The Effect of Outliers on the Within Group Least Squares Estimates For Fixed Effect Panel data Model and How to Remedy Them

Prof Habshah Midi
Faculty of Science and Institute For Mathematical Research
Universiti Putra Malaysia

INTRODUCTION

- Panel data refers to the pooling of observations on a cross-section of household, countries, firms, etc. over multiple time series.
- Panel data contains observations on multiple phenomena over multiple time periods for the same firms or individual.
- The fixed effect linear panel data model can be formulated as below:

$$y_{it} = \alpha_i + x_{it}^{'}\beta + \epsilon_{it} \qquad i = 1, \ldots, N \qquad t = 1, \ldots, T \qquad (1)$$

where i =1,2,…,N are individual units observed at time series t = 1,2,…, T. Y_{it} is the dependent variable, α_i are the unobservable time-invariant individual effects, β is K x 1 and x_{it} is the –th observation on K explanatory variables. The ε_{it} denote the error term which are assumed to be uncorrelated across time and individual units.

LAYOUT OF PANEL DATA

Individual (i)	Time (t)	y_{it}	x_{it1}	x_{it2}	- - -	x_{itp}
1	1	y_{11}	x_{111}	x_{112}	- - -	x_{11p}
	2	y_{12}	x_{121}	x_{122}	- - -	x_{12p}
	-	-		-		
	-	-		-		
	-	-		-		
	T	y_{1T}	x_{1T1}	x_{1T2}	- - -	x_{1Tp}
2	1	y_{21}	x_{211}	x_{212}	- - -	x_{21p}
	2	y_{22}	x_{221}	x_{222}	- - -	x_{22p}
	-	-		-		
	-	-		-		
	-	-		-		
	T	y_{2T}	x_{2T1}	x_{2T2}	- - -	x_{2Tp}
-	-	-		-		
-	-	-		-		
-	-	-		-		
n	1	y_{n1}	x_{n11}	x_{n12}	- - -	x_{n1p}
	2	y_{n2}	x_{n21}	x_{n22}	- - -	x_{n2p}
	-	-		-		
	-	-		-		
	-	-		-		
	T	y_{nT}	x_{nT1}	x_{nT2}	- - -	x_{nTp}

Classical Fixed Effect Panel Data

- The Classical Within Group estimator is obtained by first centering data within every time series:

$$\tilde{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} \qquad (2)$$

$$\tilde{x}_{it} = x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it} \qquad (3)$$

$$\tilde{y}_{it} = \beta' \tilde{x}_{it} + error_{it} \qquad (4)$$

- Regressing $\tilde{y}_{it}$ on $\tilde{x}_{it}$ by OLS results in the Within Group Estimator (Baltagi, 2001)

INTRODUCTION

- Like other model, the Ordinary Least Squares (OLS) method is often used to estimate the parameters of the fixed effect Panel data model (Classical Within group estimator).
- The outliers especially outliers in the X variables, known as High Leverage Points (HLP) have an adverse effect on the OLS estimates.
- Hampel et al. (1946), Hampel (2012) stated that most data even high quality data have 1-10 % outliers.
- Since the OLS is easily affected by outliers, robust methods are developed to reduce the effect of outliers on parameter estimates.
- Many robust methods are found in the literatures for linear models.
- However, only scarce literatures are available for robust methods for panel data model.

INTRODUCTION

- The damaging effect of outliers can be more crucial for the Within Group estimator.
- The classical method employs data transformation whereby data are centered within each time series by using mean centering (Baltagi,2001).
- The shortcomings of using mean centering is that it will introduce a lot more outliers into the transformed data due to the non-robust property of the mean. Data in the contaminated time series will be affected in which the values will be greatly inflated or deflated.

INTRODUCTION

- Bramati and Croux (2007) applied robust Generalized M estimator based on Minimum Volume Ellipsoid to estimates the parameters of the panel data model (Robust Within Group estimator based on GM6).

- They employed median centering instead of mean centering in order to eliminate the fixed effect.

❖ Bramati & Croux (2007) and Verardi (2010) both applied the median centering to obtain robust Within Group estimate

$$\tilde{y}_{it} = y_{it} - Med_t \quad \{y_{it}\} \tag{5}$$

$$\tilde{x}^j_{it} = x^j_{it} - Med_t \quad \{x^j_{it}\} \tag{6}$$

❖ Median centering is chosen because of it robustness, have high breakdown point, possess min max property. However, centering by median produces non-linearity to the resulting data, affects the equivariance properties of the robust parameters (Bramathi and Croux, 2007), less efficient than mean for uncontaminated data(Maronna et al. 2006)

INTRODUCTION

- Mazlina and Habshah (2015) employed Robust Within Group Estimator based on more efficient MM centering (RWGM).
- The formulation of RWGM is based on MVE to downweight the HLPs. The MVE is suffering from swamping and masking effect. The method is not very successful in identifying genuine HLPs. Hence some of good observations will be downweighted-reduce the efficiency of the estimates.
- Hence Shelan and Habshah (2018) develop a Robust Within Group Estimator based on Fast Improvised Generalized MT (WGM-FIMGT) estimator.
- This method first classify observations into good obs, vertical outliers, good and bad HLPs.
- Subsequently It will downweight only the genuine outliers, bad HLPs-improve efficiency of estimates

OBJECTIVES

- To develop fast robust WGM-FIMGT estimator to remedy problem of HLPs.
- To show that our newly developed WGM-FIMGT is more efficient than the existing estimators and less computational time.

GM estimator to reduce the effect of vertical outliers and Bad HLPs

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:

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

The weakness of this initial weight function:

1. it tends to swamp some low leverage points (Bagheri and Habshah, 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 because they increase the variability of X (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.

Determine initial weight: 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 Fast 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*.

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 $(\tilde{\beta}_R)$, residuals $(\hat{\varepsilon}_{i.R})$, hat values $(w^*_{ii.R})$,standard deviation $(\hat{\sigma}_R)$ and standard deviation with the ith case deleted $(\hat{\sigma}_{R-i})$. The Fast Improvised Generalized Studentized Residuals (FIMGT) is then defined as follows;

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

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.

Following Alguraibawi, Habshah, Imon (2015) they suggested a rule for classifying observations as follows:

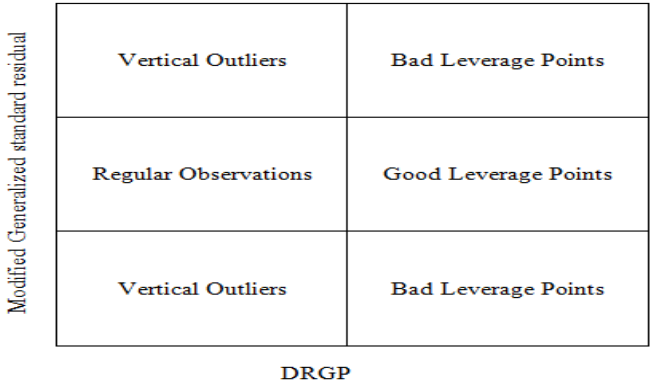
i. Regular Observation (RO): An Observation is declared as a "RO" if
$|FMGTi \leq CP_{FMGT} \text{and} p_{ii}^* \leq \text{Median}\ (p_{ii}^*) + c\ \text{MAD}\ (p_{ii}^*)$

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
$|FMGt_i| \leq CP_{FMGT} \text{and} p_{ii}^* > Median\ (p_{ii}^*) + c\ \text{MAD}\ (p_{ii}^*)$

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

Figure 1: DRGP against Fast Generalized Student zed Residuals



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, Habshah, Imon (2015) have shown that the MGT is very successful in detecting the bad high leverage points and vertical outliers. Therefore, the initial estimate of our propose GM-FIGMT is given by,

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

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

Algorithm of GM-FIGMT Estimator (Shelan and Habshah (2018))

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 $\pi_i = min\ [1, \frac{CP_{FMGT}}{FMGT}]$.

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.

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.

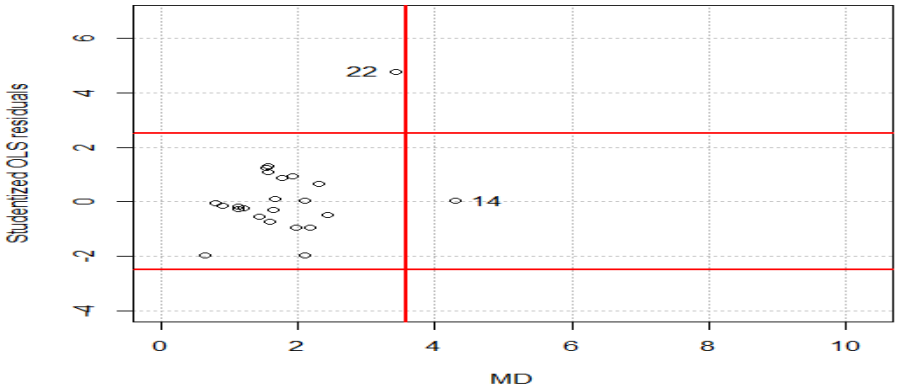


Figure 3: The Studentized OLS res. vs. MD for the Aircraft data

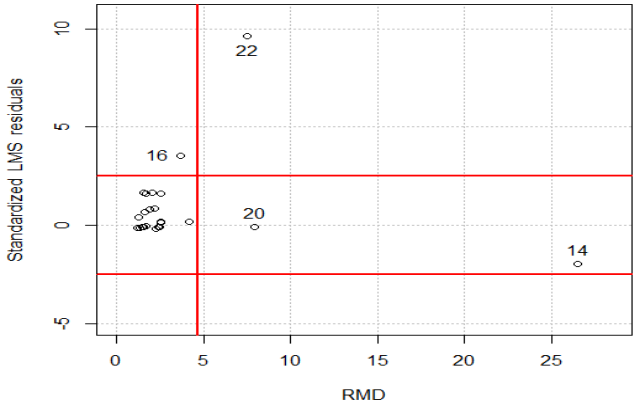


Figure 4: The Standardized LMS res. vs. RMD for the Aircraft data

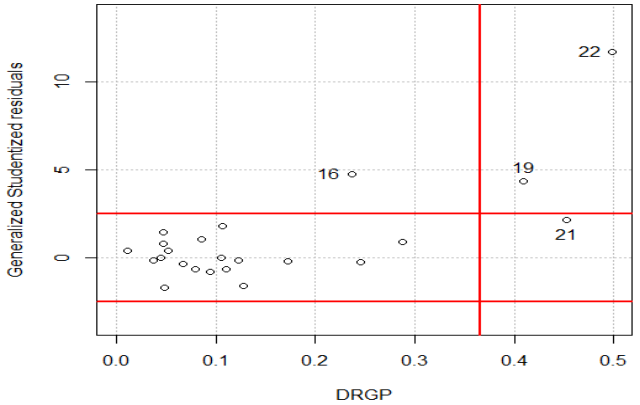


Figure 5: The Mod. Generalized studentized res. vs. DRGP for the Aircraft data

MONTE CARLO SIMULATION STUDY

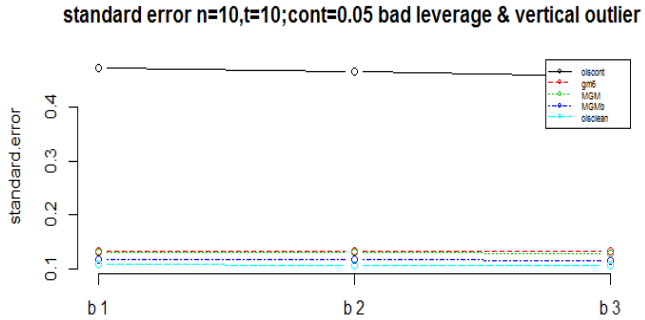
❖In this section, Monte Carlo Simulation is carried out to access the performance of our new proposed WGM-FIMGT by using MM-centering to transform data. We consider the following fixed effect linear panel data model,

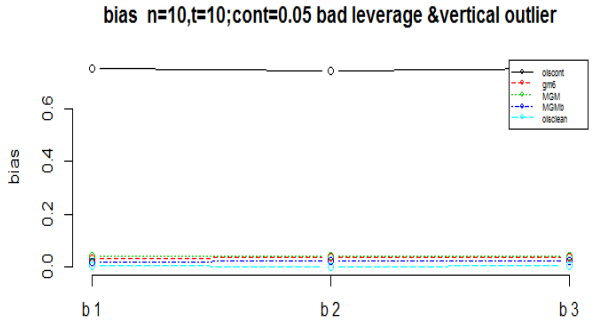
$$y_{it} = \alpha_i + x_{it}^{'}\beta + \epsilon_{it} \qquad i = 1, \ldots, N \qquad t = 1, \ldots, T \qquad (1)$$

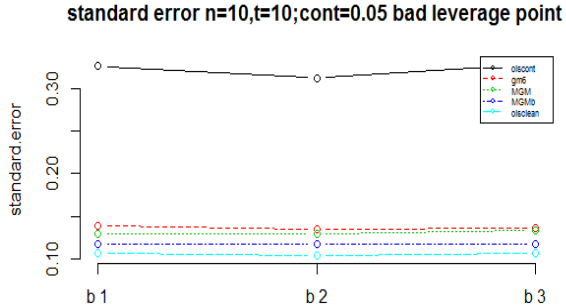
where the error distributed as N(0,1), α~U(0,1). The explanatory variables are generated from a multivariate standard normal distribution. In the simulation study, we consider panel datasets of (t=5,10,20); represent small medium large samples each with (n= 10,30), and we consider p=3

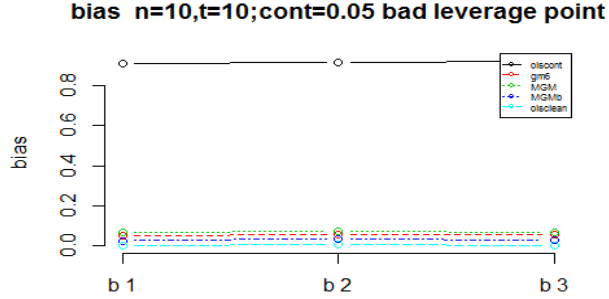
Estimated Efficiency of WOLS, WGM6,WMM,WGM-FIGMT

		Level of contamination=0.05								
		VO			BLP			VO&BLP		
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
n=10;t=5	WOLS	24.325	23.934	27.324	17.581	16.805	16.974	17.687	17.544	17.826
	WMM	63.562	61.625	66.496	56.457	56.762	57.353	59.134	61.019	60.058
	WGM6	70.283	68.344	73.693	55.702	54.668	55.573	60.384	60.913	62.542
	WGM-FIMGT	76.681	74.245	79.468	72.154	71.231	71.817	73.069	73.105	75.602
n=10;t=20	WOLS	21.903	21.434	22.679	11.084	10.782	10.858	12.256	12.313	12.294
	WMM	79.644	77.619	80.643	72.195	70.928	72.246	78.634	77.527	77.993
	WGM6	81.985	80.147	81.504	73.256	70.639	70.593	79.497	78.334	80.115
	MGM-FIMGT	91.288	89.643	91.251	89.043	85.151	88.051	89.966	89.374	91.308
n=10;t=20	WOLS	21.409	20.602	22.024	7.489	7.829	8.0486	8.099	8.344	8.417
	WMM	81.514	80.055	79.883	75.336	77.633	81.116	75.722	79.264	81.669
	WGM6	80.512	80.135	79.188	77.876	78.607	81.668	77.906	80.474	80.796
	WGM-FIMGT	92.198	92.347	91.845	91.533	93.282	93.932	90.260	91.453	94.374


standard error n=10,t=10;cont=0.05 bad leverage & vertical outlier
standard.error
0.1
0.2
0.3
0.4
b 1
b 2
b 3
olscont
gm6
MGM
MGMb
olsclean


bias n=10,t=10;cont=0.05 bad leverage &vertical outlier
bias
0.0
0.2
0.4
0.6
b 1
b 2
b 3
olscont
gm6
MGM
MGMb
olsclean


standard error n=10,t=10;cont=0.05 bad leverage point
standard.error
0.10
0.20
0.30
b 1
b 2
b 3
olscont
gm6
MGM
MGMb
olsclean


bias n=10,t=10;cont=0.05 bad leverage point
bias
0.0
0.2
0.4
0.6
0.8
b 1
b 2
b 3
olscont
gm6
MGM
MGMb
olsclean

NUMERICAL EXAMPLE

❖ In this section, we applied our method (MGM-FIMGT) to real panel data to evaluate the newly developed method.The mm-centering used to transform the data.

1-Airline data set

❑ Our first example is six airline firms, which is taken from Greene (2007) to study the efficiency in production of airline services..The data set consist three predictor variables (output, fuel price, and load factor) and response variable, production cost, over 15 yearly observations(1970-1984).

❑ Table (2) illustrate the parameter estimates for proposed method in clean data and when data are contaminated at α=0.1%in 2 scenarios (vertical outlier and bad leverage point) . Here, we contaminated 9 time series randomly, 4 bad leverage point and 5 vertical outlier shown in figure (2) . The standard errors of the estimates are obtained by bootstrapping method.

Table (2) : parameter estimates for the original and Modified Airline Data

		Mean centering	mm-centering		
		WOLS	WGM6	WMM	WGM-FIMGT
Original Data	$\hat{\beta}_0$	13.36561 (0.09842)	-0.001799 (0.00568)	-0.00511 (0.00676	-0.0015766 (0.00599)
	$\hat{\beta}_1$	0.919285 (0.0229)	0.9186127 (0.0313)	0.916500 (0.0277)	0.921756869 (0.0293735)
	$\hat{\beta}_2$	0.4175 (0.0061)	0.407768626 (0.014078)	0.410848 (0.014532)	0.41757208 (0.01224403)
	$\hat{\beta}_3$	-1.0700 (0.1914)	-1.02311091 (0.185789)	-1.076628 (0.1839966)	-0.9520885 (0.20724239)
Modified Data	$\hat{\beta}_0$	13.36561 0.11930	-0.01838107 (0.00979)	-0.01319496 (0.0083)	-0.01488057 (0.0010)
	$\hat{\beta}_1$	-0.03479 (.08131)	0.9224339 (0.052789)	0.87744 (0.02996)	0.907779152 (0.0157360)
	$\hat{\beta}_2$	-0.30806 (1.22679)	0.410947718 (0.0234966)	0.4238787 (0.03372)	0.42048643 (0.0149187)
	$\hat{\beta}_3$	8.73823 (22.04315)	-0.9855288 (0.32329049)	-0.8609513 (0.38262)	-0.9880553 (0.257718)

CONCLUSION

- The main aim of this paper is to develop a new Robust Within Group Estimator based on FIMGT and MM centering (WGM-FIMGT)
- The commonly used WOLS Estimator based on OLS and mean centering is very poor in the presence of outliers and produce inefficient estimates.
- The WGM-FIMGT is more efficient than other estimators.
- Recommend using our proposed WGM-FIMGT estimator as it is the most efficient estimator compared to some existing methods.

THANK YOU