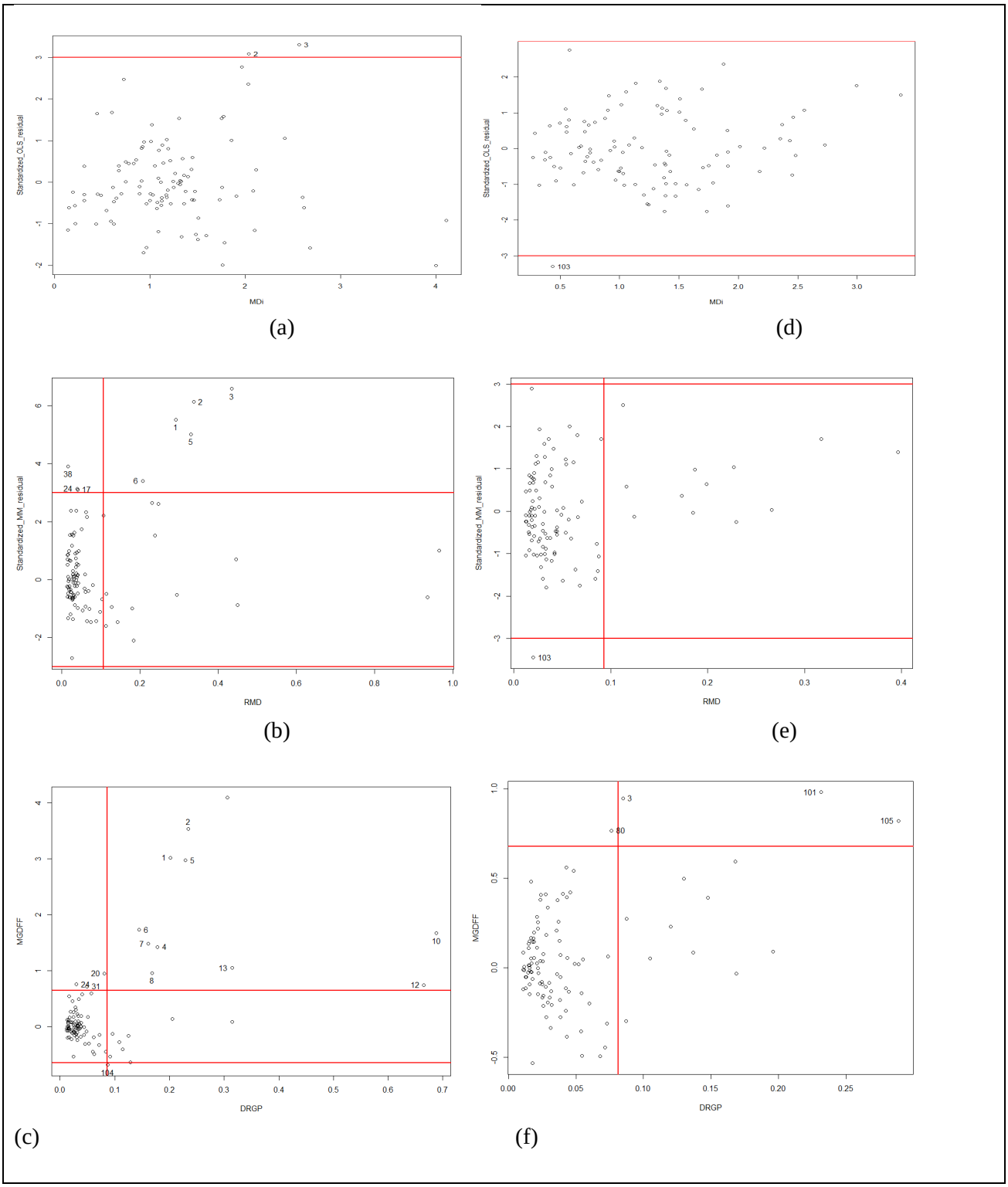

Fig. (3): The OLS-MD, MM-RMD, and MGDFF-DRGP (a, b, and c) plots for the un transformation data and the plots (d, e, and f) of transformation data.

2- Generating Dataset: In this study, one type of simulation with 4 different samples n = 50, 100, and 150 are used. We consider two independent variables and a three-parameter model given by the following relationship.

$$y_i = \beta_0 e^{\beta_1 x_{i1} + \beta_2 x_{i2}} + e_i$$

Where $\beta_0 = 2, \beta_1 = 1, \beta_2 = 1$ are the parameters and $e \sim N(0,2)$. The independent variables are distributed as $x_{i1} \sim Exp(2)$ and $x_{i2} \sim U(1,3)$. We consider these values to justify nonlinearity and non-normality assumptions in linear regression. For exponential distribution, outlier events will happen eventually. All

positive values have some probability of occurring, as is the nature of the exponential distribution(https://towardsdatascience.com/real-time-anomaly-detection-with-exponentially-distrubted-data-205e0df32096). The best results were obtained by transforming both x and y using a logarithmic function for the independent variable (x_1) and a manly function for the outcome variable (y). They show that the result is close to the linear regression assumptions like linearity, homogeneity, and normality in the data, while the assumptions of linear regression were not satisfied before the transformation.

Table 2 gives a summary of the results of the generation study. The correct numbers of detection of VO and BLPs at different levels of different sample sizes. Regardless of the number of regressor variables, the findings show that the MGDFF-DRGP plot has a superior ability to identify the correct number of BLPs than the OLS-MD and MM-RMD plots. In addition, the performance of MM-RMD plots decreases in terms of having a lesser number of correct detections. The OLS-MD plot has very bad performance, as can be seen. Before and after transformation, the MGDFF-DRGP plot consistently has a higher performance of correct detection.

Table (2): The vertical outliers and Bad leverage points in the parenthesis for the OLS-MD, MM-RMD, and MGDFF-DRGP for the original and transformed for simulation data (n= 50,100,150).

	Before transformation			**After transformation**		
n	OLS_MD	MM_RMD	DRGP_MGDFF	OLS_MD	MM_RMD	DRGP_MGDFF
50	1 (0)	2 (4)	2 (5)	0 (0)	2 (0)	4 (1)
100	1 (0)	7 (8)	11 (4)	0 (0)	0 (1)	1 (1)
150	2 (0)	6 (13)	7 (13)	0 (0)	0 (0)	1 (2)

IV. CONCLUSION

In this study, we use a diagnostic plot for both the original and transformed data to identify outliers (BLP and VO). The classical OLS-MD plot fails to correctly identify outlier. Classifying observations into 4 categories using the robust MM-RMD plot is also not successful. Furthermore, the MGDFF-DRGP plot is very successful in classifying observations into bad leverage points, good leverage points, and vertical outliers. The generation study clearly shows that the MGDFF-DRGP plot can correctly detect BLPs for both the untransformed and transformed data. The OLS-MD, MM-RMD, and MGDFF-DRGP show the number of BLP and VO decreased in the transformation data. The number of outliers is reduced due to transformation in variables.

V. REFERENCES

J. Whittaker, C. Whitehead, and M. Somers, "The neglog transformation and quantile regression for the analysis of a large credit scoring database," *J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 54, no. 5, pp. 863–878, 2005.

B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, vol. 18, no. suppl_1, pp. S105–S110, 2002.

P. J. Bickel and K. A. Doksum, "An analysis of transformations revisited," *J. Am. Stat. Assoc.*, vol. 76, no. 374, pp. 296–311, 1981.

I. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

S. Chatterjee and A. S. Hadi, "Influential observations, high leverage points, and outliers in linear regression," *Stat. Sci.*, vol. 1, no. 3, pp. 379–393, 1986.

P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," *J. Am. Stat. Assoc.*, vol. 85, no. 411, pp. 633–639, 1990.

A. Bagheri and H. Midi, "Diagnostic plot for the identification of high leverage collinearity-influential observations," *SORT-Statistics Oper. Res. Trans.*, pp. 51–70, 2015.

M. R. Norazan, H. Midi, and A. Imon, "Estimating regression coefficients using weighted

bootstrap with probability," *WSEAS Trans. Math.*, vol. 8, no. 7, pp. 362–371, 2009.

D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

P. Rousseeuw and A. Leroy, "Robust regression and outlier detection: Wiley Interscience," *New York*, 1987.

A. C. Atkinson, "Fast very robust methods for the detection of multiple outliers," *J. Am. Stat. Assoc.*, vol. 89, no. 428, pp. 1329–1339, 1994.

A. H. M. Rahmatullah Imon, "Identifying multiple influential observations in linear regression," *J. Appl. Stat.*, vol. 32, no. 9, pp. 929–946, 2005.

G. Pison and S. Van Aelst, "Diagnostic plots for robust multivariate methods," *J. Comput. Graph. Stat.*, vol. 13, no. 2, pp. 310–329, 2004.

R. T. Ahmad and S. S. Ismaeel, "A Nonlinear Transformation Methods Using Covid-19 Data in the Kurdistan Region," in *2022 International Conference on Computer Science and Software Engineering (CSASE)*, 2022, pp. 207–211.

M. S. Bartlett, "The use of transformations," *Biometrics*, vol. 3, no. 1, pp. 39–52, 1947.

G. E. P. Box and D. R. Cox, "An analysis of transformations," *J. R. Stat. Soc. Ser. B*, vol. 26, no. 2, pp. 211–243, 1964.

W. N. Venables and B. D. Ripley, "Modern applied statistics with S. 4th Springer," *New York*, vol. 118, 2002.

J. Fox and S. Weisberg, "An R companion to applied regression. Sage," *Thousand Oaks*, 2011.

S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econom. J. Econom. Soc.*, pp. 1287–1294, 1979.

Z. Yang, "A modified family of power transformations," *Econ. Lett.*, vol. 92, no. 1, pp. 14–19, 2006.

B. F. J. Manly, "Exponential data transformations," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 25, no. 1, pp. 37–42, 1976.

A. C. Atkinson, "Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis," 1985.

A. Hossein Riazoshams, B. Midi Habshah, and C. Mohamad Bakri Adam, "On the outlier detection in nonlinear regression," *World Acad. Sci. Eng. Technol.*, vol. 36, no. 12, pp. 264–270, 2009.

A. S. Hadi, "A new measure of overall potential influence in linear regression," *Comput. Stat. Data Anal.*, vol. 14, no. 1, pp. 1–27, 1992.

A. Bagheri and H. Midi, "On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations," *Math. Probl. Eng.*, vol. 2012, 2012.

A. M. Leroy and P. J. Rousseeuw, "Robust regression and outlier detection," *Wiley Ser. Probab. Math. Stat.*, 1987.

P. J. Rousseeuw, "Multivariate estimation with high breakdown point," *Math. Stat. Appl.*, vol. 8, no. 283–297, p. 37, 1985.

M. Habshah, M. R. 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. Stat.*, vol. 36, no. 5, pp. 507–520, 2009.

M. Alguraibawi, H. Midi, and A. H. M. Imon, "A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model," *Math. Probl. Eng.*, vol. 2015, 2015.

H. Midi, M. Sani, S. S. Ismaeel, and J. Arasan, "Fast Improvised Influential Distance for the Identification of Influential Observations in Multiple Linear Regression," *Sains Malaysiana*, vol. 50, no. 7, pp. 2085–2094, 2021.

G. Manimannan, M. Salomi, R. L. Priya, and R. Saranraj, "Detecting Outliers using R Package in Fitting Data with Linear and Nonlinear Regression Models," *Int. J. Sci. Innov. Math. Res.*, vol. 8, no. 4, pp. 1–13, 2020, doi: 10.20431/2347-3142.0804001.

"COVID-19: Dashboard - GOV.KRD." https://gov.krd/coronavirus-en/dashboard/ (accessed Jul. 07, 2022).

R. Taha and S. Saied, "General Letters in Mathematics (GLM) Estimating Regression Coefficients using Bootstrap with application to Covid-19 Data," vol. 12, no. 2, pp. 96–104, 2022, doi: 10.31559/glm2022.12.2.6.