Estimation of a Multiple Linear Regression Model Using Some Robust Methods

Husam Abdulrazzak Rasheed[1], Zahraa Kh. Bahez[2]

husamstat@uomustansiriyah.edu.iq, zahraakhalidbahez@gmail.com

Mustansiriyah University/ College Of Management And Economic/statistics department

[1]Corresponding Author

Husam Abdulrazzak Rasheed

Assistant Professor Doctor in Statistics

E-mail: husamstat@uomustansiriyah.edu.iq

Telephone number: +9647901181120

Article Info

Page Number: 4944 - 4954

Publication Issue:

Vol 71 No. 4 (2022)

Article History

***Article Received**: 25 March 2022*

***Revised**: 30 April 2022*

***Accepted**: 15 June 2022*

***Publication**: 19 August 2022*

Abstract

The multiple linear regression modelsare one of the important models in regression because it is used in analyzing a lot of data in various economic, medical and social fields. The relationship between the dependent variable and the interpreted variables in the form of an equation by estimating its parameters, we infer the strength and importance of this relationship. Inefficient and gives biased capabilities in the event that one of its basic hypotheses is not available. Therefore, in this research, robust methods were used instead of them because of the problem of outliers, which are observations that go out of the data pattern and that have a significant impact on the non-fulfillment of the hypothesis of a normal distribution, and this contradicts one of the basic assumptions on which multiple linear regression is based. One of the most important methods used in this research is the MM-Estimation, S-Estimation and M-Estimation method, through an applied study of a data set to study the impact of exchange rates and oil on gold prices. It was found through the results that the exchange rate variable had a more significant effect on the dependent variable than the oil price, except in the S method. It was found that the exchange rate variable did not have a significant effect on the price of gold, and based on the MSE comparison standard, it was found that the best method is the M method, followed by the S method and then the MM method.

Keywords: Regression; Robustness; Outliers;Estimations.

1- introduction:

One of the most basic objectives of studying any scientific problem or phenomenon is to find the basic equation that it represents to understand that phenomenon and determine its parameters, which is known in statistics as the modeling of phenomena.

And that this stage developed with the development of statistics, where it began long ago with one type, which is the parametric regression of its two types, linear and non-linear, and then the non-parametric and semi-parametric regression appeared .It is one of the most commonly used statistical methods in all fields of science such as business, medicine, economics and others. These parameters of this model can be estimated by traditional methods such as the OLs method, but in the event of a problem such as the problem of outliers, which makes its estimates inefficient and more efficient alternative methods must be sought.

2- Multiple Linear Regression Model

The multiple linear regression model shows the relationship between the explanatory variables and the dependent variable in the form of a mathematical equation, and the equation that includes one variable is called simple linear regression, but the equation that contains more than one variable is called multiple linear regression and the relationship can be explained as follows[6]

$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + \cdots B_k X_{ik} + u_i$,i=1, 2,… , n…(1)

Y_i :represents the dependent variable

$B_1 B_0$,:It represents the parameters of the model to be estimated

$X_1 X_{2,..} X_k$:represent explanatory variables

It can be written in matrix form

$\underline{Y} = XB + U \ldots (2)$

3- Outlier Observation

Anomalies are observations that appear inconsistent with the rest of the data, that is, they are outside the structure of the concerned group of observations. In many cases, such influential points remain hidden to the user because they do not always appear in the usual way. In 1986 (Hampel Ronchetti, Roussen & Stahel) defined them. The pattern followed by the rest of the data. The reason for this difference is due to many factors, including (sampling error, measurement error, recording error(Anomalies can be defined as those data that do not fit with the rest of the data and it is sometimes called the maximum value, and when we draw a chart, it will appear that it does not fit the pattern of the chart. Among the problems that are addressed when there are anomalous observations in the regression model data are the ones that can be classified as follows

1-The problem of the presence of outliers in the response variable (y), or it is called the dependent variable.

2-The problem of the existence of anomalous observations in the explanatory variables (x-space) and in this case it is called ((Leverage points), and there are two types, the first is (Bad Leverage), and it is understood as anomalous observations that are far from the pattern of the values of the explanatory variables (x-space) It is inconsistent with it, as it moves away from the estimated regression line, and it can be defined as those values that move away from the data center of the explanatory variables and affect the estimation of the dependent variable. The second is called (good Leverage) and its meaning is those values that are far or outside the pattern of special values with explanatory variables, but they remain close to the estimated regression line, or in other words, they do not affect the estimation of the dependent variable (y)

3-The problem of the presence of outliers in the response variable (y) as well as their presence in the explanatory variables (Leverage) together.

The most common and used methods for detecting outliers will be reviewed, as follows

1-3.Box-Whisker-Plot Method

The box graph method is one of the simplest statistical methods that are widely used in detecting outliers. The box graph is based on the interquartile range, as the quartiles can be used to form another measure of variances based on the second quartile or half of the interquartile range, which measures 50% of the average deviations of the observations.

These quartiles are used in the box graph to detect the anomalous values, and the box graph is characterized by its simple shape and is not affected by the anomalous values.Any value in the data greater than (U) or less than (L) is Outlier Observation.

$\mathrm{U} = \mathrm{Q}_3 + 1.5(\mathrm{Q}_3 - \mathrm{Q}_1) = (\mathrm{Q}_3 + 1.5 * \mathrm{IQR})\ldots$ (3)

$\mathrm{L} = \mathrm{Q}_1 + 1.5\ (\mathrm{Q}_3 - \mathrm{Q}_1) = (\mathrm{Q}_1 - 1.5 * \mathrm{IQR})\ldots$ (4)

This can be illustrated by the figure below for this method

The three vertical lines of the square refer to the three quartiles and the lines extending from the right and left are called mustaches, so that any value outside these mustaches is considered an anomalous value, as shown in the figure as

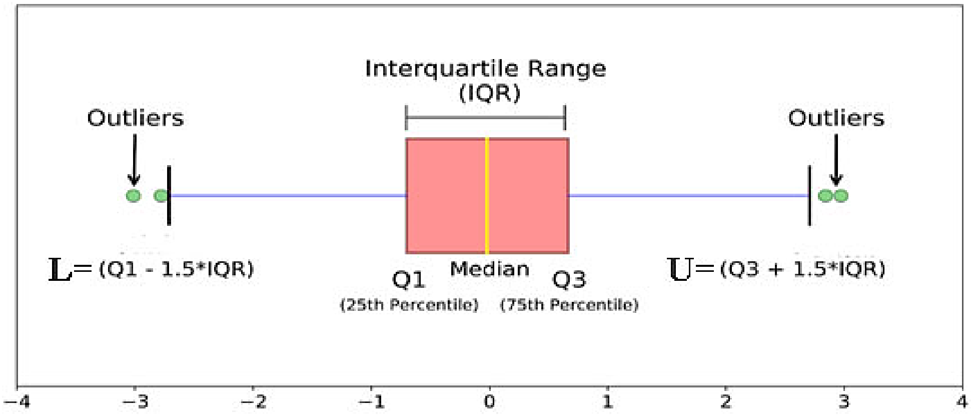

Figure (1) Outliers and Box Graph

4-Robust regression Estimation

4.1- MM Estimator.

The MM method is one of the most common and widely used estimators. It was proposed by the researcher (Yohi) in 1987. It is considered to be highly efficient in the case of a normal distribution of errors and has a high breakdown point. Its estimators can be calculated according to the following steps: [1] [3]

1-Initial values are imposed for the parameters we denote$\hat{B}_T$

2-Calculating the residual values $\mathrm{r_i}(\widehat{\mathrm{B}}_\mathrm{T}) = \mathrm{Y_i} - \mathrm{X_i}\widehat{\mathrm{B}}_\mathrm{T}$

3-Calculate the value of weights (wi) from the following formula

$$w_i = \begin{cases} \left(1 - \left(\frac{e_i/\hat{\sigma}_i}{4.685}\right)^2\right)^2 & , \ |e_i| \le 4.685 \\ 0 & , \ |e_i| > 4.685 \end{cases} \quad , \ \mathit{Tukey's\ Bisquar} \ldots (5)$$

or

$$w_i = \begin{cases} 1 & , \ |e_i| < 1.345 \\ \frac{1.345}{\left|\frac{e_i}{\hat{\sigma}_i}\right|} & , \ |e_i| \ge 1.345 \end{cases} \quad , \quad \mathit{Huber} \ldots (6)$$

4-Calculation of $(\hat{\beta}_{MM})$according to (WLS) method

5-Repeat steps (2-4) to reach close values of $(\hat{\beta}_{MM})$

4.2- M Estimator

It is an alternative hippocampal regression estimator to the method of least squares and an extension of the (MLE) method. It is widely used and sometimes called (Huber's estimates) and it is considered as strong estimates against the anomalous values in the dependent vector (Y), but it is not strong against the anomalous values in the vector (X) because of the vulnerability to points the influence.

The estimators of this method are unbiased and have the least variance, and the principle of estimating M is to reduce the residual errors

$$\tilde{\beta}_\mu = \min\rho\left(\frac{\mathrm{ei}}{\mathrm{s}}\right) \ldots (7)$$

This method is considered one of the most widespread methods of immunity and it replaces the square of the residuals $\sum ei^2$ used in the OLS method with the objective function of the residuals, while maintaining the same goal, which is to make the amount as small as possible.

The steps for obtaining the capabilities of the M method are as follows: **[1] [4]**

1-Assuming initial values for the parameters

2-Find the value of the residual errors ei

3-find value$\hat{\sigma}_M = \frac{MAD}{0.6745}$

4-Calculate the value $u_i \quad = \frac{e_i}{\hat{\sigma}_M}$

5-Calculate the weights according to the formula

$$\mathrm{w}_i = \begin{cases} \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2 & , \ |u_i \ | \leq c \\ 0 & , \ |u_i \ | > c \end{cases} \dots (8)$$

6-Finding parameters according to the WLS method

Repeat steps (2-6) to get close B_M values

4.2-S Estimator

This method is called S-estimation because it depends on the estimation of the scale of errors. The method of least squares was generalized by (Rousseeuw & yahai) to provide this new category of estimation in the framework of (S-estimation) and this method reduces the total errors to the lowest They are very robust against anomalies in the data. These estimators have similar properties to M estimators and have a high breakdown rate

Its capabilities can be obtained according to the following steps:[4]

1-Establishing initial estimates for parameters such as estimates of least squares OLS.

2-Find the value of the remaining errore_i

3-Finding a value$\hat{\sigma} = \begin{cases} \frac{MAD}{0.6745} \\ \sqrt{\frac{\sum_{i=1}^{n} \omega_i e^2{}_{i.}}{nK}} \end{cases}$

4-Finding a value$u_i = \frac{e_i}{\tilde{\sigma}}$

5-Finding the weighted values according to the formula

$$\begin{cases} \begin{cases} \left[1-\left(\frac{u_i}{c}\right)^2\right]^2 & , |u_i \ | \leq 1.547 \\ 0 & , \ |u_i \ | > 1.547 \end{cases} & iteration = 1 \\ \frac{P(u_i)}{(u_i)\text{^}2} & teration > 1 \end{cases}$$

6-Calculation of $\tilde{\beta}_s$according to the WLS formula

Repeat the steps to obtain similar capabilities

5. Applied Example

A set of monthly data was analyzed during the period (2007_2017) related to oil prices, exchange rates and gold prices. Data were collected from the Statistical Center and https://m.sa.investing.com/commodities/brent-oil-historical-data

Depending on the (Eviews) program, the gold price variable was considered the dependent variable (y), the exchange rate was the explanatory variable x1, and the oil price was the explanatory variable x2, with a sample size of (169)

The data was tested in terms of the problem of the presence of anomalous values using the box diagram and my agency

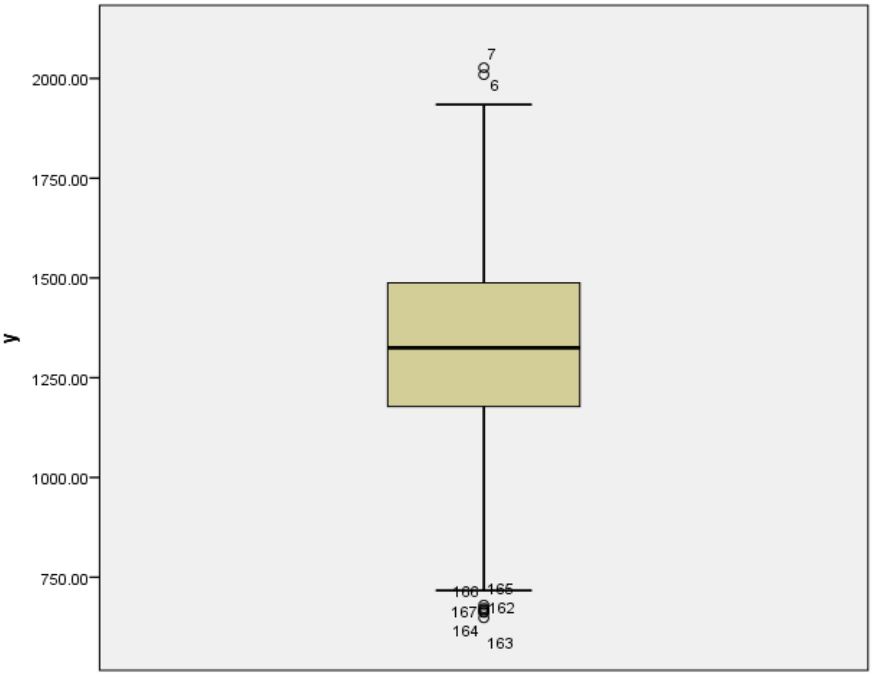

Figure (2) represents the presence of outliers in the explanatory variable Y

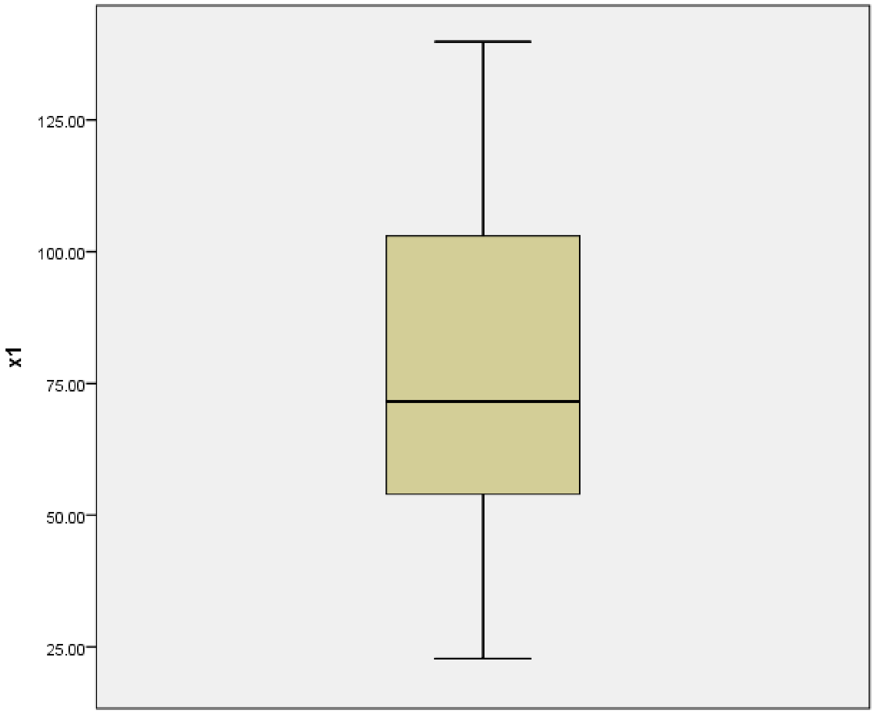

Figure (3) represents the presence of outliers in the explanatory variable X1

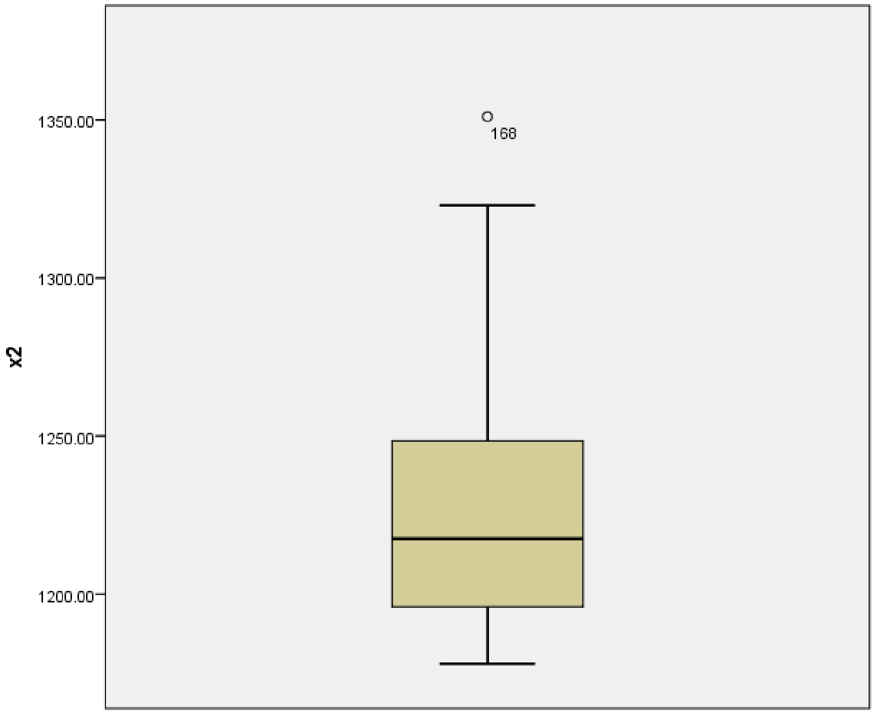

Figure (4) represents the presence of outliers in the explanatory variable x2

It was found through the box diagram that there are six outliers in the dependent variable and one outlier in the second explanatory variable

Table (1) Regression model estimation using robust methods

estimation methods	Parameters	estimating parameters	S.E.	t-value	P-value	R^2	MSE
M	β_0	6.812	0.065	104.76	0.0000	9.82	00738100.
	β_1	0.034	0.008	4.032	0.0001		
	β_2	0.013	0.006	2.297	0.0216		
S	β_0	7.011	0.048	146.132	0.0000	9.55	10660.0
	β_1	0.002	0.006	0.320	0.7490		
	β_2	0.016	0.004	3.731	0.0002		
MM	β_0	6.833	0.068	100.429	0.0000	7.85	74150.00
	β_1	0.031	0.009	3.543	0.0004		
	β_2	0.013	0.006	2.071	0.0384		

It is clear from the above table that the best method for estimating the multiple regression models under the problem of the presence of outliers is the M Estimation. Relying on the comparison criterion MSE, followed by the S Estimation, then the MM Estimation, and it was found that all the coefficients were statistically significant because the value of prob. is less than 0.05

According to the t-test, this means that each independent variable has a significant effect on the dependent variable except for the value of parameter B1 in the S Estimation is not significant.

Multiple regression model estimation according to M. Estimation

$$\hat{Y}_i = 6.812 + 0.034X_1 + 0.013X_2$$

Multiple regression model estimation according to S. Estimation

$$\hat{Y}_i = 7.011 + 0.002X_1 + 0.016X_2$$

Multiple regression model estimation according to MM. Estimation

$$\hat{Y}_i = 6.833 + 0.031X_1 + 0.013X_2$$

6.Conclusions

After studying the multiple regression models using the appropriate methods, the study concluded the following:

1- There is a significant effect by the explanatory variables on the dependent variable in the M and MM Estimation. The X1 variable had a significant effect on the dependent variable more than the X2 variable, except for the X1 variable in the S method, which did not have a significant effect.

2- The best way to estimate the model is the M Estimation, followed by the S Estimation, then the MM Estimation, based on the MSE comparison criterion.

7.Recommendations

Based on the results reached by the researcher, the following can be recommended:

1- Using the M Estimation in estimating the model because it gives the best estimates and the lowest MSE value.

2- Applying robust methods not used in the study such as LTS method and LMS method and method in estimating the multiple linear regression model

Acknowledgments: -

We would like to thank mustansiriyah university (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

References

1. Alma, Özlem Gürünlü. "Comparison of robust regression methods in linear regression." Int. J. Contemp. Math. Sciences 6.9 (2011).
2. Chen, Colin. "Paper 265-27 robust regression and outlier detection with the robustreg procedure." Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference (2002).
3. Filzmoser, Peter, and Klaus Nordhausen. "Robust linear regression for high-dimensional data: An overview." Wiley Interdisciplinary Reviews: Computational Statistics 13.4 (2021).

4. Huber, Peter J. "John W. Tukey's contributions to robust statistics." Annals of statistics (2002).
5. Lukman, Adewale F., et al. "Two-parameter modified ridge-type m-estimator for linear regression model." The Scientific World Journal (2020).
6. Tranmer, Mark; Elliot, Mark. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 2008.