10th MALAYSIA STATISTICS CONFERENCE 2023
Looking Beyond GDP: Toward Sosial Well-being and Environmental Sustainability

26th September 2023
Sasana Kijang, Bank Negara Malaysia

Statistical Communications

Generalized M-Estimator based on RFIID As a Remedy of Vertical Outliers and High Leverage Points

Habshah Midi and Hasan Talib Hend

Institute for Mathematical Research, UPM, 43400, Selangor

Abstract: Many do not realize that outliers in the *X* directions or high leverage points (HLPs), have the most detrimental effect on the computed values of various estimates which leads to misleading conclusion about the fitted regression model. Several robust methods have been proposed to remedy this problems that include the generalized – M (GM6) estimator. The GM6 uses a weighting function obtained from robust mahalanobis distance (RMD)- minimum volume ellipsoid (MVE)- based method in order to reduce the effect of HLPs. The shortcoming of this method is that it gives lower weight to HLPs irrespective of whether or not they are good (GLPs) or bad leverage points (BLPs). As such its efficiency tends to decrease as the number of GLPs increases in a data set. Moreover, the GM6 suffers from long computational running times. As a remedy to this problem, Generalized-M estimator based on fast improved MT (GM-FIMGT) which is an improvement of the GM6 is established. However, the GM-FIMGT is still not very efficient with regard to its parameter estimations and computational issues. This paper proposes a new robust GM estimator that incorporates a new weight function constructed from a new robust version of influential distance method (RFIID) which is based on Reweighted Fast Consistent and High Breakdown (RFCH) estimator. The numerical results clearly indicate that the GM-RFIID is more efficient and has less computational running times compared to some existing methods in this study.

Keywords: GM-estimator; high leverage points; index set equality; influence distance; RFCH

1.0 Introduction

The ordinary least squares (OLS) is the most widely used method in multiple linear regression due to its optimal properties and ease of computation. However, outliers have an adverse effect on the OLS estimates. In regression, outliers can be categorized as residual outliers, high leverage points and vertical outliers. Any observation that has large residual is referred to as residual outlier. Vertical outliers are those observations

that are extreme or outlying in y-coordinate. High leverage points (HLPs) not only fall far from the majority of predictor variables, but also are deviated from a regression line because they actually tilt the OLS line and their effect on OLS estimator is very large (Leroy & Rousseeuw 1987). According to Habshah et al. (2009), the detection of HLPs is very crucial due to its responsibility for misleading conclusion about the fitting of regression model, causing multicollinearity, masking/swamping of outlier etc. Therefore, the effect of HLPs should be minimized to get more efficient estimates. Nonetheless not all high leverage points have an adverse effect on the OLS estimates. Only bad leverage points, however have larger impact on the OLS estimates. Good leverage points follow the pattern of the majority of a data and hence contribute to the efficiency of an estimate (Chatterjee et al 2006).

Robust statistical methods that are less sensitive to outliers have been developed to rectify the problems of outliers. There are plenty of robust estimation methods, namely the M, MM, LMS,LTS, etc can be found in the literatures (Huber & Ronchetti 1981; Yohai 1987; Leroy & Rousseeuw 1987; Wilcox 2005). 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. It is worth mentioning that one of the goals of robust regression is to achieve a high breakdown point nearly 50%, bounded influence and high efficiency (Yohai & Ruben 1988). The M estimator has low breakdown point which is equals to (1/n) in which it can only handle outliers in the *Y* direction but not successful in handling outliers in the *X* direction. Both the S and MM estimators do not have bounded influence property but they have high breakdown and high efficiency (Hekimoğlu & Erenoglu 2013). The LTS and LMS also have high breakdown point, but they do not have bounded influence property and have very low relative efficiency which is close to 8% and 37%, respectively (Rousseeuw 1984; Rousseeuw 1993; Stromberg et al 1992). Since none of these estimators can handle high leverage point, Schweppes as described by (Hill & Paul 1977) suggested a new robust method called bounded influence Generalized M-estimator (GM-estimator) as a remedial technique for the sensitivity of M-estimator against high leverage points (see (Hill & Paul 1977; Andersen 2008)). Many types of GM-estimators were proposed in literature, such as in (Wilcox 2005; Andersen 2008) to produce good results in the presence of outliers and high leverage points. However, these methods have achieved a moderate BDP equals to 1/k, where k is the number of regression coefficients including the intercept (Simpson et al 1992). As a solution to these problems, multi-stage GM-estimators were developed.

The GM6 estimator proposed by (Coakley & Hettmansperger 1993; Wilcox 2005) is the most popular types of multi-stage GM-estimator. The GM6 estimator is based on robust mahalanobis distance (RMD) which uses minimum volume ellipsoid (MVE) as an initial of π–weight function (Rousseeuw 1985). The shortcoming of MVE is that it is not only taking a long computation running time, but tends to swamp some low leverage as high leverage points. Besides, the RMD which is based on MVE attempts to identify high leverage points without taking into consideration whether they are good or bad leverage points. Hence, the GM6-MVE considers the good leverage points as bad leverage points and its efficiency tends to decrease as the number of good leverage points increases.

The weaknesses of GM6-MVE have prompted Habshah et al. (2021) to put forward another version of GM estimator which is called the Fast GM estimator which is based on Improvised Generalized MT (FIMGT). It is denoted as GM-FIMGT estimator The Fast GM estimator utilizing high breakdown point S-estimator as an initial estimate and using π −weight function based on Fast Improvised Generalized MT (FIMGT). It has been shown that the GM-FIMGT is more efficient than the GM6 estimator. The only shortcoming of this method is that the FIMGT is based on index set equality (ISE). It is now evident that ISE is unstable as its algorithm depends on the selected initial subset,

h. Midi et al. (2020) exemplified that the final estimator of location and scatter of (ISE) is equivalent to minimum covariance determinant (MCD) if the same initial subset is employed, otherwise the results will be quite different. Moreover, the computational running times for the FIMGT still quite long.

Motivated by the fact that employing the weighting function in the GM-FIMGT algorithm has shown to be more efficient than the GM6, our primary aim is to perform some modifications to the existing GM-FIMGT algorithm by integrating a weight function based on our new approach for the detection of influential observations (RFIID) to produce more efficient estimates with less computing times. The proposed GM estimator will be based on RFIID and it is denoted as GM-RFIID. The proposed GM-RFIID estimator will be explained in detail in the following section.

2.0 Methodology.

Coakley & Hettmansperger (1993) introduced GM6 estimator which has high efficiency at normal distribution, bounded influence property and high breakdown point. It can be expressed as a solution of normal equations given by

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

where $\psi = \rho'$ is an influential function and $d_{i_i}, i = 1,2, \dots, n$ is the i^{th} initial weight function.

The GM estimators' main objective is to downweight HLPs which have large residuals. Coakley & Hettmansperger (1993) employed RMD based on MVE or MCD, using $\chi^2_{(0.95,p)} as$ cut-off points. Those detected HLPs will be assigned smaller weight while regular observations are given weight equals 1.0.

Afterwards, they defined the initial weight of the GM6 estimator as follows:

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

Bagheri & Habshah (2015) noted that this initial weight function inclines to swamp some low leverage points. Another limitation of this weight function is that, the RMD only identify HLPs (good and bad). This implies that the detected HLPs will be assigned low weight irrespective of whether they are GLPs or BLPs. Thus, as the number of GLPs increases, the GM6 efficiency tends to decrease because the precision of the parameter estimates may be contributed by GLPs as noted by Rousseeuw (1990). This is the reason why the GM6 - estimate is less efficient because both GLPs and BLPs are downweighted. The computation of GM6 estimator is very long since it uses MVE or MCD. This contributes to another weakness of GM6 estimator.

Our propose GM-RFIID estimator begins by establishing an algorithm of detecting influential observations (IOs) with the main aim of reducing their effects. According to Belsley *et al*. [18], IOs are those observations which are either done alone or together with several other observations, have an adverse effect on the computed values of various estimates. As such, to get efficient estimates, only genuine IOs are down weighted (Dhann et al. 2017). Hence, an efficient weight function should be formulated so that only genuine IOs will be assigned with smaller weight, regular observations and GLPs is assigned with weight 1. Then, the proposed weight function, d_i will be formulated as in (2).

The proposed GM-RFIID is briefly described according to the following steps:
Step I: Identify HLPs using DRGP-RFCH (see Habshah et al. 2021)
Step II: Compute Robust and Fast Improvised Influential Distance (RFIID) and cut-off point of RFIID, denoted as $CP_{RFIID}$to detect IOs (**not shown due to space constraint**).
Step III: Based on RFIID, Classify the observations into RO, GLO and IOs,

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

Step IV: Obtain the new initial estimate of our propose GM-FIID is given by

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

where, CP_{RFIID} is the cut-off point of RFIID

Step V: Based on the standardized residuals and the initial weight (Step IV), compute the bounded influence functions for IOs, $t_i = e_i/d_i$.

Step VI: Employ the weighted least squares (WLS) to estimate the parameters of the regression, $\hat{\beta} = (X^T W\, X)^{-1} X^T W\, Y$, where the weight w_i is reduced for large residuals to get good efficiency (Tukey weight function is utilised in this paper).

Step VII: Calculate the new residuals (r_i) from WLS and repeat steps (1-6) until convergence

3.0 Result:

Simulation study and two real examples are illustrated in this section to show that our proposed method is more efficient than the OLS, GM6, GM-FIMGT and GM-RFIID.Due to space limitations, the results of simulation study are not presented.

Gunst and Mason Data Set

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). Since the distribution of the GM-RFIID is intractable, bootstrap method is used to find the standard errors of its estimates. The parameter estimates and SE of the estimates (in parenthesis) of the four methods are exhibited in Table 1. The median absolute deviation for the residuals (MAD) are also presented in Table 1.

Table 6.1: The parameter estimates, SE and MAD 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**

4.0 Discussion and Conclusion:

The results of RFIID (not shown due to space limitation) reveal that the proposed RFIID technique diagnosed observations (1,4,7,25,31,41,42,45,46) as IOs while observations (3,7,20,37) as good IOs. The number of detected IOs will be utilised to determine the initial weights for GM-RFIID while the GM6s' initial weight only depends on the number of detected HLPs irrespective of whether they are GLPs or BLPs. The results of Table 1 show that the OLS gives the poor results. The results of GM6 is less efficient than the GM-FIMGT and GM-RFIID because GM6 gives smaller weight to HLPs irrespective whether they are good or bad leverage points. The good leverage points should not be down weighted because they may contribute to the precision of the estimates as their presence have no impact or less effect on the OLS estimates (see for instance; Rousseeuw and Van Zomeren, 1990; Andersen, 2008). It is very interesting to observe that the GM-RFIID is superior compared to GM-FIMGT, GM6, and OLS estimators, evident by having the smallest standard error of the estimates and NMAD. The results suggest that the GM-RFIID did remarkably well when compared to other methods and it is consistent with the results of the simulation study.

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.

Gray, J. B. 1985.Graphics for regression diagnostics. In American Statistical Association Proceedings of the Statistical Computing Section Washington, DC: *American Statistical Association*: 102-107.

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.

Hill, R.W. & Paul, W. H. 1977. Two robust alternatives to least-squares regression. *Journal of the American Statistical Association* 72(360a): 828-833.

Huber, P. J. & Ronchetti, E. M. 1981. Robust statistics. *Wiley Series in Probability and Mathematical Statistics. New York, NY, USA, Wiley-IEEE* 52: 54.

Imon, A.H.M.R. 2005. Identifying Multiple Influential Observations in Linear Regression. *Journal of Applied Statistics*, 32(9): 929-946.

Leroy, A. M., and Peter, J. R.1987.Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley*, 1987.

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.

Maronna, R. A.& Víctor, J. Y. 1976. Robust estimation of multivariate location and scatter. *Wiley StatsRef:* Statistics Reference Online.

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.

Rousseeuw, P. J. 1984. Least median of squares regression. *Journal of the American statistical association* 79(388): 871-880.

Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. *Mathematical statistics and applications* 8:283-297.

Rousseeuw, P. J.& Bert, C. V. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association* 85(411): 633-639.

Rousseeuw, P. J., and Croux, C. 1993 . Alternatives to the median absolute deviation. *Journal of the American Statistical association* 88(424): 1273-1283.

Simpson, D. G., David, R. & Raymond, J. C. 1992. On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* 87(418) : 439-450.

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*.

Yohai, V. J. 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*: 642-656.

Yohai, V. J., and Zamar R. H.1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association* 83(402): 406-413.

NOTE: THE REQUIRED NUMBER OF PAGES FOR PAPER IS SIX PAGES