Pertanika J. Sci. & Technol. 30 (4): 2315 - 2332 (2022)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Application of Groupwise Principal Sensitivity Components on Unbalanced Panel Data Regression Model for Gross Regional Domestic Product in Kalimantan

Desi Yuniarti[1,2], Dedi Rosadi[1]* and Abdurakhman[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Sekip Utara Bulaksumur 21 Yogyakarta 55281, Indonesia

[2]Department of Mathematics, Faculty of Mathematics and Natural Science, Mulawarman University,Barong Tongkok 04, Kampus Gunung Kelua, Samarinda, East Kalimantan 75123, Indonesia

ABSTRACT

Most robust estimation methods for panel data regression models do not consider the panel data structure consisting of several cross-sections and time-series units. This robust method, which does not consider the panel data structure, can completely remove all observations from a cross-section unit in trimming outlier observations. However, it can cause biased estimation results for the cross-section unit. This study determines the robust estimate for the unbalanced panel data regression model using Groupwise Principal Sensitivity Components (GPSC) by considering grouped structure data. The results were compared with Within-Group (WG) estimation and other robust estimation methods, namely Within-Group estimation with median centering (Median WG), Within-Group Least Trimmed Squares (WG-LTS), and Within Generalized M (WGM) estimators. Comparisons were made based on the Mean Squares Error (MSE) value. In this study, we applied the proposed method to the unemployed and the Gross Regional Domestic Product (GRDP) data at constant prices in Kalimantan, Indonesia. The analysis showed that GPSC was the best method with the smallest MSE value. Therefore, we can consider implementing and developing the GPSC method to detect and determine the robust estimates for the unbalanced panel data regression model because it fits the panel data structure.

Keywords: Outliers detection, robust estimation, unbalanced panel data regression

ARTICLE INFO

Article history:
Received: 08 January 2022
Accepted: 18 March 2022
Published: 21 July 2022

DOI: https://doi.org/10.47836/pjst.30.4.01

E-mail addresses:
desiyuniarti@mail.ugm.ac.id (Desi Yuniarti)
dedirosadi@gadjahmada.edu (Dedi Rosadi)
rahmanstat@ugm.ac.id (Abdurakhman)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Panel data regression analysis is a regression analysis that uses panel data. Panel data results from observations of several cross-section units, namely companies, households, or individuals, over several periods. So panel data has dimensions of space and time because it consists of several cross-sections and time-series units. Balanced panel data has the same time-series units in each cross-section unit. In contrast, unbalanced panel data has a different number of time-series units in the cross-section units (Gujarati, 2004).

Bramati and Croux (2007) stated that the outliers in panel data might lead to a biased regression estimator. Therefore, it requires a robust panel data regression estimator against outliers. Research on robust estimators against outliers in panel data has begun to develop. Bramati and Croux's (2007) research is currently being developed and has become the basis for many studies on the robust estimators of outliers for panel data regression models. Bramati and Croux (2007) discussed robust estimators for panel data regression models with a fixed-effects approach, namely the Within Groups Generalized M (WGM) estimator by Wagenvoort and Waldmann (2002) and the Within Groups MS (WMS) estimator from Maroona and Yohai (2000) applied to panel data. Several research developments offer different data-centering methods and other estimation methods, such as Aquaro and Čížek (2013), Víšek (2015), Bakar and Midi (2015), and Midi and Muhammad (2018). Another robust estimation method is by Beyaztas and Bandyopadhyay (2020), studying the impact of outlier observations on the Ordinary Least Squares (OLS) estimation method in a linear panel data model and suggested a robust alternative estimation procedure based on weighted likelihood.

The studies described above applied and developed robust methods for panel data regression models, and each method has its advantages. However, we are interested in a robust estimation method considering a panel data structure consisting of several cross-sections and time-series units. The panel data structure consisting of several cross-section units allows for any differences in the average for each cross-section unit. For instance, the cross-section unit shows an area with poverty level data. Therefore, regions with the highest or lowest poverty rates can be considered outliers. The trimming process for outliers using the robust general method can remove observations from a cross-section unit. However, this trimming process can lead to biased estimation results for the appropriate cross-section unit. For this reason, it is necessary to make a robust estimate for the regression model that considers the structure of the panel data, especially the unbalanced panel data.

This study aims to apply the Groupwise Principal Sensitivity Components (GPSC) method by Perez et al. (2013) to detect and determine robust estimates for panel data regression models. The GPSC method is an outlier detection method to obtain robust estimates for grouped data and follows a linear regression model approach with fixed group effects. This GPSC method uses a sensitivity matrix formed by the sensitivity vector of each cross-section unit. This method develops the Principal Sensitivity Components (PSC) method by Pena and Yohai (1999), adapted for grouped data. The GPSC method, which pays attention to the structure of the grouped data, is suitable for the regression model for panel data consisting of several cross-section units. This GPSC approach can identify the outliers within groups and ensure that outlier trimming does not remove more than 50% of data points from the same group. The outlyingness test of observations is based on a robust estimator from the previous analysis. The development of this GPSC method will be perfect for research on outlier detection and robust estimation for subsequent panel data regression models because this method considers the grouped data structures corresponding to panel data with several cross-section units. In the future, it is expected that research on robust estimators for panel data can be more concerned with the structure of the panel data itself.

Perez et al. (2013) studied outlier detection and robust estimation with data distributed into groups following a linear regression model with fixed group effects. They used several methods, including GPSC, the RDL1 method by Hubert and Rousseeuw (1997), and the MS method from Maroona and Yohai (2000). The results showed that GPSC could detect a high percentage of a true and small number of false outliers. This method was also capable of detecting any hidden high leverage points. In addition, this method could maintain good efficiency properties while maintaining good robustness properties. In their paper, Perez et al. (2013) also explained the deficiency of applying the robust method of M estimation, Generalized M (GM) estimation, least median of squares (LMS) method, Least Trimmed Squares (LTS) method, and Weighted Likelihood Estimator (WLE) method by Agostinelli and Markatou (1998) and Markatou, et al. (1998) on grouped data. Perez et al. (2013) stated that these methods are unsuitable for grouped data and proposed the GPSC method.

The GPSC method was developed based on the PSC method by Pena and Yohai (1999). PSC is a fast iterative procedure to estimate parameters based on a minimal robust scale. The procedure for minimizing this robust scale is obtained by eliminating possible outliers. In the study of Pena and Yohai (1999), each observation is represented by a sensitivity vector, a vector of changes in the least-squares estimate of the observations when each data point is removed. The set of possible outliers obtained as extreme points in the principal components of this vector is called the Principal Sensitivity Components (PSC), or as the set of points with large residuals. The good performance of the proposed procedure by Pena and Yohai (1999) allows the identification of outliers. Pena and Yohai (1999) explained two ways to see the outlyingness of the *i*th observation, i.e., by using an influence vector and a sensitivity vector. Pena and Yohai (1995) described an analysis based on the influence vector, and Pena and Ruiz-Castillo-Castillo (1998) used it to detect outlier groups in the regression model.

Perez et al. (2013) applied the GPSC method to income data as outcome variables, hectare, food crops, beef and lamb data as covariates variables, and state variables as grouping variables, which states the seven states where agriculture was. This application of the GPSC method does not clearly state the type of data in each group. However, based on our study, the fixed group effects regression model used by Perez et al. (2013) is more general than panel data because, in each group, there can be data from several objects in the group. If the data in each group is in the form of time series data, it means that the data is panel data. For this reason, we emphasize the application of the GPSC method to panel data, especially for unbalanced panel data. The method can be used for panel data, especially for unbalanced panel data, because the analysis method considers the unbalanced panel data structure consisting of several cross-section units with different time-series units.

This study applies the GPSC method to the GRDP data of Kalimantan, Indonesia, by looking at the effect of unemployment on the GRDP of Kalimantan. The COVID-19 pandemic has dealt a severe blow to the world economy, including Indonesia, thus providing a different pattern for economic growth data. This condition then allows the occurrence of outliers that require a robust estimation method against outliers. We also compare the GPSC estimation results with Within-Group (WG) estimation and several other robust estimation methods, namely, WG estimation with median centering (Median WG), WG Least Trimmed Squares (WG-LTS), and Within Generalized M (WGM) estimator. Finally, we will use the Mean Squares Error (MSE) value to determine the best robust estimation method.

METHODOLOGY

Unbalanced Panel Data Regression Model

This study used an unbalanced panel data regression model in Equation 1 (Baltagi, 2005):

$$Y_{it} = \alpha + \mathbf{X}_{it}^{'}\boldsymbol{\beta} + u_{it}\ ,\ \ i = 1, 2, \ldots, N\ ; \quad t = 1, 2, \ldots, T_i \qquad [1]$$

where Y_{it} is the value of the Y variable for the tth time-series unit in the ith cross-section unit, α is a constant, $\mathbf{X}_{it}' = [X_{1_{it}},\ X_{2_{it}}, \ldots, X_{K_{it}}]$ is a vector of independent variables of size $K \times 1$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_K]'$ is a vector of a parameter of size $K \times 1$, N state the number of cross-section units, T_i state the number of time-series observations in the ith cross-section unit, and the total number of all observations is $n = \sum_{i=1}^{N} T_i$. Equation 1 is an unbalanced panel data regression model because it has a different number of time-series units for cross-section units.

The panel data regression model in Equation 1 is a one-way error component model if (Equation 2) (Baltagi, 2005):

$$u_{it} = \mu_i + v_{it} \qquad [2]$$

where $v_{it} \sim IIN(0, \sigma_v^2)$ and $\mathbf{X}_{it}$ is assumed to be independent of v_{it}. If μ_i is fixed, the model in Equation 1 with the error component in Equation 2 is a one-way panel data regression model with a fixed-effects approach.

Equations 1 and 2 give the following model of Equation 3:

$$Y_{it} = \alpha_i^* + \mathbf{X}_{it}^{'}\boldsymbol{\beta} + v_{it}\ ,\quad i = 1,2,\dots,N\ ;\quad t = 1,2,\dots,T_i \qquad [3]$$

where $\alpha_i^* = \alpha + \mu_i$. Equation 3 is a panel data regression model specified for individual effects that are constant over time. For $i = 1,2,\dots,N$ we can express Equation 3 in the following vector form of Equation 4 (Hsiao, 2003):

$$\mathbf{y}_i = \alpha_i^* \cdot \mathbf{1}_{T_i} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i \qquad [4]$$

where $\mathbf{y}_i = \left[y_{i1}\ y_{i2}\ \cdots\ y_{iT_i}\right]'$, $\mathbf{1}_{T_i}$ is a vector of elements one of size $T_i \times 1$, vector $\mathbf{X}_i = [\mathbf{X}_{i1}\ \mathbf{X}_{i2}\ \cdots\ \mathbf{X}_{iT_i}]'$ of size $K \times T_i$, $\beta = [\beta_1\ \beta_2 \cdots \beta_K]'$ of size $K \times 1$, $\mathbf{v}_i = \left[v_{i1}\ v_{i2} \cdots v_{iT_i}\right]'$, and assuming $E[\mathbf{v}_i] = \mathbf{0}$, $E[\mathbf{v}_i\mathbf{v}_i'] = \sigma_v^2\mathbf{I}_{T_i}$, $E\left[\mathbf{v}_i\mathbf{v}_j'\right] = \mathbf{0}$ for $i \neq j$, where $\mathbf{I}_{T_i}$ is the identity matrix of size $T_i \times T_i$.

We could then obtain the estimation of Equation 4 by the Ordinary Least Squares (OLS) method. The OLS estimator of α_i^* and β was obtained by minimizing $S = \sum_{i=1}^{N} \boldsymbol{v}_i'\boldsymbol{v}_i$. The within-group covariance matrix of variables X and the within-group covariance vector between the variables X and Y, respectively, are as in Equation 5 (Perez et al., 2013):

$$\mathbf{S}_{XX,i} = \frac{1}{T_i}\sum_{t=1}^{T_i}(\mathbf{X}_{it} - \overline{\mathbf{X}}_i)\,(\mathbf{X}_{it} - \overline{\mathbf{X}}_i)'$$

$$\mathbf{S}_{XY,i} = \frac{1}{T_i}\sum_{t=1}^{T_i}(\mathbf{X}_{it} - \overline{\mathbf{X}}_i)\,(y_{it} - \bar{y}_i) \qquad [5]$$

for $i = 1,2,\dots,N$, where $\overline{\mathbf{X}}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{X}_{it}'$ is the mean of the variable $X_k, k = 1,2,\cdots,K$ in the *i*th cross-section unit, and $\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it}$ is the average of the Y variables in the *i*th cross-section unit. The combined covariance matrix $\mathbf{S}_{XX}$ and the combined covariance vector $\mathbf{S}_{XY}$ are as in Equation 6 (Perez et al., 2013):

$$\mathbf{S}_{XX} = \sum_{i=1}^{N}\frac{T_i}{n}\mathbf{S}_{XX,i} \quad ; \quad \mathbf{S}_{XY} = \sum_{i=1}^{N}\frac{T_i}{n}\mathbf{S}_{XY,i} \qquad (6)$$

The least squares (LS) estimator of β and α_i^* respectively are stated as Equations 7 and 8 (Perez et al., 2013):

$$\widehat{\boldsymbol{\beta}} = \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$$

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T_i}(\mathbf{X}_{it} - \overline{\mathbf{X}}_i)\,(\mathbf{X}_{it} - \overline{\mathbf{X}}_i)'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T_i}(\mathbf{X}_{it} - \overline{\mathbf{X}}_i)\,(y_{it} - \bar{y}_i)\right] \qquad (7)$$

and

$$\hat{\alpha}_i^* = \bar{y}_i - \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}} \quad ; \quad i = 1, 2, \cdots, N \tag{8}$$

Within-Groups Estimator of Unbalanced Panel Data Regression Model

Within-Groups (WG) estimator for the unbalanced panel data regression model uses the within transformation by forming a $\mathbf{Q} = \mathrm{diag}(\mathbf{E}_{T_i})$ matrix where $\mathbf{E}_{T_i} = \mathbf{I}_{T_i} - \bar{\mathbf{J}}_{T_i}$, $\mathbf{I}_{T_i}$ is an identity matrix of size $T_i \times T_i$, $\bar{\mathbf{J}}_{T_i} = \frac{1}{T}\mathbf{J}_{T_i}$, $\mathbf{J}_{T_i}$ is a one-element matrix of size $T_i \times T_i$. The matrix elements of $\mathbf{Q}$ are the deviations from the individual mean. This transformation is applied to the vector form of Equations 1 and 3, giving the following Equations 9 and 10:

$$\mathbf{y} = \alpha\boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad [9]$$

and

$$\mathbf{u} = \mathbf{Z}_\mu\boldsymbol{\mu} + \mathbf{v} \qquad [10]$$

where $\mathbf{y}$ is a $n \times 1$ vector, α is a scalar, $\boldsymbol{\iota}_n$ is a vector of ones of size $n \times 1$, $\mathbf{X}$ is a matrix of independent variable of size $n \times K$, $\boldsymbol{\beta}$ is a parameter vector of size $K \times 1$, $\mathbf{u}$ is an error vector of size $n \times 1$, and $\mathbf{Z}_\mu = \mathrm{diag}(\boldsymbol{\iota}_{T_i})$ is a square diagonal matrix with the vector element $\boldsymbol{\iota}_{T_i}$ on the main diagonal.

Thus, we could get within transformation of Equation 9 to Equation 11 (Baltagi, 2005):

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{v}} \qquad [11]$$

where $\tilde{\mathbf{y}} = \mathbf{Qy}$, $\tilde{\mathbf{X}} = \mathbf{QX}$, and $\tilde{\mathbf{v}} = \mathbf{Qv}$. Applying the least-squares method to Equation 11, we could get the WG estimator as Equation 12 (Baltagi, 2005):

$$\tilde{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y} \qquad [12]$$

provided $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ exists.

Groupwise Principal Sensitivity Components for Unbalanced Panel Data Regression Model

This section explains the GPSC method corresponding to the unbalanced panel data regression model in Equation 1. In applying GPSC to the model of Equation 3, the cross-section unit in the panel data corresponds to the group in the model of GPSC method by Perez et al. (2013). Perez et al. (2013) developed the GPSC algorithm we applied in this study. The GPSC method goes through two stages: the first stage is to determine the clean set from outliers and then determine the initial robust estimate of the clean set. In stage 1, we formed the sensitivity matrix $\mathbf{R}_i$ for *i*th cross-section unit. Given that $\hat{y}_{it(ij)}$ is the estimated value of y_{it} if we remove the *ij*th observation $\left(y_{ij},\ \mathbf{X}'_{ij}\right)$ as Equation 13 (Perez et al., 2013):

$$\hat{y}_{it(ij)} = \hat{\alpha}_{i(ij)} + \mathbf{X}'_{it(ij)}\widehat{\boldsymbol{\beta}}_{(ij)} \qquad [13]$$

We determined the estimated value of $\widehat{\boldsymbol{\beta}}_{(ij)}$ based on Equation 7 and $\hat{\alpha}_{i(ij)}$ based on Equation 8 using all observations except the *ij*th observation. Furthermore, for each y_{it} observation in the *i*th cross-section unit, the sensitivity vector is the change value vector if every point in the *i*th cross-section unit is deleted as Equation 14 (Perez et al., 2013):

$$\mathbf{r}_{it} = \left(\hat{y}_{it} - \hat{y}_{it(i1)}, \hat{y}_{it} - \hat{y}_{it(i2)}, \cdots, \hat{y}_{it} - \hat{y}_{it(iT_i)}\right)' \qquad [14]$$

$\mathbf{r}_{it}$ vector is the sensitivity vector of the *i*th cross-section unit. Then, we formed the sensitivity matrix $\mathbf{R}_i$ of the *i*th cross-section unit as in Equation 15:

$$\mathbf{R}_i = \left[\mathbf{r}_{i1}\ \mathbf{r}_{i2} \cdots \mathbf{r}_{iT_i}\ \right]' \qquad [15]$$

To avoid any different modeling as much as T_i, we could determine the elements of the matrix $\mathbf{R}_i$ based on the leverage and residuals of the following least square model in Equation 16 (Perez et al., 2013):

$$\hat{y}_{it} - \hat{y}_{it(ij)} = \frac{h^i_{tj} e_{ij}}{1 - h^i_{jj}} \qquad [16]$$

where h^i_{jj} is the leverage effect of the jth observation of the ith cross-section unit as in Equation 17:

$$h^i_{jj} = \frac{1}{T_i} + \left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i\right)'(n\boldsymbol{S}_{XX})^{-1}\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i\right) \qquad [17]$$

Thus, the sensitivity matrix for the ith cross-section is presented in Equation 18 (Perez et al., 2013):

$$\mathbf{R}_i = \mathbf{H}_{ii}\mathbf{W}_i \qquad [18]$$

where,

$$\mathbf{H}_{ii} = \begin{bmatrix} h^i_{11} & h^i_{12} & \cdots & h^i_{1T_i} \\ h^i_{21} & h^i_{21} & \cdots & h^i_{2T_i} \\ \vdots & \vdots & \ddots & \vdots \\ h^i_{T_i1} & h^i_{T_i2} & \cdots & h^i_{T_iT_i} \end{bmatrix} \text{and } \mathbf{W}_i = \begin{bmatrix} \frac{e_{i1}}{1-h^i_{11}} & 0 & \cdots & 0 \\ 0 & \frac{e_{i2}}{1-h^i_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{e_{iT_i}}{1-h^i_{T_iT_i}} \end{bmatrix}$$

Next, we form the matrix as in Equation 19 (Perez et al., 2013):

$$\mathbf{M}_i = \mathbf{R}_i'\mathbf{R}_i \qquad [19]$$

and determined the non-zero eigenvalues of the $\mathbf{M}_i$ matrix and the set of eigenvectors $\{\mathbf{v}_q^i, \quad q = 1, 2, \cdots, K+1\}$ corresponding to the non-zero eigenvalues of the $\mathbf{M}_i$ matrix. The maximum eigenvalue of $\mathbf{M}_i$ expressed as λ_1^i is a measure of the global influence from the observation of the *i*th cross-section unit on the predicted values of observations in that cross-section unit. The eigenvector v_1^i corresponding to the eigenvalue λ_1^i is the direction of the maximum sensitivity of the observations at the *i*th cross-section unit. Eigenvector $\{\mathbf{v}_q^i,$ $q = 1, 2, \cdots, p+1\}$ is the orthogonal direction in which the joint effect of deleting multiple data points from the *i*th cross-section in the estimated value is maximized. Therefore, the projection of Equation 20 is as follows:

$$\mathbf{z}_q^i = \mathbf{R}_i\mathbf{v}_q^i \qquad [20]$$

in the direction of $\mathbf{v}_q^i, q = 1, 2, \cdots, K+1$ detects high leverage points with high mutual influence in the *i*th cross-section unit. This projection is the principal component of the sensitivity vector. According to Pena and Yohai (1999), the group of points that together have a leverage effect in the *i*th cross-section unit is expected to have extreme coordinates at least one of $p+1$ PSC $\{\mathbf{z}_q^i, q = 1, 2, \cdots, K+1\}$. Furthermore, for each principal component of q, a different data set is formed, namely the first set containing all observations from each cross-section unit and the second set, deleting 50% of observations with the largest coordinates in the vector (Equation 21):

$$\mathbf{d}_q^i = \left|\mathbf{z}_q^i - \mathrm{med}\left(\mathbf{z}_q^i\right)\right|; \quad q = 1, 2, \cdots, K+1 \qquad [21]$$

The two sets for each of the *ı*th cross-section units were combined. Furthermore, other small but potentially clean data sets were also formed with the smallest number of eigenvalues provisions. Then, the LS estimates for each of these sets were determined. Based on the results of this LS estimation, we chose the LS estimate that minimized the *s*-scale estimator. In this study, the robust scale estimate used was the median absolute deviation (MAD) as in Perez et al. (2013). Next, all observations were deleted as in Equation 22:

$$|r_{it}| \geq C_1 \cdot s_i \qquad [22]$$

for $C_1 = 2$ and s_i is the Median Absolute Deviation (MAD) robust scale for the *i*th cross-section unit. Iterations were performed for all remaining observations. For example, $\boldsymbol{\gamma}^{(r)} = \left(\left(\boldsymbol{\beta}^{(r)}\right)^T, \alpha_1^{(r)}, \alpha_2^{(r)}, \cdots, \alpha_N^{(r)}\right)^T$ is the estimator obtained by minimizing robust scale on the *r*th iteration. The iteration will end when $\boldsymbol{\gamma}^{(r+1)} = \boldsymbol{\gamma}^{(r)}$ and then $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^{(r+1)} = (\boldsymbol{\beta}^{*T}, \alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*)^T$ is the initial robust estimator.

Based on the results of stage 1 analysis, we obtained a data set that may be clean because we removed observations that could be outliers. Furthermore, in stage 2, we tested these potential outliers using a robust t-test. Finally, we returned the observations not rejected by this robust t-test to the sample data and used them to determine the final estimator. The steps in stage 2 include determining the residuals from the initial robust estimator and eliminating observations by using Equation 23:

$$|r_{it}| > C_2 \cdot s_i \quad ; \quad i = 1, 2, \cdots, N \qquad [23]$$

for $C_2 = 3$. Let n^* be the total number of deleted observations. Then, the LS estimator for the remaining $n - n^*$ observations is calculated and expressed as $\tilde{\alpha}_i, i = 1, 2, \cdots, N$ and $\widetilde{\boldsymbol{\beta}}$. Also, the standard error $\tilde{\sigma}$ using the residuals of these remaining observations and the corresponding leverage $\tilde{h}_{tt}^i$ based on Equation 17 was calculated. The outlyingness test of each n^* observation used the following robust t-test statistics in Equation 24:

$$t_{it} = \frac{y_{it} - \tilde{\alpha}_i - \mathbf{X}_{it}'\widetilde{\boldsymbol{\beta}}}{\tilde{\sigma}\sqrt{1 + \tilde{h}_{tt}^i}} \qquad [24]$$

We eliminate every *n* observation with $|t_{it}| > C_3$, where $C_3 = 3$ (Perez et al., 2013).

Other Robust Estimators for Unbalanced Panel Data Regression Model

In this study, the median WG, WG-LTS, and WGM estimator were other analytical methods used to determine robust estimators. Initially, we determined the centering of variables (dependent and independent) on the median (med) as follows (Equation 25) (Bramati & Croux, 2007):

$$\tilde{Y}_{it} = Y_{\boldsymbol{it}} - \underset{t}{\text{med}}\, Y_{it} \text{ and } \tilde{X}_{K_{it}} = X_{K_{it}} - \underset{t}{\text{med}}\, X_{K_{it}} \qquad [25]$$

for $1 \leq i \leq N, 1 \leq t \leq T_i$, $n = \sum_{i=1}^{N} T_i$ and $1 \leq k \leq K$, $X_{K_{it}}$ is the *k*th independent variable, $k = 1, 2, \dots, K$ measured at the *t*th time-series unit in the *i*th cross-section unit.

We then obtained the median WG estimator by doing the median centering based on Equation 25 first and determined the WG estimator using the variables $\tilde{Y}_{it}$ and $\tilde{X}_{K_{it}}$ based on Equation 12. Aquaro and Čížek (2013) have also used this robust method before.

WGM estimator is one of Bramati and Croux's (2007) methods, and the WG-LTS estimator is the initial estimator for the WGM estimator. This study applied the WGM method to determine a robust estimator for the unbalanced panel data regression model in Equation 1. After doing median centering of variables, Bramati and Croux (2007) regressed $\tilde{Y}_{it}$ against $\tilde{X}_{K_{it}}$ using a robust regression method, namely the LTS method that minimizes the sum of h, the smallest residual squared as in Equation 26:

$$\tilde{\beta}_{LTS} = \arg\min_{\beta} \sum_{j=1}^{h} \left[\left(\tilde{Y}_j - \tilde{\mathbf{X}}'_j \boldsymbol{\beta} \right)^2 \right]_{j:n} \quad [26]$$

for $n = \sum_{i=1}^{N} T_i, h = 3/4\, NT$ is the truncation value. For the estimator β with median centering, we got Equation 27 (Bramati & Croux, 2007):

$$\tilde{\alpha}_i(\beta) = \underset{t}{\operatorname{med}}(Y_{it} - \mathbf{X}'_{it}\boldsymbol{\beta}),\, i = 1, 2, \dots, N \quad [27]$$

The WGM estimator extends the LTS within-group estimator to improve statistical efficiency while maintaining robustness. To determine the WGM estimator, we formed a diagonal matrix of $\mathbf{W}_r$ of size $n \times n$ to reduce observations' weight with a large residual value from a robust initial LTS fit regression model. The loss function used Tukey's biweight function so that the diagonal element $\mathbf{W}_r$ became Equation 28 (Bramati & Croux, 2007):

$$(W_r)_{it} = \begin{cases} 0, & if \ \left| \frac{r_{it}}{\hat{\sigma}_{LTS}} \right| \geq c \\ \left(1 - \left(\frac{r_{it}}{c\hat{\sigma}_{LTS}} \right)^2 \right)^2, if & \left| \frac{r_{it}}{\hat{\sigma}_{LTS}} \right| < c \end{cases} \quad [28]$$

where $r_{it} = \tilde{Y}_{it} - \tilde{\mathbf{X}}'_{it}\hat{\boldsymbol{\beta}}_{LTS}$ is the residual of the WG-LTS model, $\hat{\sigma}_{LTS}$ is the robust scale estimate of residual, $\hat{\sigma}_{LTS} = c\frac{1}{h}\sum_{j=1}^{h}\left[\left(\tilde{Y}_j - \tilde{\mathbf{X}}'_j \boldsymbol{\beta} \right)^2 \right]_{j:n}$, and c = 4.685 according to Wagenvoort and Waldmann (2002). Furthermore, we formed a $\mathbf{W}_x$ matrix of size $n \times n$. The diagonal elements of the $\mathbf{W}_x$ matrix are presented in Equation 29 (Bramati & Croux, 2007):

$$(W_x)_{it} = \min\left(1, \frac{\sqrt{\chi^2_{K,0.975}}}{RMD_{it}} \right) \quad [29]$$

where $\chi^2_{K,0.975}$ is the upper 97.5% quantile of a Chi-Squared distribution with K degrees of freedom. Robust distance RMD_{it} is a robust of Mahalanobis distance computed for every $\tilde{X}_{it}$ as in Equation 30:

$$RMD_{it} = \sqrt{\left(\tilde{X}_{it} - \hat{\mu}\right)' \hat{V}^{-1} \left(\tilde{X}_{it} - \hat{\mu}\right)}, \qquad i = 1, 2, \cdots, N, \qquad t = 1, 2, \cdots, T_i \quad [30]$$

where $\hat{\mu}$ and $\hat{V}$ are the robust location estimates and covariate estimates of the centered independent variables, calculated by applying the S-multivariate location and scale estimator, respectively.

Thus, we could determine WGM estimator for the one-way panel data regression model with a fixed-effects approach as follows (Equation 31) (Bramati & Croux, 2007):

$$\hat{\boldsymbol{\beta}}_{WGM} = \left(\tilde{\mathbf{X}}' \mathbf{W}_x \mathbf{W}_r \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}' \mathbf{W}_x \mathbf{W}_r \tilde{\mathbf{y}} \quad [31]$$

Research Data

The COVID-19 pandemic has hit the economy of the world, including Indonesia. Various problems then occur because of the COVID-19 pandemic, such as business closures and staff reduction, leading to an increase in the unemployment rate and a decrease in people's purchasing power. In the end, it also ultimately affects Indonesia's economic growth, which has decreased. The government is attempting to reduce the impact of COVID-19 on the economy through the National Economic Recovery (NER) program. This program aims to protect, maintain, and improve the economic capacity of business actors in running their businesses during the COVID-19 pandemic.

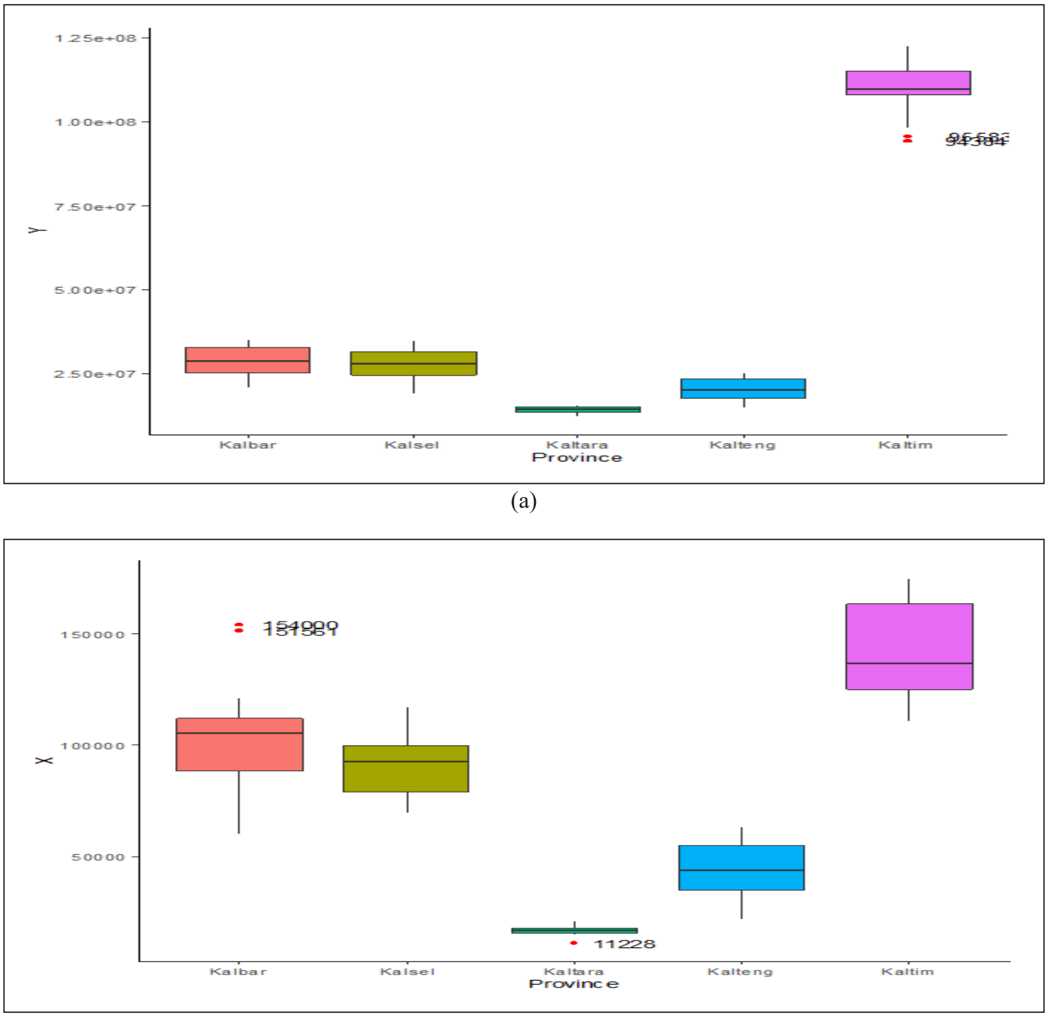
The government's economic restoration efforts are faced with challenges such as the lack of people's purchasing power and a reduction in employees in various businesses. Okun (1962) described the relationship between economic growth (output) and unemployment (input), also known as Okun's Law. Okun's law explains a negative relationship between economic growth and unemployment: when unemployment increase, the economic growth decrease, and vice versa.

This study investigated the effect of unemployment on economic growth using data on the number of unemployed people and the Gross Regional Domestic Product (GRDP) at constant prices (millions of rupiah), showing the economic growth. It used panel data with a cross-section unit covering five provinces on Kalimantan Island, Indonesia, including West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, and North Kalimantan. In addition, the study used quarterly data from 2010 to 2021 as a time-series unit from the Badan Pusat Statistik or Central Bureau of Statistics (BPS) website for each region.

We used some research variables and cross-section units, as shown in Tables 1 and 2. The complete time-series unit referred to quarterly data from 2010 to 2021. Because there were incomplete quarterly data for each variable and unavailable data due to the formation of a new province, causing a different number of time series for each cross-section unit, the data in this study are unbalanced panel data.

Table 3 summarizes data from the variable for each cross-section unit. The average GRDP of East Kalimantan and West Kalimantan provinces were the two highest regions, while the lowest was North Kalimantan, the youngest province on the Kalimantan Island. The highest average unemployment data was in East Kalimantan province, followed by West Kalimantan. However, the average unemployment rate in West and East Kalimantan was not much different, though GRDP was significant. As shown in Table 3, we suspected an outlier in variable X for the province of West Kalimantan and an outlier in variable Y for the province of East Kalimantan. For this reason, we investigated outlier observations through each cross-section unit's boxplots of each research variable.

Figure 1 shows the boxplot of variables Y and X. Based on Figures 1(a) and 1(b), the observations outside the boxplot indicate a presence of outlier observations. Figure 1(a)



(a)

(b)

Figure 1. (a) Boxplot of variable Y; (b) Boxplot of variable X shows two outliers on variable Y in the East Kalimantan Province, 94,384,716 in quarter 1 of 2010 and 95,583,067 in quarter 3 of 2010. Variable X in Figure 1(b) has outlier observations in West Kalimantan Province, 151,562 in quarter 3 of 2020 and 154,000 in quarter 1 of 2021. Also, North Kalimantan province has outlier observations of 11,228 in quarter 1 of 2016.

Table 3 and Figure 1 indicate the presence of outlier observations, necessitating an analytical method to produce their robust parameter estimates. The analysis for robust parameter estimation against outliers is discussed in the next section.

RESULTS AND DISCUSSION

This section determined the robust estimation of the panel regression model for GRDP data in Kalimantan, Indonesia, using the GPSC method. This method considers the structure of grouped data to provide better estimation results for panel data consisting of cross-section and time-series units. The results obtained were compared with WG, median WG, WG-LTS, and WGM estimators based on the smallest MSE value. All analyses were conducted using software R. Moreover, the WG estimator was determined using the PLM package, while median WG, WG-LTS, and WGM were performed using R software following the analysis steps described by Bramati and Croux (2007). Finally, the GPSC analysis was conducted using R software based on the syntax given by Perez et al. (2013).

Based on Equation 1 to Equation 3, the one-way unbalanced panel data regression model with a fixed-effects approach for GRDP of Kalimantan data is given as in Equation 32:

$$Y_{it} = \alpha_i^* + \beta X_{it} + v_{it}, \qquad i = 1, 2, \dots, 5 \qquad [32]$$

where $\alpha_i^* = \alpha + \mu_i$. The i-index shows the ith cross-section unit according to Table 2. The t-index is the index for the tth time-series unit, which shows quarterly data from 2010 to 2021. The data in this study possessed many different time-series units for each cross-section unit. Variables Y and X are following Table 1. The estimation model of Equation 33 is:

$$Y_{it} = \hat{\alpha}_i^* + \hat{\beta} X_{it}\,, \qquad \hat{\alpha}_i^* = \hat{\alpha} + \hat{\mu}_i\ , \qquad i = 1,2,\dots,5 \qquad [33]$$

Outliers in the GRDP and the unemployment data required a robust estimation model in Equation 33 against outliers.

Table 1
Research variable

Variable	Description	Measure
Y	Gross Regional Domestic Bruto (GRDB) at 2010 Constant Market Prices	Million Rupiahs
X	Number of Unemployment	People

Table 2
Cross-section units

i-index	Province	Abbreviation
1	West Kalimantan	Kalbar
2	Central Kalimantan	Kalteng
3	South Kalimantan	Kalsel
4	East Kalimantan	Kaltim
5	North Kalimantan	Kaltara

Table 3
Data summary of each research variable

Variable	Province	Min	Q_1	Mean	Q_3	Max	NA
Y	Kalbar	2,0760,144	24,714,642	28,357,183	32,700,599	34,995,845	0
	Kalteng	14,942,800	17,896,000	20,575,264	23,970,710	25,414,100	2
	Kalsel	19,181,665	24,593,119	27,717,131	31,463,387	34,989,964	0
	Kaltim	94,384,716	107,887,123	110,379,865	115,089,682	122,535,328	0
	Kaltara	12,360,709	13,504,097	14,271,830	15,237,585	15,584,110	12
X	Kalbar	59,884	88,397	103,097	112,081	154,000	1
	Kalteng	21,838	34,994	44,894	54,995	63,309	3
	Kalsel	69,537	79,227	90,845	99,816	117,209	0
	Kaltim	110,574	125,024	141,546	163,517	174,807	0
	Kaltara	11,228	16,079	16,735	17,290	20,867	10

Estimation Robust of Gross Regional Domestic Product of Kalimantan using Groupwise Principal Sensitivity Components

This section introduces the use of the GPSC method on unbalanced panel data. The method has been proposed to detect and determine robust estimates for linear regression models with fixed group effects corresponding to panel data with several cross-section units.

The first stage of the GPSC method determines the sensitivity matrix for each *i*th cross-section unit based on Equation 18 and the $\mathbf{M}_i$ matrix based on Equation 19. The $\mathbf{M}_i$ matrix is $T_i \times T_i$ for $i = 1, 2, \cdots, 5$, where $\mathbf{T}_i$ is the number of time-series units in the *i*th cross-section unit. Table 4 shows the number of $\mathbf{T}_i$ for the *i*th cross-section unit. The application of GPSC to unbalanced panel data began by eliminating incomplete observations, so the number of observations used in this study was 99.

Tables 5 and 6 show the analysis results from stage 1 using the GPSC method. Table 5 shows the parameter estimation results in stage 1 with an estimated robust *s*-scale of 3.082×10^6. Table 6 shows the observations suspected of being outliers at stage 1. Based on Table 6, we could get six observations suspected of being outliers, namely the first and second observations from the first cross-section unit (West Kalimantan), the 43rd observation from the third cross-section unit (South Kalimantan), and the 66th, 67th, and 68th observations from the fourth cross-section unit (East Kalimantan).

Table 4
The number of time-series units of each cross-section unit

i-index	Number of time-series units (T_i)
1	22
2	20
3	23
4	23
5	11
Total (n)	99

Table 5
Parameter estimation of stage 1

Parameter	Parameter Estimation
α_1	3.392×10^7
α_2	2.044×10^7
α_3	2.988×10^7
α_4	1.172×10^8
α_5	1.460×10^7
β	–19.741
MAD	3,081,709

Table 6
Outliers observation of stage 1

ith Observation	Cross-section Unit	Number of Outliers
1, 2	1	2
43	3	1
66, 67, 68	4	3

Table 7
Outliers observation of stage 2

ith Observation	Cross-section Unit	t_{it}	Decision
66	4	-3.749	Outlier
67	4	-3.471	Outlier

Furthermore, in stage 2 of the GPSC method, we tested the outlyingness of the observations as in Table 6. Based on the initial robust estimator, we determined the robust s-scale estimation of each ith cross-section unit and deleted the observations according to Equation 23. Based on the analysis results for each cross-section unit, we found that the 66th and 67th observations in the fourth cross-section unit were the potential outliers. Next, the LS estimator for the remaining observations was determined, and an outlyingness test using the robust test statistic based on Equation 24 was performed. Table 7 shows the results of the robust t-test, and we can conclude that the 66th and 67th observations were the outliers. The final step was to determine the final robust estimate in the second stage based on the LS estimation from the remaining observations without the 66th and 67th observations. The final robust estimate for GRDP data using the GPSC method is shown in Table 8.

The Comparison of Robust Estimates for Gross Regional Domestic Product of Kalimantan

We compared the robust results obtained using GPSC with the WG estimation method and several other robust methods for panel data regression, namely Median WG, WG-LTS, and WGM estimators. The comparison was based on the MSE value, as shown in Table 8.

Table 8
Comparison of robust estimates for GRDP in Kalimantan

Parameter	Parameter Estimation				
	WG	Median WG	WG-LTS	WGM	GPSC
α_1	3.156×10^7	3.157×10^7	2.462×10^7	2.856×10^7	2.889×10^7
α_2	2.158×10^7	2.149×10^7	1.834×10^7	2.013×10^7	2.042×10^7
α_3	3.023×10^7	2.989×10^7	2.547×10^7	2.797×10^7	2.789×10^7
α_4	1.143×10^7	1.140×10^8	1.049×10^8	1.096×10^8	1.121×10^8
α_5	1.474×10^7	1.488×10^7	1.386×10^7	1.444×10^7	1.430×10^7
β	-27.701	-24.217	35.710	1.756	-1.899
MSE	2.530×10^{13}	2.416×10^{13}	3.342×10^{13}	2.436×10^{13}	2.038×10^{13}

Table 9
Intercept estimation of each cross-section unit of the GPSC model

Province	Intercept Estimation
West Kalimantan	2.889×10^7
Central Kalimantan	$2,042 \times 10^7$
South Kalimantan	2.789×10^7
East Kalimantan	1.121×10^8
North Kalimantan	1.430×10^7

In Table 8, the GPSC method gave the lowest MSE value, meaning the GPSC method was the best estimate of Kalimantan's GRDP data. These results were expected because the GPSC method considers a data structure that matches the panel data structure. To generate a robust estimate for outlyingness test statistics, eliminating observations suspected of being outliers consists of several stages. First, this method ensures that the deletion does not exceed 50% of the observations in each cross-section unit, thereby preventing the elimination of the observations in a single cross-section unit and providing an appropriate estimation result for each cross-section unit.

Based on Equation 33 and Table 8, the estimation model of the GPSC method is written as Equation 34:

$$GRDP_{it} = \hat{\alpha}_i^* - 1.899UE_{it} \qquad [34]$$

$GRDP_{it}$ is the Gross Regional Domestic Product of Kalimantan for the tth time-series unit in the ith cross-section unit, UE_{it} is the number of unemployed for the tth time-series unit in the ith cross-section unit. The value of $\hat{\alpha}_i^*$ in Table 9 represents the estimated intercept for each ith cross-section unit. The cross-section units can be seen in Table 2 and the time-series units for the quarterly period.

The model in Equation 34 means that every additional unemployed person will reduce Kalimantan's GRDP by Rp. 1,890,000.00. This model shows a negative relationship between unemployment and GRDP, appropriate with Okun's Law. Therefore, the government should create more job opportunities to fulfill their lives. Income increases people's purchasing power and the economic growth of Kalimantan in particular and Indonesia in general.

Perez et al. (2013) also concluded that the GPSC method is better if the group means differ significantly. This conclusion is consistent with Kalimantan's GRDP data, where East Kalimantan had a much higher average GRDP and unemployment rate than other regions. Thus, the GPSC method is suitable for determining robust estimates for Kalimantan GRDP data. Perez et al. (2013) compared the RDL1, M-S, and GPSC methods. The GPSC and MS methods by Moronna and Yohai (2000) gave almost similar estimation results to other methods, indicating their credibility. The comparison results in Table 8 showed that the estimated intercept parameters were not much different for the WGM and GPSC methods, but the slope differed in sign. The slope sign for the WGM estimation results did not follow Okun's Law, and the MSE value was still higher.

CONCLUSION

This study applies the GPSC method in detecting outliers and determining robust estimates for unbalanced panel data regression models. We intend to emphasize that GPSC can be applied to panel data regression models, especially unbalanced panel data, because this method considers grouped data structures. So that this method is suitable for an unbalanced panel data structure consisting of several cross-section units with a different number of time-series units. We use unbalanced panel data from data on unemployment and the GRDP at constant prices in Kalimantan, Indonesia. We compare the robust estimation results using GPSC with the WG, Median WG, WG-LTS, and WGM estimation methods. Based on the analysis results, we conclude that the GPSC estimation method provides the best robust estimation results for data of GRDP in Kalimantan.

Based on the results and discussion, we suggest developing a method that considers the panel data structure in detecting outliers and determining robust estimation for the panel data regression model, particularly for the unbalanced data panel. Therefore, we can consider implementing and developing the GPSC method because this method is very suitable for an unbalanced panel data structure consisting of several cross-section units with different time-series units.

ACKNOWLEDGEMENTS

The authors thank the Lembaga Pengelola Dana Pendidikan (LPDP) for its financial support in this doctoral study.

REFERENCES

Agostinelli, C., & Markatou M. (1998). A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. *Statistics & Probability Letters, 37*(4), 341-350. https://doi.org/10.1016/S0167-7152(97)00136-3

Aquaro, M., & Čížek, P. (2013). One-step robust estimation of fixed-effects panel data models. *Computational Statistics and Data Analysis, 57*, 536-548. https://doi.org/10.1016/j.csda.2012.07.003

Bakar, N. M. A., & Midi, H. (2015). Robust centering in the fixed effect panel data model. *Pakistan Journal of Statistics, 31*(1), 33-48.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data* (3rd Ed.). John Wiley & Sons Inc.

Beyaztas, B. H., & Bandyopadhyay, S. (2020). Robust estimation for linear panel data models. *Statistics in Medicine, 39*(29), 4421-4438. https://doi.org/10.1002/sim.8732

Bramati, M. C., & Croux C. (2007). Robust estimators for the fixed effects panel data model. *Econometrics Journal, 10*, 521-540. https://doi.org/10.1111/j.1368-423X.2007.00220.x

Gujarati, D. (2004). *Basic Econometrics* (4th Ed.). McGraw-Hill Companies, Inc.

Hsiao, C. (2003). *Analysis of Panel Data* (2nd Ed.). Cambridge University Press.

Hubert, M., & Rousseeuw, P. J. (1997). Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference, 57*(1), 153-163. https://doi.org/10.1016/S0378-3758(96)00041-9

Markatou, M., Basu, A., & Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association, 93*(442), 740-750. https://doi.org/10.2307/2670124

Maroona, R. A., & Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference, 89*(1-2), 197-214. https://doi.org/10.1016/S0378-3758(99)00208-6

Midi, H., & Muhammad, S. (2018). Robust estimation for fixed and random effects panel data models with different centering methods. *Journal of Engineering and Applied Sciences*, *13*(17), 7156-7161.

Okun, A. M. (1962). *Potential GNP, its measurement and significance*. https://milescorak.files.wordpress.com/2016/01/okun-potential-gnp-its-measurement-and-significance-p0190.pdf

Pena, D., & Ruiz-Castillo, J. (1998). The estimation of food expenditures from household budget data in the presence of bulk purchases. *Journal of Business* & *Economic Statistics, 16*(3), 292-303. http://dx.doi.org/10.1080/07350015.1998.10524768

Pena, D., & Yohai, V. J. (1995). The detection of influence subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society, 57*(1), 145-156.

Pena, D., & Yohai, V. J. (1999). A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association, 94*(446), 434-445. https://doi.org/10.2307/2670164

Perez, B., Molina, I., & Pena, D. (2013). Outlier detection and robust estimation in linear regression models with fixed group effects. *Journal of Statistical Computation and Simulation, 84*(12), 2652-2669. https://doi.org/10.1080/00949655.2013.811669

Víšek, J. Á. (2015). Estimating the model with fixed and random effects by a robust method. *Methodology and Computing in Applied Probability, 17*, 999-1014. https://doi.org/10.1007/s11009-014-9432-5

Wagenvoort, R., & Waldmann, R. (2002). On B-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics, 106*(2), 297-324. https://doi.org/10.1016/S0304-4076(01)00102-6