GM Estimator Based on RFIID As a Remedy of Vertical Outliers and HLPs

Habshah Midi and Hasan Hendi
Institute For Mathematical Research
Universiti Putra Malaysia

PRESENTATION OUTLINE

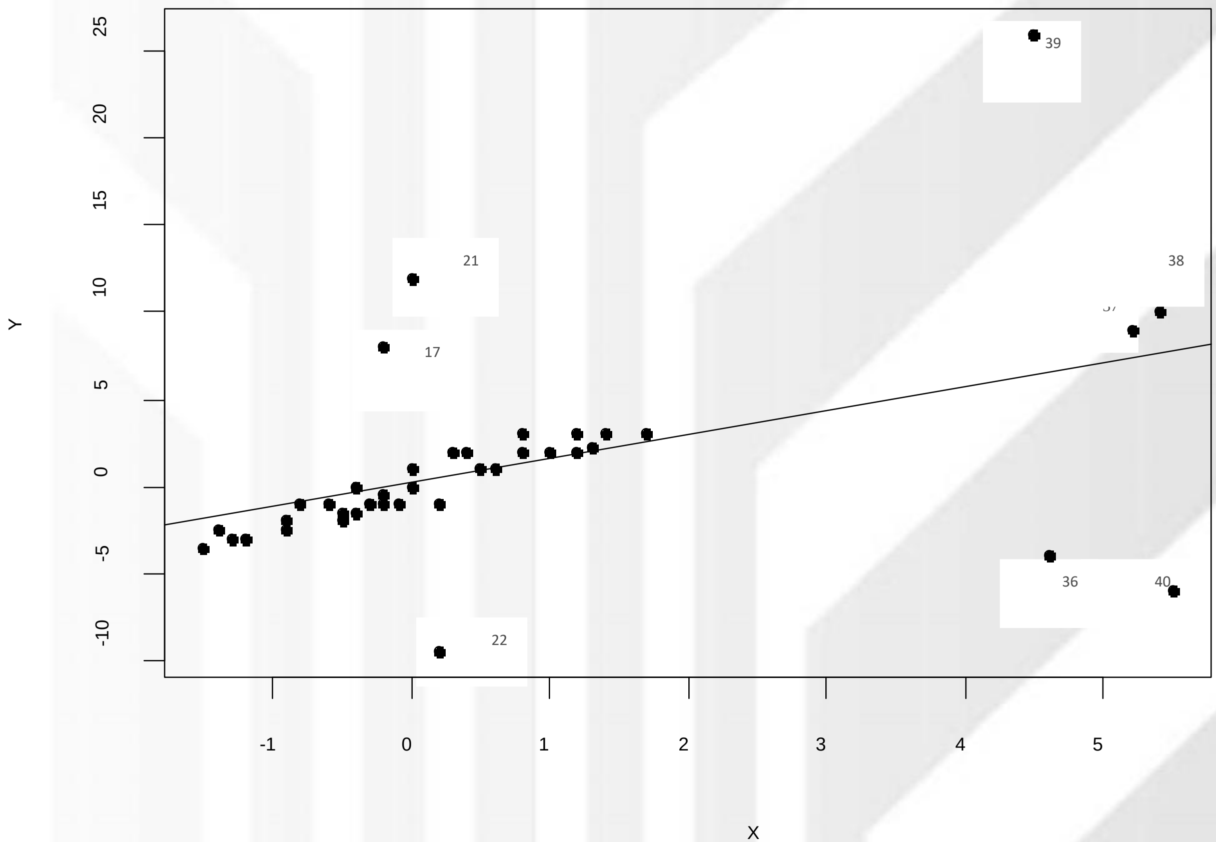
1. Introduction
2. Objectives
3. Outliers in Regression
4. Methodology
5. Simulation Study and Real Example
6. Conclusions
7. References

OUTLIERS IN REGRESSION

- In statistical Data Analysis-Only one type of outlier.

- But in Regression, several versions of outliers;

- residual outliers –observations with large residuals
- vertical outliers –observations outlying in y-coordinate
- high leverage points-observations outlying in x-coordinate



Y
X
25
20
15
10
5
0
-5
-10
-1
0
1
2
3
4
5
39
21
38
17
22
36
40

INTRODUCTION

- The ordinary least squares(OLS) method is the most popular technique in regression analysis due to its optimal properties and ease of computation.
- However, many do not realized that the OLS estimates are much affected by outliers.
- Among the three type of outliers, the HLPs, outlying observations in the X direction, have the most detrimental effect on the computed values of various estimates.
- Relying on the OLS method may give inefficient estimates and inaccurate predictions and causing uncertainties in predicting future outcomes.
- As An alternative, we may use robust statistical method that try to reduce the effect of outliers.

INTRODUCTION

- Many robust methods such as M, MM, LMS, LTS are available in the literature (Huber and Ronchetti, 1981; Yohai, 1987; Leroy and Rousseeuw, 1987).
- Simpson et al. (1992) pointed out that even though some of them have high efficiency and possess high breakdown point (BDP), they do not have bounded influence properties in the sense that they are unable to reduce the effect of HLPs.
- Schweppes as described by Hill and Paul (1977) proposed a new robust method call Generalized M estimator that can handle HLPs.
- The GM6 which is based on Robust Mahalanobis Distant (RMD) which uses MVE and MCD to obtain the initial estimates has several shortcomings: long computational running times, swamping, downweigt both good and bad HLPs, efficiency tends to decrease as the no of good leverage points increases.

INTRODUCTION

- As a solution, Habshah et al. (2021) proposed GM-FIMGT which is based on Improvised Generalized MT (FIMGT).
- It has been shown that GM-FIMGT more efficient than GM6. However, it is based on ISEwhich is unstable because its algorithm depend on the selected initial subset, h.
- Midi et al. (2020) showed that the final estimator of location and scatter of ISE is equivalent to MCD if same initial subset is used. Otherwise, results will be different. Moreover, computational running times still quite long.
- Hence, a more efficient GM estimator is needed to remedy these problems.

Objectives

- To develop a new GM estimator (GM-RFIID) by integrating an initial weight function based on robust and fast method of the identification of influential observations (IOs).
- To compare the proposed method with some existing methods.
- To apply the proposed method to real data.

.

METHODOLOGY

- Belsley et al. (2004) noted that influential obs (Ios) are those obs which either alone or together with several other observations have a detrimental effect on the computed values of various estimates.
- It is generally believe that Ios are outlying obs in X or Y –space.
- However according to Chatterjee and Hadi (1986), IOs are not always HLPs and vice versa.
- When establishing an approach to determine IOs, both the dependent and independent variables should be taken into consideration. According to Rahmatullah Imon (2002) and Rousseeuw and Leroy (1987), failing to do that may result in inaccurate detection of IOs and will lead to misleading interpretation.

Methodology

- Three steps is proposed to analyse a dataset for multiple linear regression.
- Step1 : Identify the existence of HLPs using DRGP-RFCH
- Step 2: Identify the existence of Ios using Robust and Fast Improvised Influential Distance (RFIID) aaand cut-off point RFIID. Based on RFIID, classify observations into RO, GLO and IOs
- Step 3: Obtain the initial weight function for the solution of normal equation of the GM estimator.

Step 1:Diagnostic Robust Generalized Potential based on RFCH to Detect HLP

❖Midi et al. (Pertanika Journal of Sc & Tech, 2021), see alsoLim and Habshah (Computational Statistics,2016) (see also Habshah, Norazan et al. (2009), J. of Applied Stat., Mazlina & Habshah (2015), Pak. J of Statistics) formulated RMD- Reweighted Fast Consistent and High Breakdown (RFCH) to detect multiple high leverages. It consists of two steps.

Step i) suspect high leverage points are determined by the robust Mahalanobis Distance based on Index Set Equality:

$$RMD_i = \sqrt{(X - T_R(X))^T C_R(X)^{-1} (X - T_R(X))} \qquad i = 1, 2, \ldots, n$$

where $T_R(X)$ and $C_R(X)$ are robust locations and shape estimtes of the RFCH, respectively. A set of 'good' cases 'remaining' in the analysis denoted by *R* and deleted by D

- Step ii) Diagnostic Approach used to confirm the suspected groups

$$p^*_{ii} = \begin{cases} w_{ii}^{(-D)} & for\ i \in D \\ \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} & for\ i \in R \end{cases}$$

- Where $w_{ii}^{(-D)} = X_i^T (X_R^T X_R)^{-1} X_i$

- An observation is considered as HLps if $\boldsymbol{p^*_{ii}}$ is large :

$$p^*_{ii} > Median\ (p^*_{ii}) + c\ MAD\ (p^*_{ii})$$

- Where c can be taken as a constant value of 2 or 3.

Step 2: Identify IOs

The RFIID can be summarized as the following:

RFGSR	Influential Observation (IO)	Influential Observation (IO)
	Regular Observations (RO)	Good Leverage Observation (GLO)
	Influential Observation (IO)	Influential Observation (IO)
	RFGLV	

Since it is not easy to proof the distribution of $RFID_i^*$, confident bound type of cutoff point is again utilized as in Habshah et al. (2009) and Rashid et al. (2022)

$$CPRFID_i^* = \text{median}\ (RFID_i^*) + 3\text{MAD}(RFID_i^*)$$

Step 3: The Proposed GM estimator based on RFIID

For the general linear regression model with the usual assumptions, the GM estimator is defined as a solution of normal equations which is given by,

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - x_i^t \hat{\beta})}{\hat{\sigma}\pi_i}\right) x_i = 0$$

Where $\psi = \rho'$ is a derivative of redescending function (weight function) and $\pi_i, i = 1,2,\ldots,n$ is the initial weight element of the diagonal matrix *W*, $\hat{\sigma}$ is the scale estimate, and $\hat{\beta}$ is the vector of parameters estimates.

Coakley and Hettmansperger (1993) proposed GM6 estimator which employs Robust Mahalanobis Distance (RMD) based on Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD) to identify high leverage points and subsequently initial weight of this GM estimator is formulated based on RMD which is given by:

$$\pi_i = \min\left[1, \left(\frac{\chi^2_{(0.95,p)}}{RMD^2}\right)\right], i = 1,2,\ldots,n$$

The weakness of this initial weight function

1. it tends to swamp some low leverage points (Bagheri and Habshah, - Transaction in Statistics,2015), some of good leverages (GLPs) will be given low weights. Hence, the efficiency of the GM6 estimator tends to decrease with the presence of good leverage points. GLPs have no effect or have very little effect on parameter estimates and may contribute to the precision of parameter estimation(Rousseeuw, and Van Zomeren, 1990). On the other hand, BLPs have high impact on the regression estimates. This is the reason why the GM6-estimate is less efficient.

2.GM6 estimator takes too much computing time. -

Hence, we will propose a relatively easy and fast method based on the detection of Ios using RFIID. Then only minimize the weights of IOs.

$$d_i = min[\quad 1, (\frac{CP_{RFIID}}{RFIID})], i = 1,2, \ldots, n$$

Algorithm of the Proposed GM-RFIID Estimator

The proposed GM-RFIID estimator is similar to that of Dhhan Habshah Sohel (Journal of Appl Stat. 2016). The only different is the calculation of the initial weight function. The algorithm of our proposed GM estimator is summarized as follows:

Step 1: Use the LTS method as an initial estimator to achieve a high breakdown of 50% with a $n^{-1/2}$ rate of convergence, and calculate the residuals (r_i).

Step 2: Based on the residuals in Step 1, compute the estimated scale $(\sigma)$of the residuals, $s = (1.4826)(the\ median\ of\ the\ largest\ (n-p) of\ the\ |r_i|)$.

Step 3: Using the estimated residuals $(r_i)$and the estimated scale (s), find the standardized residuals (e_i), where, $e_i = r_i/s$

Step 4: Compute the initial weight based on FMGT (4), where $\pi_i = min\ [1, \frac{\mathrm{CP_{FMGT}}}{FMGT}]$.

Step 5: Employ the initial weight (step 4) and the standardized residuals (step 3) to achieve a bounded influence function for bad leverage points, $t_i = e_i/w_i$.

Step 6: Use the weighted residuals (t_i) in first iteration WLS to estimate the parameters of the regression based on$\hat{\beta} = (X^T W\ X)^{-1} X^T W\ Y$, where the weight w_i is small for large residuals to get good efficiency (Tukey weight function is used in this chapter).

Step 7: Calculate the new residuals (r_i) from WLS and repeat steps (2-6) until the parameters converge.

Real examples and Simulation Study

A real examples and Simulation Study are carried out in this section to assess the performance of our proposed method.

Simulation Study

Consider linear regression model;

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

Where the error terms ε distributed as N(0,1). The *X variables are generated* from N(0,1). The contamination is created by randomly replaced some good observation in variable x1 (for GLPs and BLPs) and in y1 for vertical outliers, with arbitrarily large number equal to 100 with different percentage levels.

Table1 : Efficiency (%) and bias (parenthesis), 5% and 10% of vertical outliers (VOs)

VOs	n	method			
		OLScont.	GM6	GM-FIMGT	GM-RFIID
5%	50	6.1021 (4.9731)	90.2981 (0.0102)	**96.8012 (0.0121)**	**96.8810 (0.0119)**
	100	4.0372 (8.9114)	83.0918 (0.1404)	92.3823 (0.0110)	95.091 (0.0791)
	150	4.083 (6.1021)	85.2013 (0.1611)	94.218 (0.0101)	94.910 (0.0091)
	200	3.1940 (5.8105)	82.0132 (0.1184)	94.781 (0.0201)	95.9182 (0.0091)
	300	4.2910 (7.3091)	80.1820 (0.1302)	95.6812 (0.0391)	96.0172 (0.0192)
10%	50	5.3810 (9.2876)	84.2017 (0.2339)	92.010 (0.0192)	92.9123 (0.0282)
	100	3.0915 (10.971)	81.8120 (0.2091)	91,381 (0.0401)	92.912 (0.094)
	150	3.1910 (5.321)	86.2819 (0.1001)	93.912 (0.0192)	94.010 (0.011)
	200	3.2971 (5.0190)	79.231 (0.2320)	93.2918 (0.0981)	94.991 (0.0401)
	300	2.91081 (8.0110)	80.1201 (0.8891)	94.9180 (0.1029)	95.1810 (0.0912)

Table 2: Efficiency (%) and bias (parenthesis), 5% and 10% of Good Leverage points (GLPs) and Vertical Outliers (VOs)

GLP & VOs	n	method			
		OLScont.	GM6	GM-FIMGT	GM-RFIID
5%	50	15.342 (3.837)	90.032 (0.1501)	98.710 (0.016)	98.821) (0.0152)
	100	6.1021 (3.1231)	89.001 (0.1010)	99.1540 (0.093)	99.2620 (0.0987)
	150	9.4132 (3.0412)	90.891 (0.1172)	97.9243 (0.0791)	98.320 (0.0921)
	200	10.366 (2.2130)	92.001 (0.141)	98.0112 (0.0521)	99.001 (0.012)
	300	9.320 (3.1721)	94.891 (0.1891)	99.3901 (0.0435)	100.102 (0.001)
10%	50	16.2712 (3.326)	92.0124 (0.0812)	99.0012 (0.0110)	99.201 (0.0081)
	100	14.238 (4.910)	91.321 (0.0932)	98.0129 (0.0523)	98.981 (0.010)
	150	10.1293 (3.324)	90.991 (0.0890)	97.345 (0.0107)	98.791 (0.005)
	200	11.291 (2.923)	92.532 (0.0931)	98.981 (0.0867)	99.012 (0.0087)
	300	12.642 (3.781)	90.2171 (0.1039)	99.001 (0.099)	99.811 (0.010)

Table 3: Efficiency (%) and bias (parenthesis), 5% and 10% of Bad Leverage Points.

BLPs	n	method			
		OLScont.	GM6	GM-FIMGT	GM-RFIID
5%	50	25.0271 (1.1401)	95.6201 (0.0020)	92.3980 (0.0219)	92.427 (0.0212)
	100	14.9302 (1.0297)	92.9301 (0.0280)	90.5231 (0.0928)	91.213 (0.0107)
	150	12.9801 (1.0198)	93.612 (0.0112)	94.3932 (0.0181)	95.002 (0.0041)
	200	13.1981 (1.2101)	95.0012 (0.0019)	95.128 (0.0012)	95.976 (0.0023)
	300	11.9301 (1.1230)	95.3983 (0.0109)	95.6310 (0.0018)	96.104 (0.0013)
10%	50	28.3012 (1.0691)	93.256 (0.0138)	93.029 (0.0291)	93.058 (0.0293)
	100	15.0289 (1.0297)	92.086 (0.0192)	91.128 (0.0207)	91.205 (0.0201)
	150	12.0380 (1.9012)	90.0921 (0.0231)	92.417 (0.0126)	92.326 (0.014)
	200	13.1902 (1.1941)	91.3803 (0.0239)	92.530 (0.0117)	93.026 (0.0091)
	300	11.987 (1.1190)	92.498 (0.0281)	93.960 (0.0163)	94.102 (0.011)

Table 4: Efficiency (%) and Bias (parenthesis), 5% and 10% of Good and Bad Leverage point

GLPs & BLPs	n	method			
		OLScont.	GM6	GM-FIMGT	GM-RFIID
5%	50	20.361 (2.023)	93.104 (0.030)	98.3601 (0.0180)	98.823 (0.0128)
	100	18.341 (1.634)	91.436 (0.0741)	99.028 (0.0127)	100.081 (0.0039)
	150	16.3061 (1.136)	94.361 (0.0126)	99.305 (0.0019)	99.518 (0.0017)
	200	13.310 (1.037)	94.621 (0.011)	100.012 (0.0028)	100.280 (0.0012)
	300	12.936 (1.318)	93.497 (0.024)	100.031 (0.0014)	100.105 (0.0010)
10%	50	21.274 (2.0346)	94.783 (0.016)	97.457 (0.0112)	98.036 (0.0103)
	100	19.297 (1.513)	93.267 (1.063)	99.0362 (0.0135)	100.046 (0.0083)
	150	16.215 (1.153)	91.403 (0.0045)	100.036 (0.0051)	100.188 (0.0021)
	200	12.938 (1.0491)	93.304 (0.054)	101.231 (0.0030)	103.031 (0.0013)
	300	13.873 (1.318)	92.361 (1.073)	102.345 (0.00971)	104.136 (0.0053`)

Real examples

Gunst and Mason Data

The Gunst and Mason data set is our first example taken from Gunst and Mason (1980). This data set contains 49 observations, i.e. name of countries (Selected Demographic Characteristics of Countries of the World) and six independent variables (INFD, PHYS, DENS, AGDS, LIT, HIED) with response variable (GNP). The classification plot for the detection of Ios, SE of the estimated parameters and MAD are presented in table.

PLOT OF THE REAL DATA

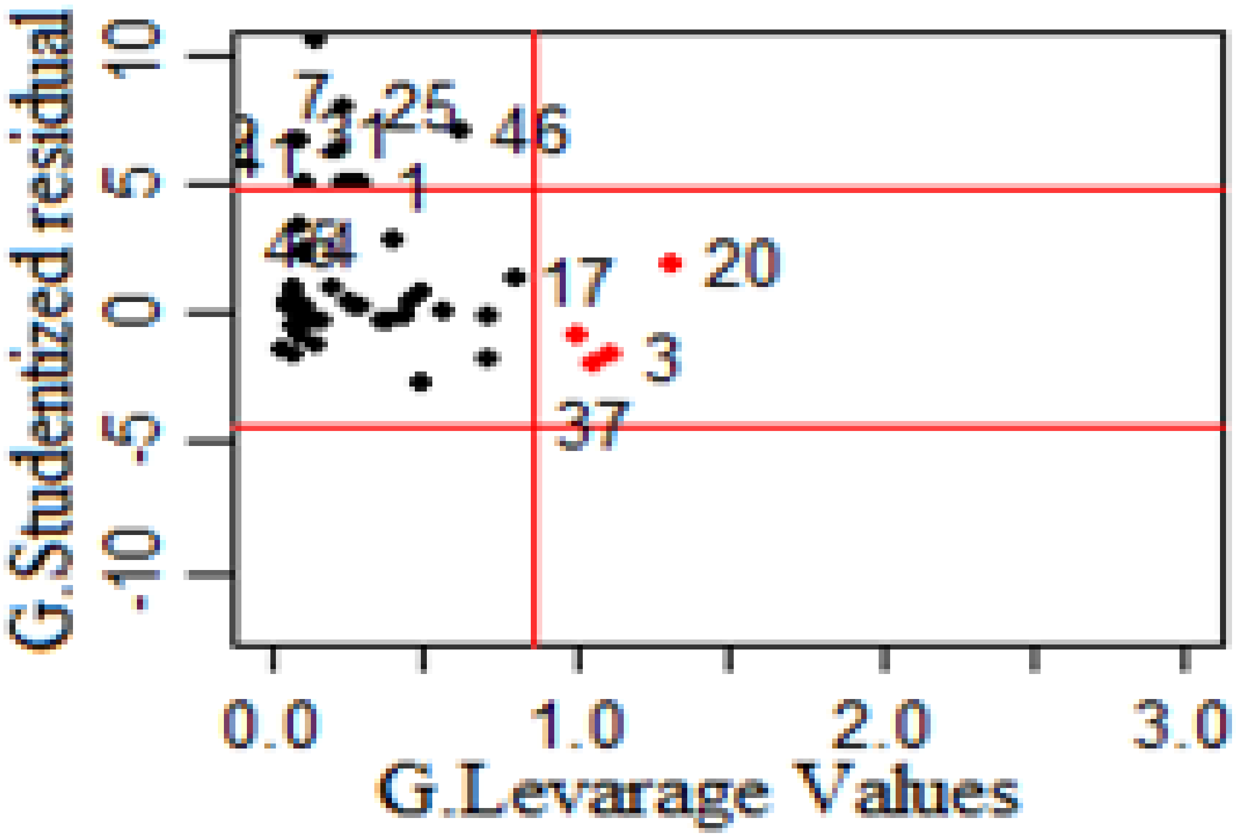


Table5: The results based on different regression methods for for Gunst and Mason data Set.

Variables	Methods			
	OLS	GM6	GM-FIMGT	GM-RFIID
	Parameter (Boot.SE)	Parameter (Boot.SE)	Parameter (Boot.SE)	Parameter (Boot.SE)
(INFD)	-0.2323 (0.2541)	-0.4765 (0.2460)	-0.1946 (0.1679)	-0.1635 (0.1107)
(PHYS)	-0.0063 (0.2127)	0.0158 (0.2070)	-0.0196 (0.0990)	-0.0199 (0.0871)
(DENS)	-0.1640 (0.6173)	-0.2959 (0.5303)	-0.1209 (0.4918)	-0.0542 (0.3843)
(AGNS)	0.1483 (0.5035)	1.0451 (0.9125)	0.0725 (0.4153)	0.0136 (0.4012)
(LIT)	0.2473 (0.2380)	0.0658 (0.1826)	0.2269 (0.1095)	0.1950 (0.1302)
(HIED)	0.4565 (0.2346)	0.4191 (0.2213)	0.2634 (0.1058)	0.2905 (0.0975)
Intercept (GNP)	0.0032 (0.2932)	0.0327 (0.2263)	-0.1724 (0.1775)	-0.2205 (0.1745)
MMAD	**0.5498**	**0. 4731**	**0.4574**	**0.3943**

Conclusion

- The main aim of this presentation is to show that the OLS gives the poor results when IOs are present in the data.
- The GM6 is not that efficient when GLPs are present in the data.
- The proposed GM-RFIID outperformed the other methods in the presence of IOs and good leverage points.

REFERENCES

Andersen, R. 2008. Modern methods for robust regression. *The United States of America: Sara Miller McCune. SAGE publications* 152.

Bagheri, A. & Midi, H. 2016. Diagnostic plot for the identification of high leverage collinearity-influential observations. *SORT-Statistics and Operations Research Transactions* 39(1): 51-70.

Chatterjee, S., Hadi, A. S. & Price, B. 2006. Simple linear regression. *Regression Analysis by Example*, Fourth Edition: 21-51.

Coakley, C. W. & Thomas, P. H. 1993. A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* 88(423): 872-880.

Dhhan,W., Rana,S. and Midi,H. 2017. A high breakdown, high efficiency and bounded influence modified GM estimator based on support vector regression, *Journal of Applied Statistics,* vol. 44, no. 4, pp. 700-714, 2017.

Habshah, M., Norazan, M. R. & Imon, A. H. M.R.2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics* 36(5): 507-520.

Hekimoğlu, S. & Erenoglu, R. C. 2013. A new GM-estimate with high breakdown point. *Acta Geodaetica et Geophysica* 4(48): 419-437.

REFERENCES

Lim, H. A.& Midi, H. 2016. Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics* 31(3): 859-877.

Midi, H., Talib, H., Arasan, J. & Uraibi, H.S. (2020). Fast and Robust Diagnostic Technique for the Detection of High Leverage Points. *Journal of Science and Technology. 28* (4).1203-1220.

Riazoshams, H. & Midi, H.2016. The Performance of a Robust Multistage Estimator in Nonlinear Regression with Heteroscedastic Errors. *Communications in Statistics-Simulation and Computation* 45(9): 3394-3415.

Stromberg, A. J., Ola, H. & Hawkins, D. M. 2000. The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association* 95(451): 853-864.

Wilcox, R. R. 2005. Introduction to robust estimation and hypothesis testing (Statistical Modeling and Decision Science). *Academic Press*.