Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 1863 – 1878

Suggested Method for Treatment of Outliers in Panel Data

Muysser Salahaldin Hussein[1], Amir Fadil Tawfiq[2]

Abstract

The research deals with the study of the effect of outliers on the results of estimating panel data models, and aims to find ways to diagnose and treat outliers where the box Plot method is used to detect outliers in all variables (independent and dependent) In addition, to confirm the previous diagnostic method, hat Matrix elements have been used to detect outliers in independent variables as well as Mahalanobis distance, while in the dependent variable the outliers have been detected by the Studentized Deleted Residual, After considering several ways to diagnose outliers, it is necessary to measure their effect on the estimated dependent variable values, model coefficients and standard errors as well as measure the effect on all regression coefficients using some scientific methods including (DFFITS,DFBEATS, COOKS DISTANCE, COVRATIO) ،To eliminate the effect of outliers on panel data models, where used a Suggested method for treatment outliers, efficient and easily applicable outliers treatment method was used to detect the extent to which outliers affect panel data model tests before and after treatment.

Keywords: *Panel data, outliers, Suggested method, Akaki information Criterion, Corrected determination coefficient, Mean square error.*

Introduction

The regression model is one of the most important statistical methods in the prediction process, where the regression model is used in the study of the relationship between a dependent variable which relies on one or more independent variables to explain its behavior ،Where he Establish statistical models to estimate the relationship between variables research gives us a model in the form of an equation that explains the relationship between these variables in terms of predicting future behavior, which helps to plan and make the right decisions. More recently, panel data models have emerged as one of the most important regression models because they take into account the effect of Cross Section Data as well as the effect of time (time series) at the same time. Panel data models are characterized by features that differ if the time data is used alone or the sectional data alone and their features contain greater degrees of freedom and contain more information content and reflect the accuracy of estimates where it gives estimates with higher confidence, and the problem of correlation between variables is less. It is understood that estimation in statistical methods depends on a set of important hypotheses to obtain an accurate panel data model and that the probability distribution of the data is one of the most important hypotheses under study, which is often normal distributed. Sometimes the data takes a different pattern and may not be represented by a certain pattern of distributions and the reason may be due to the presence of outliers in the data studied, which leads to the non-realization of the basic assumptions of the least squares and thus affect their Estimation and make them inaccurate and then lose their properties and good advantages. This

research sought to provide ways to detect outliers and a Suggested Method to address their effect and extent on panel data models and tests.

Many researchers in this field have touched upon them: (Algamal, 2012) The study examined the use of panel data models and the use of three model methods (pooled regression model, fixed effects model, random effects model) and how to choose in reconciling the three models. It depend on two criteria (corrected determination coefficient and Akaki information Criterion) to test the best partial model of the general model by studying four independent variables, and through the Fisher Test and Husman Test, which showed that the appropriate model of study data was the fixed effects model.

Either study (Cousineau & Chartier, 2010) The researchers study aims to review some techniques to detect outliers because their existence aims to distort the results of the study, where these techniques are divided into two categories, the first category specializes in single-variable data and the other category specializes in multi-variable data to show which values are extreme and which is normal non-extremist. As well as study (Alkathemy, 2018) The research aims to study the effect of the presence of outliers on the multiple linear regression model and the accuracy of prediction by studying its effect on the parameters of the model. By employing some methods and tests for the diagnosis and treatment of outliers, the box plot was used to detect outliers, Hat Matrix elements were used to detect outliers in independent variables and the Studentized Deleted Residual method was used to detect outliers in the dependent variable and then several measures were used to detect the effect of outliers including a scale (DFFITS, DFBEATS, COOKS DISTANCE, COVRATIO) and Outliers were treated in two ways (deletion method, and Trimmed Mean) where it depend on some before and after treatment indicators (determination coefficient (R^2), adjusted determination coefficient (R_a^2), statistic (F), mean error absolute and others). The researcher concluded that outliers have a significant effect on the multiple linear regression model and concluded that treatment by the Trimmed Mean method is preferable to the deletion method.

Research objective

The objectives of the research are as follows:

1- purification of study data from outliers through the use of a treatment method Suggested by the researcher.

2- Measure the relative efficiency of the Suggested method of treatment outliers by modeling Panel data.

3- comparison is done before treatment outliers and after treatment in the Suggested method through Panel data modeling .

Panel data models

During the last decade, Panel Data has become widely used in econometric analysis, as well as in many social sciences, including (medical, economics...etc) because it focuses on the effect of change in time as well as the effect of change in cross-sectional observations in It is one, and thus provides a more attractive structure for data analysis. Panel data is defined as cross-sectional observations measured in certain time periods, or as data that combines time-series data with cross-sectional data (Frees, 2007). So that it combines the characteristics of both cross-sections and time-series at the same time. It is used in the analysis of historical events in the sense that it studies cross-sectional data and its movements during a certain period of time. It should be noted that there are panel data called “balanced” if the cross sections do not contain missing observations, then it is called

unbalanced panel data if one or some of the cross sections contain missing observations during the time period and our research agrees with the Balanced Panel Data.

Individual Effects

Panel data is important because it takes into account the (unobserved heterogeneity) of the sample units, whether cross-sectional or temporal. In this study, we will take into account those differences or individual effects for each of the three basic models according to their cross-sections and review them as follows:

1 - If the individual effect is similar in all cross-sections, then the model is a pooled regression model.

2 - If there is a difference in the individual effect across the "ai" cross sections, the model is divided into two basic models:

A- Fixed Effect Model.

b- Random Effect Model.

Basic Models for Analyzing Panel Data

The basic formulation of the panel data models is three models, depending on the difference in individual effects for each cross section. It is assumed that this effect is constant over time and for each cross section. So that we have N cross-sections measured in T of time periods, and the panel data model is according to the following formula:

$$y_{it} = \beta_{0(i)} + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \epsilon_{i,t} \qquad \text{i=1,2,3,…..N} \quad \text{t=1,2,3….T} \qquad (1)$$

y_{it}: represents the value of the dependent variable in observation i at time period t. $\beta_{0(i)}$:represents the point of intersection value of the cross section i. β_j :Represents the value of the slope of the regression l in, $x_{j(i,t)}$:Represent the value of the independent variable j in observation i at the time period t. $\epsilon_{i,t}$:represents the error value of the observation (i) at the time interval (t).

Pooled Regression Model

This form is one of the simplest types of Panel Data models where all transactions ($\beta_{0(i)}$, β_j) are fixed for all cross-sections and for all time periods (i.e. neglecting the influence of time from the model), where the size of observations of the model represents the product of cross-sectional data N multiplied by the time periods T ie it is in the form (N * T) observations. It becomes clear that the pooled regression model represents a multiple linear regression model that has the same assumptions and by rewriting the model in equation (1) we get the pooled regression model according to the following formula:

$$y_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \epsilon_{i,t} \qquad \text{i=1,2,3,…..N} \quad \text{t=1,2,3….T} \qquad (2)$$

$$E(\epsilon_{i,t}) = 0 \cdot var(\epsilon_{i,t}) = \sigma_\epsilon^2$$

The Ordinary Least Squares (OLS) method is used in estimating the model parameters in equation (2) after arranging the values with independent variables and the dependent variable (dependent)

according to (cross-sections or according to time) starting from the first observation until reaching the last observation for all cross-sections, we will get (n*T) observation.

The Ordinary Least Squares (OLS) method provides consistent and efficient estimators for both the constant term and the regression parameters.

Fixed Effect Model

The Individual effect in the fixed effects model is a fixed set of boundaries for each cross section, and the goal is to know the behavior of each cross section individually by making the parameter (β_0) vary from section to another and the stability of the slope coefficients (β_i) For each cross section. That is, we will deal with the case of heterogeneity of variance between sections, and therefore the fixed effects model is in the following form:

$$y_{it} = \beta_{0(i,t)} + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \epsilon_{i,t} \qquad i=1,2,3,\ldots..N \quad t=1,2,3\ldots.T \qquad (3)$$

Thus, we are in the process of a one-way model that contains one set of dummy variables that we take for the vocabulary of the cross sections only, or a two way model that contains two sets of dummy variables that we take for the vocabulary of the cross sections as well as for time. The one-way model is the most common among researchers, which we will focus on a one-way model, i.e. (the first case) and we mean that the parameter (β_0). They change in cross-sectional data sets (countries, establishments, etc.) for the purpose of estimating the parameters of the fixed effects model and often use dummy variables of (N-1) in order to avoid falling into the problem of complete multilinearity between the independent variables, and this study will be based on Least Square Dummy Variable Method. After adding the dummy variables D in equation (3), the model is as follows (Muatee and Belhawisel, 2019) (Greene, 2012):

$$y_{it} = \alpha_1 + \sum_{d=2}^{N} \alpha_d D_d + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \epsilon_{i,t} \qquad (4)$$

$\alpha_1 + \sum_{d=2}^{N} \alpha_d D_d$ represents the change in the cross-sectional sums of the parameter β_0

Random Effect Model

The random effects model assumes random parameter of the cross section and $\beta_{0(i)}$ is treated as part of the error term, similar to ϵ_i except for each section. That is, the cross-sectional and temporal effects are independent random variables with a mean equal to zero and a specific variance σ_ϵ^2. They are added as random components in the random error limit of the model. This indicates that the random effects model is appropriate when the cross-section units are randomly selected from a large population, and therefore the model regression parameters represent the entire population. In fact, the use of dummy variables in the FGM model is an alternative method due to the lack of knowledge of the real model and it is expressed through the amount of error ϵ_i, hence the idea of the random effects model was proposed.

Based on this, the random effects model (REM) is more comprehensive than the fixed effects model (FEM), because it assumes that each cross section within the time effect as a case that differs in its random limit from the rest of the cross sections. Thus, fixed effects are seen as a special case of random effects, because Error Component Model (ECM) Combines the difference within each

section over time periods, as well as the difference between sections (Al-Tamimi, 2020) (Baltagi, 2012) (Hiestand, 2005).

The random effects model (REM) will treat the coefficient $\beta_{0(i)}$ as a random variable that has a rate value "μ"

$$\beta_{0(i)} = \mu + V_i \qquad i=1,2,3,\ldots,N \qquad (5)$$

By replacing equation No. (5) in the main formula of the panel data model, a random effects model (REM) is obtained, as follows:

Random effects (REM) as follows:

$$y_{it} = \mu + V_i + \sum_{j=1}^{k} \beta_j x_{j(i,t)} + \epsilon_{i,t} \quad i=1,2,\ldots,N \quad t=1,2,\ldots,T \qquad (6)$$

Where, μ :indicates the average constant of the units of the cross-sections

V_i: represents the effect of the cross-section i, which is a constant compound over time, $\epsilon_{i,t}$ represents the effect of each of the cross-section (i) and time (t), which is the common error compound between the cross-sections and time.

The random effects model (REM) contains the following mathematical assumptions:

$$E(V_i) = 0 \quad \text{Var}(V_i) = \sigma_v^2 \quad E(\epsilon_{i,t}) = 0 \quad \text{Var}(\epsilon_{i,t}) = \sigma_\epsilon^2$$

The common or compound error of the model is as in the following formula:

$$w_{i,t} = V_i + \epsilon_{i,t} \qquad (7)$$

So the amount of compound error $w_{i,t}$consists of two parts; V_i which represents the error component of the cross-sections or the definition of items. The volume $\epsilon_{i,t}$ represents the error component resulting from merging the time series with the cross-sections. That is, the model is called the Error Components Model because the compound error amount is composed of two or more error components (Gujarati, 2004).

Tests To Determine The Appropriate Model For Panel Data

To determine the best and most suitable model for study data from among the three models, according to the tests to determine the appropriate model. The mechanism used in choosing the best model will be presented in two ways: First: We use the (Fisher) restricted test to compare between the pooled regression model and the fixed effects model, or we use the Lagrange Multiplier Test to compare between the pooled regression model and the random effects model. In the event that the pooled regression model is suitable, we stop moving to the second method.

But in the event that the pooled regression model is not appropriate, we move to the second method: - which is the comparison between the two models of fixed effects and random effects, where we use the test of the best model between them to represent the studied data through the Hausman Test (panchanan, 2019) (Al-Jammal, 2012).) (Al-Tamimi, 2020).

Fisher's restricted statistic test

It is a test that is conducted to identify the fundamental difference between the two models of fixed effects and pooled regression in order to reach the appropriate model, as the hypotheses of this test are formulated as follows:

H_0: $\beta_{0,1}=\beta_{0,2}=\ldots=\beta_{0,N}=0$ A suitable pooled regression model

H_1: $\beta_{0,1}\neq 0$ Fixed effects suitable model

The value of the restricted Fisher statistic is calculated according to the following formula:

$$F=\frac{(R^2_{FEM}-R^2_{PM})/(N-1)}{(1-R^2_{FEM})/(NT-N-K)}\sim F_{(N-1,NT-N-K)} \quad (8)$$

Where N: number of sections T: length of time K: number of estimated parameters.

R^2_{FEM} represents the coefficient of determination when estimating a fixed-effects model. R^2_{PM} represents the coefficient of determination when estimating the pooled regression model.

If the calculated F-statistic value is greater than its tabular value, it would mean that the p-value is less than or equal to the specified significance level (0.05). The null hypothesis is rejected and the alternative hypothesis is accepted, which states that the fixed effects model (FEM) is the appropriate model and vice versa in the case of If the calculated F value is less than its tabular value, the null hypothesis is accepted, i.e. PRM is the appropriate model.

Lagrange Multiplier Test

The researchers devised (Breusch and Pagan, 1980) the Lagrange test, where its statistic includes tracking ($\chi^2_{(1)}$) distribution with one degree of freedom, and the test (LM) depends on the residual estimation of the Ordinary Least Squares (OLS) method. If the hypothesis of this test is as follows:

H_0:$\sigma_\epsilon=0$ A suitable pooled regression model

H_1: $\sigma_\epsilon\neq 0$ The suitable random effects model

The value of the Lagrange multiplier (LM) statistic is calculated according to the following formula:

$$LM=\frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}[\sum_{t=1}^{T}\epsilon_{i,t}]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{i,t}^2}-1\right]^2=\frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}[T^2\overline{\epsilon_i}]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{i,t}^2}-1\right]^2 \quad (9)$$

$$LM=\frac{NT}{2(T-1)}\left[\frac{T^2\bar{\epsilon}\bar{\epsilon}^{\prime}}{\epsilon\epsilon^{\prime}}-1\right]^2\sim\chi^2_{(1)} \quad (10)$$

ϵ residual vector regression of the panel data model between cross sections and time series.

$\bar{\epsilon}$: represents a vector of degree (n×1) for the averages of the residual residuals of the panel data model between cross-sections and time series.

Hausman Test

The idea of this test is based on which of the effects is more suitable for model estimation, whether it is a fixed effects model (FEM) or a random effects model (REM). this is done through the

use of Hasuman test statistic that following the x-distribution with a degree of freedom k, the statistical hypothesis is formulated as follows:

H_0: The random effects suitable model

H_1 The static effects suitable model.

Thus, the formula for the Haussmann test is as follows:

$$H= (\hat{b}_{OLS/FEM} - b_{GLS/REM})'[var(\hat{b}_{OLS/FEM} - b_{GLS/REM})]^{-1}(\hat{b}_{OLS/FEM} - b_{GLS/REM}) \quad (11)$$

$(\hat{b}_{\underline{OLS}} - b_{\underline{GLS}})$ represents the difference between the estimators of fixed effects and random effects.
$\overline{FEM}$ $\overline{REM}$

$var\left(\hat{b}_{\underline{OLS}} - b_{\underline{GLS}}\right)$ represents the difference between the estimators of fixed effects and random effects.
$\overline{FEM}$ $\overline{REM}$

The statistical decision is: If the calculated H statistic value is greater than $\chi^2_{(k)}$ tabular value, meaning that the p-value is less than or equal to the specified significance level (0.05), the null hypothesis is rejected and the alternative hypothesis is accepted, which states that the effects model fixed (FEM). This is the appropriate model, and vice versa. If the calculated F value is less than its tabular value, the null hypothesis is accepted, i.e., the random effects model (REM) is the appropriate model.

Methods for Detecting Outlers Observation

We often find that the data taken from any phenomenon may contain a number of outlier observations, and the outlier observations, as previously mentioned, are a small group of observations whose values are far removed from the rest of the other observations in the sample. Outliers may be in all variables or one of the variables regardless of whether their presence is in the dependent variable or the independent variables. (Kleinbaum, 1988) showed that the presence of outliers in the regression model data is worse in the observations of the dependent variable or the independent variables affects the model parameter estimates and the statistics and associated tests.

In this part, we will focus on the methods of diagnosing outliers and methods of treating them and determining the extent of the effect of outliers on the estimation of panel data models, And we'll go through some of these ways very briefly.

- Box Plot (Tukey Method)
- Hat Matrix Detecting The Outliers in The Independent Variables
- Studentized Deleted Residual Detecting The Outliers in The Dependent Variable

$$d_i^* = e_i \left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{\frac{1}{2}} \sim t_{(n-p-1)} \quad (12)$$

To measure the effect of outliers on estimated values, the following measures were used:

(COVRATIO،COOKS DISTANCE،DFBEATS،DFFITS)

Suggested Method

Assuming the availability of a number of independent variables, let it be k and a dependent variable, if any of the variables include outliers, then to treat them. We suggest an example to illustrate the following method:

In the absence of outlier values in the independent variable (x_1) to be treated in the Suggested Method will be as follows:

$$L=\frac{\log(x_{1i})-\log(x_{min})}{\log(x_{max})-\log(x_{min})} \quad (13)$$

Where, x_i . : represents the variable that has outliers, x_{min} : represents the smallest element in the variable. x_{max} : is the largest component of the variable

Application

Test data moderation (Kolmogorov -Smirnov Test & Shapiro Wilk Test)

The two tests (Kolmogorove-Smirnov testing & Shapiro Wilk test) are used to see if the data follows a normal distribution by testing the data of the variable (y) and the following is explained the decision of the data before treatment outliers

Table 5

The decision to test data moderation before processing outliers

Test of Normality			
Sig	df	statistic	Testing
0.000	55	0.629	Shapiro-Wilk
0.000	55	0.264	Kolmogorov-Smirnov

The table above shows that the probability value p-value is less than the significance level (0.05) in both Tests, so we reject the null hypothesis and accept the alternative hypothesis that the data is not moderate, i.e. the data is affected by the presence of outliers.

Detect outliers in data before processing them

Box Plot (Tukey Method)

Using Box Plot to detect outliers :

It is a box plot that includes the limits of normal values if the values exceed these limits are considered extreme values and this box plot depends in drawing on the Quartiles represented by the median (2^{nd} Quartile) Q_2 and the lower Quartile (1^{st} Quartile) Q_1 and the Upper Quartile (3^{rd} Quartile) Q_3 and is added value to the third Quartile, which is(Q_3-Q_1) Subtract this value from the lower Quartile and thus we are the box institute, and the drawing of both variables can be obtained separately directly from the R program. The figures showed the discovery of an outliers in the independent variables (x_1) and (x_2) as well as in the dependent variable (Y), where the independent variable x_1 includes a number of outliers including (x_{34}, x_2, x_{28}, x_6, x_{31}, x_{27}, x_{30}) and The

Independent Variable x_2 contains outliers (x_{23}, x_{24}, x_{27}, x_{26}, x_{28}, x_{29}) as for the dependent variable also includes outliers (y_2, y_{34}, y_{31}, y_{21}, y_{29},y_{25}).

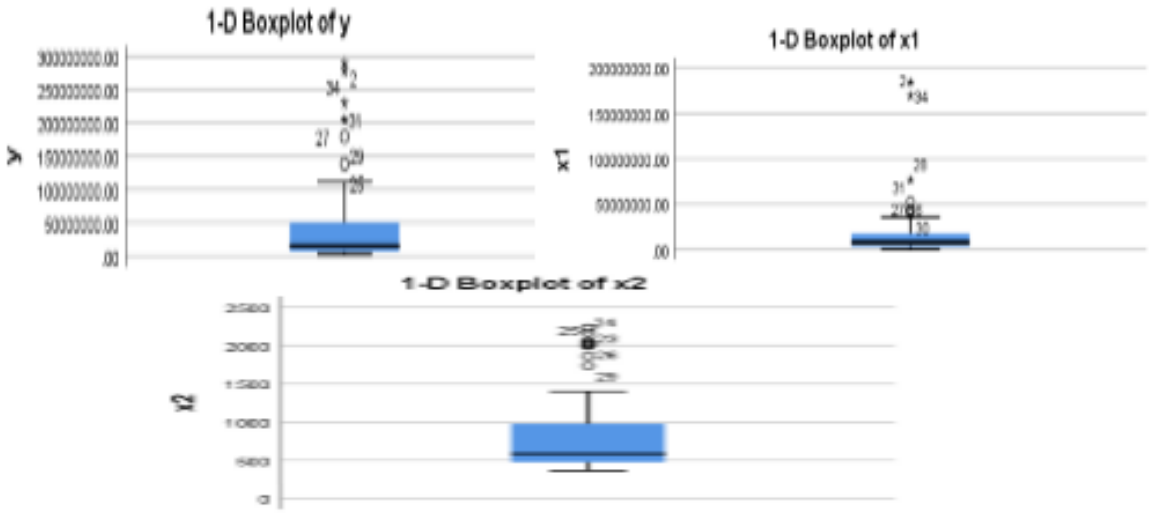

Figure 1 shows the distribution of outliers and extrem value in independent and dependent variables.

Use of Hat Matrix in detecting outliers in independent variables

By the program R the hat Matrix values were then extracted by some calculations by the program R according to the Belsley, John Fox, and Mahalanobis distance estimated their value respectively (0.1455) ،(0.1636) ،(12.6893)، Where differences were found between them and the values of the Hat matrix, it is clear that there are two extreme values of outliers in the independent variables as shown in Table no. (4-9) and (4-10) in the main study tables :thus the outliers in the independent variables are cases 2, 34.

Use Studentized Deleted Residual to detect outliers in the dependent variable

The outliers in the dependent variable are found using the Studentized Deleted Residual of the formula by means of a program R, as we indicated in the main study tables no. (4-11), where the absolute value of the Studentized Deleted Residual was compared with the value of (t), which amounted to (2.0076) by finding the difference between them and then determining the positive value, which represented . outliers case, which is case 27, 28, 29, 31.

Measuring the effect of outliers on estimated values before processing

The effect of these values was studied using the scales of DFFITS, DFBETAS, Cook scale, COVRATIO, and through the application of the R program it was shown that outliers affect the scales as in Table no. (4-12) and (4-13) attached in the main study tables. Also includes results of many outcome metrics 2،23 ،25 ،27 ، 28 ،29 ،31 ،32،34 ،affected by outliers, where we note the depth of the effect to the transactions, especially case 2 which we will be deep effect after treatment later in improvement transactions and various indicators.

Analysis of panel data models before treatment outliers

In this paragraph, what has been mentioned in the theoretical aspect will be applied to estimate the parameters of the three panel data models (Pooled regression model, Fixed effects model, Random

effects model). Based on the R language program, the three panel data models were estimated and the table showing the results of the estimate.

From Table No. (2) the equation of each (estimated Pooled regression model, estimated Fixed effects model, and estimated Random effects model) can be written in the following order:

$$\hat{y}_{it} = -4733700 + 1.4402x_{1(i,t)} + 25338x_{2(i,t)}$$

$$y_{it} = 23440000 + 1.653\, x_{1(i,t)} - 32840\, x_{2(i,t)} - 7679000 \text{ company B} + 87970000 \text{ company C} + 6486000 \text{ company D} + 10520000 \text{ company E}$$

$$\hat{y}_{it} = -4733700 + 1.4402\, x_{1(i,t)} + 25338\, x_{2(i,t)}$$

Tests to determine the appropriate model for panel data

After estimating the three models, the test is carried out which of the three models is the best according to three Tests as shown in the table above. From what has been mentioned in the theoretical aspect, we see that the appropriate model of our studied data is the pooled regression model where we observed through the restricted F test as well as from the Lagrange test that the probability value (p-value) is greater than 0.05, which means that the pooled regression model excels in the F-restricted test on the fixed effect model as well as in the Lagrange test outperforms the random effects model. As for the Husman test, where we see that the probability value (p-value) is less than 0.05, i.e. acceptance of the alternative hypothesis, this indicates that the Fixed effects model is better than the random effects model.

Table 2

shows the results of estimating panel data models before treatment outliers

Statistical indicators of the parameters of the random regression model	**Estimated model**	**Statistical indicators of the parameters of the fixed regression model**	**Estimated model**	**Statistical indicators of the parameters of the pooled regression model**	**Estimated model**
-4.7337e+06	$\hat{\beta}_0$	2.344e+07	$\hat{\beta}_0$	-4.7337e+06	$\widehat{\boldsymbol{\beta}}_0$
8.4852e+06	Std-Error	1.665e+07	Std-Error	8.4852e+06	**Std-Error**
-0.5579	z-value	1.407	t-value	-0.5579	**t-value**
1.4402e+00	$\hat{\beta}_1$	1.653e+00	$\hat{\beta}_1$	1.4402e+00	$\widehat{\boldsymbol{\beta}}_1$
1.5668e-01	Std-Error	1.746e-01	Std-Error	1.5668e-01	**Std-Error**
9.1923	z-value	9.466	t-value	9.1923	**t-value**
2.5338e+04	$\hat{\beta}_2$	-3.284e+04	$\hat{\beta}_2$	2.5338e+04	$\widehat{\boldsymbol{\beta}}_2$
8.8725e+03	Std-Error	2.345e+04	Std-Error	8.8725e+03	**Std-Error**
2.8557	z-value	-1.401	t-value	2.8557	**t-value**
148.654	Chisq	5.831e-14	F	74.3269	**F**
0.74085	R^2	0.7747	R^2	0.74085	$\boldsymbol{R^2}$
0.73088	R^2_{adj}	5856.62	MSE	1.24925e+15	**MSE**
1.24925e+15	MSE	37.65759	AIC	37.65220	**AIC**
		-7.679e+06	factor(Co. B)		
		8.797e+07	factor(Co. C)		
		6.486e+06	factor(Co. D)		
		1.052e+07	factor(Co. E)		

Table (3) shows the test to determine the best model of panel data

Sig	**Test value**	**Type of test**
0.1435	1.8042	**Fisher Test**
0.9083	-1.3303	**Lagrange Multiplier Test**
0.02753	7.1847	**Husman Test**

After estimating the three models, the test is carried out which of the three models is the best according to three Tests as shown in the table above. From what has been mentioned in the theoretical aspect, we see that the appropriate model of our studied data is the pooled regression model where we observed through the restricted F test as well as from the Lagrange test that the probability value (p-value) is greater than 0.05, which means that the pooled regression model

excels in the F-restricted test on the fixed effect model as well as in the Lagrange test outperforms the random effects model. As for the Husman test, where we see that the probability value (p-value) is less than 0.05, i.e. acceptance of the alternative hypothesis, this indicates that the Fixed effects model is better than the random effects model.

Through the above tests we note that the best model is the pooled model and in the Haussmann test that the fixed effects model is better than the model of random effects and that outliers affect panel data models and the difference will be observed after the treatment of outliers and the estimation of panel data models and here we will address three ways to treat outliers.

Treatment of outliers in the Suggested Method

After conducting tests to detect outliers and measure their effect, it was found that all variables in the study data include outliers, which will be treated in the Suggested Method according to the following equation:

$$L=\frac{\log(x_{ii})-\log(x_{min})}{\log(x_{max})-\log(x_{min})}$$

After Treatment the outliers in the statistical program R and cleaning the data from the outliers, where the detection methods of the data will be used in Annex (4) the main study data after treatment as in the first phase, and then the re-estimate and find the best model of panel data.

To detect whether the study data includes extreme value or outliers, refer to the box plot method as follows:

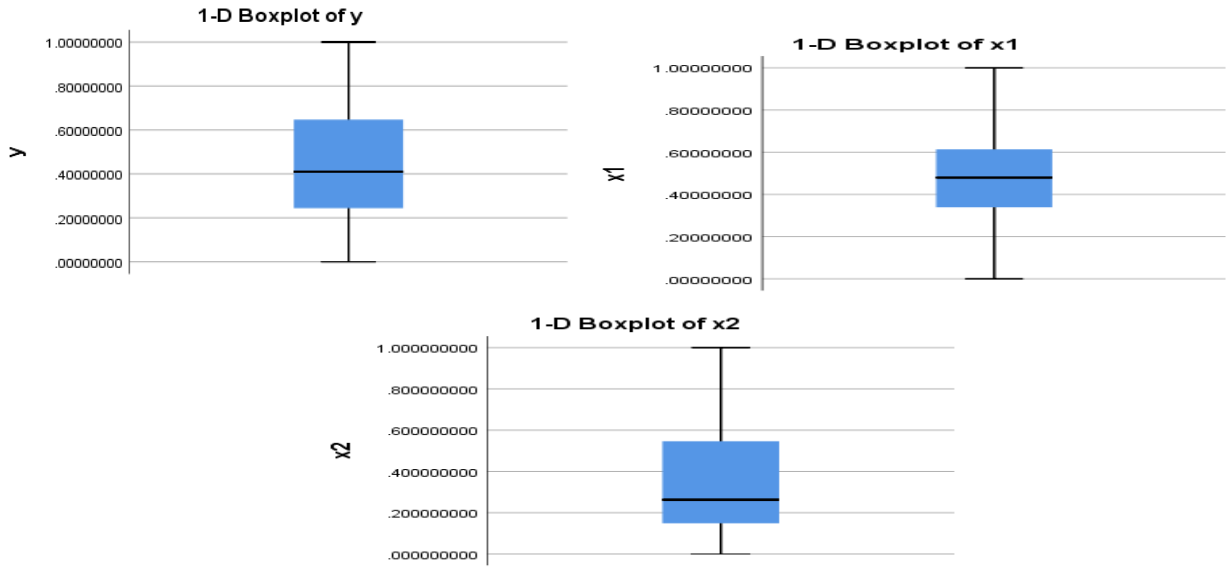

Figure 2 shows that independent variables and dependent variables are free of outliers and after Treatment

Test data moderation (Kolmogorov -Smirnov Test & Shapiro Wilk Test)

From the two tests (The Kolmogorov -Smirnov & Shapiro Wilk) it is clear whether the y variable data follows a normal distribution, and the data decision is explained after the outliers are treated in the Suggested Method.

Table 4

The decision to test data moderation after treatment outliers in the Suggested Method

Test of Normality			
Sig	df	statistic	**Testing**
0.09992	55	0.96415	**Shapiro-Wilk**
0.2	55	0.089	**Kolmogorov-Smirnov**

Seen from the above table (4) that the p-value probability value is higher than the significance level (0.05) in both tests, we accept null hypothesis and reject the alternative hypothesis which states moderating the data to the data distribution normal and free from the influence of outliers.

Analysis of panel data models after treatment outliers

In this paragraph panel data models will be estimated after treatment outliers in the Suggested Method and according to the following.

Table 5

The results of estimating panel data models after treatment outliers

Statistical indicators of the parameters of the random regression model	**Estimated model**	**Statistical indicators of the parameters of the fixed regression model**	**Estimated model**	**Statistical indicators of the parameters of the pooled regression model**	**Estimated model**
-0.00039589	$\hat{\beta}_0$	-0.006409	$\hat{\beta}_0$	-0.00039589	$\widehat{\boldsymbol{\beta}}_0$
0.04263344	Std-Error	0.066113	Std-Error	0.04263344	**Std-Error**
-0.0093	z-value	-0.097	t-value	-0.0093	**t-value**
0.80845051	$\hat{\beta}_1$	0.798096	$\hat{\beta}_1$	0.80845051	$\widehat{\boldsymbol{\beta}}_1$
0.11036204	Std-Error	0.123036	Std-Error	0.11036204	**Std-Error**
7.3254	z-value	6.487	t-value	7.3254	**t-value**
0.18131034	$\hat{\beta}_2$	0.149642	$\hat{\beta}_2$	0.18131034	$\widehat{\boldsymbol{\beta}}_2$
0.07706689	Std-Error	0.144836	Std-Error	0.07706689	**Std-Error**
2.3526	z-value	1.033	t-value	2.3526	**t-value**
156.254	Chisq	24.64	F	78.1272	**F**
0.75031	R^2	0.7549	R^2	0.75031	$\boldsymbol{R^2}$
0.7407	R^2_{adj}	0.3709447	MSE	0.017795	**MSE**
0.017795	MSE	-1.010927	AIC	-1.137941	**AIC**
		0.014127	factor(Co. B)		
		0.051966	factor(Co. C)		
		0.044915	factor(Co. D)		
		0.005283	factor(Co. E)		

From Table (5) the equation of each of the (estimated pooled regression model, estimated fixed effects model, and estimated random effects model) can be written in the following order:

$$\hat{y}_{it} = -0.00039589 + 0.80845051\ x_{1(i,t)} + 0.18131034\ x_{2(i,t)}$$

$$\begin{aligned} y_{it} = {} & -0.006409 + 0.798096\ x_{1(I,t)} + 0.149642 x_{2(I,t)} + 0.014127 \text{company B} \\ & + 0.051966 \text{company C} + 0.044915 \text{company D} + 0.005283 \text{company E} \end{aligned}$$

$$\hat{y}_{it} = -0.00039589 + 0.80845051\ x_{1(i,t)} + 0.18131034\ x_{2(i,t)}$$

Tests to determine the appropriate model for panel data after treatment outliers

Table (6)

The test to determine the best model of panel data

Sig	Test value	Type of test
0.9241	0.22334	**Fisher Test**
0.9121	-1.3538	**Lagrange Multiplier Test**
0.9468	0.10942	**Husman Test**

After estimating the three models, the test is carried out which of the three models is the best according to three Tests as shown in the table above. From what has been mentioned in the theoretical aspect we see that the appropriate model for our studied data is the pooled regression model where we observed through the restricted F test Also from the Lagrange test, the probability value (p-value) is greater than 0.05, which means accepting the null hypothesis, i.e. the pooled regression model outperforms on the fixed effect model in the F-restricted test as well as in the Lagrange test that outperforms the random effects model. As for the Husman test, where we see that the probability value (p-value) is greater than 0.05, i.e. acceptance of the null hypothesis, this indicates that the random effects model is better than the fixed effects model. In the above tests, we note that the best model is the pooled model, and in the Haussmann test, the random effects model is better than the fixed effects model. Before Treatment the outliers, we observed that the fixed effects model is better than the random effects model, meaning that if the percentage of outliers in the data studied increases, it may change the method of panel data models as a whole.

Comparison of the results of the estimated models to choose the best model according to the pooled model

After the best model of panel data was selected according to the Tests mentioned earlier and it turned out that the pooled regression model is better than other models where its results are relied on in the comparison before and after treatment, the following table shows the comparison of statistical indicators before and after treatment and shows the extent to which the strength of statistical indicators improved after treatment outliers in the Suggested Method.

Table (7)

Comparison of statistical indicators before and after treatment

After treatment	Before treatment	Statistical indicators
0.75031	0.74085	R^2
0.017795	1.24925e+15	**MSE**
-1.137941	37.65220	**AIC**

The table above (7) shows that there is a clear effect of outliers on the model parameter estimates and on the criteria, and to compare before and after treatment results in the Suggested Method and determine the extent of the outliers effect, three criteria are used (determination coefficient (R^2), Mean Square Erorr MSE, and Akaike Information Criterion AIC).

The results in the table above indicate that the value of the determination coefficient has increased to R^2 = 0.75031 after treatment the Suggested Method, meaning that the independent variables included in the model explain 75% of the changes in the dependent variable (y) and the remaining 25% are due to other factors, including random error, For the MSE we note their decreased after treatment where their value MSE= 0.017795, Akaki's Information Criterion was AIC= -1.137941, which represents the least missing information for the model. From the above, we conclude that outliers have significantly influenced the estimation of panel data models.

Conclusions and recommendations

1- Outliers affect panel data models substantially by reducing the value of the determination coefficient (R^2) and amplifying the value of the mean squares error MSE and the AIC Akaki's Information Criterion.
2- the presence of outliers in certain proportions affects the Haussmann test, so if they increase by a greater percentage it may affect all tests.
3- The proposed method proved highly efficient in treatment outliers.
4- The proposed method showed a significant improvement in the indicators of panel data models, which indicates that panel data models were significantly affected by outliers.
5- The need to diagnose outliers before conducting any study on the data under study.
6- Apply new methods of detection and treatment of outliers to other statistical models.
7- We recommend using the proposed method of treatment outliers for panel models.

References

1. Algamal, Y Zakariya. (2012). Selecting Model in Fixed and Random Panel Data Models. IRAOI JOURNAL OF STATISTICAL SCIENCES, 12(21).
2. Alkathemy, abdulaziz saad, (2018), “Breach of the assumption that data are free of outliers and its effect on the use of multiple linear regression analysis in forecasting”, Comprehensive Electronic Journal, No (5).

3. AL-Tameemi. Hazim, (2020) , “Detection and Measurement the Effect of the Outliers on the Estimation of the Panel Data Model with Application”, Unpublished master's thesis, Benha University, Faculty of commerce.
4. Baltagi, B. H. (2005). Econometric analysis of panel data 3rd Edition England JW &Sons.
5. Cousineau, D., & Chartier, S. (2010). “Outliers detection and treatment: a review”. International Journal of Psychological Research, 3(1), PP58-67.
6. Frees. A, Kim, (2007) , “ longitudinal and panel data ”, University of Wisconsin Madison.
7. Greene, W., H., (2012). “Econometric Analysis” 7th ed.,Pearson Education, Inc.,NJ.
8. Hiestand, T. (2005). Using pooled model, random model and fixed model multiple regression to measure foreign direct investment in Taiwan. International Business & Economics Research Journal (IBER), Vol.4, No.12.
9. Kleinbaum, J, et al, (1988),"Applied Regression Analysis and other Multivariable Methods", PWS-Kent Publishing Company, Boston, second edition, p:210.
10. Muatee, S., & Balhuwaisl, M., (2019), “Using Panel Data in Modeling the Fluctuation of Foreign Trade Variables with Economic Groth in Yemen during (2006 – 2013)”, Al Rayyan Journal of Humanities and Applied Sciences, 2, No (1).