Advances in Social Science, Education and Humanities Research, volume 292

1st Aceh Global Conference (AGC 2018)

The Applications of Robust Estimation in Fixed Effect Panel Data Model

[1,2,*]Nor Mazlina Abu Bakar, Habshah Midi[2]
[1] Centre of Management Sciences, Faculty of Economics and Management Sciences,
Universiti Sultan Zainal Abidin, Terengganu, Malaysia
[2] Institute of Mathematical Research, Universiti Putra Malaysia, Serdang, Malaysia
[*]Corresponding author: normazlina@gmail.com

***Abstract*-High leverage points (HLPs) are known to have significant effects on parameter estimation of linear fixed effect regression. Their presence causes panel data to become heavily contaminated which in turn leads to biasness and wrong analysis. Thus, robust regression estimators are introduced to provide resistant estimates towards HLPs. In this study, two Robust Within Group (RW) estimators are applied to a few economics and finance real world data. The study is aimed to estimate the usefulness and efficiency of robust methods in contaminated panel data. Results show the advantage of using robust estimation to reduce the influence of HLPs on panel data over the Ordinary Least Square (OLS).**

Keywords--panel data; fixed effect; regression; GM-estimator; MM-estimator, robust.

I. INTRODUCTION

Panel data analysis plays an important role in modern econometrics because its grouping structure can provide important information rather than simpler forms of data. In particular, the structure can be used to estimate models with complicated forms of heterogeneity across units or entities. For the past decade, there has been an increasing trend on the use of this type of data in the research of economics and finance. The uproar of Industry Revolution 4.0 makes it more essential to evaluate large panel data in order to gain critical and upfront information. Fixed effect linear regression is one of the popular methods in the econometrics. Commonly, OLS is very popular due to its universal acceptance, computational simplicity with the best linear unbiased statistical properties. Unknown parameters for the fixed effect linear regression are estimated using ordinary least square (OLS) by minimizing the sum of the squares of the differences between the observed values in panel data and the predicted linear values given by the explanatory variables (Greene, 2017). However, the method depends on a number of undeniably restrictive and unrealistic assumptions. Among the assumptions are the normality of error distribution, independency of the explanatory variables and error terms with constant variance for all observations or homoscedasticity (Kutner et al., 2004; Baltagi, 2013; Greene, 2017). In the existence of the outliers, the assumption of independent and identically distributed (i.i.d) errors for linear regression is completely violated set, resulting in bias and unreliable estimates of the model parameters (see Montgomery et al., 2001 and Chatterjee and Hadi, 2006).

In addressing the problem, modern robust methods are researched to find highly efficient estimators which mimic least square estimates in the absence of outliers; highly advanced robust methods have been developed for linear regressions (Hampel et al., 2001; Chatterjee and Hadi, 2006). Typically, intensive computer simulations are required in this type of research which are now widely accessible; for example R computing by R Core Team (2014). Somehow, only limited investigations are done for regression of panel data (Croux et al., 2003; Verardi, 2010; Aquaro and Cizek, 2013). The effects of outliers can be crucial for fixed effect model especially when multiple x-outliers or high leverage points (HLPs) occur concentrated in the time series (Bramati and Croux, 2007). Any atypical observation can cause panel data to become highly contaminated due to data transformation by non-robust centering procedure. Thus, median centering and MM-Centering are introduced to provide robust alternatives for the fixed effect panel data (Bramati and Croux, 2007; Verardi and Wagner, 2010; Abu Bakar and Habshah, 2015). Nonetheless, robust estimator(s) need to be employed to replace OLS and provide unbiased parameter estimates for contaminated panel data. Bramati and Croux (2007) have proposed robust Generalized M-estimator (RWGM) under median centering and achieved moderate breakdown points. However, the existing GM-estimators; such as the RWGM, heavily relies on the efficiency of robust Mahalanobis distance (RMD) as part of its

weighting scheme. RMD is known to swamp tremendously where many inliers are detected as outliers (Habshah and Abu Bakar, 2015). The efficiency of the RWGM-estimator can be improved vastly by introducing more robust weights; determined by robust outlier detection techniques (Hashah et al., 2009). Abu Bakar and Habshah (2015) have proposed the Robust Within Group MM-estimator under the MM-centering and the estimator is found to provide more efficiency than RWGM.

This study is therefore aimed to achieve two main objectives. The first objective is to apply RWGM and RWMM on numerical examples and investigate the effects of HLPs on the classical and robust fixed effect parameter estimations. The second objective is to compare the performances of the robust methods with the OLS in contaminated data using illustrations and graphs. The paper is organized as follows. Section 2 discussed the existing robust estimators for fixed effect panel data model and their algorithms. The performances of the robust estimators evaluated in Section 3 using numerical examples from Greene (2017) and be compared with conventional non-robust method. The conclusion of the study is given in Section 4.

II. THE EXISTING ROBUST WITHIN GROUP ESTIMATORS

A. Robust Within Group GM-Estimator

For panel data, robust Within Group GM-estimators (RWGM) are solutions to the normal equation:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\pi_{it}\psi\left(\frac{y_{it}-x_{it}^{'}\beta}{s\pi_{it}}\right)$$

for $i=1,2,\ldots N$ unit and $t=1,2,\ldots T$ time series. Robust Mahalanobis Distance (RMD) is incorporated in GM-estimator to down weigh the effects of HLPs. The algorithm is as follows:

Step 1: Data are robustly transformed around the MM-estimate of location to obtain the fixed data, $\tilde{x}_{it}$ and $\tilde{y}_{it}$

Step 2: Least Trimmed Square (LTS) estimator is performed to provide initial estimate and compute standardized residuals, $r_{it}=\tilde{y}_{it}-\tilde{x}_{it}^{'}\hat{\beta}_{LTS}$ and scale estimate, $\hat{\sigma}_{LTS}^{2}$.

Step 3: Tukey's Biweight function is selected as the first weighting scheme where any observation with large residual is down weighted. The diagonal elements, W_r can be simplified as:

$$W_r=\begin{cases}0 & \text{if} \quad \left|\dfrac{r_{it}}{\hat{\sigma}_{LTS}}\right|\le \text{c}\\ \left(1-\left(\dfrac{r_{it}}{c\hat{\sigma}_{LTS}}\right)^2\right)^2 & \text{if} \quad \left|\dfrac{r_{it}}{\hat{\sigma}_{LTS}}\right|> \text{c}\end{cases}$$

c is chosen to be 4.685 to provide a balance between efficiency and robustness (Wagenvoort and Waldmann, 2002).

Step 4: Robust Mahalanobis Distance is calculated and the diagonal elements of the second weighting matrix W_x is rewritten as, $W_x=\min\left(1,\sqrt{\chi^2_{K,0.975}/\text{RMD}_{it}}\right)$

Step 5: The final weight for each observation is determined by utilizing the Tukey's biweight function.

Step 6: Compute a one-step reweighted least square method as convergence approach. Under the two weighting schemes, the weighted least square estimation is applied to obtain an estimate of high breakdown RWGM-estimate, denoted by $\hat{\beta}_{RWGM}$ and simplified as:

$$\hat{\beta}_{RWGM}=\left(\tilde{X}^{'}W_xW_r\tilde{X}\right)\left(\tilde{X}^{'}W_xW_r\tilde{Y}\right)$$

B. Robust Within Group MM-Estimator

Step 1: Panel data is transformed by the MM-centering procedure to obtain fixed $\tilde{x}_{it}$ and $\tilde{y}_{it}$.

Step 2: The initial high breakdown coefficient, $\hat{\beta}_0$ and scale estimates, $\hat{\sigma}_s$ are computed by using S-estimates. The estimates can be obtained using the fast algorithm by Salibian-Barrera and Yohai (2006) to provide increased computational speed.

Step 3: The second step involved the computation of the residuals, r_{it} based on the initial estimates where $r_{it} = \tilde{y}_{it} - \tilde{x}_{it}'\beta_0$. The M-estimate of scale, $\hat{\sigma}$ with high break down point can now be obtained as a solution of:

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_0\left(\frac{r_{it}}{\hat{\sigma}}\right) = b$$

where $\frac{b}{a} = 0.5$ and $a = \max \rho_0$

Step 4: In the last stage, the M-estimate is computed such that $\hat{\beta}_{RWMM}$ is the solution of:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\psi_1\frac{r_{it}}{\hat{\sigma}}x_{it} = 0$$

where $\psi_1 = \rho_1'$ to achieve high efficiency. A Huber or bisquare function is employed to assign weights for the residuals. The MM-estimator is very well developed and its regression function is also readily available in S-Plus as lmRobMM and in R as lmrob (R Core Team, 2014). By default, the function gives a highly robust and highly efficient estimate to the fixed effect panel data.

III. NUMERICAL EXAMPLES

In this section, two numerical examples are taken from Green (2017) to evaluate the performance of the existing RWGM and RWMM under the influence of high leverage values. Both original and contaminated data are studied to compare their performances. Contamination is presented in each data by introducing leverage values using simple random selection. Data must first be transformed under the classical mean centering or the robust MM-centering for fixed effects. Dare are then regressed either by the classical OLS, RWGM or RWMM. Results are recorded for the beta coefficients and their standard errors.

A. Investment Data

The first example is an artificial investment data with profit values as the independent variable. The panel data are taken from 3 firms over the period of 10 years. Observations number 2 and 28 are randomly selected to be contaminated and become high leverage points (HLPs). Once contaminated, data need to be transformed and then regressed. Results are reported in Table I with standard errors of the beta coefficients written in the parentheses. The results indicate that the robust estimators are able to provide resistant and efficient results even though HLPs are introduced into the data set. The robust estimators are also able to provide similar results to the estimation by OLS in the original, uncontaminated data. On the other hand, OLS only provides best estimates for the original data but bias and wrong results are given for the contaminated data. OLS is highly influenced by the HLPs and the influence became more intense in the presence of HLPs.

TABLE I. BETA ESTIMATES OF THE ORIGINAL AND MODIFIED INVESTMENT DATA WITH STANDARD ERRORS

Estimate	***OLS***	***RWGM***	***RWMM***
	Mean Centering	***MM-Centering***	
Original data, $\hat{\beta}_1$	1.1022(0.0489)	1.1914(0.0436)	1.1128(0.0754)
Modified data, $\hat{\beta}_1$	3.1426(0.1917)	1.2164(0.0437)	1.1394(0.0955)

Figure I(A) and I(B) are the plots of mean-centred data and MM-centred data; respectively. The leverage values are excluded from the plot for simplicity. The plots indicate that the non-robust, mean-centred data caused the resulting data to become less linear but data linearity is maintained under MM-centering. Figure II shows the scatter plot of the investment data under MM-centering. The solid line represents the estimated OLS regression line which is similar to the estimated RWMM. On the other hand, the dotted

line is the OLS regression line for the contaminated data which is pulled towards the HLPs on the far right hand side. This shows the bad influence of HLPs on the regression by OLS; translated by the values of the beta coefficient in Table I for the modified data.

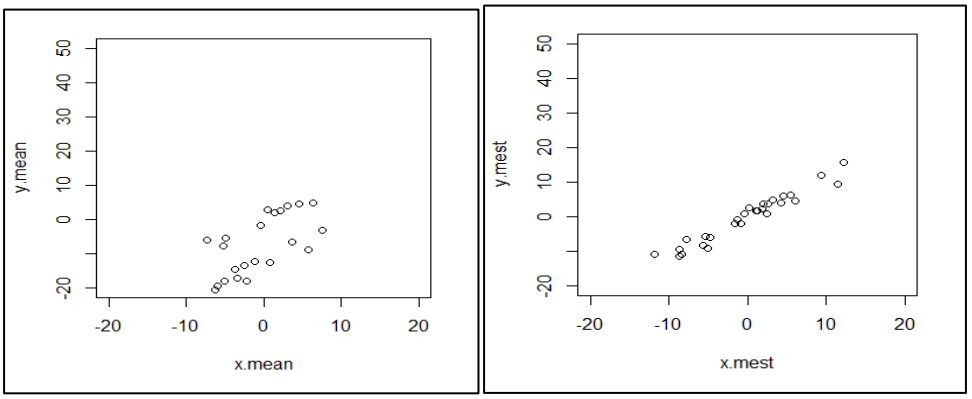

FIGURE I(A) FIGURE I(B)

FIGURE I. SCATTER PLOT OF CONTAMINATED INVESTMENT DATA (A) UNDER MEAN CENTERING (B) MM-CENTERING

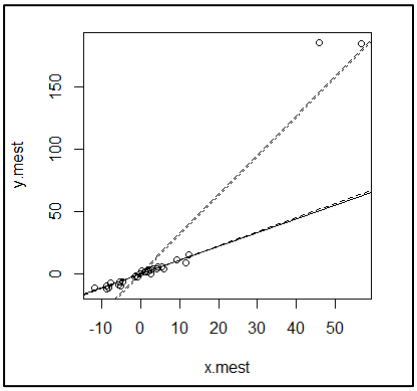

FIGURE II. SCATTER PLOT OF INVESTMENT DATA UNDER MM-CENTERING FOR CONTAMINATED DATA

B. Airline Data

The second example is a panel data consisting of 90 yearly observations taken from six airline companies. From Table II, beta coefficients values in the original data, derived by the classical OLS estimates are considered as the best linear unbiased estimates. Beta coefficients provided by the robust estimators are found to be very close to the OLS estimates under the central model. On the other hand, beta coefficients derived by OLS for the contaminated cases are badly affected by the presence of the HLPs. Both β_1 and β_3 have negative signs, indicating a negative influence of the independent variables when originally they held positive relationships. Further destructions on the parameter estimates are observed when extra HLP is introduced into the dataset. On the other hand, both robust estimators are found to provide robust beta estimates with low standard errors for the modified data. This indicates highly resilient estimates provided by RWGM and RWMM in the presence of HLPs.

Figures II(A) and II(B) show the scatter plots of Mahalanobis Distance (MD) and its robust counterpart, Robust Mahalanobis Distance (RMD); respectively. Both measures are used to evaluate the distance of each data point from the mass of central data. Ideally, inlying values will lie close to the mass of central data and have low MD values whereas outliers will have large MD values. Figure II(A) shows that more outliers are found in a mean-centred data which is due to the non-robust data transformation by the highly infected arithmetic mean in the first step of OLS. Furthermore, MD uses non-robust arithmetic mean and covariance in its formulation which eventually leads to the event of masking and/or swamping. Both events leads to false identification of the outliers. On the other hand, RMD provides resilient distance which is unaffected by the outlying values. Figure II(B) clearly shows the distinct

distance of the HLPs from the rest of the data set. From the RMD plot, we can easily identify two HLPs which were introduced in the panel data set.

TABLE II. PARAMETER ESTIMATES OF THE COST EQUATION WITH FIXED EFFECTS FOR AIRLINE

Estimate	*OLS*	*RWGM*	*RWMM*
	Mean Centering	*MM-Centering*	
Original Data			
$\hat{\beta}_1$	0.9193(0.0229)	0.9183(0.0267)	0.9172(0.0171)
$\hat{\beta}_2$	0.4175(0.0061)	0.4042(0.0127)	0.4104(0.0145)
$\hat{\beta}_3$	-1.0700(0.1914)	-0.8922(0.1714)	-1.0613(0.2248)
Modified Data			
$\hat{\beta}_1$	3.3716(0.3641)	0.9098(0.0282)	0.9119(0.0187)
$\hat{\beta}_2$	3.6680(0.1817)	0.4126(0.0136)	0.4150(0.0153)
$\hat{\beta}_3$	4.8126(0.2406)	-0.9274(0.1782)	-1.0553(0.2199)

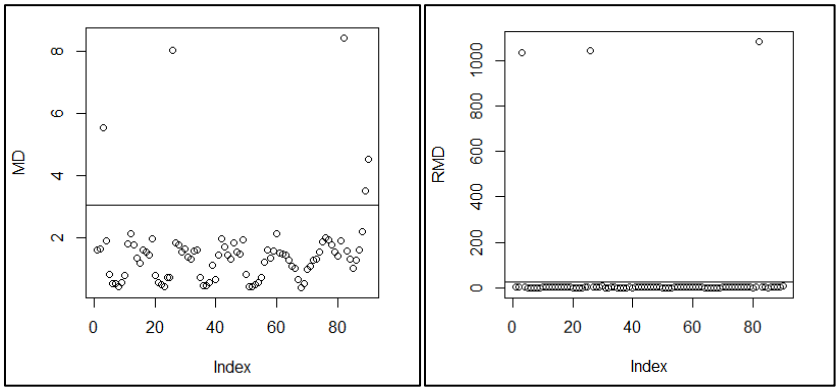

FIGURE III(A) FIGURE III(B)

FIGURE III. SCATTER PLOT OF CONTAMINATED AIRPLANE DATA (A) MAHALANOBIS DISTANCE (B) ROBUST MAHALANOBIS DISTANCE

IV. CONCLUSION

The classical OLS parameter estimation is known to be highly influenced by the presence of HLPs especially when they are of leverage type. More outliers may be introduced into the dataset when non-robust data transformation is applied for the fixed effect by the arithmetic mean. As a consequence, the performance of OLS is vastly reduced in the contaminated cases. In this study, we applied two robust within group estimators called RWGM and RWMM to two panel data from Green (2017) under robust MM-centering. Results indicate that the robust estimators are able to provide resistant estimates despite the presence of HLPs. In RWGM and RWMM, any unusually large residual will be assigned the weight of "0" to completely eliminate their effects. As a result, the efficiency of MM-estimator almost coincides with OLS in a clean dataset (Maronna et al., 2006).

REFERENCES

Abu Bakar, N.M. and Habshah, M. (2015). Robust Centering of the Fixed Effect Panel Data Model. *Pakistan Statistics Journal*, 31(1).

Aquaro, M. and Cızek, P. (2013). One-Step Robust Estimation of Fixed-Effects Panel Data Models. *Computational Statistics & Data Analysis,* 57(1), 536–548.

Bagheri, A., Habshah, M. and Imon, A.H.M.R. (2009). Robust Estimations as a Remedy for Multicollinearity Caused by Multiple High Leverage Points, *Journal of Mathematics and Statistics*, 5(4), 311-321.

Baltagi, B.H. (2013) *The Econometrics of Panel Data*. John Wiley and Sons, New York. Econometric Analysis of Panel Data, 5th Edition. ISBN: 978-1-118-67232-7.

Bramati, M.C and Croux, C. (2007). Robust Estimators for the Fixed Effects Panel Data Model. *Econometrics Journal*, 10(3), 521–540.

Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Example*. 4th edition. New York: Wiley.

Croux, C., Dhaene, G. and Hoorelbeke, D. (2003). *Robust Standard Errors for Robust Estimators*. Research Report, Dept. of Applied Economics, K.U. Leuven.

Djauhari M. (2010). A Multivariate Process Variability Monitoring Based on Individual Observations. *Modern Applied Science*, 4(10).

Greene, W.H. (2017). *Econometric Analysis*. Prentice Hall, New York.

Habshah. M. and Abu Bakar, N.M. (2015). The Performance of Robust-Diagnostic F in the Identification of Multiple High Leverage Points. *Pakistan Journal of Statistics*, 31(5)

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, 507-520.

Hampel, F.R. (2001). Robust Statistics: A Brief Introduction and Overview. *Seminar for Statistics.*

Imon, A.H.M.R. (2002). Identifying Multiple High Leverage Points in Linear Regression. *Journal of Statistical Studies 3*, 207–218.

Kutner, M.H., Nachtsheim, C.J. and Neter, J. (2004). *Applied Linear Regression Models*. 4th Edition, McGraw Hill, New York, ISBN: 978-0256086010.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. John Wiley, New York.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rousseeuw, P. and Leroy, A.M. (2003). *Robust Regression and Outlier Detection*. John Wiley, New York.

Salibian-Barrera, M. and Yohai, V.J. (2006). A Fast Algorithm for S-regression Estimates. *Journal of Computational and Graphical Statistics*, 15(2), 414-427.

Simpson, J. R. and Montgomery, D. C. (1998). The Development and Evaluation of Alternative Generalized M-Estimation Techniques. *Communications in Statistics - Simulation and Computation*, 27, 999–1018.

Verardi, V. and Wagner, J. (2010). Robust Estimation of Linear Fixed Effects Panel Data Models with an Application to the Exporter Productivity Premium. *SSRN eLibrary*.

Wagenvoort, R. and Waldmann, R. (2002). On B-robust Instrumental Variable Estimation of the Linear Model with Panel Data. *Journal of Econometrics,* 106, 297-324.