ERURJ, 2024, 3,4, 1888-1902

Review

The Performance of Robust Regression Estimators in Presence of Outliers

Amer Ibrahim Amer[1*], Rowida Ali Abd Albary [1], Salem Adel Salem[1], Ahmed Ibrahim Abd-Elftah[1]

[1]*Department of Business Technology, Faculty of Management, Economics, and Business Technology, Egyptian Russian University, Badr City, Cairo, Egypt.*

*Corresponding author(s): Amer Ibrahim Amer, E-mail: amer-ibrahim@eru.edu.eg, Tel: +201118378256

Received 25th March 2024, Revised 27th April 2024, Accepted 1st May 2024

DOI:10.21608/erurj.2024.279277.1137

ABSTRACT

Linear regression models are common, powerful statistical methods that are used to build a model between dependent variable and one or more independent variables to explain and validate the relationship between the dependent variable and the independent variables the parameters of the linear regression model are unknown. Estimators are derived to estimate those parameters. Ordinary Least Square (OLS) is one of the most common estimates for the linear regression parameters since its best linear unbiased estimators (BLUE) under certain assumptions. The occurrence of outliers in the data leads OLS to have a poor fit and misleading results. Robust estimates are designed to handle the presence of outliers by different methods, among many robust estimates which developed across the years, the most common and efficient estimates are discussed. The results indicate that among the different estimates MM estimate had superiority over OLS and other robust estimates, leading to the conclusion that the presence of outliers could lead to many consequences, checking for their presence and handling them appropriately is a most for efficient fitting.

Keywords: linear regression, ordinary least squares (OLS), M estimate, S estimate, MM estimate.

1-Introduction

Linear regression models are common, powerful statistical methods that used to build a model between dependent variable and one or more independent variables to explain and validate the relationship between the dependent variable and the independent variables the parameters of the linear regression model are unknown. Many estimators are derived to estimate those parameters. Ordinary Least Square (OLS) is one of the most common estimates for the linear regression parameters since its Best Linear Unbiased Estimator (BLUE) under certain assumptions. The OLS relies on assumptions to be BLUE, such as are linear it assumes that the relationship between the dependent variable and the independent variables must be linear.

$$Y = X\beta + \varepsilon \tag{1.1}$$

Where Y the dependent variable $(n \times 1)$ vector, X is $(n \times k)$ matrix of the independent variables, β is $(k \times 1)$ vector of coefficients, and ε is $(n \times 1)$ vector of errors.

zero mean OLS assumes that the error term have mean of zero $E(\varepsilon) = 0$, Homoscedasticity assumed that the variance of the error term across all the values of the independent variables must be equal $E(\varepsilon\varepsilon') = \sigma^2 I_n$, normality the distribution of the error term must be normal with zero mean and one standard deviation $\varepsilon \sim N(0,1)$, exogeneity (independent) the error term are independent from the independent variables no correlation between them, no autocorrelation the error term observations are independent from each other's, and no multicollinearity assumed that the independent variables are not perfectly correlated of each other. When these assumptions are achieved the OLS will be best linear unbiased estimator (BLUE), best indicating to the minimization criteria of the OLS which make the OLS the most efficient estimator among all linear unbiased estimators, linear as the coefficients estimated through OLS are naturally linear, unbiased OLS estimators are unbiased since they are on average accurate estimates of the true population parameter, and estimator since the OLS is estimation method for the linear regression unknown population parameters.

The formula of OLS estimator in matrix form

$$\hat{\beta} = (X'X)^{-1}(X'Y). \tag{1.2}$$

Where $\hat{\beta}$ is the estimated parameter vector, X' is the transpose of the independent variables matrix, $(X'X)^{-1}$ represents the inverse of the product of the matrix of the independent variables and its transpose.

The violation of OLS assumptions leads to many consequences biased estimates, inefficient estimates, invalid hypotheses tests, inaccurate predictions, inflated or deflated standard errors, and incorrect inferences. The presence of outliers in the dependent or independent variables violates several OLS assumptions, such as linearity, error zero mean, normality of errors, homoscedasticity, and independence of errors. Robust regression is one of the best remedies when handling violations that include the presence of outliers since it has high resistance against outliers. Unlike classical methods, this chapter will discuss different types of outliers in regression, detection techniques, and remedies while focusing on robust regression.

2. Methodology

2.1 Outliers

As Barnett and Lewis (1) defined, outliers are observations that appear inconsistent with the rest of the data set. They can have a strong influence on the regression analysis. There are various types of outliers in regression. It is common to classify them as follows:

- **Residual outlier:** A point that has a large standardized or studentized residual when it is used in the sample of n observation to fit the model.
- **X-space outlier:** An observation that is remote in one or more x coordinates, some types of robust regression techniques are ineffective against that type of outlier, also referred to as leverage points.
- **Y-space outl1ier:** An observation that is unusual in its y coordinate. The effect that the observation has on the regression model depends on its x coordinate, and one of the general dispositions of the other observations in the sample.
- **X-space and Y-space outlier:** An outlying observation in both x and y coordinates. This type of point depends on the disposition of the other sample observations.

Figure 1 shows the different types of outliers. The ellipse defines most of the data. Points A, B, and C are outliers in Y-1space since their y value is signif1icantly dif1ferent from the rest of the data and residual outliers. Points B, C, and D are outliers in X-space since their x value is unusual, also referred to as leverage points. D is an outlier in X-space but not a residual outlier. Points B and C are leverage points and residual outliers. Point A is an inlier in X-space but a residual outlier. Point E is an inlier in Y-space but a residual outlier.

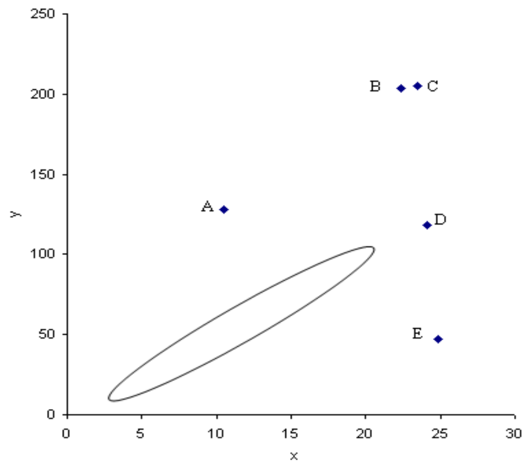

Figure 1: Scatter plot for the different type of outlying observations**.**

Source*:* (2)

Classic estimation methods for regression can be misleading in many cases as the presence of outliers in the data, violation in the regression assumptions and the error distribution is not normal. One of the most used estimation methods is ordinary least squares (OLS). However, this method gives misleading and insignificant results in the presence of outliers, on the other hand, the robust estimators can achieve the desirable characteristics for the estimators when there are violations in the assumptions. As follows a more complete review of existing robust estimators.

2.2. Evaluating Regression Estimators

Evaluation of regression estimators can be done by using many different criteria. Some of them are more significant than others.

2.2.1 The Breakdown Point

The breakdown point (BDP), Donoho & Huber (2) introduced one of the most important quantitative characteristics of robustness. Breakdown point is defined as "The minimum proportion of observations in data set that need to be changed to make the resulting arbitrarily far from the estimate based on the original data" (3). BDP is a global assessment robustness estimator, where the high BDP point methods are considered as one primary goal in most research in robustness. There are two types of BDP: Addition breakdown point (ABDP) and replacement breakdown point (RBDP).

1- ABDP of an estimator is defined as "The minimum addition fraction which could drive the estimator beyond any bound."

Let $(X^n) = \{X_1, \dots, X_n\}$ be a sample of size n. the finite sample ABDP of an estimator θ at X^n is defended as:

$$ABDP(\theta, X^n) = \min\left\{\frac{m}{m+n} : sup_{Y^m}\|\theta(X^n \cup Y^m) - \theta(X^n)\| \sim \infty\right\}, \quad (2.1)$$

Where Y^m denotes the dataset from size m with arbitrary values, and $X^n \cup Y^m$ denotes the contaminated sample by adjoining Y^m to X^n.

2- RBDP of an estimator θ is defined as “the minimum replacement fraction which could drive the estimator beyond any bound.”

The finite RBDP of an estimator at X^n is formulated as

$$RBDP(\theta, X^n) = min\left\{\frac{m}{n} : \sup_{X_m{}^n}\|\theta(X_m{}^n) - \theta(X^n)\| \sim \infty\right\}, \quad (2.2)$$

Where $X_m{}^n$ denotes the contaminated sample from X^n by replacing m points of X^n with arbitrary values, sup denotes supremum function, where supremum is taken over all the samples, and || || denotes the norm.

2.2.2 Influence function

Influence function (IF) is a critical quantitative assessment of robustness Hampel (4) defined IF as for sample $z = (z_1, \dots, z_n)$.

$$IF(z, T, F) = \lim_{\in \to 0}\left(\frac{T(F_\in) - T(F)}{\in}\right), \quad (2.3)$$

where T(F) is a functional that defines the estimator $T(F^{(n)})$, $F^{(n)}$ is the empirical distribution function $F_\in = (1-\in)F + \in \Delta_z$, and Δ_z is the distribution that puts all its mass at z. The IF measures the effect on the estimate of an infinitesimal contamination at point z, standardized by the amount of contamination.

There are properties of an influence function which grant it with desirable performance are (4):

- Gross-error sensitivity (G.E.S) $G.E.S = \sup_{x \in X}|IF(x; F, T)|$ Where G.E.S represent the highest value of the influence function.

- Local shift sensitivity (L.S.S) $L.S.S = \sup_{x \neq y} \frac{|IF(x)-IF(y)|}{|x-y|}$ where (L.S.S) is the highest possible effect that can happen from adjusting the observations from the values of the estimated parameter.
- Rejection points $\rho^* = \rho^*(T, F) = \inf\{r > 0 : IF(x: T, F) = 0, \forall |x| > r\}$ these measures how large an observation must be before the estimator ignores it completely. If sizeable observations are almost gross errors, it is good for ρ^* to be finite.

2.3. Robust Regression Estimators

2.3.1 M-Estimator

M-estimators are one of the most common methods of robust regression (5). This class estimators can be regarded as a generalization of maximum-likelihood estimation. M-estimators are solutions of normal equation with appropriate weight function; they are resistant to outliers in Y coordinate, but sensitive to outliers in X coordinate (leverage points) M-estimate defined by replacing the least-square criteria $\sum_{i=1}^{n}(y_i - x_i^t \beta)^2$ with robust criteria:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{\rho}\left(\frac{y_i - x_i^t \boldsymbol{\beta}}{\hat{\boldsymbol{\sigma}}}\right), \tag{2.4}$$

Where $\rho(\cdot)$ is a robust loss function, $\hat{\sigma}$ (median absolute deviation) is an error scale estimate, differentiating the objective function and setting the derivative to 0 produce:

$$\sum_{i=1}^{n} \boldsymbol{\psi}\,(\boldsymbol{y_i} - \boldsymbol{x_i^t b})\boldsymbol{x_i} = \mathbf{0}, \tag{2.5}$$

The BDP of M-estimate is $\frac{1}{n}$ and IF is:

$$\boldsymbol{\psi_c(t) = \rho'(t)} = \mathbf{max}\{-\boldsymbol{c}, \mathbf{min}(\boldsymbol{k}, \boldsymbol{t})\}, \tag{2.6}$$

Huber (6) recommended using k = 1.345. This choice produces a relative efficiency of approximately 95% when the error density is normal.

The iteratively Reweighted least squares (IRLS) method is used to solve the M-estimates nonlinear normal equations. The following iterative algorithm summarize this (7):

1. Start by estimating OLS as initial estimate of β and the estimate $\hat{\sigma}$.
2. Calculate the weights, w_i.
3. Calculate a new estimate using equation (2.4).

4. Repeat step 2 and 3 until the algorithm converges. In the end, the formula of M-estimator is:

$$\boldsymbol{\beta_M} = (\boldsymbol{x^t w x})^{-1}(\boldsymbol{x^t w x}); \boldsymbol{w} = \boldsymbol{diag}\{\boldsymbol{w_i}\}, \qquad (2.7)$$

Table 1 Objective and Weight Functions of the Estimation Methods

Method	Objective function	Weight function
OLS	u^2	1
Huber ($k = 1.345$)	$\begin{cases} \frac{1}{2}u^2 \; for \; \lvert u \rvert \le k \\ k\lvert u \rvert - \frac{1}{2}k^2 \; for \; \lvert u \rvert > k \end{cases}$	$\begin{cases} 1 \; for \; \lvert u \rvert \le k \\ \frac{k}{\lvert u \rvert} \; for \; \lvert u \rvert > k \end{cases}$
Bisquare ($k = 4.685$)	$\begin{cases} \frac{k^2}{6}\left\{1 - [1 - (\frac{u}{k})^2]^3\right\} \; for \; \lvert u \rvert \le k \\ \frac{k^2}{6} \; for \; \lvert u \rvert > k \end{cases}$	$\begin{cases} \left[1 - \left(\frac{u}{k}\right)^2\right]^3 \; for \; \lvert u \rvert \le k \\ 0 \; for \; \lvert u \rvert > k \end{cases}$

Source: (7)

The ψ-function of Huber estimator is constant-linearly. A re-descending ψ-function increases the weight assigned to an outlier until a specified distance and then decrease the weight to 0 as the outlying distance get larger. Montgomery et al. (8) introduced two types of re-descending ψ-function: soft re-descender and hard re-descender. Alamgir et al. (9) proposed a new re-descending M-estimator, called Alamgir re-descending M-estimator known as (ALARM) the ψ-function is defined as:

$$\boldsymbol{\psi(e)} = \begin{cases} \dfrac{\boldsymbol{16xe^{-2\left(\frac{e}{b}\right)}}}{\left(\boldsymbol{1 + e^{-\left(\frac{e}{b}\right)}}\right)^{4}} & \boldsymbol{if \; \lvert e \rvert \le b} \\ \boldsymbol{0} & \boldsymbol{if \; \lvert e \rvert > b} \end{cases}, \qquad (2.8)$$

Where e denotes the error and b is a tuning constant. Tukey's Bisquare functions $\psi_c(t) = t\left\{1 - \left(\frac{t}{c}\right)^2\right\}^2$ and $k = 4.685$ produces 95% efficiency. Bai et al. (10) also applied Huber's ψ-

function and Tukey's function to provide robust fitting of mixture regression models. If $\rho(t) = |t|$, then least absolute deviation (LAD) estimates are achieved by minimizing the sum of the absolute value of the residuals:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^t \beta|, \tag{2.9}$$

LAD is known as l_1 estimate, it had a low efficiency of 0.64 when the errors are normally distributed.

2.3.2 S-Estimator

Considering the problem of low BDP of M-estimators Rousseeuw and Yohai (11) proposed a scale estimator of M-estimator (S-estimator). S-estimates are aims to find the smallest possible dispersion of the residuals, that minimize the variance of the residuals to attain high BDP but with low efficiency defined as:

$$\hat{\beta} = \arg\min_{\beta} \hat{\sigma}\big(e_1(\beta), \dots, e_n(\beta)\big), \tag{2.10}$$

Where $e_i(\beta) = y_i - x_i^t \beta$ and $\hat{\sigma}\big(e_1(\beta), \dots, e_n(\beta)\big)$ is the scale M-estimate. Which is defined as the solution of $\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i(\beta)}{\hat{\sigma}}\right) = \delta$, where δ is taken to be $E_{\Phi}[\rho(e)]$. It can attain a high BDP of 0.5 and has an asymptotic efficiency of 0.29 under the assumption of normally distributed errors.

2.3.3 MM-Estimator

Another regression estimator that should be mentioned is the MM-estimator introduced by Yohai (12), The modified version of M-estimates has been popular and is one of the most used robust regression techniques. The MM-estimate can be found by a three-stage procedure:

1. Compute an initial consistent estimate $\hat{\beta}_0$ with high BDP but possibly low normal efficiency by using LMS or S-estimate with Huber or Bisquare function.
2. Compute a robust M-estimate of scale $\hat{\sigma}$ of the residuals based on the initial estimate $\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{e_i(\hat{\beta})}{\hat{\sigma}}\right) = 0.5$, when $\rho_0(e) = \rho_1\left(\frac{e}{k_0}\right)$, $\rho(e) = \rho_1\left(\frac{e}{k_1}\right)$, assume ρ function is bounded and $k_0 = 1.56$ ensures that the estimator has the asymptotic BDP =0.5.
3. Find M-estimate $\hat{\beta}$ starting at $\hat{\beta}_0$ by $L(\beta) = \sum_{i=1}^{n} \rho\left(\frac{e_i(\hat{\beta})}{\hat{\sigma}}\right)$.

The BDP of the MM-estimate depends only on k_0 and the asymptotic variance of the MM-estimate depends only on k_1. The larger the k_1 is, the higher efficiency the MM-estimate can become normal distribution. Maronna et al. (13) provided the values of k_1 with the corresponding efficiencies of the biweight $\rho - function$. The following table shows more details:

Table 2 corresponding efficiencies of the biweight -function

Efficiency	0.80	0.85	0.90	0.95
$\boldsymbol{k_1}$	3.14	3.44	3.88	4.68

Source: (14)

Yohai (12) indicates that MM-estimate with larger values of k_1 are more sensitive to outliers than the estimates corresponding to smaller values of k_1. In practice, an MM-estimate with Bisquare function and efficiency 0.85 (k_1= 3.44) starting from a Bisquare S-estimate is recommended.

3. Application

This application demonstrates the comparison between the robust regression estimators applied on datasets containing different types of outliers. The dataset was provided from properti website which provides datasets for real estate in several countries. The used data was collected 2014 from Mexico City the data contains three independent variables and a dependent variable the dataset consists of 2100 observations.

The following table provides a summary statistic for the dataset.

Table 3: Descriptive statistics for the variables.

Variable	**Mean**	**Median**	**Min**	**Max**	**Standard deviation**
$price\ USD$	133618	113714	33193	458268	79832.01
$Latitude$	20.77	19.64	15.75	32.67	2.736219
$Longitude$	-98.82	-99.20	-117.05	-86.77	4.813789
$area\ per\ m^2$	173	160	60	385	79.73616

Table 3 presents the descriptive statistics for various variables. For the 'price USD' variable, the median is substantially lower than the mean, suggesting a right-skewed distribution and potential presence of high-value outliers. Conversely, for the 'Latitude' variable, the median is relatively close to the mean, indicating a more symmetric distribution and suggesting fewer outliers. The 'Longitude' variable, despite a median seemingly close to the mean, requires careful interpretation. The broader range compared to its standard deviation suggests variability that might not be captured by just comparing these two metrics alone, hinting at possible outliers. Lastly, the $'area\ per\ m^2'$ shows a mean close to the median, suggesting a symmetric distribution and indicating that outliers are less likely to be present. The subsequent box plots should help in confirming these deductions by visually depicting the distribution spread and highlighting the presence of outliers for each variable.

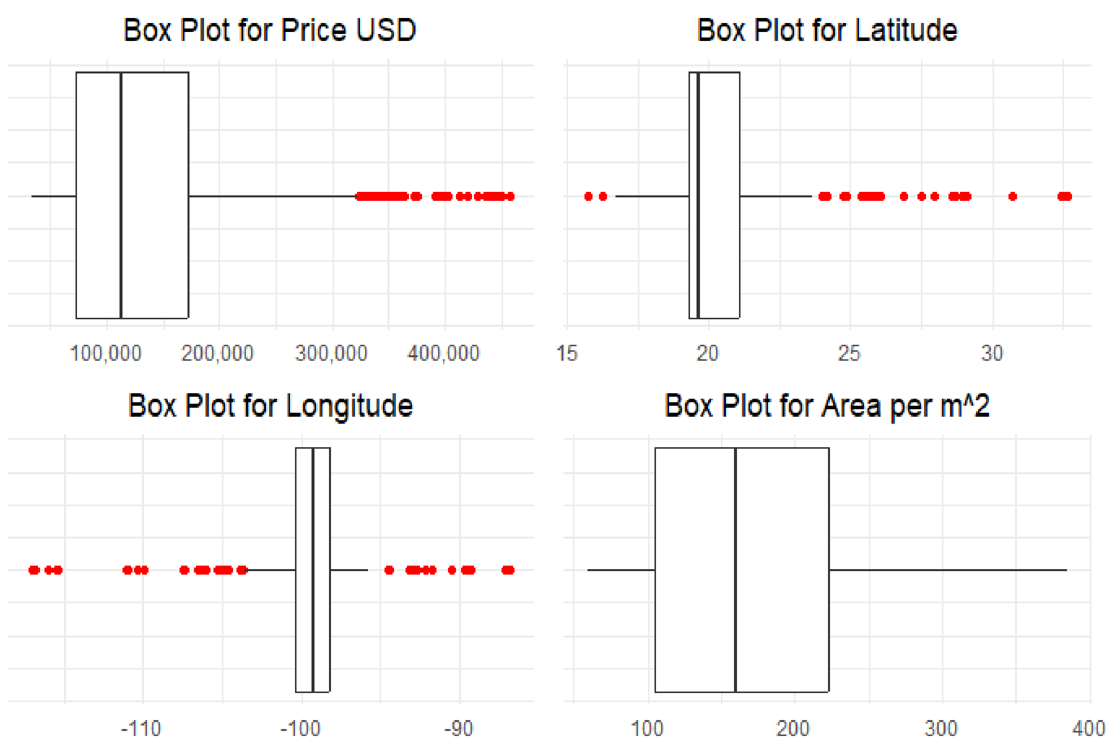

Figure 2: Box plots for the model variables

The box plots verify the presence of outliers in all the variables except for the area per m^2 while dependent variable price USD seems to have the highest number of outliers.

The following figure provides a scatter and correlation matrix for all the variables in the dataset and on the diagonal the histogram for each variable.

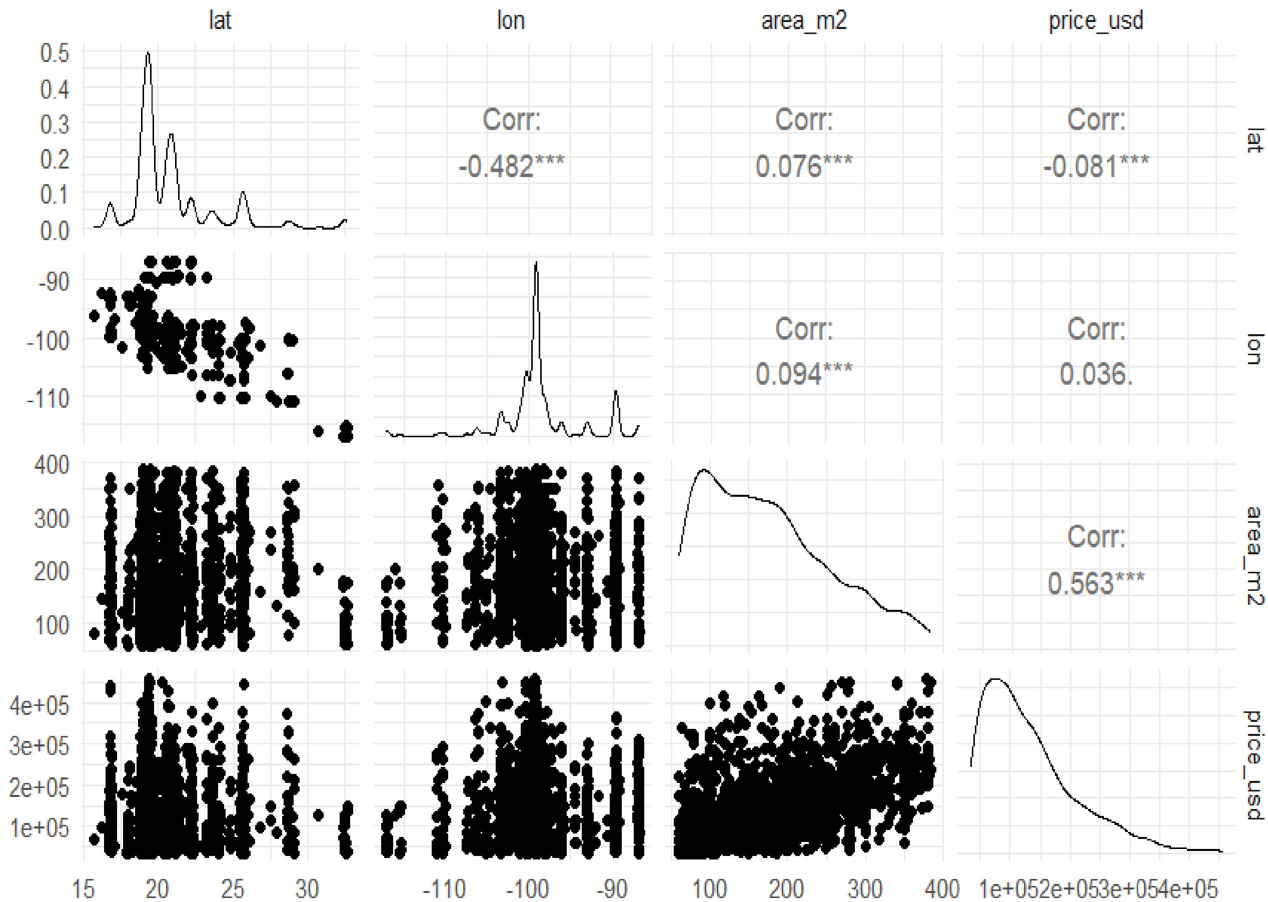

Figure 3: scatter and correlation matrix for the model variables

The following part demonstrates the comparison between the robust regression estimates and OLS estimate.

The following table illustrates the results from OLS estimate:

Table 4: OLS estimate Summary.

Coefficients	***Estimate***	***Standard error***	***t value***	***p − value***
Intercept	-30679.43	30042.65	-1.021	0.307
Latitude	-5113.83	596.64	-8.571	0.0000
Longitude	-1710.88	339.64	-5.037	0.0000
area per m^2	586.41	18.02	32.540	0.0000
Goodness of fit				
F − statistic	359.9		***p − value***	0.0000
R^2	$Adj\ R^2$	***RMSE***	***AIC***	***BIC***
0.34	0.339	64900	52504.32	52532.57

The following table illustrates the results from M estimate:

Table 5: M estimate Summary

$Coefficients$	$Estimate$	$Standard\ error$	$t\ value$	$p - value$
Intercept	-34335.0188	24025.4429	-1.4291	0.153153
$Latitude$	-4313.6304	477.1420	-9.0406	0.0000
$Longitude$	-1481.7537	271.6158	-5.4553	0.0000
$area\ per\ m^2$	592.6405	14.4119	41.1217	0.0000
Goodness of fit				
$F - statistic$	365.6		$p - value$	0.0000
R^2	$Adj\ R^2$	$RMSE$	AIC	BIC
0.38	0.379	46750	52300.97	52323.22

The following table illustrates the results from S estimate:

Table 6: S estimate Summary

$Coefficients$	$Estimate$	$Standard\ error$	$t\ value$	$p - value$
Intercept	-26269.62	21866.19	-1.201	0.23
$Latitude$	-3760.45	441.86	-8.511	0.0000
$Longitude$	-1254.82	254.61	-4.928	0.0000
$area\ per\ m^2$	585.38	17.23	33.980	0.0000
Goodness of fit				
R^2	$Adj\ R^2$	$RMSE$	AIC	BIC
0.483	0.4823	43550	52245.97	52271.22

The following table illustrates the results from MM estimate:

Table 7: MM estimate Summary

$Coefficients$	$Estimate$	$Standard\ error$	$t\ value$	$p - value$
Intercept	-46210.29	22134.55	-2.088	0.0369
$Latitude$	-2589.11	441.90	-5.859	0.0000
$Longitude$	-1082.50	254.45	-4.254	0.0000
$area\ per\ m^2$	591.58	14.57	40.603	0.0000
Goodness of fit				
R^2	$Adj\ R^2$	$RMSE$	AIC	BIC
0.787	0.7867	42550	52023.86	52048.42

The above table shows that all the independent variables are statistically significant for all the models while as noted the standard error for the robust estimation are smaller than OLS, all the models are overall significant while the MM estimate goodness of fit results demonstrates that it's the best estimate among the other estimates the following chart illustrates the differences in the evaluation criteria for each model.

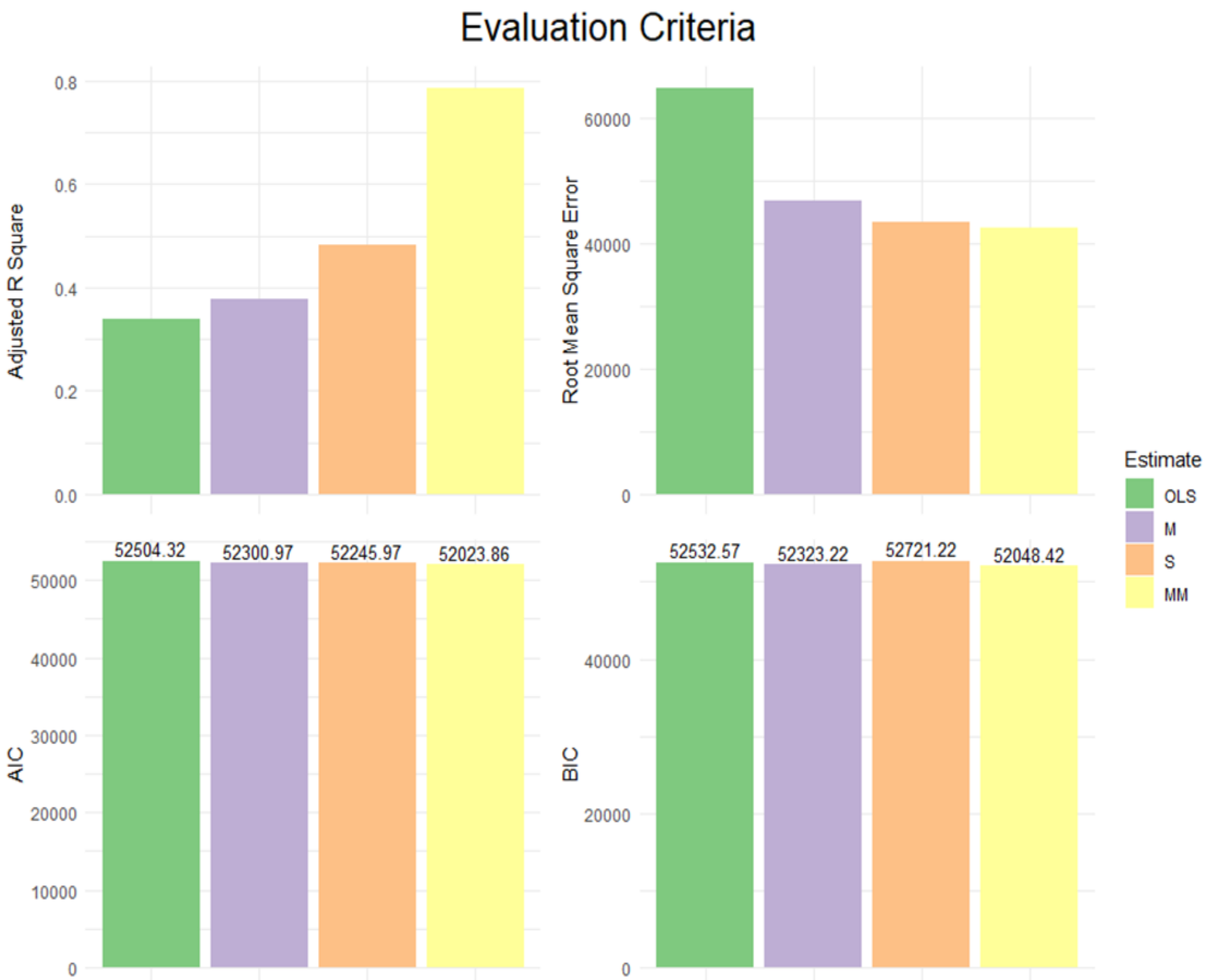

Figure 5: Column charts for Estimates evaluation Criteria.

4. Conclusion

In this paper, several robust estimates were reviewed in the presence of outliers, alongside OLS. The estimates were compared through an application with real data. The results obtained were favorable for MM estimate by having lower AIC, BIC, RMSE, and higher Adjusted R square. The results lead us to conclude that using robust estimators in the presence of outliers is a most since the outliers are part of everyday data.

Conflict of Interest

The Authors declare no conflict of interest.

5. References

1. Barnett V, Lewis T. Outliers in statistical data New York: Wiley; 1994.
2. Adnan R, Mohamad MN, Setan H. Multiple Outliers Detection Procedures in Linear Regression. Matematik. 2003;: 29-45.
3. Donoho DL, Huber PJ. The notion of breakdown point. In Peter JB, Kjell D, J.L. HJ. A Festschrift For Erich L. Lehmann. California: Wadsworth; 1983. p. 157-184.
4. Sakata S, White H. An Alternative Definition of Finite Sample Breakdown point with Applications to Regression Model Estimators. Journal of the American Statistical Association. 1995; 90(431): 1099-1106.
5. Hampel FR. The influence curve and its role in robust estimation. Journal of the American Statistical Association. 1974; 69(346): 383-393.
6. Hoaglen DC, Mosteller F, Tukey JW. Understanding Robust and Exploratory Data Analysis. New York: Wiley; 1983.
7. Huber P. Robust version of a location parameter. Annals. 1964; 35(1): 73-101.
8. Huber PJ. Robust statistics New York: Wiley; 1981.
9. Fox J, Weisberg S. An R Companion to Applied Regression. 3rd ed. London: Sage; 2019.
10. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 5th ed. New York: John Wiley & Sons; 2012.
11. Alamgir AA, Khan SA, Khan DM, Khalil U. A New Efficient Redescending M- Estimator: Alamgir Redescending M-estimator. Research Journal of Recent Sciences. 2013; 2(8): 79-91.
12. Bai X, Yao W, Boyer JE. Robust fitting of mixture regression. Computational Statistics and Data Analysis. 2012; 56(7): 2347-2359.
13. Rousseeuw P, Yohai VJ. Robust Regression by Means of S-Estimators. Springer Lecture Notes in Statistics. 1984; 26: 256-272.
14. Yohai V. High Breakdown-Point and High Efficiency Robust Estimates for Regression. The Annals of Statistics. 1987; 15(20): 642-656.

15. Maronna RA, Martin RD, Yohai VJ. Robust Statistics Theory and Methods (with R). 2nd ed. New York: John Wiley; 2019.

16. Yu C, Yao W. Robust Linear Regression: A Review and Comparison. Communication in Statistics- Simulation and Computation. 2017; 46(8): 6261-6282.

17. Adnan R, Mohamad MN, Setan H. Multiple Outliers Detection Procedures in Linear Regression. Matematik. 2003; 19: 29-45.