PRIME MINISTER'S DEPARTMENT
DEPARTMENT OF STATISTICS MALAYSIA

Prediction Uncertainties: Dealing with Mulcollinearity Caused by High Leverage Collinearity Enhancing Observations in Regression Analysis

Prof. Dr. Habshah Midi

Department of Mathematics and Statistics, Faculty of Science and Institute For Mathematical Research Universiti Putra Malaysia

BEAUTIFUL MALAYSIA

PRESENTATION OUTLINE

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

INTRODUCTION

- Research related to business often uses regression analysis to predict future outcome. The ordinary least squares(OLS) method is the most popular technique in regression analysis due to its optimal properties and ease of computation.
- However, the OLS estimates are much affected when multicollinearity (when two or more independent variables are highly correlated) is present in a data.
- Its result in wrong sign problem, produce large standard errors of regression estimates (Midi, Bagheri and Imon., 2012, Comp. Stat &Simu.).
- Relying on the OLS method may give inefficient estimates and inaccurate predictions and causing uncertainties in predicting future outcomes.
- VIF commonly used diagnostic method to identify multicollinearity.VIF<5 indicate no multicollinearity. VIF between 5 and 10 moderate and severe multicollinearity, VIF>10.

INTRODUCTION

- Many are not aware that high leverage point which fall far from majority of the explanatory variables, can induce or disrupt multicollinearity pattern in a data. Observations responsible for this are known as high leverage collinearity influential observations (HLCIO) (Bagheri, Midi and Imon, 2012, Comp. Statistics, Simu.& Comp., 2011, Math. Prob in Engineering).
- HLP that induce multicollinearity are referred as HLC-Enhancing Observations while those that reduce multicollinearity in their presence are called HLC-reducing observations(Midi and Bagheri, 2011, Math Prob in Engineering; Bagheri, Midi and Imon, 2012).
- When multicollinearity is due to highly correlated predictor variables, ridge regression, latent root regression and Jackknife ridge regression can be used to remedy this problems (Hoerl and Kennard, 1970, Singh et al. , 1986).

INTRODUCTION

- Bagheri, Midi and Imon, (2012), Comp. Stat. Simul & Comp. pointed out that when multicollinearity is caused by HLCEO, those suggested method is inappropriate.
- Not much research is done on the remedy of HLCEO. Imon and Khan (2003) suggested deletion of suspected HLPs from the analysis using Generalized potential (GP) method.
- According to Midi et al. (2009), the GP is not successful in detecting genuine HLPs.
- Since HLPs is the caused of multicollinearity, to remedy the problem of HLCEO, the HLPs first need to be correctly identified and their effect on the parameter estimates, need to be reduced.
- Hence, robust methods which are known to be resistant to HLPs need to be employed.

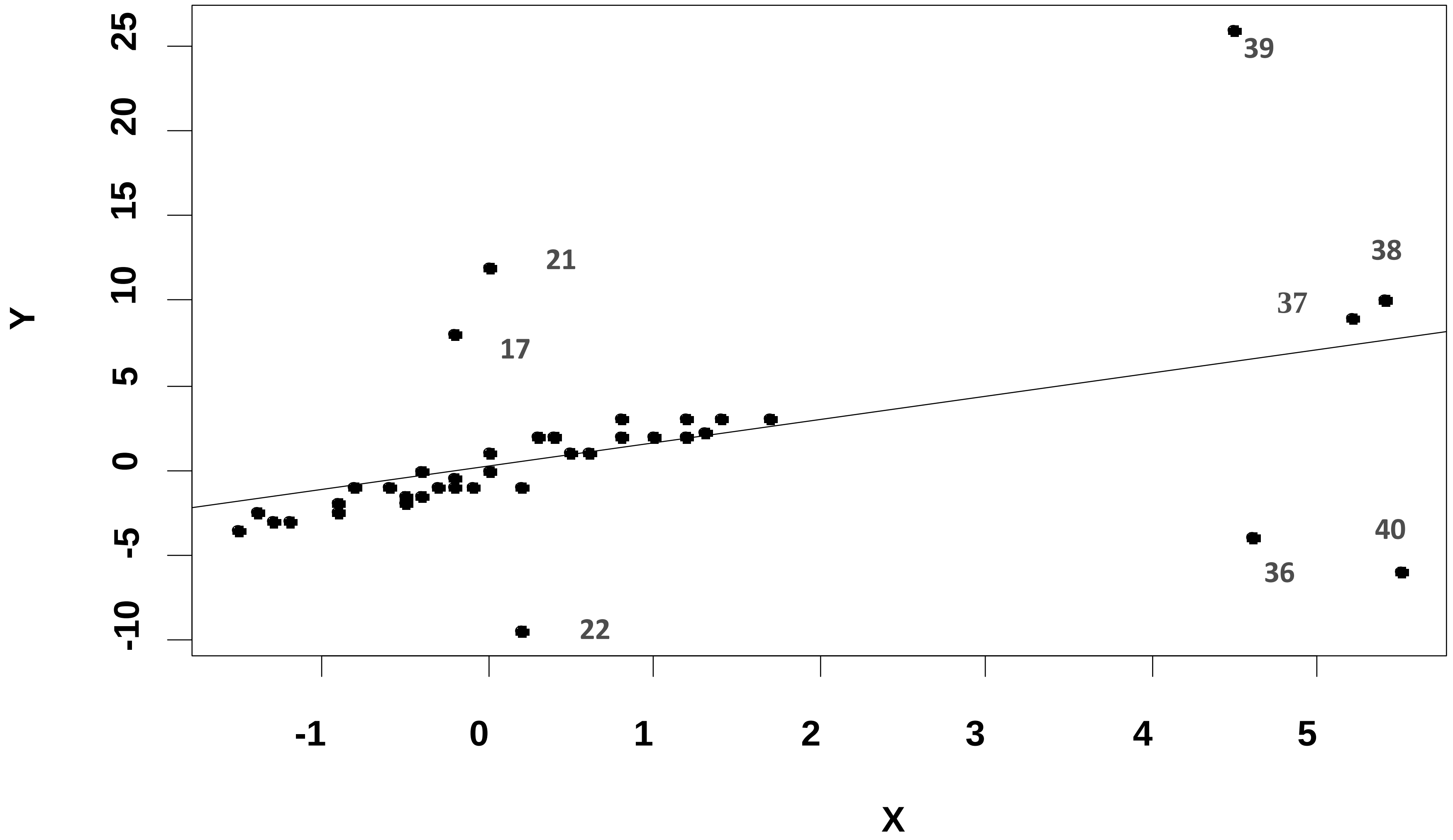
OBJECTIVES

- To develop a method of identification of HLCIO that can change multicollinearity pattern of data: HLC Enhancing and HLC Reducing Observations.
- To establish a robust method that is resistant to HLPs.
- To propose correct estimation method to remedy the problem of HLCEO so that future outcome can be predicted with at least 95% certainty.
- To apply the proposed method to real data.

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
-10
-5
0
5
10
15
20
25
-1
0
1
2
3
4
5
39
21
38
37
17
40
36
22

METHODOLOGY

- Two steps is proposed to analyse a dataset for multiple linear regression.
- Step1 : Identify the existence of HLCEO
- Step 2: Apply Generalized-M estimator based on fast improvised GMT estimator for data having HLCEO.

Step 1:DEVELOP COLLINEARITY INFLUENTIAL OBSERVATION DIAGNOSTIC MEASURE BASED ON A GROUP DELETION APPROACH

- High leverage points can induce or disrupt multicollinearity.
- Observations responsible for this problem are generally known as collinearity-influential observations.
- Development of collinearity-influential observation diagnostic measures has not been reported extensively in the literature (Hadi, 1988; Sengupta and Behimasankaram, 1997; Bagheri and Midi, 2012, Comp. Stat., Bagheri et al., Math Prob in Eng, 2012, Midi&Bagheri, 2015, Statistics& Operation Research J.).The weakness of Hadi and SengupataBehimasankaram (lack of symmetry and no cutoff points) motivated us to propose another measure, .
  - The proposed high leverage collinearity-influential measures based on DRGP (HLCIM (DRGP)), denoted as $\delta_i^{(D)}$ *and d*efined and summarized as:

$$\delta_i^{(D)} = \begin{cases} log\dfrac{k_{(D)}}{k_{(D-i)}} & if\ i \in D \quad n(D) \neq 1 \\ log\dfrac{k_{(D)}}{k} & if\ i \in D \quad n(D) = 1 \\ log\dfrac{k_{(D+i)}}{k_{(D)}} & if\ i \in R \end{cases}$$

where D is the group of multiple hlps diagnosed by DRGP(ISE), n(D) is the size of the D group.$k_{(D-i)}$ indicate the condition number of the X matrix without the entire group of D minus the ithhlps where i belongs to D group.

Diagnostic Robust Generalized Potential based on Index Set Equality (ISE) to Detect HLP

❖Lim and Habshah (Computational Statistics,2016) (see also Habshah, Norazan et al. (2009), J. of Applied Stat., Mazlina & Habshah (2015), Pak. J of Statistics) formulated fast diagnostic robust generalized potential (DRGP-ISE) to detect multiple high leverages. It consists of two steps.

Step 1) suspect high leverage points are determined by the robust $i = 1, 2, \ldots, n$

Robust Mahalanobis Distance based on Index Set Equality:

$$RMD_i = \sqrt{\left(X - T_R(X)\right)^T C_R(X)^{-1}\left(X - T_R(X)\right)}$$

where $T_R(X)$ and $C_R(X)$ are robust locations and covariance estimtes of the ISE, respectively.

A set of ‘good’ cases ‘remaining’ in the analysis denoted by R and deleted by D

Index Set Equality (ISE)

The running time of MVE ,MCD and even Fast MCD is still very long. To reduce the running time, we proposedusing Index Set Equality (ISE) which is
another new technique from fast MCD. Denoting $I_{old} = \{\pi_{(1)}^{old}, \pi_{(2)}^{old}, ..., \pi_{(h)}^{old}\}$ as the index set that correspond to the sample items in H_{old} when their Mahalanobis Distance squares are arranged in increasing order and $I_{new} = \{\pi_{(1)}^{new}, \pi_{(2)}^{new}, ..., \pi_{(h)}^{new}\}$ the index set that correspond to the sample items in H_{new}. The ISE is summarized
as follows;

Step 1: Select an arbitrarily a subset H_{old} containing h different observations.
Step 2: Calculate the average vector $\bar{T}_{Hold}$ and covariance matrix C_{Hold} of all observations belong to H_{old}.
Step 3: Compute $d_{old}^2(i) = (t_i - \bar{T}_{Hold})' C_{Hold}^{-1}(t_i - \bar{T}_{Hold})$ for $i = 1,2,..., n.$
Step4: Arrange $d_{old}^2(i)$ for $i = 1,2,..., n.$ in increasing order
$d_{old}^2(\pi(1)) \le d_{old}^2(\pi(2)) \le .... \le d_{old}^2(\pi(n))$ where π is permutation equal to {1,2,…,n}.
Step5: Construct $H_{new} = \{t_{\pi(1)}, t_{\pi(2)}, ..., t_{\pi(h)}\}$
Step 6: If $I_{new} \neq I_{old}$ let $H_{old} := H_{new}$ and $C_{Hold} := C_{Hnew}$,compute $\bar{T}_{Hnew}$ and let $\bar{T}_{Hold} := \bar{T}_{Hnew}$
then go to step(3) .Otherwise, the process is stopped .
The running time of ISE is much faster than fast MCD because ISE only involves a comparison of two index sets.

- Step 2) 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 3:The proposed high leverage collinearity-influential measures based on DRGP (HLCIM (DRGP)), denoted as $\delta_i^{(D)}$ *and d*efined and summarized as:

$$\delta_i^{(D)} = \begin{cases} log\,\dfrac{k_{(D)}}{k_{(D-i)}} & if\ i \in D \quad n(D) \neq 1 \\ log\,\dfrac{k_{(D)}}{k} & if\ i \in D \quad n(D) = 1 \\ log\,\dfrac{k_{(D+i)}}{k_{(D)}} & if\ i \in R \end{cases}$$

where D is the group of multiple hlps diagnosed by DRGP(ISE), *n*(D) is the size of the D group.$k_{(D-i)}$ indicate the condition number of the X matrix without the entire group of D minus the ith hlps where i belongs to D group.

The well-known Hawkins, Bradu, and Kass (1984) data is used to show the merit of our proposed method.This artificial three-predictor data set contains 75 observations with 14 high leverage points (cases 1-14); cases 1-10 bad hlp, cases 11-14 good hlp.

Table 1. Collinearity-influential measures for Hawkins-Bradu-Kass data

Index	$k_{(t)}$	δ_t	l_t	$\delta_t^{(D)}$
		(-.008)	(-.004)	(-.019)
		(-0.048)	(-0.021)	(-.022)
1	13.221	-0.027	-0.012	-0.228
2	13.183	-0.03	-0.013	-0.241
3	13.289	-0.022	-0.01	-0.234
4	13.18	-0.03	-0.013	-0.254
5	13.188	-0.029	-0.013	-0.248
6	13.185	-0.03	-0.013	-0.24
7	13.166	-0.031	-0.014	-0.248
8	13.237	-0.026	-0.011	-0.227
9	13.235	-0.026	-0.011	-0.242
10	13.327	-0.019	-0.008	-0.226
11	13.06	-0.039	-0.017	-0.29
12	13.424	-0.012	-0.005	-0.272
13	13.035	-0.041	-0.018	-0.319
14	17.125	0.26	0.101	-0.391
15	13.67	0.006	0.003	-0.005
16	13.752	0.012	0.005	0.01
17	13.644	0.004	0.002	0.002
18	13.589	0	0	-0.003
19	13.669	0.006	0.003	-0.002
20	13.708	0.009	0.004	-0.005
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
70	13.582	0	0	-0.004
71	13.611	0.002	0.001	-0.006
72	13.608	0.002	0.001	-0.003
73	13.584	0	0	-0.002
74	13.611	0.002	0.001	-0.002
75	13.651	0.005	0.002	0.009

Multicollinearity Diagnostic Measures

Classical Variance Inflation Factor

Marquardt [11] developed a diagnostics method which is known as variance inflation factor (CVIF) to detect multicollinearity in a data. The CVIF is the most popular method to identify multicollinearity and it is given by:

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1, 2, \ldots, p \qquad (1)$$

where R_j^2 is the coefficient of multiple determination when is regressed on other $X_{(p-1)}$ variables in the model, using the Ordinary Least Squares (OLS) method.

$VIF_{max} \in (5, 10)$ moderate multicollinearity among all of predictors

$VIF_{max} \geq 10$ severe multicollinearity (Belsley et al.[1]).

Table 2. Collinearity diagnostics for Hawkins-Bradu-Kass data

Diagnostics	Status	1	2	3
Pearson correlation coefficient	Original data	r_{12}= 0.946	r_{13}= 0.962	r_{23}= 0.979
	Without observations 1 –14	r_{12}= 0.044	r_{13}= 0.107	r_{23}= 0.127
VIF > 5	Original data	13.432	23.853	33.432
	Without observations 1 – 14	1.012	1.017	1.027
Condition index of X matrix > 10	Original data	13.586	7.839	1.00
	Without observations 1 – 14	3.275	2.946	1.00

Step 2: Proposed GM estimator to remedy HLCEO

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 $\pi_i, i = 1,2,\dots,n$ 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:

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

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, Midi, Shelan et al. (2021) propose a relatively easy and fast method to detect bad leverage points and outliers. Then only minimize the weights of bad leverage points and outliers.

The propose method is based on the procedure of constructing diagnostic plot of Alguraibawi, Midi and Imon (Math Problem Engineering, 2015)(see also Midi & Bagheri, 2015, Statistics& Operation Research J) for classifying observations into outliers, good and bad leverage points, with slight modification to make it fast by employing RMD based on Index Set Inequality (ISE).

The proposed GM-FIMGT estimator is almost similar to that of Dhhan, Midi, Sohel (Journal of Appl Stat. 2016). The only different is the calculation of the initial weight function. Their weight is based on support vector regression method. 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 $\pi_i = min\ [1, \frac{\mathrm{CP}_{\mathrm{FMGT}}}{FMGT}]$.

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.

The algorithm of the classification of observations into outliers and bad high leverage points is summarized as follows:

Classification Step I: Identify the suspected vertical outliers by using the robust Reweighted Least Squares (RLS) based on Least Median of Squares (LMS).
Denote these suspected outliers by *L* set.

Classification Step II: Identify the suspected high leverage points (HLP) by using Diagnostic Robust Generalized Potential based on Index Set Inequality (DRGP (ISE)) proposed by Lim and Midi (2015).

whereby, the Robust Mahalanobis Distance that they employed is based on Index Set Inequality (ISE).
Denote this set of suspected HLPs by *H set*.

Rohayu (2013) has proved that the ISE provides the same final location and scale estimator as that obtained by using MCD if the same initial subset is employed.

It has been shown by Lim and Midi (2015) that ISE is much faster than the commonly used method, namely MVE or MCD.

Classification Step III: From steps 1 and 2, observations that correspond to the union of *L* set and *H* set will be considered as deletion group/set, *D* and the remaining data are labeled *as R set.*

Classification Step IV: Fit the remaining R set using OLS method to estimate the regression coefficients $(\hat{\beta}_R)$, residuals $(\hat{\varepsilon}_{i.R})$, hat values $(w^*_{ii.R})$ standard deviation $(\hat{\sigma}_R)$ and standard deviation with the i^{th} case deleted $(\hat{\sigma}_{R-i})$. The Fast Improvised Generalized Studentized Residuals (FIMGT) (a slight modification of Imon's MGT (2005) is then defined as follows;

$$\mathrm{FIMGt}_i = \begin{cases} \dfrac{\hat{\varepsilon}_{i.R}}{\hat{\sigma}_{R-i}\sqrt{1 - w^*_{ii.R}}} & \text{for } i \in R \\ \dfrac{\hat{\varepsilon}_{i.R}}{\hat{\sigma}_{R}\sqrt{1 + w^*_{ii.R}}} & \text{for } i \notin R \end{cases}$$

The observations are declared as vertical outliers if they have values of FIMGT greater than its cutoff point (CP_{FMGT}). The CP_{FMGT} is defined as follows:

$$CP_{FMGT} = \text{Median}(FMGT) + c\ \text{MAD}(FMGT_i)$$

where c is equals to 2 or 3.

Alguraibawi et al. (2015) then suggested a rule for classifying observations as follows,

i. Regular Observation (RO): An Observation is declared as a "RO" if $|FMGTi \leq CP_{FMGT} \text{and} p^*_{ii} \leq \text{Median}\ (p^*_{ii}) + c\ \text{MAD}\ (p^*_{ii})$

ii. Vertical Outlier (VO): An Observation is declared as a "VO" if $|FMGt_i| > CP_{FMGT} \text{and} p^*_{ii} \leq \text{Median}\ (p^*_{ii}) + c\ \text{MAD}\ (p^*_{ii})$

iii. GLPs: An Observation is declared as a GLP if $|FMGt_i| \leq CP_{FMGT} \text{and} p^*_{ii} > \mathit{Median}\ (p^*_{ii}) + c\ \text{MAD}\ (p^*_{ii})$

iv. BLPs: An Observation is declared as a BLP if $|FMGt_i| > CP_{FMGT} \text{and} p^*_{ii} > \mathit{Median}\ (p^*_{ii}) + c\ \text{MAD}\ (p^*_{ii})$

Figure 1: DRGP against Fast Generalized Student zed Residuals

Modified Generalized standard residual

Vertical Outliers	Bad Leverage Points
Regular Observations	Good Leverage Points
Vertical Outliers	Bad Leverage Points

DRGP

It is clearly seen from the above table, that the vertical outliers and bad leverage points are detected based on our proposed FIGMT method. Alguraibawi et al. (2015) have shown that the MGT is very successful in detecting the bad high leverage points and vertical outliers. According to Dhhan, Sohel, Midi (2016) and Rashid, Midi, Dhann (2021;IEEE Access), the effective of the weight function depends on the efficient diagnostic method of identifying outliers. Therefore, the initial estimate of our propose GM-FGMT is given by,

$$\pi_i = \min\left[1, \left(\frac{\mathrm{CP_{FMGT}}}{FIGMT}\right)\right], i = 1,2, \ldots, n$$

where $\mathrm{CP_{FMGT}}$ was defined above.

Example: Aircraft data set

❖ The Aircraft dataset, which is taken from Gray (1985) is used to illustrate the merit of our proposed plot. This dataset contains 23 cases with four predictor variables (Aspect ratio, Lift-to-drag ratio, Weight of the plane, and Maximal thrust) and the response variable is the Cost.

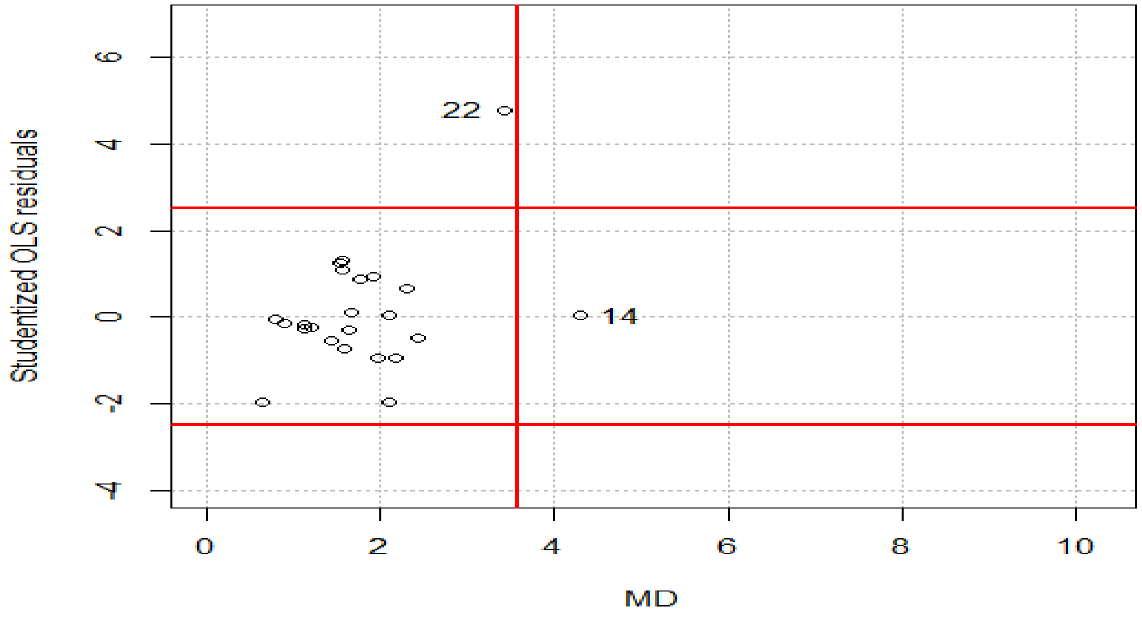


Figure 1: The Studentized OLS res. vs. MD for the Aircraft data

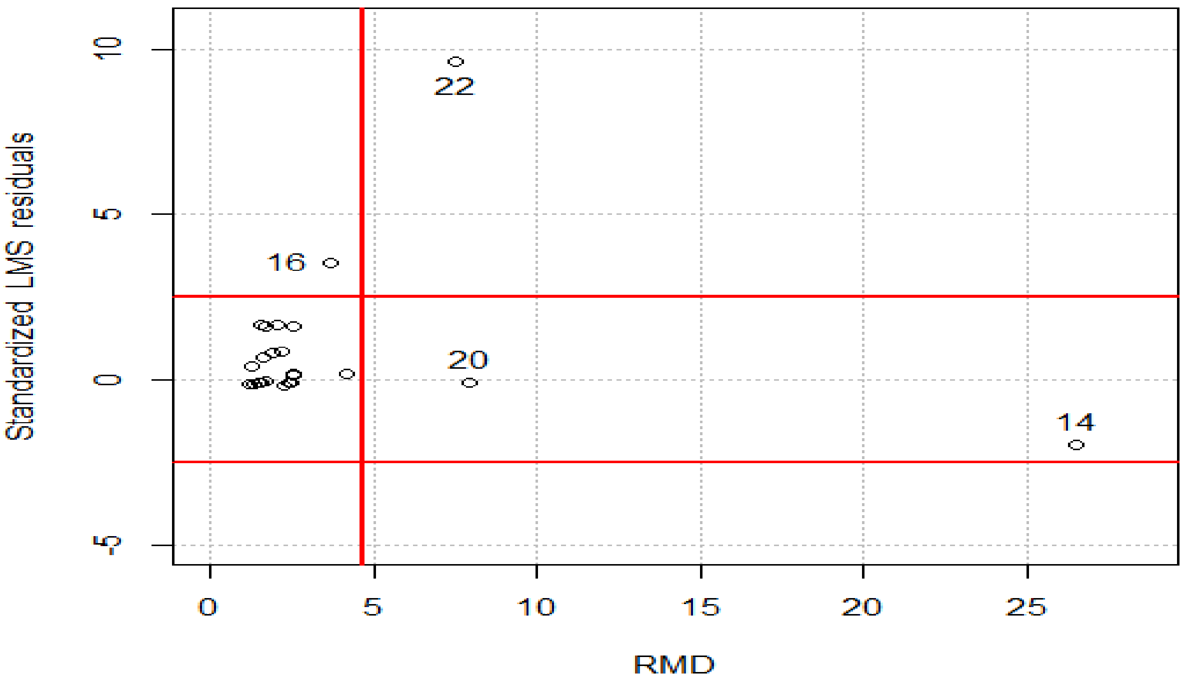


Figure 2: The Standardized LMS res. vs. RMD for the Aircraft data

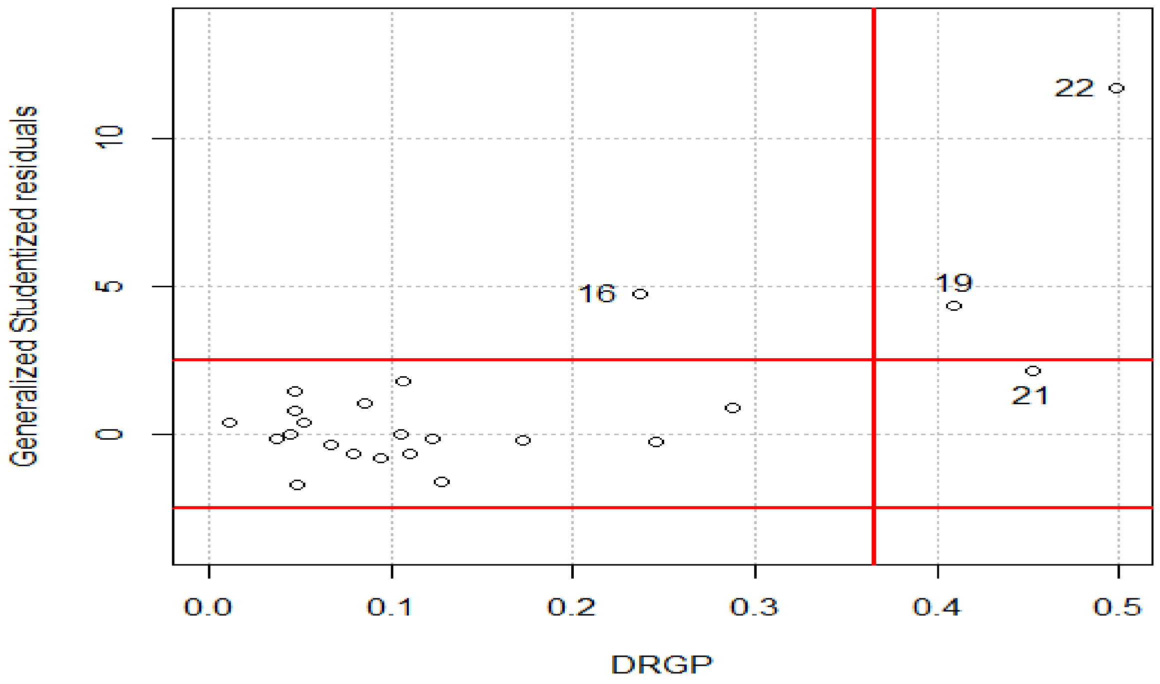


Figure 3: The Mod. Generalized studentized res. vs. DRGP for the Aircraft data

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). In order to create good and bad leverage points, certain clean observations are replaced with contaminated data. To create bad leverage points, the first 100(α/2) % for both *X* and *Y* variables are replaced by contaminated observations generated from N(1,10). To create good leverage points, the last 100(α/2) % observations of *X's* variables are replaced with contaminated observations distributed as N(1,10).

REMEDY HLCEO-SE and Ratio of the estimated Ridge, GM6, MM and GM-FIMGT for clean generated data set

n	Coef.	VIF	Ridge		GM6		MM		GM-FIMGT	
			S.E	Ratio	S.E	Ratio	S.E	Ratio	S.E	Ratio
20	β_1	1.14	0.7662	94.96	0.7472	97.38	0.7352	98.97	0.7355	98.93
	β_2	1.11	0.6953	94.61	0.6784	96.96	0.6656	98.83	0.6655	98.84
	β_3	1.12	0.6812	94.47	0.6649	96.78	0.6504	98.94	0.6522	98.67
40	β_1	1.05	0.4432	95.17	0.4363	96.68	0.4275	98.67	0.4279	98.57
	β_2	1.06	0.4012	95.29	0.3916	97.63	0.3883	98.45	0.3877	98.61
	β_3	1.05	0.4911	96.95	0.4905	97.06	0.4835	98.47	0.4851	98.14
100	β_1	1.03	0.3072	94.56	0.2985	97.32	0.2932	99.08	0.2921	99.45
	β_2	1.02	0.2979	95.37	0.2936	96.76	0.2883	98.54	0.285	99.68
	β_3	1.02	0.2494	94.23	0.2407	97.63	0.2367	99.28	0.2351	99.96
200	β_1	1.01	0.2165	95.94	0.2133	97.37	0.2088	99.47	0.208	99.86
	β_2	1.01	0.2127	96.90	0.2087	98.75	0.2066	99.76	0.2069	99.61
	β_3	1.01	0.2145	96.69	0.2138	97.01	0.2083	99.57	0.2078	99.81

REMEDY HLCEO SE and Ratio of the estimated OLS, Ridge, GM6, MM and GM-FIMGT for contamination generated data

A	Coef.	VIF	OLS		Ridge		GM6		MM		MGM	
			S.E	Ratio	S.E	Ratio	S.E	Ratio	S.E	Ratio	S.E	Ratio
n = 20												
0.05	β_1	9543	1.4597	31.27	1.4866	30.7076	0.6198	73.65	0.5876	77.69	0.5241	87.1
	β_2	9510	1.5408	28.72	1.5934	27.97	0.6196	71.42	0.5693	77.72	0.5142	86.06
	β_3	9756	1.4962	29.04	1.5232	28.52	0.6247	69.55	0.5694	76.3	0.4969	87.44
0.1	β_1	9913	1.5144	33.1	1.641	30.55	0.7293	68.74	0.6308	79.47	0.5978	83.86
	β_2	9540	1.5088	31.89	1.502	32.04	0.7354	65.43	0.6304	76.32	0.5664	84.96
	β_3	9732	1.4917	33	1.7322	28.42	0.7106	69.28	0.6122	80.42	0.5702	86.34
n = 40												
0.05	β_1	4357	1.1962	22.49	1.2232	22.01	0.3607	74.58	0.3415	78.78	0.2771	97.08
	β_2	4489	1.2128	20.91	1.3571	18.69	0.406	62.46	0.3217	78.83	0.2656	95.5
	β_3	4349	1.2211	22.21	1.2778	21.22	0.4053	66.92	0.3328	81.49	0.2832	95.75
0.1	β_1	8394	1.2393	21.87	1.2782	21.2	0.4119	65.79	0.3616	74.94	0.329	82.38
	β_2	8644	1.2607	20.34	1.2997	19.73	0.4185	61.27	0.3457	74.16	0.3226	79.49
	β_3	8608	1.2572	20.19	1.2663	20.04	0.43	59.02	0.3511	72.28	0.3219	78.85
n = 100												
0.05	β_1	4030	1.0894	12.3	1.2616	11.62	0.1778	75.38	0.1756	76.31	0.1487	90.15
	β_2	4161	1.0715	11.63	1.0686	11.66	0.1715	72.65	0.1687	73.89	0.1387	89.82
	β_3	4114	1.102	12.39	1.2262	11.14	0.185	73.83	0.1834	74.44	0.1517	90
0.1	β_1	8085	1.0946	11.71	1.2544	10.22	0.2498	51.33	0.196	65.42	0.1884	68.06
	β_2	8195	1.1476	10.95	1.2192	10.31	0.2348	53.53	0.206	61.01	0.1793	70.11
	β_3	8156	1.0947	11.09	1.1518	10.54	0.2553	47.55	0.1893	64.15	0.1805	67.28

Real examples

Commercial Properties Dataset

This dataset is taken from Neter et al. (2004). This data set is non collinear and contained 81 observations with three explanatory variables, namely, age , operating expenses and taxes. The dependent variable is the rental rates. Neter et al. (2004) noted that this data set has 19 HLPs. However, this dataset is not HLCIO. In order to investigate the effect of HLPs on non-collinearity pattern among the explanatory variables, we created severe multicollinearity in this dataset by adding HLCIOs. The first observations for each of the two explanatory variables was kept fixed with values 300.

Table 1: VIF values and Person correlation of coefficients (r) for original and modified Commercial properties dataset

Data set	r	VIF		
		x_1	x_2	x_3
Original Data	$r_{1.2} = 0.387$ $r_{1.3} = 0.226$ $r_{2.3} = 0.366$	1.187	1.302	1.1668
Modified data	$r_{1.2} = 0.982$ $r_{1.3} = 0.025$ $r_{2.3} = 0.009$	29.595	29.577	1.007

Table 2: Standard deviations of the estimates of Original (modified) Commercial dataset

Parameter	OLS		Ridge		GM6		GM-FIMGT	
	Est.	SE	Est.	Boot.	Est.	Boot	Est.	Boot
$\hat{\beta}_0$	11.61(14.89)	0.67(0.28)	9.77(14.46)	0.73(0.31)	11.09(11.15)	0.48(0.42)	11.23(11.80)	0.42(0.15)
$\hat{\beta}_1$	-0.12 (-0.13)	0.03(0.03)	-0.13(-0.14)	0.03 (.03)	-0.11(-0.11)	0.02(0.01)	-0.12(-0.13)	0.02(0.01)
$\hat{\beta}_2$	0.44(0.13)	0.07(0.03)	0.63(0.16)	0.07(.03)	0.04(0.47)	0.05(0.04)	0.46(0.41)	0.04(0.02)
$\hat{\beta}_3$	2.44(0.18)	1.21(1.29)	2.48(0.59)	1.19(1.36)	1.21(-0.32)	1.20 (1.29)	3.57(3.41)	0.69 (0.67)

Conclusion

- The main aim of this presentation is to show that HLPs can change the multicollinearity pattern of data (HLCIO): HLCEO and HLCRO.

- HLCIOM diagnostic measure can be used to detect HLCIO.

- The OLS method performs poorly for data having HLCEO.

- Ridge regression, latent root regression and Jackknife Ridge regression are incorrect remedial measure for multicollinearity problem caused by HLCEO.

- In this regard using either OLS, Ridge regression, latent root regression and Jackknife Ridge regression will give inefficient estimates, inaccurate prediction, misleading conclusion, and hence lead to prediction uncertainties.

- Suggest to use GM-FIMGT for data having HLCEO to avoid prediction uncertainties.
- For HLCEO suggest using GM-FIMGT, HLCRO suggest using Jackknife Ridge FIMGT to avoid incorrect interpretation, incorrect remedial measures and misleading conclusion.

References

Habshah, M., Norazan, M.R. and 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.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). Introduction to linear regression anlaysis, 3rd Ed. John Wiley and Sons, New York..

Lim, H.A. and Midi, H, (2016), Diagnostic Robust Generalized Potential based on Index Set Equality for identification of High Leverage Points in linear regression. *Computational Statistics 3(31), 859-877*.

Rashid, A.M., Midi, H, Dhnn, S.W. and Arasan, J. (2022). The Detection of Outliers in high dimensional data using nu support vector regression, *Journal of Applied Statistics*. 49 (10), 2550-2569.

Bagheri, A., Midi, H.,& Imon, R. H. M. R. (2012). *A Novel Collinearity-Influential Observation Diagnostic Measure Based on A Group Deletion Approach*. Communications in Statistics-Simulation and Computation, 41(8), 1379-1396.

References

Hadi, A. S. (1988). *Diagnosing Collinearity-Influential Observations*. Computational Statistics & Data Analysis, 7(2), 143-159.

Sengupta, D. and Bhimasankaram, P. (1997).*On the Roles of Observations in Collinearity in the Linear Model*. Journal of American Statistical Association. 92:1024-1032.

Bagheri, A. and Habshah, M. (2012a). *On the performance of the measure for diagnosing multiple high leverage collinearity-reducingobservations*. Mathematical Problems in Engineering. vol. 2012, Article ID 531607, 16 pages, 2012. doi:10.1155/2012/531607

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. and Van Zomeren, B. (1990). *Unmasking Multivariate Outliers and Leverage Points*. Journal of American Statistical Associations. 85: 633-639.

References

Hoerl, A. E., & Kennard, R. W. (1970). *Ridge Regression: Biased Estimation for Non-orthogonal Problems*. Technometrics, 12(1), 55-67.

Batah, F. S. M., Ramanathan, T. V., & Gore, S. D. (2008). *The Efficiency of Modified Jackknife and Ridge Type Regression Estimators: A Comparison*. Surveys in Mathematics and its Applications, 24(2), 157-174.

Midi, H., M., Mohammed, A. M., Imon A. H. M. R. and Sohel R. (2015). *A New Classification Scheme for the Identification of Bad Influential Observations.* Mathematical Problems in Engineering.

Midi, H, Ismaeel, S,S.Arasan, J. And Mohamed. M. (2021). Simple and Fast Generalised M estimator and its applications to real data. *Sains Malaysiana*. 50(3), 859-867.

Coakley C.W. and Hettmansperger, T.P. (1993). A bounded-influence high breakdown efficient regression estimator. *Journal of the American Statistical Association. 88. 872-880.*

Kamarruzzaman, M.D. and Imon, A.H.M.R. (2002). High leverage point another source of multicollinearity. *Pakistan Journal of Statistics. 18(3), 435-448.*

References

Neter, J. Kutner, M.H. Wasserman, W. and Nachtsheim, SC.J.(2004). Applied Linear Regression model. New York: Mcgraw Hill.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). Robust Statistics. New York: John Wiley.

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

Imon, A.H.M.R. and Khan, M.A.I. *(2003).* A solution to the problem of multicollinearity caused by the presence of multiple high leverage points . *Journal of Statistical Sciences. 2, 37-50.*

THANK YOU

StatsMalaysia

www.dosm.gov.my

9TH MALAYSIA STATISTICS CONFERENCE

Department of Statistics Malaysia

4TH OCT. 2022 (VIRTUAL) & 5TH OCT. 2022 (ILSM, SUNGKAI, PERAK)

Dealing with Uncertainties: Unearthing Measures for Recovery

Organised by: PRIME MINISTER'S DEPARTMENT DEPARTMENT OF STATISTICS MALAYSIA, BANK NEGARA MALAYSIA, MALAYSIA INSTITUTE OF STATISTICS