The Study of Robust Estimators on Panel Data Regression Model for Data Contaminated with Outliers

Mia Amelia[1], Kusman Sadik[2], Bagus Sartono[3]
{mia.amelia.0515@gmail.com[1], kusmansadik@gmail.com[2], bagusco@gmail.com}

Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia[1],
Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia[2],
Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia[3]

Abstract. Outliers can cause biased parameter estimators and deviate from the actual values. This research studies robust estimators on panel data regression model. The robust estimators used are least trimmed squares (LTS) and within-group generalized M (WGM). This research aims to study robust estimator method in estimating panel data regression parameter on simulation data with various kinds of outliers and outlier proportions. This research utilizes primary data taken from the results of simulation data designed based on fixed effects of the panel data regression. The variety of overall simulation data in this study contains 16 types of contamination. The result shows that the within estimation method is not robust against outliers. Based on the absolute relative bias and RMSE, the WGM method produces a small variety of estimators and high accuracy of estimators for various types of outliers and levels of outlier contaminations.

Keywords: outliers, panel data, regression, robust estimators

1 Introduction

Panel data are combined data between cross section and time series data. The same cross-section in panel data are observed repeatedly over a period of time. For two decades, analysis using panel data has developed well and is widely used in various fields. The panel data regression model is often applied to micro-level data, which often contains data that are contaminated with outliers. In general, the regression estimator used in the panel data model is the least squares (LS) method. This method is sensitive to data contaminations and outliers [1], [2].

Outliers indicate an observation that deviates considerably far from other observations [3]. The existence of outliers in the regression model can cause biased parameter estimations and deviate from the actual values leading to invalid conclusions. Outliers happen due to various reasons. Outliers in panel data can be found in vertical or leverage [4], [5]. Outliers in panel data are often found concentrated in several block-concentrated outliers [6].

One of the alternatives that can be used to deal with outliers is the robust estimator method [7]. Robust estimators involve systematic procedures to investigate model deviations caused by outliers. Apart from the outlier effects on the classical estimator method, there are only a few robust estimator methods that can be used in the fixed effects panel data regression model.

The research on robust estimators in panel data regression model have been carried out by Aquaro and Čížek[1],Bramati and Croux[6], Bakar and Midi[7], Čížek[8], Verardi and

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290517

Wagner [9]. The least trimmed squares (LTS) estimator method is one of the robust parameter estimation methods of outliers that has a high breakdown point (BDP) compared to other parameter estimation methods [1], [9]. In panel data regression with a fixed effects model, Bramati and Croux[6] also Aquaro and Čížek[1] used within-group generalized M (WGM) estimators as an alternative to classical estimators in the within-groupmethod. When there is no outlier in the panel data, the results of the two robust estimators are very close to the within estimator on the fixed effects model.

The research on robust estimators for panel data regression model have not been done that many in Indonesia. Therefore, this research contains robust estimator studies on panel data regression model. Robust estimators that will be used are LTS and WGM. Based on the background described, this research aims to examine the robust estimator method in estimating panel data regression parameters on simulation data with various types of outliers and levels of outlier contaminations.

2 Materials

This research utilizes type of primary data. The primary data in this study came from the results of simulation data designed based on the fixed effects of panel data regression, namely:

$$\mathbf{y}_{it} = \alpha_i + \mathbf{X}_{it}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{it}, i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, t$$

with y_{it} as the variable response for the individual-i and the time-t, α_i as the intercept coefficient which is scalar, β is the coefficient vector of the explanatory variable with the size kx1, and x_{it} is the observation-i vector and time-t in the explanatory variable k.

Simulation data were given contamination in two different dimensions, namely y-axis (vertical outliers) and x-axis (leverage). The variety of overall simulation data in this study contained 4 types of contamination, namely vertical outliers, leverage points, concentrated vertical outliers, and concentrated leverage points, with outright proportions of 0% (without outliers), 5%, 10%, and 20%.

3 Methods

The simulation was conducted to examine the effect of combined outlier types and levels of outlier contaminations on parameter estimators in the fixed effects of panel data regression model. The steps taken in this simulation are as follows:

1. Set the number of cross-individuals (N) as 3 and the number of times (t) as 5.
2. Generate data with the following steps:
   a. Determine the parameters for the population in the panel data regression model $(\beta_1 \text{ and } \beta_2)$ which are determined as 10 and 5 respectively.
   b. Generate the response variable (y_{it}) in the fixed effects panel data model by generating $\varepsilon_{it} \sim$ Normal (0,1)*0.1, $\alpha_i \sim$ Uniform (0,20).
   c. Generate explanatory variable $(X_{1,it} \text{ and } X_{2,it})$ that follows the Normal distribution (0,1) .

d. Generate values for the response variable (y_{it}) which is randomly selected from several time periods (t) to produce vertical outliers through the Normal distribution (30,1).
e. Generate leverage points obtained by replacing the values in the explanatory variable $(X_{1,it}$ and $X_{2,it})$ derived from the Normal distribution (10,1).
f. Determine the proportion of outliers in the data, namely 0%, 5%, 10%, and 20%.
g. Repeat stage a to stage f as many as 10000 times.

3. Estimate the β parameter of each replication with the following steps:
   a. Least Trimmed Squares(LTS) method
      1) Calculate the median per time for $med_t(y_{it})$ and $med_t(x_{k,it})$ with k as the number of explanatory variables. In this simulation k = 1, 2.
      2) Calculate $\tilde{y}_{it} = y_{it} - med_t(y_{it})$ and $\tilde{x}_{k,it} = x_{k,it} - med_t(x_{k,it})$.
      3) Calculate $\hat{\beta}_{LTS}$ for each replication by minimizing $\sum_{k=1}^{h}\left[(\tilde{y}_k - \tilde{X}_k\beta)^2\right]_{k:NT'}$, where $h = 0.75 \times NT$.
      4) Calculate the absolute relative bias and root mean square error (RMSE) from $\hat{\beta}_{LTS}$

$$brm_{LTS} = \frac{1}{M}\sum_{j=1}^{M}\left|\frac{\hat{\beta}_{LTS} - \beta}{\beta}\right| x100\%$$

$$RMSE_{LTS} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left|\left|\hat{\beta}_{LTS} - \beta\right|\right|^2} \times 100\%$$

   b. Within-Group GeneralizedM(WGM) method
      1) Calculate $\hat{\beta}_{LTS}$
      2) Calculate $r_{LTS} = \tilde{y}_{it} - \hat{\beta}'_{LTS}\tilde{x}_{it}$.
      3) Calculate $\hat{\sigma}^2_{LTS} = c_{LTS}\frac{1}{h}\sum_{k=1}^{h}(\tilde{y}_k - \tilde{x}'_k\hat{\beta}_{LTS})^2_{k:NT'}$
      4) Calculate W matrix with the size NT x NT with the main diagonal $\rho'\left(\frac{r_{it}}{\hat{\sigma}_{LTS}}\right) / \left(\frac{r_{it}}{\hat{\sigma}_{LTS}}\right)$ where ρ is the function of Tukey's biweight:

$$\rho = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} \text{ jika } |x| \le c \\ \frac{c^2}{6} \text{ jika } |x| > c \end{cases}$$

with this option, so the main diagonal of W_r became

$$(W_r)_{it} = \begin{cases} 0 & \text{jika } \left|\frac{r_{it}}{\hat{\sigma}_{LTS}}\right| \ge c \\ \left(1 - \left(\frac{r_{it}}{c\hat{\sigma}_{LTS}}\right)^2\right)^2 & \text{jikajika } \left|\frac{r_{it}}{\hat{\sigma}_{LTS}}\right| < c \end{cases}$$

where c = 4,685.

      5) Calculate robust distance (RD_{it}) for each $\tilde{x}_{it}$: $RD_{it} = \sqrt{(\tilde{x}_{it} - \hat{\mu})\hat{V}^{-1}(\tilde{x}_{it} - \hat{\mu})'}$
      6) Calculate matrix of W_x with the size NT x NT

$$(W_x)_{it} = \min\left(1, \frac{\sqrt{\chi^2_{K,0,975}}}{RD_{it}}\right)$$

with degree of freedom = K, K is the number of explanatory variables

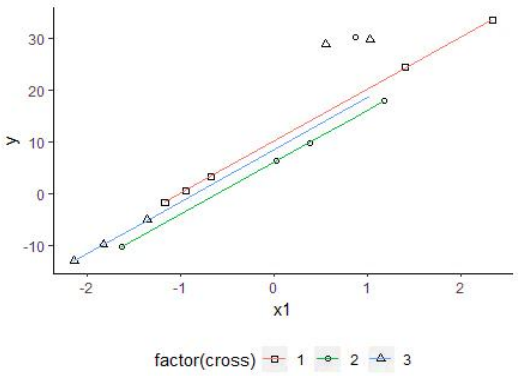
7) Calculate $\hat{\beta}_{WGM}$ for each replication with $\hat{\beta}_{WGM} = \left(\tilde{X}' W_x W_r \tilde{X}\right)^{-1} \tilde{X}' W_x W_r \tilde{y}$
8) Calculate the absolute relative bias and root mean square error (RMSE) from $\hat{\beta}_{WGM}$

$$brm_{WGM} = \frac{1}{M}\sum_{j=1}^{M}\left|\frac{\hat{\beta}_{WGM} - \beta}{\beta}\right| x100\%$$

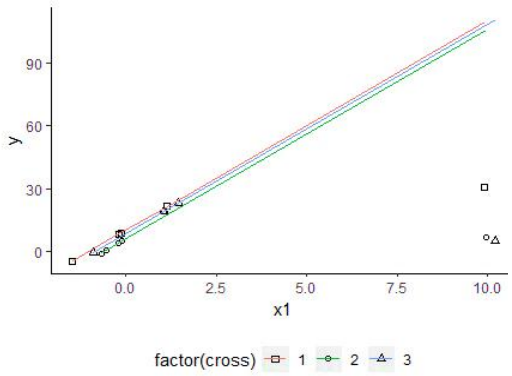
$$RMSE_{WGM} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left|\left|\hat{\beta}_{WGM} - \beta\right|\right|^2} \times 100\%$$

4Results and Discussion

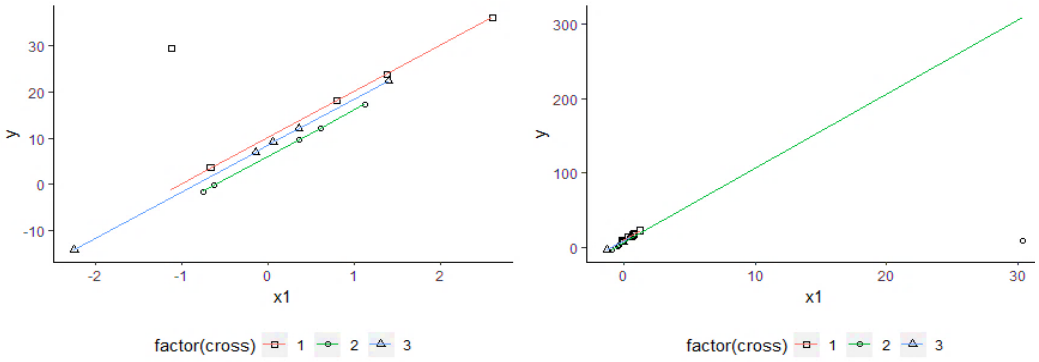
The scattered plot between the explanatory variable and the response variable (x_{it}, y_{it}) from one of the simulation data sets is displayed with a regression line $y = \alpha_i + \beta x$ for each individual cross i = 1, 2, 3(**Figure 1**).



a) vertical outliers



b) leverage points



c) concentrated vertical outliers d) concentrated leverage points

Fig. 1. Simulation data distribution plot y_{it} and x_{it} for various types of outliers with 20%outliers proportion

Figure 1 shows a linear line pattern for vertical outliers, leverage points, concentrated vertical outliers, and concentrated leverage points. In vertical outliers, the remote observation is on the y-axis, but not isolated on the x-axis. Conversely, the type of outliers that are remote observations on the x-axis are known as leverage points. Meanwhile, concentrated outliers (both vertical outliers and leverage points) are the most remote observations that tend to be contaminated over a period of time.

The existence of outliers is detected by observation points outside the main line pattern so that the linear regression line model is suitable to be used in this simulation. The scatter plot (x_{it}, y_{it}) from the simulation data for 20% of outliers proportion in various types of outliers is shown in Figure 1.As an illustration, this study used 3 individual-crosses and 5 time-crosses so that the overall observation is 15. If the proportion of outliers used in this study was 20%, then the remote observation in the x-axis or y-axis is 3. Determining the position of observations in this study was defined randomly

Furthermore, the number of remote observations concentrated in several time periods were obtained from the multiplication of outlier proportions with the number of time periods in an individual cross. If an individual cross was observed for 5 time periods, the proportion of 20% outliers implies that the observations on individual cross were remote at 1 time.

4.1 Parameter Estimation β_1

The parameter estimation was done by using 2 robust estimator methods, those were the LTS and WGM methods. Determining the best robust estimator method was evaluated using absolute relative bias and RMSE. The absolute relative bias values of parameter estimation ($\beta_1 = 10$) with the *within* estimation method and robust estimation method are shown in **Table 1**.

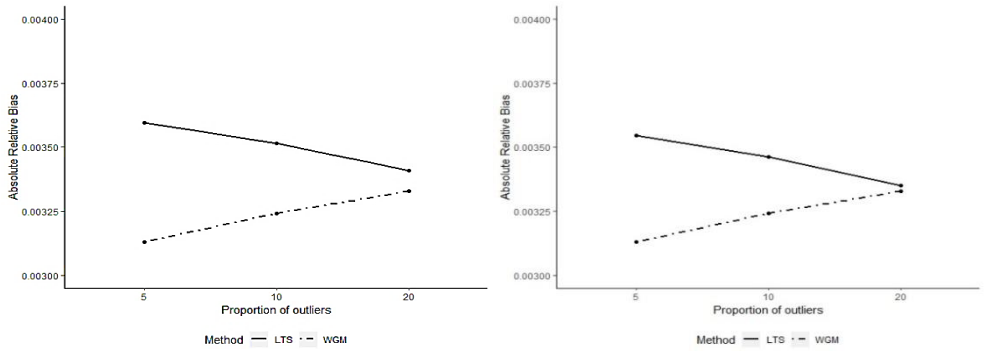
Table 1. The absolute relative bias value of parameter estimation ($\beta_1 = 10$)

Types of outliers	Proportion of outliers	Within method	Robust estimator method	
			LTS	WGM
Without outliers	0%	0.0027	0.0041	0.0034

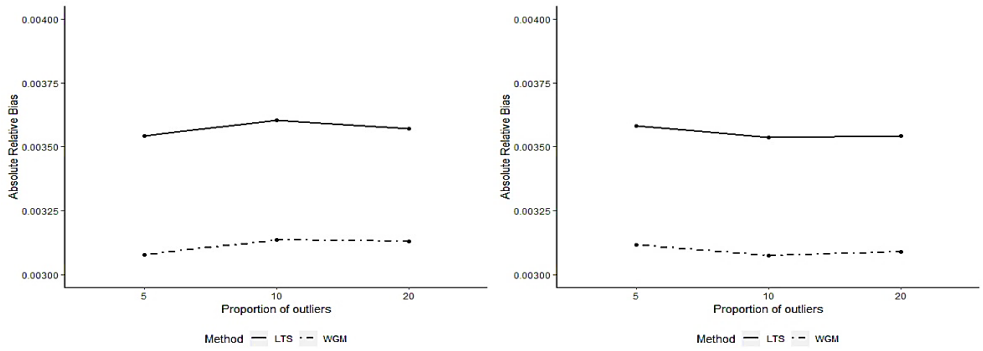
Vertical outliers	5%	0.1511	0.0036	0.0031
	10%	0.2309	0.0035	0.0032
	20%	0.2953	0.0034	0.0033
Leverage points	5%	0.7045	0.0035	0.0031
	10%	0.7483	0.0035	0.0032
	20%	0.7715	0.0034	0.0033
Conc. vertical outliers	5%	0.1483	0.0035	0.0031
	10%	0.1510	0.0036	0.0031
	20%	0.1529	0.0036	0.0031
Conc. leverage points	5%	0.7451	0.0036	0.0031
	10%	0.7450	0.0035	0.0031
	20%	0.7415	0.0035	0.0031

In the data group with 0% outliers (without outliers), parameter estimation β_1 is almost the same for each method (both within method and the robust estimator method).Based on the absolute relative bias values, the parameter estimation value $\beta_1$obtained from the withinmethod deviated far from the actual value when the data began to be contaminated with outliers, especially the data containing the types of leverage points and concentrated leverage points. The existence of leverage points has a significant effect on parameter estimation performed by MKT [4]. Based on the explanation above, the within method is a parameter estimation method that is not robust againstoutliers.The greater the level of outlier contaminations in each type of outlier, the greater the absolute relative bias values produced. The absolute relative bias value for parameter estimator β_1 with the robust method is shown in **Figure 2**. Based on the types of outliers, leverage points produce the maximum values of absolute relative bias compared to other types of outliers. This shows that the types of leverage points are very risky if they are found in the tested data.

Based on the absolute relative bias value, the parameter estimation using the WGM method had the closest value to zero when the level of outlier contaminations increased from 5% to 20%. The WGM method also produced the lowest absolute relative bias value compared to that of LTS method. This means that the WGM method is the best robust estimator method because the estimated value was almost the same as the value of the parameters at various levels of contamination and types of outliers.



a) vertical outliers b) leverage points

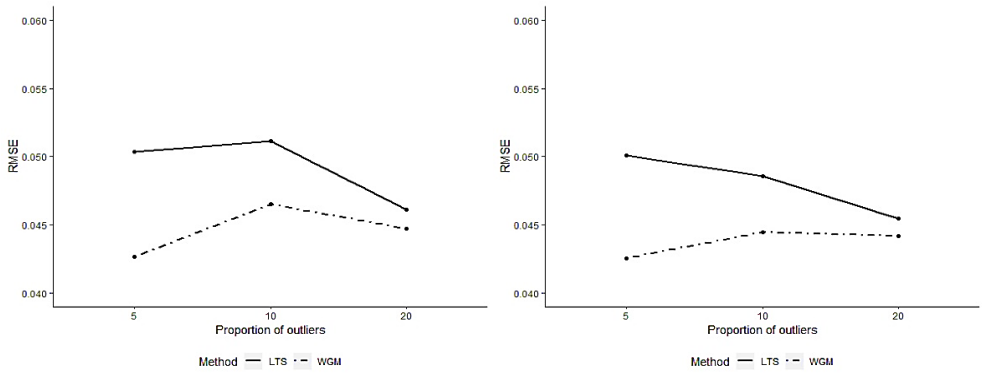


c) concentrated vertical outliers d) concentrated leverage points

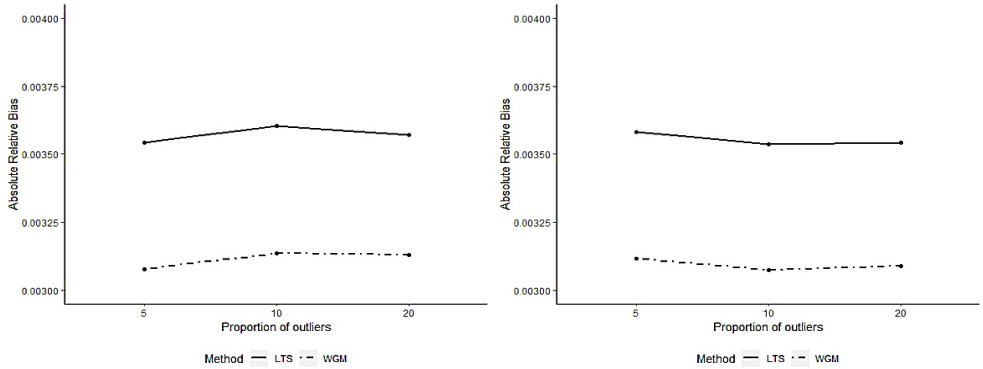
Fig. 2. The absolute relative bias value for parameter estimator (β_1 = 10) with robust method: a) vertical outliers, b) leverage points, c) concentrated vertical outliers, d) concentrated leverage points

The value of the mean square error (MSE) contains 2 components, namely the variety of the estimator (precision) and the bias (accuracy). Estimators with good MSE properties are the ones that control variety and bias. The large value of root meansquare error (RMSE) indicates a large variety of estimators so that it is more risky on the estimation results, which results in the lower accuracy of estimation.

The RMSE value for estimating the $\beta_1$parameter is shown in **Figure 3**.



a) vertical outliers b) leverage points



c) concentrated vertical outliers d) concentrated leverage points

Fig. 3. RMSE value for parameter estimator($\beta_1 = 10$) with robust method: a) vertical outliers, b) leverage points, c) concentrated vertical outliers, d) concentrated leverage points

The RMSE value obtained also gives the same depiction with the absolute relative bias value in Figure 2, which is the greater the level of outliers contaminating the data, the greater result of the RMSE value is.In estimating the parameter β_1 with various types of outliers and levels of outlier contaminations, the WGM method produces the lowest RMSE value compared to that of LTS method. This means that the WGM method produces a small variety of estimators of the parameter β_1 and the high accuracy of the parameter estimator β_1 for various types of outliers and levels of outlier contaminations. Based on the evaluation criteria, the absolute relative bias value and RMSE, the WGM method is better at dealing with data contaminated outliers.

4.2Parameter Estimation β_2

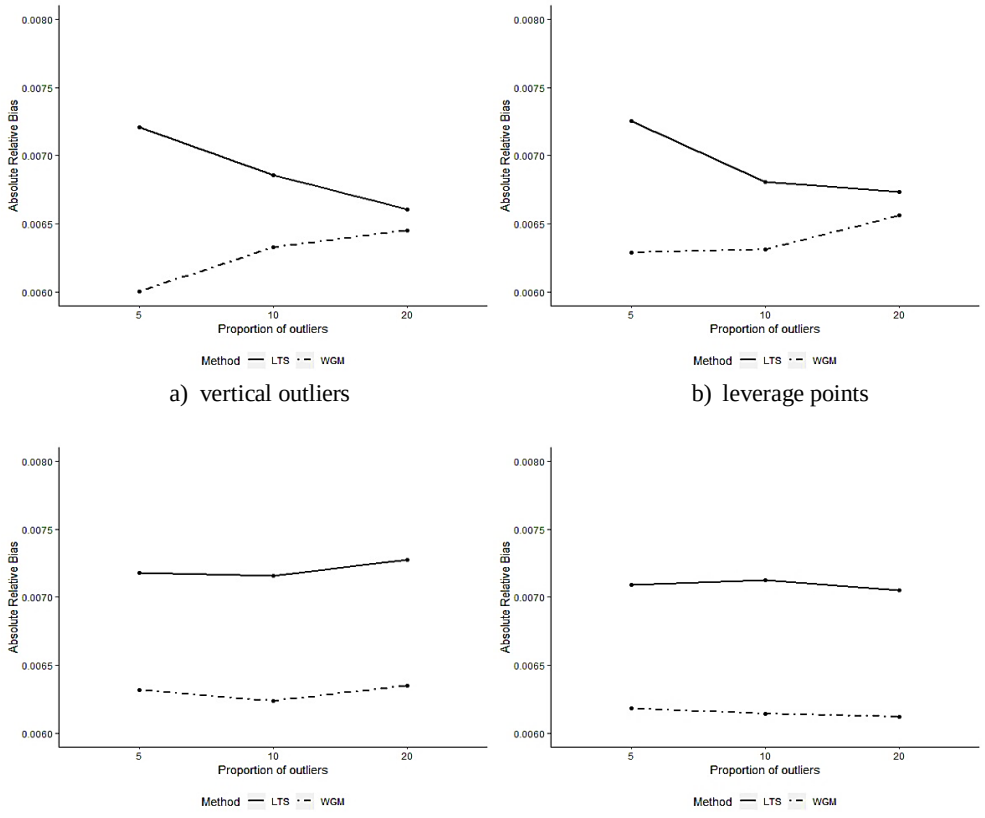
The absolute relative bias value of the parameter estimation ($\beta_2 = 5$) with the within estimator method and robust estimator method is shown in **Table 2**.The absolute relative bias value obtained by the parameter estimation β_2 indicates that the within estimator method and the robust estimator method in the data group without outliers gives almost the same predictive results. The difference in the absolute relative bias value between the within method and the robust estimator method occurs when the data contains various types of outliers and has an increase in outlier contaminations from 0% to 5%, 10%, and 20%. When the data contaminated by types of leverage points outliers and concentrated leverage points, the estimated values of the parameter β_2 obtained from the method of within deviates far from the actual values. This shows that the within method is not robust against outliers.

Table 2. The absolute relative bias value of parameter estimation($\beta_2 = 5$)

Types of outliers	Proportion of outliers	Within method	Robust estimator method	
			LTS	WGM
Without outliers	0%	0.0052	0.0079	0.0064
Vertical outliers	5%	0.3052	0.0072	0.0060
	10%	0.4410	0.0069	0.0063

	20%	0.5323	0.0066	0.0065
Leverage points	5%	1.4071	0.0073	0.0063
	10%	1.4095	0.0068	0.0063
	20%	1.3966	0.0067	0.0066
Conc. vertical outliers	5%	0.2943	0.0072	0.0063
	10%	0.2976	0.0072	0.0062
	20%	0.3009	0.0073	0.0063
Conc. leverage points	5%	1.4902	0.0071	0.0062
	10%	1.4911	0.0071	0.0061
	20%	1.4980	0.0070	0.0061

The absolute relative bias value produced is directly correlated with the level of outlier contamination in the data being tested. This means that the greater the level of contamination outliers given, the greater the absolute relative bias value produced. Based on the evaluation of the absolute relative bias value, the WGM method is the best estimator for various outliers and levels of outlier contaminations (**Figure 4**).In this method, the absolute relative bias value produced is lower than the absolute relative bias value produced by other robust estimator methods. This indicates that the parameter estimation $\beta_2$with the WGM method produces a value close to the actual parameter value for various types of outliers and levels of outlier contaminations.



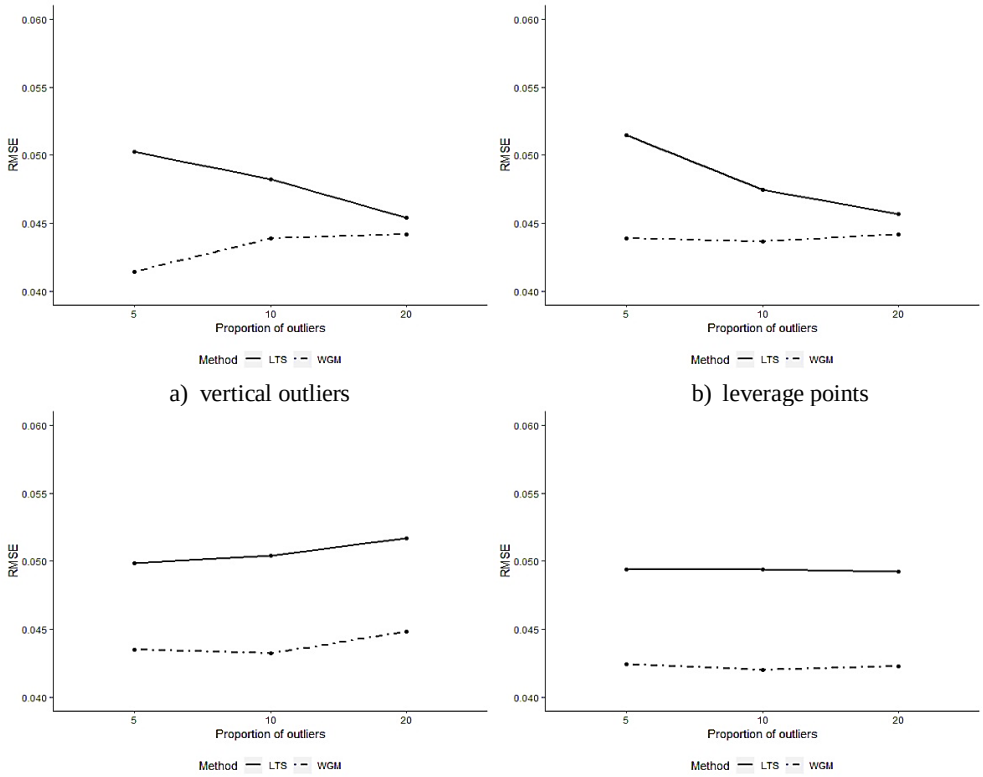
a) vertical outliers b) leverage points

c) concentrated vertical outliers d) concentrated leverage points

Fig. 4. The absolute relative bias value for parameter estimator (β_2 = 5) with robust method: a) vertical outliers, b) leverage points, c) concentrated vertical outliers, d) concentrated leverage points

A good estimator should have a small variety and bias. Therefore, the estimation methods that can control bias and the range of estimators are very necessary to produce high precision of estimation. The variety of estimators and biases can be shown through the root mean square error (RMSE) value produced.

The RMSE value for the parameter estimation β_2 in various types of outliers and the levels of outlier contaminations is shown in **Figure 5**. When it is viewed based on the types of outliers, the RMSE value tends to increase with increasing levels of outlier contaminations in the data.Based on the RMSE value, the WGM method produces the lowest RMSE parameter estimation β_2 compared to other robust estimator methods. This shows that the WGM method produces a small variety of estimation and the high accuracy of estimation for various types of outliers and levels of outlier contaminations. In other words, the WGM method is more robust to outliers than what appeared on the data being tested.



a) vertical outliers b) leverage points

c) concentrated vertical outliers d) concentrated leverage points

Fig. 5.RMSE value for parameter estimator(β_2 = 5) with robust method: a) vertical outliers,b) leverage points, c) concentrated vertical outliers, d) concentrated leverage points

5Conclusion

Based on the estimation evaluation of the parameters (β_1 and β_2), the absolute relative bias and RMSE value of the WGM method have a steady absolute relative bias value and the lowest RMSE value compared to that of LTS method. Thus, the WGM method is more robust against outliers that appear on the data being tested compared to the LTS method.

References

[1]M. Aquaro and P. Čížek: One-step robust estimation of fixed-effects panel data models. Computational Statistics and Data Analysis. Vol. 57, No. 1, pp. 536–548 (2013).
[2]S. Li, K. Wang, and Y. Ren: Robust estimation and empirical likelihood inference with exponential squared loss for panel data models. Economics Letters.Vol. 164, pp. 19–23 (2018).
[3]Y. Lyu: Detection of Outliers in Panel Data of Intervention Effects Model Based on Variance of Remainder Disturbance. Mathematical Problems in Engineering. pp. 1–12 (2015).
[4]B. H. Baltagi and G. Bresson: Robust Panel Data Methods and Influential Observations," in The Oxford Handbook of Panel Data. pp. 418–450, Oxford University Press, US (2015).
[5]A. Al Sayed, A. Aljarah, S. S. Kun, and Z. Isa: Robust Estimation and Outlier Detection on Panel Data: an Application to Environmental Science. Book of Abstract: International Conference on Robust Statistics,2017, p. 56 (2017).
[6]M. C. Bramati and C. Croux: Robust estimators for the fixed effects panel data model. Econometrics Journal. Vol. 10, No. 3, pp. 521–540 (2007).
[7]N. M. A. Bakar and H. Midi: Robust Centering in the Fixed Effect Panel Data Model. Pakistan Journal of Statistics. Vol. 31, No. 1, pp. 1–3 (2015).
[8]P. Čížek: Reweighted least trimmed squares: An alternative to one-step estimators. CentER Discussion Paper. No. 2010-100, pp. 1–34 (2010).
[9]V. Benáček and E. Michalíková: The Factors of Growth of Small Family Businesses. A Robust Estimation of the Behavioural Consistency in Panel Data Models.Prague Economic Papers. Vol. 25, No. 1, pp. 85–98 (2016).