Zanco Journal of Human Sciences (ZJHS)

TYPE: Original article
DOI https://doi.org/10.21271/zjhs.29.2.14

OPEN ACCESS

*Corresponding author
Ismat Mousa Ibrahim
ismat.mousa@dpu.edu.krd

Comparison between Linear Regression and Robust Regression Models by Using Error Criteria and Information Criteria Applied to Human Sample

Ismat Mousa Ibrahim /Technical College of Duhok, Department of Dental Technology, Duhok Polytechnic University, Duhok, Iraq

RECEIVED :29 /07/2024
ACCEPTED :13/10/ 2024
PUBLISHED :15/04/ 2025

Keywords:
Robust Method,
Least Trimmed Square,
Maximum Likelihood.

Abstract

Hypertension is a common and serious disease, and for this reason, a sample of patients was chosen from Azadi Teaching Hospital in Duhok. In this study a comparison was made between Ordinary Least Squares (OLS) with two robust methods Least Trimmed Square Estimator (LTS) and the Modified Maximum likelihood type estimator (MM), they were evaluated using two types of criteria represented by error criteria (MSE, MAPE, MSAE, SAZ1, SAZ2,) and information criteria (AIC, BIC). Criteria have played a fundamental role in the statistics field and at the same time have applied to obtain the lowest rate of errors as well as to get the optimal solutions. The Modified Maximum likelihood type estimator method showed high efficiency in calculating values by most criteria, with exception of the SAS2 criteria, which recorded as the best value using the Least Trimmed Square Estimator method, and the (MSE) criteria, which showed best value using the Ordinary least squares method.”

1.Introduction

The estimate of parameters is a crucial aspect of regression analysis, therefore selecting a suitable estimation technique is aid in our understanding of the population that the sample under study is drawn from. The estimate method's description of the kind and strength of the link between the response variable and the explanatory factors explains this. Although there are several established estimating techniques, the Ordinary Least Squares technique is the most often applied (Lee and Han, 2024). Since its estimators offer so many advantages, especially the normal distribution of the random errors, Ordinary Least Squares has long dominated approaches for estimating linear regression parameters (Isazade *et al.*, 2023). Researchers have discovered that these techniques lose their effectiveness, though, when one or more of the assumptions are not supported by the data. For example, the aberrant distribution of random errors resulting from an outlier's observations that considerably deviate from the majority of observations. They frequently come from mixture distributions or heavy-tailed distributions. Natural reasons or mistakes in reading, counting, or recording might be the source of their irregularity. They may lever the fit line to its tip, which can have an impact on the Ordinary Least Squares estimators. The estimations of Ordinary Least Squares may be affected by the issue of auto-correlation or multicollinearity among the explanatory factors (if any) within the data, as well as the issue of error departure from the normal distribution (Rousseeuw and Yohai, 1984). The solution of these issues is through estimate the linear model parameters using more objective techniques. As a result, scientists are working to identify substitute techniques that are even more effective and insensitive to departures from the assumptions of the linear regression model (Sahu, 2023). These are known as Robust Methods, and the estimators that come from them are known as Robust Estimators. Whether the distribution of errors is normal or abnormal, the parameters exhibit acceptable properties. Put differently, they address the issues of auto-correlation and multicollinearity and are less susceptible to outliers. In this study a comparison was made between Ordinary least squares (OLS) with two robust methods, Least Trimmed Square Estimator (LTS) and Modified Maximum likelihood type estimator (MM), they were evaluated using two types of criteria represented by error criteria which include: MSE, MAPE, MSAE, SAZ1, SAZ2, and information criteria that including: AIC and BIC. The purpose of this comparison is to find the best estimation method of parameters of linear regression model. The Modified Maximum likelihood type estimator method showed high efficiency of MAPE, MSAE, SAZ1, AIC and BIC criteria based on a sample of 102 individuals were used in this study, it has been found that six factors: age, gender, blood pressure, serum cholesterol, sugar status, and smoking have an impact on the heart rate at rest. Data were optained from a sample of hypertensive patients in azadi teaching hospital in Duhok city.

2- Problem of study

To compare the classical method Ordinary Least Squares (OLS) and the robust methods represented by the Least Trimmed Square Estimator (LTS) and Modified Maximum likelihood type estimator (MM). Obtaining the best estimation of linear regression model with the use of information criteria (AIC and BIC) and error criteria (MSE, MAPE, MSAE, SAZ1 and SAZ2).

3- Aim of study

The aim of this study is to compare Ordinary Least Squares (OLS) with two robust methods: the Least Trimmed Squares (LTS) estimator and the Modified Maximum Likelihood (MM) estimator. The evaluation was conducted using two sets of criteria: error criteria (MSE, MAPE, MSAE, SAZ1, SAZ2) and information criteria (AIC, BIC). Both types of criteria were employed to assess the performance of these estimation methods**,** the following software is used: S-plus, and SPSS.

4- Hypothesis of study

According to the study, six factors (Age, Gender, Blood pressure, Serum cholesterol, Sugar status and Smoking) have effect on the Heart Rate at rest, from a sample consisting of 102 patients. When utilizing robust approaches to obtain the estimated value instead of the ordinary least squares method, we obtain the greatest results when employing multiple linear regression.

5- Linear regression models

Statistical models called linear regression are used to determine how one or more independent variables and a dependent variable are related. The dependent variable and the independent variables are assumed to have a linear relationship by the model (DeForest et al., 2023). Finding the best-fitting line that depicts the connection between the independent and dependent variables is the aim of a linear regression model. By estimating the coefficients that reduce the discrepancy between the values predicted by the model and the actual values, this line may be found. **(**Shi, 2023**)**.

5-1 Simple linear regression

Simple Linear Regression, when making predictions about a dependent variable, simple linear regression models use just one independent variable. According to (Sahu, 2023), a straight line can be used to describe the relationship between the independent variable (X) and the dependent variable (Y). A common way to express the equation of a basic linear regression model is:

$$Y_i=\beta_0+\beta_1X_i+e_i \qquad (1)$$

$i=1, 2,\dots, n$ and n is sample size.

Where

***Y*:** the variable that is dependent.

***X*:** the variable that is independent.

β_0**:** is the line's y-intercept or the value of Y when X equals zero.

β_1 **:** represents the line's slope, or the change in Y for a unit change in X.

e_i **:** is the error term, which shows how the values predicted by the model and the observed values of ***Y*** differ.

5-2 Multiple linear regression

Multiple linear regression models by using two or more independent variables into the prediction of a dependent variable, multiple linear regression expands upon the capabilities of simple linear regression. The model now incorporates numerous predictors rather than simply one, based on the assumption that the dependent variable's relationship with the independent variables is linear (Wang et al., 2023). If you have p independent variables in your multiple linear regression model, you may write its equation as:

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\dots+\beta_pX_{ip}+e_i \qquad i=1,\dots\dots,n. \qquad (2)$$

Where

Y_i **:** depended variable

X_i **:** independent variables

(β_0 → Intercept; β_1 → **,…,** β_P → slope coefficients) are the regression parameters

denoted by β_0 →, β_1 →**,…,** and β_P .

***e_i* :** The random error

***P*:** The number of explanatory variables

6- The ordinary least squares method

The ordinary least squares Technique When fitting a linear regression model, one way to estimate the unknown parameters is using the ordinary least squares (OLS) approach. It ranks well among linear regression analysis methods. A line that minimizes the sum of the squared residuals is considered to be the best-fitting line to a set of data points; this notion is the foundation of the ordinary least squares (OLS) approach. What remains after subtracting the regression line's projected values from the dependent variable's actual values is known as the residuals. The linear regression model's coefficients the line's slope and the intercept can be estimated using the ordinary least squares (OLS) approach (Lee & Han, 2024). For every one-unit shift in the independent variable, the slope coefficient shows how much the dependent variable shifts. When the independent variable is set to zero, the intercept coefficient shows the dependent variable's value. One easy and effective way to estimate linear regression model parameters is the ordinary least squares (OLS) method (Isazade *et al.*, 2023).

For the OLS method to be legitimate, the following conditions must be satisfied:

1-There is no correlation between the independent variables.

2-A regularly distributed dependent variable is given.

3-There is no correlation between the mistakes.

4-The dispersion of the mistakes is consistent.

The OLS method's accuracy is dependent on not violating any of these assumptions.

6.1- Assumptions of least square method

In the linear regression model there are some of assumptions based on the study usual method of Ordinary Least Squares:

A-Assumptions about the error:

1- Mathematically expectation of the random variable is equal to zero $E(e_i)=0$ $i=1,2,\dots,n$

It follows that: $E(Y_i) = E(\beta_0+\beta_1 X_i+e_i)$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_i + E(e_i)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Therefore, the regression function for model (1) is:

$E(Y) = \hat{\beta}_0 + \hat{\beta}_1 X$, since the regression function relates the means of the probability distributions of *Y* for any gave *X* to the level of *X* (Alma, 2011).

2- Contrast the values of a random variable to be constant in each period of time $Var(e_i) = E(e_i^2) = \sigma^2 I_n$ $,i=1, 2,\dots,n$. This is called homoscedasticity of error variation (Rousseeuw, 1984). It therefore follows that the divergence of the response variable Y_i is: $\sigma^2(Y_i)= \sigma^2$, since $\sigma^2(\beta_0+\beta_1 X_i+e_i)= \sigma^2(e_i)= \sigma^2$, thus model (1) supposes that the probability distribution of *Y* have the same divergence σ^2.

3- A random variable e_i distributed as normal distribution with $e_i \sim N(0, \sigma^2)$, $i= 1, 2,\dots, n$

4- If for any $(i\neq j)$, e_i,e_j are independent, then there is no autocorrelation between e_i and e_j , hence the outcome in any one test has no effect on the error term for any other test as to whether it is positive or negative or small or large since the error term e_i and e_j are uncorrelated, this means that the covariance between them is equal to zero $Cov(e_i,e_j) = 0,$ $(i \neq j = 1,2,\dots n)$ (Sanford, 2005).

5- Independence between the illustrative variables and random variables, $E(e_i,X_{ij})=0$,that is to say e_i independent of X_{ij} for all different values of *i*, this means there is no problem of multicollinearity.

B- Assumptions on the distribution of the response variable *Y*:

1-Average Y_i is a function of straight line

$\overline{Y} = E(Y_i) = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p$, $i=1,2,...,n$. where $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ are estimates of regression parameter, since the regression function relate the means of the probability distribution of *Y* for any given *X* to the level of *X* (Wilcox, 2005).

2-Variance of Y_i, *Var* (Y_i) has one value for any value of i *Var* $(Y_i)=\sigma^2$, $i=1, 2, ..., n$.

3-Response variable Y_i distributed as normal distribution with mean μ and variance σ^2,i. e., $Y_i \sim N(\mu,\sigma^2)$.

4- Any two observation Y_i and Y_j are uncorrelated, this implies that,

Cov $(Y_i , Y_j)=0$ for all $i \neq j = 1,2,...n$.

5- The relationship between X_i's, $\hat{Y}$ be a linear relationship is the equation of straight line

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} \qquad ,i = 1,2,...,n\,.$$

7- Robust regression

One type of regression analysis is robust regression, aims to enhance the reliability of estimates in scenarios featuring outliers, influential data points, or other deviations from standard regression assumptions. Unlike ordinary least squares (OLS) regression, which can be significantly influenced by such irregularities, robust regression techniques, such as M-estimation or robust regression with iteratively reweighted least squares, mitigate or disregard the impact of outliers to enhance the overall model fit (Rousseeuw and Van, 1999). This approach can yield more precise parameter estimates and inferences, rendering robust regression an invaluable asset for analyzing real-world data that may deviate from the assumptions of conventional regression models.

7-1 Least trimmed square estimator

The least trimmed squares (LTS), a p-vector, was first proposed by Rousseeuw in 1984. Here is the expression of this method:

$$\hat{\beta}_{LTS} = argmin\; Q_{LTS}\,(\beta)\; where\;\; Q_{LTS}\,(\beta) \sum_{i=1}^{n} e_i^2$$

$$e_1^2 \le e_2^2 \le e_3^2 \le \cdots \le e_n^2 \quad are\ the\ ordered\ squared\ residuals$$

$$e_1^2 = (Y_i - \acute{X}_i\,\beta\,)^2\ , \quad i = 1, 2, \ldots, n \quad and\ h\ is\ defined\ in\ the\ range$$

$$\frac{n}{2} + 1 \le h \le \frac{3n + p + 1}{4}$$

According to Chen's nomenclature, which uses the sample size (n) and the number of parameters (p), the procedure entails removing the sum of the greatest squared residuals. Based on the values of n and the configuration of outlier data, this exclusion allows for the total eradication of outlier data points. An efficient way to find outliers is the least trimmed squares method. There is discussion of the breakdown value in relation to the least trimmed squares estimation in (Bai, 2010). In order to determine the estimator $\hat{\beta}_{LTS}$ it be taken into consideration (n-h + 1) from the subsequent subsample. Where (h) is a parameter that determines the number of smallest residuals to be included in the sum.

$$\{X_1\,,\ X_2\,, \ldots, X_h\}$$

$$\{X_2\,,\ X_3\,, \ldots, X_{h+1}\}$$

$$\{X_3\,,\ X_4\,, \ldots, X_{h+2}\}$$

$$. \quad . \quad . \quad .$$

$$. \quad . \quad . \quad .$$

$$\{X_{n-h+1}\,,\ X_{n-h+2}\,, \ldots, X_n\}$$

There are *h* items in each subgroup; these make up the infectious half. Then, it does the following to get the means for each subset:

$$\overline{X}_1 = \frac{1}{h}\sum_{i=1}^{h} X_i$$

$$\overline{X}_2 = \frac{1}{h}\sum_{i=2}^{h+1} X_i$$

.

.

$$\overline{X}_{n-h+1} = \frac{1}{h}\sum_{i=n-h+1}^{n} X_i$$

It also calculates the sum of squares for every subsample.

$$SQ_{(1)} = \sum_{i=1}^{h}(X_i - \overline{X}_1)^2$$

$$SQ_{(2)} = \sum_{i=2}^{h+1}(X_i - \overline{X}_2)^2$$

.

.

$$SQ_{(n-h+1)} = \sum_{i=n-h+1}^{n}(X_i - \overline{X}_{n-h+1})^2$$

Also, it adds up all the squares for each subsample. According to (Muhlbauer, et al., 2009), if the mean that corresponds to the smallest square to the equation is also the least trimmed squares estimator $\hat{\beta}_{LTS}$, then the two will be equivalent.

This method achieves computational parity with OLS when the precise amount of outlier data points is removed. On the other hand, if there are more outliers than reduced, its efficiency will decrease. In contrast, crucial data points could be removed from the calculation due to over-trimming. As a high breakdown approach, LTS is defined as having a breakdown point of 50%.

7-2 Modified maximum likelihood type estimator

One very reliable class of estimators in the world of linear models is the Modified Maximum Likelihood type, which was first introduced by (Yohai, 1987). This estimator type combines efficiency with the features of high breakdown value estimating. There is a three-step process that Yohai's Modified Maximum Likelihood estimators adhere to. The first step of the calculation is to use an influence function with an S-estimate.

$$\rho(X) = \begin{cases} 3(\frac{X}{c})^2 - 3(\frac{X}{c})^4 + (\frac{X}{c})^6 & if\,|X| \le c \\ 1 & otherwise \end{cases} \qquad (3)$$

It is determined that the tuning constant *c* is 1.548.

Rousseeuw and Yohai presented the S-estimate in (1984) it is a high-breakdown-value, robust regression approach. Reducing the residuals' dispersion is its primary objective. The aim of the function is $\min\ s\,(e_1(\beta), e_2\,,(\beta), \dots, e_n(\beta))\ where\ e_i\ (\beta),$ is the *i-th* residuals for candidate $\boldsymbol{\beta}$. The solution is how this objective function is expressed: $\frac{1}{n-p}\sum_{i=1}^{n} X(\frac{Y_i - \hat{Y}_i}{s}) = k$ where k is a constant (Heritier and Copt, 2006). In the second stage, the MM parameters are computed to achieve the minimum value of $\sum_{i=1}^{n}\rho(\frac{Y_i - X'_i\hat{\beta}_{MM}}{\hat{\sigma}_0})$ where $\hat{\sigma}_0$ is the scale estimate from the first step (standard deviation of the residuals) and $\rho(X)$ is the influence function used in the first stage with tuning constant 4.687. (Bianco, et al., 2003). The MM-estimate of scale is computed in the final step by solving $\frac{1}{n-p}\sum_{i=1}^{n}\rho(\frac{Y_i - X'_i\hat{\beta}}{s}) = 0.5$ (Schumann).

8- Investigation of the accuracy of estimation multiple linear regression

Examining the precision of estimate various linear regression models in order to determine which model is the best, practical statisticians constantly employ criteria. Numerous types of criteria exist, including information criteria and mistakes criteria. This study made use of these two varieties. The accuracy of estimate methods, which show how well they can predict outcomes, is usually judged by how well they are able to do so. Some of the criteria that are utilized include:

8-1 Criteria SAZ1 and SAZ2

We used these functions in our work: In addition to the mean square error, Wasfi Taher Kahwachi's **SAZ1** and **SAZ2** criteria can be used to determine the amount of error caused by utilizing particular models. Mean Squared Error, **SAZ1**, and **SAZ2** were assessed for each selected model. We can get a sense of the data error behavior from these two functions (Shareef and Ibrahim, 2020). Its behavior is similar to that of the mean square error, as observed in the application; that is, **SAZ1** and **SAZ2** have the same direction as the mean square error value is larger, and a criterion that measures the average of the error ratios is the product of their studies of the error observations ratio divided by its number. We can deduce that they have potential applications beyond the mean square error metric (Ibrahim, 2016). Here are the functions:

$$\text{SAZ1} = \frac{\left|\sum_{i=1}^{n-1} \frac{\mathrm{e}_i}{e_{i+1}}\right|}{n-1} \qquad (4)$$

and

$$\text{SAZ2} = \frac{\left|\sum_{i=1}^{n-1} \frac{\mathrm{e}_{i+1}}{e_i}\right|}{n-1} \qquad (5)$$

where

e_i: Prior random error

e_{i+1}: the subsequent random error

n : the number of the sample

8-2 Mean squares error

When analyzing the accuracy of a prediction model, one frequent metric to utilize is the Mean Squared Error (MSE). The mean squared deviation from the actual value is what it measures. When doing regression analysis, MSE is a common tool for measuring the discrepancy between the model's projected values and the data's actual values. When calculating MSE, the formula is (Hodson et al., 2021):

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p-1} = \frac{\sum_{i=1}^{n}(e_i)^2}{n-p-1} \qquad (6)$$

Where:

$$e_i = Y_i - \hat{Y}_i \qquad (7)$$

e_i: the random error
n : the number of the samples
p : the number of the parameters

The goal in predictive modeling is usually to minimize the MSE. A lower MSE indicates a model that predicts more accurately the observed data. However, one should consider that MSE values depend on the scale of the target variables, making it sometimes hard to interpret the magnitude of the error without context. Also, being in squared units of the target variable, it gives more weight to large errors.

8-3 Mean absolute percentage error

One metric used to assess a forecasting model's accuracy is the Mean Absolute Percentage Error or MAPE. The mean of the absolute percentage forecast errors is computed (Nabillah and Ranggadara, 2020). The MAPE calculation is:

$$MAPE = \frac{\sum_{i=1}^{n}\left|PE_i\right|}{n} \qquad (8)$$

Where:

$$PE_i = \frac{(Y_i - \hat{Y}_i)}{Y_i} \qquad (9)$$

MAPE gives a relative measure, making it easier to compare forecast accuracy between different time series data sets regardless of the scale of the data. Lower MAPE values indicate better predictive accuracy.

8-4 Mean sum of absolute error

One way to measure how well a model or forecasting method works is by looking at its MSAE, or mean sum of absolute errors. The mean absolute deviation (MSAE) between expected and actual values is calculated. Calculating MSAE is as follows (Shareef and Ibrahim, 2020):

$$MSAE = \frac{\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|}{n} = \frac{\sum_{i=1}^{n}\left|e_i\right|}{n} \qquad (10)$$

MAE offers a clear and easy approach to comprehending the average magnitude of prediction mistakes. To get it, just add up all the absolute mistakes that have occurred between the actual and anticipated numbers (Yang, 2020).

9- Information Criteria

9-1 Akaike's Information Criteria

In statistics, the quality of fit of several models is compared using a metric called Akaike's Information Criterion (AIC). It strikes a balance in the trade-off between the model's complexity and its ability to fit the data well (Bozdogan, 1987). For AIC, the formula is:

$$AIC = -2log(L) + 2k \qquad (11)$$

Where:
L: stands for the model's maximal likelihood, which gauges how well the model conforms to the data.

K: denotes the total number of parameters in the model, including the intercept and any extra variables.

AIC seeks to identify the model that best explains the data. The best model out of the set being compared is the one with the lowest AIC, since lower values suggest a better fit (Acquah, 2010).

9-2 Bayesian information criteria

The BIC is defined as (Bollen et al.,2014):

$$BIC = -2log(L) + klog(n) \qquad (12)$$

where:

L : the maximum likelihood of the model

K : the number of parameters there are in the model

n : represent the number of observations.

More parameterized models are punished by the BIC, which is a penalized likelihood criterion. This is so that they can less effectively generalize to new data since more complicated models have a higher tendency to overfit the data. The BIC is often used to select between competing statistical models. It is often accepted that the model that has the lowest BIC is the optimal model. However, the BIC should be used with caution, as it can be misleading in some cases. For example, the BIC can favor models that are too simple, and it can also be sensitive to the choice of prior distributions.

10- Practical side

In this section Ordinary least squares with two robust methods: Least Trimmed Square Estimator and Modified Maximum likelihood type estimator were compared, by using two types of criteria represented by error criteria which include: MSE, MAPE, MSAE, SAZ1, SAZ2, and information criteria that including: AIC and BIC. S-plus, and SPSS were used as software to find the parameters and tables.

10-1 Heart rate line model for (OLS)

By using SPSS (version 22), and by applying the Ordinary least squares (OLS) Estimator Method the fitted linear model for heart Rate line is as the following:

$\hat{Y}_{i} = 471.246-34.203X_{i1} - 25.980X_{i2} + 11.908X_{i3} -15.879X_{i4} -76.545Xi5-14.680Xi6$

The following are the regression analysis findings for the OLS estimator that are displayed in tables (1) and (2):

Table (1) represents the coefficients of regression for heart rate parameters by using OLS method

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model	**B**	**Std. Error**	**Beta**	**t**	**Sig.**	**Tolerance**	**VIF**
1 (Constant)	471.246	123.128		3.827	.000		
Age	-34.203	20.057	-.285	-1.705	.091	.327	3.058
Gender	-25.980	28.048	-.106	-.926	.357	.694	1.441
Blood_pressure	11.908	29.836	.055	.399	.691	.486	2.058
Serum_cholester	-15.879	35.249	-.055	-.450	.653	.609	1.643
Sugar_status	-76.545	25.637	-.453	-2.986	.004*	.397	2.517
Smoking	-14.680	34.007	-.049	-.432	.667	.720	1.390

a. Dependent Variable: Heart_Rate_at rest

Table (1) shows only one parameter which is sugar status in the Hypertensive patient's data has a significant effect on the Heart Rate at rest. In contrast, age, gender, blood pressure, serum cholesterol, and smoking, show no significant impact on the heart rate at rest. All the parameters chosen in our study have affected the heart rate in direct proportion. The results in table (1) show parameters that are affected in reverse proportion, which does not match the model, which means in future studies the sample must be increased. The results from the same table illustrate whether or not the multicollinearity trouble exists by using the VIF test, the results declare that all VIF values are less than 5 which means that multicollinearity trouble doesn't exist between the independent variables.

Table (2) shows the analysis of variance for heart rate by using OLS method

ANOVA[a]

Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**
1	**Regression**	199569.330	6	33261.555	2.397	.034[b]
	Residual	1318407.581	95	13877.975		
	Total	1517976.912	101			

a. Dependent Variable: Heart_Rate_at rest

b. Predictors: (Constant), Smoking, Serum cholesterol, Gender, Sugar status, Blood pressure, Age

Table (2): Displays an ANOVA table which is showing that the F-value in the Hypertension line is significant due to the F-calculated value of 2.397 is greater than the F-tabulated value under the significant level of α = 0.05 and the degrees of freedoms are df1 = 6 and df2 = 95, resulting in F (0.05,6,95) = 2.195, or the p-value is less than 0.05.

Table (3) shows the test of Durbin-Watson for heart rate by using OLS method

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.363a	.131	.077	117.805	2.003

a. Predictors: (Constant), Smoking, Serum_cholesterol, Gender, Sugar_status, Blood_pressure, Age

b. Dependent Variable: Heart_Rate_at rest

On the other hand, the SPSS program was used to test whether autocorrelation trouble existed by using the Durbin-Watson test. The result of the Durbin-Watson test was 2.003 ≅ 2.00 which demonstrates the optimum value of Durbin-Watson and confirms that there is no autocorrelation trouble in the study sample.

10-2 Heart Rate line model for LTS estimator

Using the Least Trimmed Square (LTS) Estimator Robust Method, S-plus (8.0 professional), and the fitted linear model for the heart Rate line is as follows:

$\hat{Y}_i = 219.654 – 17.939X_{i1} – 30.547X_{i2} + 22.052X_{i3} – 28.158X_{i4} – 60.5130Xi5 + 71.539Xi6$

```
Coefficients:

Intercept Age Gender Blood.pressure Serum.cholesterol Sugar.status Smoking
 219.654 -17.939 -30.547   22.052    -28.158          -60.5130    71.539

Scale estimate of residuals: 95.4

Robust Multiple R-Squared: 0.1937

Total number of observations:  102

Number of observations that determine the LTS estimate:  91

Residuals:

     Min.    1st Qu.     Median    3rd Qu.      Max.

 -123.80746  -48.32067  -20.42441   88.42933  387.10544
```

```
Weights:
 0  1
 8 94
```

Table (4) Regression coefficients for Heart Rate parameters by Robust LTS estimator

The regression analysis shows the coefficients, residuals and weights of Least Trimmed Squares Robust estimator represented in the table (4) which can be evulated that includes a significant parameter which is sugar status in the dataset which is a significant effect on the dependent variable (Heart Rate at rest) whereas the others variables do not effect on the Heart Rate at rest significantly.

10.3 Heart Rate line model for MM estimator

Following the Modified Maximum Likelihood Type Estimator Robust Method and S-plus (8.0 professional), the fitted linear model for the heart Rate line is as follows:

$$\hat{Y}_i = 90.3071 - 1.5000X_{i1} + 9.31931X_{i2} + 8.5428X_{i3} - 8.0420X_{i4} - 1.1982Xi5 - 0.2243Xi6$$

Table (5) Regression coefficients for Heart Rate parameters by Robust ***MM*** method

```
Coefficients:

                Value Std. Error  t value Pr(>|t|)
     (Intercept)  90.3071  31.8066     2.8393   0.0055
             Age  -1.5000   5.1565    -0.2909   0.7718
          Gender   9.3193   7.0491     1.3221   0.1893
  Blood.pressure   8.5428   7.7737     1.0989   0.2746
Serum.cholesterol  -8.0420   8.4684    -0.9496   0.3447
    Sugar.status  -1.1982   5.9794    -0.2004   0.8416
         Smoking  -0.2243   8.9661    -0.0250   0.9801


Residuals:

   Min     1Q Median    3Q   Max
 -56.51 -8.026  6.084 138.8 450.5
Residual scale estimate: 31.03 on 95 degrees of freedom


Proportion of variation in response explained by model: 0.01589

Test for Bias


           Statistics P-value
 M-estimate       4.38   0.735
LS-estimate       3.80   0.802

The seed parameter is : 1313
```

Table (5): Demonstrates the results of regression analysis, coefficients, residuals and Proportion of variation of Modified Maximum likelihood type robust estimator.

Table (6) Estimation Of coefficient by *OLS*, *LTS* and *MM* Methods

Variables	***Coefficients***	***Coefficient Estimation Using Three Techniques***		
		OLS	***MM***	***LTS***
Constant	B_0	470.468	90.3071	219.6543
X1	B_1	-34.410	-1.5000	-17.9397
X2	***B2***	-25.9801	9.3193	-30.5471
X3	***B3***	11.061	8.5428	22.0527
X4	***B4***	-13.361	-8.0420	-28.1589
X5	***B5***	-77.059	-1.1982	-60.5130
X6	***B6***	-14.344	-0.2243	71.5391
SAZ1		0.506249502	1.457166961	0.122516088
SAZ2		0.82198184	0.354358158	1.896310543
MSE		13877.97453	22062.80517	14924.27478
MAPE		0.670893428	0.323334468	0.539052954
MSAE		92.15741588	83.90866144	88.0822287
AIC		1270.622	981.6363	1027.324
BIC		1302.121	1000.012	1045.699

11- Comparison

The results of the analysis can be compared after evaluating the multiple linear regression model coefficients of the three methods. Several measures were used to evaluate the estimated accuracy of the regression equation in order to discuss the results of the robust methods and compare them with the results of the Ordinary Least Squares method. This shows why it is important to highlight the advantages of each method for determining parameters.

Table (6) shows that LTS method was better than the two other methods MM and OLS according to SAZ1 criteria in value 0.122516088. In addition, OLS method was better than the two other methods MM and LTS depending on MSE criteria which recorded a value of 13877.97453. Moreover, the MM method was better than the two other methods OLS and LTS methods according on SAZ2, MAPE, MSAE, AIC, BIC criteria, where their values were (0.354358158, 0.323334468, 83.90866144, 981.6363, 1000.012) respectively. In conclusion, the best criteria in the first method (OLS) and third method (LTS) is SAZ1 which showed the lower values (0.506249502, 0.122516088) respectively. While, the optimal criteria in the second method (MM) was MAPE with value of 0.323334468 which is the lowest value in compare with the other criteria.

12- Conclusion

The following conclusions are drawn from the practical aspect of our study's results:

From the results extracted from Table (6) it is clear that the MM method is better than its counterparts LTS and OLS in recording the lowest values according to the criteria such as (SAZ2, MAPE, MSAE, AIC and BIC) used . It can be seen from the same table that the robust methods in general are better than the OLS method in calculating the lowest ratios, whether they are error criteria or information criteria. This is due to several considerations, the most important of which is that the OLS method is affected by outliers.

13- Acknowledgment

My wonderful assistant professor Drs. Jiyar M. Naji and Hamid M. Daabo, for their gracious assistance, direction, and encouragement over the course of this study, have my deepest thanks and appreciation.

References

-Acquah, H. de-G. (2010) “Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship”.

- Alma, O. G. (2011): Comparison of Robust Regression Methods in Linear Regression. Int. J. Contemp. Math. Sciences, Vol. 6. Mugla University, Turkey.

- Bai, X. (2010): Robust Linear Regression. B. S., Mathematics and Applied Mathematics China. 5/11/2013.

- Bianco, A. M. & Ben, M. G. & Yohai, V. J. (2003): Robust Estimation for Linear Regression with asymmetric Errors with applications to log-gamma regression. Universidad de Buenps Aires.

-Bollen, K. A. , Harden, J. J. , Ray, S., and Zavisca, J. (2014) “BIC and alternative Bayesian information criteria in the selection of structural equation models,” Struct. Equ. Model. a Multidiscip. J., vol. 21, no. 1, pp. 1–19.

-Bozdogan, H. (1987) “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions,” Psychometrika, vol. 52, no. 3, pp. 345– 370,

- DeForest, D.K., Ryan, A.C., Tear, L.M. and Brix, K.V., 2023. Comparison of multiple linear regression and biotic ligand models for predicting acute and chronic zinc toxicity to freshwater organisms. Environmental toxicology and chemistry, 42(2), pp.393-413.

- Heritier, S. & Copt, S. (2006): Robust MM-Estimation and Inference In Mixed Linear Models.Universite de Geneve.

-Hodson, T.O., Over, T.M. and Foks, S.S., (2021). Mean squared error, deconstructed. Journal of Advances in Modeling Earth Systems, 13(12), p. e2021MS002681.

- Isazade, V., Qasimi, A.B., Dong, P., Kaplan, G. and Isazade, E., 2023. Integration of Moran’s I, geographically weighted regression (GWR), and ordinary least square (OLS) models in spatiotemporal modeling of COVID-19 outbreak in Qom and Mazandaran provinces, Iran. Modeling Earth Systems and Environment, 9(4), pp.3923-3937.

- Ismat mousa Ibrahim, “Using New Criteria to Compare between Some Robust Method and Ordinary Least Squares in Multiple Regression with Application on Wheat Data in Iraq”, Sofia University "ST. KLIMENT OHRIDSKI", 2016.

- Kahwachi, Wasfi (2014). Some Statistical Functions for Model Accuracy Measuring, Unpublished papar. Salahaddin University.

- Lee, M., & Han, C. (2024). Ordinary least squares and instrumental-variables estimators for any outcome and heterogeneity. The Stata Journal, 24(1), 72-92. https://doi.org/10.1177/1536867X241233645.

- Muhbauer, A. & Spichtinger. & Lohmann, U. (2009): Application and Comparison of Robust Linear Regression Methods for Trend Estimation. Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.

-Nabillah, I. and Ranggadara, I., (2020). Mean absolute percentage error untuk evaluasi hasil prediksi komoditas laut. Journal of Information System, 5(2), pp.250-255.

- Rousseeuw, P.J. & Van Driessen, K. (1999): A Fast Algorithm for the Minimum Covariance Determinat Estimator, Technometrics, No. 41, pp. 212 – 223.

- Rousseeuw, P. J. (1984): Least Median of Squares Regression. Journal of the American Statistical Association 79: 871-880.

- Rousseeuw, P. & Yohai, V. (1984): Robust Regression by Means of S-estimates. In Robust and nonlinear time series, J. Franke, W. Hardle and R. Martin (eds.). Lecture Notes in Statistics,26, 256 - 272.

- Sahu, S.K., 2023. Simple Linear Regression Model. In Introduction to Probability, Statistics & R: Foundations for Data-Based Sciences (pp. 361-404). Cham: Springer International Publishing.

- Sanford, W. (2005): Applied Linear Regression, 3rd edition, John Wiley & Sons, Inc. Hoboken, New Jersey, Canada.

- Schumann, D.: Robust Variable Selection. Raleigh, North Carolina.http://repository.lib.ncsu.edu/ir/bitstream/1840.16/4764/1/etd.pdf 4/2/2014.

-Shareef, A.A. and Ibrahim, I.M., (2020). Comparison of mistakes criteria using multiple linear regressions applied to cotton data. Periodicals of Engineering and Natural Sciences, 8(1), pp.231-241.

- Shi, Y., 2023. Application of Improved Linear Regression Algorithm in Business Behavior Analysis. Procedia Computer Science, 228, pp.1101-1109.

- Wang, Y., Guo, Z., Zhang, Y., Hu, X. and Xiao, J., 2023. Iron Ore Price Prediction Based on Multiple Linear Regression Model. Sustainability, 15(22), p.15864.

-Yang, D., Yang, H.M., Wang, P. and Li, S.J., (2020). MSAE: a multitask learning approach for traffic flow prediction using deep neural network. In Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceedings of the 15th International Conference on IIH-MSP in conjunction with the 12th International Conference on FITAT, July 18-20, Jilin, China, Volume 1 (pp. 153-161). Springer Singapore.

- Yohai, V. J. (1987): High Breakdown Point and High Efficiency Robust Estimates for Regression. Annals of Statistics 15, 642-656.

بەراوردکردنی مۆدێلی پاڵپەرانی هێڵی و پاڵپەرانی بەهێز بە بەکارهێنانی مەرجەکانی هەڵە و زانیاری. جێبەجێکراو لەسەر نموونەی نەخۆشەکان بەکارهێنراوە

عسمه ت موسا ئیبراهیم
بەشی جیکرنا ددانا, کۆلێژا تکنیکی دهوک,
زانکۆیا بولیتکنیک دهوک- دهوک،
هەرێما کوردستانی،عیراق.
ismat.mousa@dpu.edu.krd

پوختە

بەرزی فشاری خوێن نەخۆشییەکی باو و مەترسیدارە، هەر لەبەر ئەم هۆکارەش نمونەیەک لە نەخۆشەکان لە نەخۆشخانەی فێرکاری ئازادی لە دهۆک هەڵبژێردران. لەم لێکۆڵینەوەیەدا بەراوردێک لە نێوان کەمترین چوارگۆشەی ئاسایی (OLS) لەگەڵ دوو شێوازی بەهێزی خەمڵێنەری (LTS) و خەمڵێنەری (MM) ئەنجامدرا، ئەوان بە بەکارهێنانی دوو جۆر پێوەر کە بە پێوەرەکانی هەڵە نیشان دراون (MSE، MAPE، MSAE، SAZ1، SAZ2)، و پێوەرەکانی زانیاری (AIC، BIC). پێوەرەکان ڕۆڵێکی بنەڕەتییان لە بواری ئاماردا گێڕاوە و لە هەمان کاتدا بەکارهێنراون بۆ بەدەستهێنانی کەمترین ڕێژەی هەڵەکان و هەروەها بۆ بەدەستهێنانی چارەسەرە گونجاوەکان. شێوازی خەمڵێنەری (MM) کارایی بەرزی نیشان دا لە حیسابکردنی بەهاکان بە زۆربەی پێوەرەکان، جگە لە پێوەرەکانی SAS2، کە وەک باشترین بەها تۆمارکرا بە بەکارهێنانی شێوازی خەمڵێنەری (LTS)، و پێوەرەکانی (MSE)، کە باشترین بەهای نیشان دا بە بەکارهێنانی شێوازی (OLS).

وشە سەرەکییەکان: خەمڵێنەری OLS، خەمڵێنەری LTS، خەمڵێنەری MM.

مقارنة بين الانحدار الخطي المتعدد والانحدار الحصين بأستخدام معايير الاخطاء ومعايير المعلومات و تطبيقها على عينة من المرضى

عصمت موسى أبراهيم
ق صداعة الاسدان , لة دهك الق ة,
جامعة بـ لـ دهك-دهك, ع اق
ismat.mousa@dpu.edu.krd

ملخص

ارتفاع ضغ الـ م هـ مـ ض شائع وخ ولها الـ ت اخ ار ع ة م الـ ضى مـ مـ فى آزاد الـ عل ي في دهك. في هـ ه الـ راسة ت إجـ اء مقارنة بـ الـ عات الـ غ العادة (OLS) مع ق مـ الـ ق الـ ة الـ لة قـ ر الـ عات الـ غ الـ ذمة (LTS) ومقـ ر ذ ع الاحـ ال الأق ى الـ ع ل (MM)، وت تق ها اسـ ام ذ ع مـ الـ عاي الـ لة عاي الاخـ اء (MSE، MAPE، MSAE، SAZ1، SAZ2)، ومعاي الـ عل مات (AIC، BIC). لعـ الـ عاي دورًا أساسًا في مـ ال الإحـ اء, وفي نف الـ ق ت ت قها للـ ل على أق ة للأخـ اء و لـ للـ ل على الـ الام . أ هـ ت قة مقـ ر ذ ع الاحـ ال الأق ى الـ ع لة فاءة عالة في حـ اب الـ م خلال مع الـ عاي ، اسـ اء مـ ار SAS2 والـ سـ أف ة له اسـ ام قة مقـ ر الـ عات الـ غ الـ ذمة، و لـ مـ ار (MSE) الـ دون أف ة له اسـ ام قة الـ عات الـ غ العادة.

الكلمات المفتاحية : مقدر OLS ومقدر LTS ومقدر MM.