Al-Nahrain Journal of Science
ANJS, Vol. 27(4), Special Issue – October, 2024, pp. 40 – 46

Detecting Outliers and Using Robust Methods in Linear Panel Data Model

Haneen Namah Jaseem*, Lekaa Ali Mohammad

[1] Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq

Article's Information

Received: 07.07.2024
Accepted: 24.08.2024
Published: 03.10.2024

Keywords:
Fixed effects model
Random effects model
s- method
Weighted Likelihood Estimator method

Abstract

The increasing use of panel data across various fields necessitates robust estimation methods that can resist the influence of outliers, which often lead to biased and ineffective estimates when using traditional methods like least squares. This research investigates two robust estimation techniques within fixed and random effects models for panel data, comparing their performance using the mean square error. Through a simulation experiment with varying sample sizes and contamination levels, the results for the fixed effects model indicate that the Weighted Likelihood Estimator consistently outperformed other methods across all sample sizes at a 10% contamination rate, while the S method excelled at a 20% contamination rate. For the random effects model, the former was most effective with a sample size of 200, while the latter proved superior at a sample size of 800, regardless of contamination levels.

http://doi.org/10.22401/ANJS.27.4.07

* Corresponding author: hanin.nimah2101m@coadec.uobaghdad.edu.iq

1. Introduction

Currently, Panel data are widely used in many fields, including epidemiology, economics, and statistics, due to their ease of analysis. However, the presence of outliers may greatly affect the process of analyzing the results and thus lead to giving biased estimates, and it is difficult to use traditional estimation methods, which prompts us to resort to methods that are not significantly affected by outliers, and among these methods are robust methods [1]. In our research, we dealt with both the fixed and random effects model in the presence of outliers. The fixed effects model has fixed parameters, that is, they do not change during the analysis. While all or some of the parameters of the random effects model are variable. Therefore, abnormal values may be formed as a result of an error in the measurements or the occurrence of an error during the data entry process, and in most cases they are observations that are far from the truth, as they distort parameter estimates and affect the results of the analysis [2].To reduce the effect of outliers. We use robust methods because they are more resistant to outliers than traditional estimation methods, including ordinary least squares. The most important estimation methods are (M), (S), (Huber), trimmed least squares, and others [3] [4]. This research aims to determine the most effective method for estimating the parameters of panel data models (fixed and random) in the presence of extreme values, by conducting a simulation experiment using samples of different sizes and different contamination ratios.

2. Outliers

Outliers in panel data represent those distant observations in the data that deviate from the rest of the data, as their presence leads to a negative impact on the data analysis process [5].It can appear in response variables(y), called outliers, or in explanatory variables(x), called leverage or high inflection points which are called leverage or high inflection points. Its presence in both cases affects the estimation process if it produces ineffective estimates [6].

3. Panel Data Models

Panel data have been widely used in econometrics and statistics because they contain two dimensions, the spatial dimension represented by cross sections and the temporal dimension represented by time series data [7]. Data model for panel is as follows

$$y_{it} = \beta_0 + x'_{it}\beta + \varepsilon_{it} \qquad \dots(1)$$

where
i=1, 2... N represents cross-sectional
t = 1, 2,... T represents the time series
y_{it}: the dependent or response variable
x_{it}: independent variables
β_0: regression constant
β: regression coefficient
ε_{it}: the error term and, it $\varepsilon_{it}^{idd} \sim$(N $(0,\sigma_\varepsilon^2$)

4. Fixed Effects Model:

In a fixed effects model, individual effects are represented by including dummy variables for each individual in the regression equation. This approach allows for unobserved time-invariant variance to be controlled by separating out the individual effects, making it suitable for analyzing the effect of time-varying independent variables on the dependent variable [8] This model assumes that individual effects are correlated with the independent variables but uncorrelated with the error term [9].The fixed effects model has the form:

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it} \qquad \dots(2)$$

$$For\ i = 1,2, \dots\dots N \quad ; \quad t = 1,2, \dots T$$

In this context, the term "individual-specific effect" denoted by (α_i) represents unobserved heterogeneity across individuals, potentially correlated with the regressors. The subscript (i) indicates the individual, ranging from (1 to N). The model includes one independent variable x_{it} and an error term (ε_{it}) with mean zero and variance (σ_ε^2) The coefficient (β) signifies the impact of the independent variable on the dependent variable [10].

4.1. Random Effects Model

The random effects model is a statistical model commonly used in panel data analysis, particularly when dealing with unobserved heterogeneity among individuals or entities. In contrast to fixed effects models, which assume that each individual has a specific constant effect that is correlated with the explanatory variables, random effects models treat these individual-specific effects as random variable [11]. In the random effects model, the individual-specific effects are assumed to be uncorrelated with the explanatory variables, but correlated with the error term ، the effects of individuality cannot be observed directly, however [12].

The random effects model are as follows [10]:

$$y_{it} = \beta_o + x'_{it}\beta + v_i + \varepsilon_{it} \qquad ..(3)$$

where

y_{it} it's dependent variable for individual i at time t.
x_{it} it's explanatory variable for individual i at time t.
v_i represents individual-specific effect, it's be a random variable with mean zero and variance σ_v^2 .
β_o and β it's model parameters.
ε_{it} it's represents the error term, it's follows a normal distribution distributed with mean zero and variance σ_ε^2.

whereas:

$$\alpha_i = \beta_o + v_i$$

In summary, the random effects model treats cross-sections as random variables, which leads to understanding the various changes in the data without ignoring the individual variation of each cross-section, as it can have a significant impact on longitudinal data [13].

5. Estimation Methods

There are several methods for estimating the fixed effects model, the most important and famous of which are the dummy variables method, called the least squares method for the dummy variable (LSDV), and the within-group transformations method. In our research, we will discuss the within-group conversion method because it is used when the number of individuals (sample size) is large [5]. The method for within-group transformations is to remove individual effects by subtracting each observation from its mean. [2].

For model (2) we find the average of each observation over time, holding the term parameter constant [4].

$$\bar{y}_i = \alpha_i + \beta\bar{x}_i + \bar{\varepsilon}_i \qquad \dots(4)$$

where

$$\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$$

$$\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$$

$$\bar{\varepsilon}_i = T^{-1}\sum_{t=1}^{T} \varepsilon_{it}$$

Subtracting equation (2) from equation (4), Hence, the equation becomes

$$\ddot{y}_{it} = \ddot{x}_{it}{}'\beta + \ddot{\varepsilon}_{it} \qquad \ldots(5)$$

Equation (5) represents the transformed model after removing individual effects between groups. Therefore, the parameters of the within-group effects model are as follows [14].

$$\hat{\beta}_{(fe)} = (\ddot{x}'_{it}\ddot{x}_{it})^{-1}(\ddot{x}'_{it}\ddot{y}_{it}) \qquad \ldots.(6)$$

For the random effects model, we most often use the semi-decompositional transformation method to estimate its parameters, as this method consists of a simple adjustment to the data after removing individual effects, and it will be similar to the method of estimating the parameters of the fixed effects model[15].So this method subtracts the time averages with a weight (λ) from each observation. The conversion process is as follows [2].

$$\tilde{y}_{it} = y_{it} - \lambda\overline{y_i}$$

$$\tilde{x}_{it} = x_{it} - \lambda\overline{x_i}$$

$$\lambda = \left[1 - \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_v^2}\right]^{1/2} \qquad \ldots(7)$$

where $0 \leq \lambda \leq 1$

The transformed random effects model is represented by the formula [16].

$$\tilde{y}_{it} = \tilde{x}'_{it}\beta + \tilde{w}_{it} \qquad \ldots(8)$$

Therefore, the parameters of the random fluctuations model become as follows [14].

$$\hat{\beta}_{re} = \quad (\tilde{x}_{it}\tilde{x}'_{it})^{-1}(\tilde{x}'_{it}\tilde{y}_{it}) \qquad \ldots(9)$$

Equation (9) represents the estimation of the parameters of the random effects model by the transformation method.

6. Robust estimation methods

Robust estimation techniques provide alternative approaches to traditional methods such as ordinary least squares (OLS) method, especially when outliers exist within a set of data [17]. The ordinary least squares (OLS) can become unstable in the presence of outliers, using alternative methods that are more resistant to outliers. Robust regression methods aim to enhance stability by reducing the weight or influence of outliers. These methods include different estimators, some of which we will explain here [18].

6.1. S-Estimator

One of the most widely used robust methods, as its working principle is based on minimizing the sum of squares of error. Introduced by(Rossio and Yahoo) in (1984) [19]. It can be known through the following equations:

$$\hat{\beta} = arg\, min_\beta\, \hat{\sigma}_s(e_{1t}, e_{2t}, e_{it}, \ldots\ldots e_{nt}.) \quad \ldots\ldots(10)$$

where the minimum robust estimator for ($\hat{\sigma}_s$) is determined by[20].

$$min \sum_{i=1}^{n} p(\frac{\boldsymbol{y_{it}} - \sum_{i=1}^{n} \boldsymbol{x_{it}\beta}}{\hat{\sigma}_s}) \qquad \ldots.(11)$$

whereas

$$\hat{\sigma}_s = \sqrt{\frac{1}{nK}\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}e_i^2} \qquad \ldots(12)$$

where k is a constant equals to (0.199), and the initial estimate is $w_i = \omega_\sigma = \frac{p(U_i)}{U_i^2}$

$$\hat{\sigma}_s = \frac{\text{median}\,|\text{e}_{\text{it}} - \text{median}(\text{e}_{\text{it}})|}{0.6754}$$

where p is a Tukey's bi weight function .

$$P_{(x)} = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2^{c^2}} + \frac{x^6}{6c^4} & if|x| \leq c \\ \frac{c^2}{6} & if|x| > c \end{cases}$$

where c is a constant it is equal (1.547). Equation number (11) after deriving function (p) is in the following formula [21].

$$\sum_{i=1}^{N} x_{itj}\varphi\left(\frac{y_{it} - \sum_{j=0}^{k} \boldsymbol{X'_{it}\beta}}{\hat{\sigma}_s}\right) \qquad \ldots(13)$$

6.2. Weighted Likelihood Estimator (WLE)

The WL method is considered one of the most important estimation methods used, as it is characterized by giving more efficient estimates and resistance to outliers, as its collapse point reaches 50% [22]. The process of this method begins by giving lower weights to distant observations, which reduces their influence and enhances the accuracy of the estimates. The method is based on the probability density function of errors that follows a normal distribution. The estimator can be found using the following formula [23].

$$\sum_{i=1}^{N} w\left(\varepsilon_{it}(\hat{B}); M_B, \hat{F}_N\right) s_{\sigma_v}(\varepsilon_{it},(\beta);\sigma_v) = 0 \quad \dots (14)$$

The weight function [24]

$$w\left(r_i(\hat{B}); M_B, \hat{F}_N\right) = min\left\{1, \frac{\left[A\left(\delta\left(r_i(\hat{\beta})\right)\right)+1\right]^{+}}{\delta\left(r_i(\hat{\beta})\right)+1}\right\} \quad (15)$$

where $[.]^{+}$indicates the positive integer part.

$$A(\delta) = 2\left[(\delta+1)^{1/2} - 1\right] \quad \dots (16)$$

Represents the residual adjustment function (RAF), used to prune outliers. It takes positive values, $\delta\left(r_i,(\hat{\beta})\right)$ represents the Pearson coefficient of the residuals [25].

$$\delta\left(\varepsilon_{it},(\hat{\beta})\right) = \frac{f^{*}(\varepsilon_{it}(\hat{\beta}))}{m_{\beta}^{*}(\varepsilon_{it},(\hat{\beta});\hat{\sigma}_v)} - 1 \quad \dots (17)$$

$f^{*}\left(\varepsilon_{it}(\hat{\beta})\right) == \int k\left(r_i,(\hat{\beta})\right); t,h) d\hat{F}_N(t\,)$ is a kernel density estimator, and

$$m_{\beta}^{*}(\varepsilon_{it},(\hat{\beta});\hat{\sigma}_v) = \int k\left(r_i,(\hat{\beta})\right); t,h) dM_{\,\beta}(\,t;\hat{\sigma}_v)$$

is the smoothed model density

7. Data Simulation and Comparisons of the Methods

In our research, we simulate data that is closest to the truth, based on two sample sizes and two pollution percentages, and compare robust methods used with ordinary least squares to show the best way to estimate parameters in the case of pollution data for fixed values. And random effects models, using different sample sizes represented by (n=200; 800) cross sections and (t=8) representing time series, assuming different pollution rates (10%, 20%) and experiment repetitions (1000). Once for each sample assuming the number of variables (p = 3). The results were as follows:

Table 1. The mean square error values for all sample sizes and all pollution percentages for the fixed effects model.

Estimates	*Samples*	*outliers*	*Mean square error (mes)*	*Outiers*	*Mean square error (mes)*
Ols	200	10%	1.052355976	20%	1.05623163
S	200	10%	0.991009575	20%	0.986885517
Wle	200	10%	0.984369987	20%	0.996669421
Ols	800	10%	0.997104367	20%	0.995134905
S	800	10%	0.937896683	20%	0.942803375
Wle	800	10%	0.817286877	20%	0.968045225

It is clear from the table that the weighted likelihood estimation (WLE) method showed superior performance when the sample size was 200 and 800, and with a contamination rate of 10%. However, as the contamination rate increased, the (s) method emerged as an optimal choice for sample sizes of 200 and 800. It is clear that as the contamination rate of outliers increases, the effectiveness of the (s) method in estimating fixed effects parameters also increases.

Table 2. The mean square error values for all sample sizes and all pollution percentages for the random effects model

Estimates	*Samples*	*outliers*	*Mean square error (mes)*	*Outlers*	*Mean square error (mes)*
Ols	200	10%	1.0523434	20%	1.053793249
S	200	10%	0.99123491	20%	0.990893789
Wle	200	10%	0.937231471	20%	0.905318845
Ols	800	10%	0.997169063	20%	0.994885686
S	800	10%	0.939095222	20%	0.942490378
Wle	800	10%	0.968986807	20%	0.995916259

From the results of the table above, it appears to us that the (S) method was the most efficient at a sample size of 800 and with contamination rates of 10% and 20%, while the WLE method was the best at a sample size of 200 and with contamination rates of 10% and 20%. It is clear to us that the (S) method is the best with increasing sample size and with different pollution levels when estimating the parameters of the random effects model.

8. Conclusion

I. In the fixed effects model, the WLE method outperformed all sample sizes at (10%) contamination, while the s-method performed better at (20%) contamination across all sample sizes.

II. In the random effects model, the WLE method was most effective at a sample size of (200) and contamination rates of 10% and (20%), while the s-method outperformed at a sample size of (800) for the same contamination rates.

III. In the fixed effects model, the s- method becomes more preferable as the contamination rate increases.

IV. In the random effects model, the s-method is preferable as the sample size increases.

V. We propose alternative robust estimation methods to improve the estimation of data models.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

[1] Rasheed, H.A.; Bahez, Z.K.; "Estimation of a Multiple Linear Regression Model Using Some Robust Methods," Mathematical Statistician and Engineering Applications, 71(4): 4944-4954, 2022.

[2] Nwakuya, M.T.; Biu, E.O.; "Comparative Study of Within-Group and First Difference Fixed Effects Models," American Journal of Mathematics and Statistics 9(4): 177-181, 2019.

[3] Beyaztas, H.B.; Bandyopadhyay, S.; "Robust estimation for linear panel data models," Statistics in medicine 39(29): 4421-4438, 2020.

[4] Beyaztas, B.H.; Bandyopadhyay, S.; "Data driven robust estimation methods for fixed effects panel data models," Journal of Statistical Computation and Simulation, 92(7): 1401-1425, 2022.

[5] Alma, Ö. G.; "Comparison of robust regression methods in linear regression," Int. J. Contemp. Math. Sciences 6(9): 409-421, 2011.

[6] Baltagi, B. H.; "Econometric Analysis of Panel Data". John Wiley&Sons Ltd. West Sussex, England, 2005.

[7] Bălă, R. M.; Prada, E. M.; "Migration and private consumption in Europe: a panel data analysis," Procedia Economics and Finance 10: 141-149, 2014.

[8] Zhang, L.; "The use of panel data models in higher education policy studies," in Higher Education: Handbook of Theory and Research 25: 307-349, 2010.

[9] Bhargava, A.; Franzini, L.; Narendranathan, W.; "Serial correlation and the fixed effects model," The review of economic studies 49(4): 533-549, 1982

[10] Aquaro, M.; Čížek, P.; "One-step robust estimation of fixed-effects panel data models," Computational statistics and Data Analysis 57(1): 536-548, 2013.

[11] Laird, N. M.; Ware, J. H.; "Random-effects models for panel data," Biometrics: 963-974, 1982.

[12] Beyaztas, B. H.; Bandyopadhyay, S.; Mandal, A.; "A robust specification test in linear panel data models," arXiv preprint arXiv:2104.07723, 2021.

[13] Nwakuya, M.T.; Ijomah, M.A.; "Fixed Effects Versus Random Effects Modeling in a Panel Data Analysis: A Consideration of Economic and Political Indicators in Six African Countries," International Journal of Statistics and Applications 7(6): 275-279, 2017.

[14] Bresson et al., "Heteroskedasticity and random coefficient model on panel data," Working Papers ERMES 0601, ERMES, University Paris 2, 2006.

[15] Allison, P.D.; "Fixed Effects Regression Models". SAGE Publications, 2009.

[16] Laird, N. M.; Ware, J. H.; "Random-effects models for panel data," Biometrics: 963-974, 1982.

[17] Bahez, Z.K.; Rasheed, H.A.; "Comparing Some of Robust the Non-Parametric Methods for Semi-Parametric Regression Models Estimation," Journal of Economics and Administrative Sciences 28(132): 105-117, 2022.

[18] Almongy, H.; Almetwaly, E.; "Comparison between methods of robust estimation for reducing the effect of outliers," The Egyptian Journal for Commercial Studies 4: 1-23, 2018.

Al-Nahrain Journal of Science
ANJS, Vol. 27(4), Special Issue – October, 2024, pp. 40 – 46

[19] Susanti, Y.; Pratiwi, H.; Sulistijowati, S.; Liana, T.; "M estimation, S estimation, and MM estimation in robust regression," International Journal of Pure and Applied Mathematics 91(3): 349-360, 2014.

[20] Bramati, M.C.; Croux, C.; "Robust estimators for the fixed effects panel data model," The Econometrics Journal 10(3): 521-540, 2007.

[21] Susanti, Y.; Qona'ah, N. ; Ferawati, K.; Qumillaila, C.; "Prediction modeling of annual parasite incidence (API) of Malaria in Indonesia using Robust regression of M-estimation and S-estimation," in AIP Conference Proceedings 2296(1): November 2020.

[22] Warm, T.A.; "Weighted likelihood estimation of ability in item response theory," Psychometrika 54(3): 427-450, 1989.

[23] Xue, X.; Lu, J.; Zhang, J.; "Item-Weighted Likelihood Method for Measuring Growth in Panel Study With Tests Composed of Both Dichotomous and Polytomous Items," Frontiers in Psychology 12: 580015, 2021.

[24] Agostinelli, C.; Markatou, M.; "A one-step robust estimator for regression based on the weighted likelihood reweighting scheme," Statistics and Probability Letters 37(4): 341-350, 1998.

[25] Agostinelli, C.; "A package for robust statistics using weighted likelihood." Porting R to Darwin/X11 and Mac OS X 1: 32, 2001.