UMT PRESS

Journal of Mathematical Sciences and Informatics
Journal Homepage: https://journal.umt.edu.my/index.php/jmsi
eISSN: 2948-3697
DOI: http://doi.org/10.46754/jmsi.2024.06.003

MATHEMATICAL MODELLING OF FACTORS FOR MEDICAL INSURANCE COST IN THE UNITED STATES USING ROBUST REGRESSION

FARAH AYUNI ABDUL HALIM AND NORIZAN MOHAMED*

Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

Corresponding author: norizan@umt.edu.my

ARTICLE INFO

Article History:
Received 26 SEPTEMBER 2023
Accepted 26 APRIL 2024
Published 5 JUNE 2024

Section Editor: Che Mohd Imran Che Taib

Keywords:
Robust Regression;
Outliers;
LTS-estimator;
MM-estimator;
S-estimator.

ABSTRACT

The rising cost of medical insurance in the United States requires a thorough understanding of the factors that influence it including age, sex, BMI, smoking habits and number of children. Problems arise when analysing data that contain outliers, as individual observations can have a large impact on results. Robust regression is one of the useful methods in decreasing the effect of outliers in modelling. Hence, this paper aims to determine the best among three estimators and to test the robustness of the best estimator when the data are contaminated with outliers. We then applied to the dataset collected from the US Census Bureau published by Brett Lantz in 2013. Using the LTS-estimator, MM-estimator and S-estimator to test the hypothesis, only one factor, BMI, had no significant effect on medical insurance data, while the remaining factors had significant effects. The findings showed that R^2 of LTS-estimator, MM-estimator and S-estimator were 0.9813, 0.6735, and 0.9728 respectively. When the data were contaminated with 10%, 20% and 30% of outliers, the R^2 values of the LTS-estimator were 0.9399, 0.9030, and 0.8678. Thus, it can be concluded that the LTS-estimator can help in producing results that are resistant to outliers.

2020 Mathematics Subject Classification:

Introduction

Of all the types of insurance available in the U.S., medical insurance is the most vital. Accidents and illnesses do not know who the victims are, therefore when they occur, one should seek medical help. Moreover, health insurance systems are investigating predictive modelling as a means of enhancing business operations and services. Many factors can affect medical insurance cost. Research shows that the inflated cost of medical insurance can result in under-insurance, where individuals cannot afford out-of-pocket expenses even with adequate insurance coverage. According to the Unhealthy Debt report [15], nearly one third of insured adults in the U.S. were under-insured, indicating that they had health insurance but still faced high out-of-pocket costs.

Rousseeuw *et al*. [16] introduced the Least Trimmed Squares (LTS) technique to insurance that can withstand most of the effects of outliers in regression analysis. The LTS estimator is a robust regression estimator used in statistical analysis. It aims to minimise the influence of outliers on the regression model [1]. When compared to the least squares (LS) technique, the LTS method provides fair estimates to the majority of the population and works well even for very small datasets. The National Science Foundations of Belgium supported Rousseeuw's research. Indeed, the LTS method is a superior strategy that can withstand high outlier effects because the LS regression methodology is not suitable for those with outliers. A study by [17], states that the S-estimator is a useful 'tool'

due to its tendency to fit the majority of the data points even though it does not tend to fit the minority of the data points. The asymptotic properties of the S-estimator are favourable in many situations. Sakata and White have shown that the S-estimator is not less efficient than the ordinary least squares (OLS) estimator in general. The proof is done by performing asymptotic and Monte Carlo simulations. Furthermore, in general, the stochastic limit of two different minimum estimator scales is the difference of the estimator itself. Empirical examples of the use or application of the S-estimator in financial time series have also been highlighted, where the use of the S-estimator can also be linked to the study of exchange rate forecasting.

Variables in the dataset used for modelling medical insurance cost include age, sex, body mass index (BMI), smoking habits and number of children. The dataset is known to have outliers, therefore, causing disturbances in the model. Important statistical methods such as regression estimation are used to analyse financial data, and data is analysed using OLS regression methods. The presence of outliers and influential observations makes LS regression analysis susceptible to the data and even produces misleading results. The robust alternatives, such as LTS estimation, Least median square (LMS) estimation, and M estimation, can withstand a certain level of contamination [1]. Consequently, the problem focuses on the use of RR to find the best estimator among the estimators, and to test the robustness of the estimator in the presence of outliers.

Data and Methodolody

The data used in this study is sourced from the US Census Bureau published by [14]. This data can be found at https://www.kaggle.com/datasets/mirichoi0218/insurance or https://github.com/stedy/Machine-Learning-with-R-datasets. Dataset consists of 1,338 records. The variables considered for this study are the dependent variable (y) and the independent variable (). The dependent variable is charges. The independent variable consisted of five variables, namely age (x_1), sex (x_2), BMI (x_3), smoker (x_4) and children (x_5). The description of the dataset is presented in Table 1.

Table 1: Dataset description

Cod	Variables	Description
x_1	Age	Age is a crucial factor in healthcare.
x_2	Sex	Gender (Male = 1, Female = 0)
x_3	Body mass index (BMI)	Body weight (kg/m^2)
x_4	Children	Number of children covered by health insurance
x_5	Smoker	Smoking status (Smoker = 1, Non-smoker = 0)
y	Charges	Individual medical costs billed by health insurance

The technique used in this study is robust regression. Although there are four estimators in robust regression, the main focus here is on the three estimators which are the LTS-estimator, MM-estimator, and S-estimator

Regression Analysis

Regression analysis is a valuable statistical method for modelling the functional relationship between dependent and independent variables, provided certain regularity conditions are satisfied [2]. There are several types of regression analysis but the most widely used model is linear regression. The linear regression can be expressed in terms of matrices and vectors as

$$y = X\beta + \varepsilon \tag{1}$$

In this model, y is the response vector X while is the design matrix, and β is the vector of parameters with error vector. All these notations can be expressed as

$$Y_i = \beta_0 + \beta_1 Xi_1 + \beta_2 Xi_2 + \cdots + \beta_j X_{ij} + \varepsilon_i, \text{ for } j=1,2,...,p; i=1,2,...,n \quad (2)$$

for any observations i.

Basically, linear regression can be classified into two types, simple linear regression and multiple linear regression. The difference between those two linear regressions is based only on the independent variable. Simple linear regression examines the relationship between only one independent variable and one dependent variable, while multiple linear regression examines the relationship between two or more independent variables and one dependent variable.

The OLS is the most widely used for estimating coefficients of linear regression equations. The best linear unbiased estimators have always been OLS estimators of parameters but if the dataset contains outliers, the LS estimates may be affected [3]. Thus, we used a robust approach to overcome the weaknesses of the OLS regression model [4].

Outlier Analysis

Outliers, which are observations that are inconsistent and deviate significantly from the majority of observations in the data, pose a serious threat to the regression model and its estimated coefficients, resulting in misleading results that should be properly handled. Data points that are far apart from other data points are known as outliers. In short, they are unusual values in a dataset. Outliers in a dataset are quite common due to some factors such as data entry error, sampling problems, or instrument error. Outliers are also often unrecognised because most of the data are processed by computers these days without careful inspection or screening [5]. Outlier identification is a part of data screening that should be performed on a regular basis before statistical analysis [6]. Identifying data contaminated by outliers is very crucial since it will affect the validity of the results. There are three diagnostic methods to identify outliers, the Robust distance, the Mahalanobis distance, and the Cook's distance. The diagnostic is crucial for identifying outliers and providing resistant outcomes in the presence of outliers [7].

Robust Regression

Mathematical modelling provides a systematic approach to understanding the relationships between these factors and medical insurance cost. Robust regression is a statistical technique used to analyse data in a way that is resistant to outliers and other sources of noise. By employing robust regression, researchers can develop models that capture the nonlinear relationships and interactions between factors, while minimising the influence of extreme values. This technique ensures more reliable and accurate estimates of the factors contributing to medical insurance cost. When there are predicted outliers that affect the model or whenever the residual distribution is not normal, robust regression is implemented [8]. To model medical insurance cost in this study, we applied the LTS-estimator, MM-estimator, and S-estimator.

(a) LTS-estimator

First estimator, introduced by Rousseeuw, was the LTS-estimator where this method also tries to eliminate the possible outliers. According to [9], the LTS will be similar to OLS if the exact number of outlying data points are trimmed. The data will be excluded from the computation if there is more trimming than the outlying data points. Otherwise, this method is not as efficient. According to [10], LTS-estimator can be defined as

$$\hat{\beta}_{LTS} = \hat{\beta} \sum_{i=1}^{h} r_i^2 \tag{3}$$

with the h value obtained from

$$h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right] \tag{4}$$

where;

r_i^2 : the ordered square residuals from smallest to largest

$$r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$$

n : sample size

p : the number of parameters

Below is the algorithm of the LTS-estimator [18]:

1: By using OLS, estimate the regression coefficients on the data

2: Compute the residuals, $r_i = y_i - \hat{y}_i$

3: Compute r_i^2 and value of $h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$

4: Compute the estimated value of $\hat{\beta}_{LTS}$

5: Estimate the parameter of from $b_{new(i)}$ from $h_{new(i)}$

6: Determine from r_i^2 from $h_{new(i)}$

7: Compute the estimated value of $\hat{\beta}_{LTS(new)}$

8: Repeat steps 5 to 7 until convergent value of $\hat{\beta}_{LTS}$ is obtained

(b) MM-estimator

Next estimator is the MM-estimator. This method is frequently used in robust regression, where it is the combination of two estimators – M-estimator and S-estimator – in order to achieve high efficiency and high robustness. The goal of estimation is to produce estimates with a high breakdown value that are more efficient [8]. MM-estimator is also called a modified M-estimator and can be defined as

$$\sum_{i=1}^{n} \rho'(u_i) x_{ij} = \sum_{i=1}^{n} \rho' \left(\frac{y_i - \sum_{j=0}^{k} x_{ij} \hat{\beta}_j}{S_{MM}} \right) x_{ij} = 0 \tag{5}$$

with $y_i - \sum_{j=0}^{k} x_{ij} \beta_j$ is the regression model's residual from the estimation of its parameters

where stands for the standard deviation obtained from S estimator's residual, while refers to the objective function of Tukey's Biweight:

$$\rho(u_i) = \{ \frac{{u_i}^2}{2} - \frac{{u_i}^4}{2c^2} + \frac{{u_i}^6}{6c^2}, \ -c \leq |u_i| \leq c \ \frac{c^2}{6} u_i < -c \ or \ u_i > c \tag{6}$$

Below is the algorithm of MM-estimator based on Susanti:

1: By using OLS, estimate the regression coefficients on the data

2: Test the assumptions of the classical regression model

3: Determine whether the data contain outliers

4: Compute the residuals, $e_i = y_i - \hat{y}_i$ of S-estimation

5: Compute the value of $\hat{\sigma}_i = \hat{\sigma}_{sn}$

6: Compute the value $u_i = \dfrac{e_i}{\hat{\sigma}_i}$

7: Compute the weighted value, w_i

$$w_i = \{\left[1 - \left(\frac{u_i}{4.685}\right)^2\right]^2, \qquad |u_i| \leq 4.685\ 0, \qquad |u_i| > 4.685$$

8: Compute $\hat{\beta}_M M$ using WLS method with weighted

9: Repeat steps 5 to 8 until convergent value of $\hat{\beta}_M M$ is obtained

10: Check whether the independent variables significantly affect the dependent variable

(c) S-estimator

S-estimator is used to obtain robust predictions as an alternative to the common LS method. It was introduced by [11] and it is related to M-estimator since S-estimator minimised the residual scale of M-estimator. According to [8], M-estimator only uses the median as the weighted value, which would be a problem since it will cause a lack of consideration for the data distribution, and it is not the overall data function. S-estimator can be defined by
with determining the minimum robust scale estimator $\hat{\sigma_s}$ and satisfying [8]

$$min \sum_{i=1}^{n} \quad \rho\left(\frac{y_i - \sum_{i=1}^{n} \quad x_{ij}\beta}{\hat{\sigma}_s}\right) \tag{7}$$

where

$$\hat{\sigma}_s = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} \quad w_i {e_i}^2} \tag{8}$$

$K = 0.199, w_i = w_\sigma(u_i) = \dfrac{\rho(u_i)}{{u_i}^2}$, and the initial estimate is

$$\hat{\sigma}_s = \frac{median|e_i - median(e_i)|}{0.6745}$$

Next, we get the solution by differentiating to β, so that

$$\sum_{i=1}^{n} \quad x_{ij}\psi\left(\frac{y_i - \sum_{i=1}^{k} \quad x_{ij}\beta}{\hat{\sigma}_s}\right) = 0, j = 0,1, \ldots, k \tag{9}$$

ψ is the derivative of ρ:

$$\psi(u_i) = \rho(u_i) = \{u_i\left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, \qquad |u_i| \leq c\ 0, \qquad |u_i| > c$$

where w_i is an Iteratively Reweighted Least Square (IRLS) weighted function:

$$w_i(u_i) = \{u_i \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, \quad |u_i| \leq c\ 0, \quad |u_i| > c$$

$u_i = \frac{e_i}{\sigma_s}$ and $c = 1.547$. By using the IRLS method, we can solve the equation (9).

Below is the algorithm of S-estimator based on Susanti:

1: Compute $\hat{\beta}^0$ with OLS

2: Compute the residuals, $e_i = y_i - \hat{y}_i$

3: Compute the value of $\hat{\sigma}_i$

$$\hat{\sigma}_i = \{\frac{median|e_i - median(e_i)|}{0.6745}, \quad iteration = 1 \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i {e_i}^2}, \quad iteration > 1$$

4: Compute the value e $u_i = \frac{e_i}{\sigma_s}$

5: Compute the weighted value, w_i

$$w_i = \{\{\left[1 - \left(\frac{u_i}{1.547}\right)^2\right]^2, \quad |u_i| \leq 1.547\ 0, \quad |u_i| > 1.547, \quad \&iteration = 1\ \frac{\rho(u)}{u^2}, \quad \&iteration > 1$$

6: Compute $\hat{\beta}_i$ with the IRLS method with weighted w_i

7: Repeat steps 2 to 5 until the convergent value of $\hat{\beta}_s$ is obtained

Breakdown Point

Breakdown point is one of the most crucial characteristics for robust regression estimators. The breakdown point is the point or limit of the percentage of contamination in data at which any test statistic becomes swamped for the first time [9]. The larger the sample size n, the smaller the breakdown point because the smallest breakdown point is 1/n which tends to be 0%. The robustness of an estimator is proportional to its breakdown point. The highest possible breakdown point is 50% which is also referred as the percent of outliers in the dataset. It means that if a robust estimation technique has a 50% breakdown point, then 50% of the data might have outliers but the coefficients would still be usable [3].

R-squared

R-squared is computed to measure the goodness-of-fit for linear regression models. R-squared is also known as the coefficient of determination, or multiple correlation coefficient. It is similar to the correlation coefficient since it is the only square value of the correlation coefficient as has been proven [12]. It can take any value between 0 to 1 while the correlation coefficient can take any value from -1 to 1. The coefficient of determination can be used for multiple explanatory variables where it is not limited to cases with only one explanatory variable as has been proven [13]. The formula of R-squared is shown below

$$R^2 = \frac{SS_{regression}}{SS_{total}} \tag{10}$$

where
$SS_{regression} = \sum_{i=1}^{n} \left(\underline{y} - \hat{y}_i{}^2\right)$ and $SS_{total} = \sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2$

The higherst value of R-squared indicates a better fit for the model.

Result and Discussion

This part discusses the results of the best estimator robust regression and tests on the robustness of the best estimator when the data are contaminated with outliers.

(a) LTS-estimator

Table 2 below illustrates the parameter estimates for the LTS-estimator, including the estimated value of the parameters, standard error, and Pr > ChiSq value using SAS software. These results demonstrate that only four independent factors were significant with a p-value of less than 0.05, which are age (x_1) , sex (x_2) , number of children (x_4) and smoker (x_5). BMI (x_3) was not significant since the p-value was greater than 0.05. The parameter estimates ranged from -435.2790 to 32939.4800, except for the intercept estimate value. The standard error values of these variables were 1.4355, 40.2158, 3.4321, 16.5633, and 70.5631 respectively.

Based on the data in Table 2, a fitted model for LTS-estimator with the R-squared value, 0.9814 was obtained as follows:

$$\hat{y} = -3769.0300 + 267.9288x_1 - 435.2790x_2 - 0.1583x_3 + 440.5335x_4 + 32939.4800x_5$$

Table 2: Parameter estimates of LTS-estimator

Parameter Estimates			
AIC			
BIC			
Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-3769.0300	117.4967	< 0.0001
Age	267.9288	1.4355	< 0.0001
Sex_Code	-435.2790	40.2158	< 0.0001
BMI	-0.1583	3.4321	0.9632
Childr en	440.5335	16.5633	< 0.0001
Smoker_Code	32939.4800	70.5631	< 0.0001

Since the BMI is not significant, it is removed. Table 3 shows the parameter estimates of the LTS-estimator after BMI was removed.

Table 3: Parameter estimates of LTS-estimator (BMI was removed)

Parameter Estimates			
	AIC		
	BIC		
Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-3769.0200	66.3133	< 0.0001
Age	267.8745	1.4468	< 0.0001
Sex_Code	-446.0450	40.7832	< 0.0001
Children	442.8682	16.8046	< 0.0001
Smoker_Code	32980.0700	69.6090	< 0.0001

After BMI was removed, the R-squared value changed to 0.9813. The robust LTS-estimator regression model is obtained as follows:

$$\hat{y} = -3769.0200 + 267.8745x_1 - 446.0450x_2 + 442.8682x_4 + 32980.0700x_5$$

From Figure 1, it was found that 272 out of 1,338 observations were identified as outliers by LTS-estimator since the standardised robust residual exceeded the cut off value. It also shows that there were 299 leverage points in the data because the standardized robust residual exceeded the cut off value from Table 4.

Table 4: Diagnostics summary of LTS-estimator

Diagnostic Summary		
Observation Type	**Proportion**	**Cut Off**
Outlier	0.2033	3.0000
Leverage	0.2235	3.0575

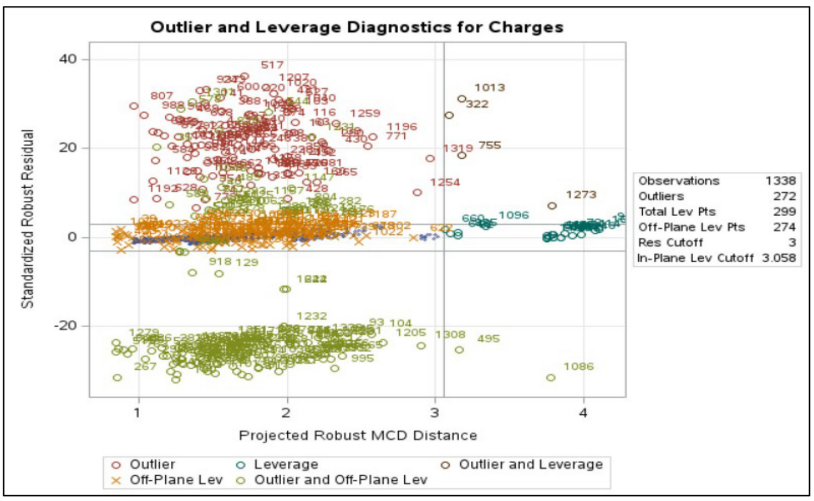

Figure 1: Outlier and leverage diagnostics for charges (LTS-estimator)

The Q-Q plot presented in Figure 2 illustrates that the residual has a heavy tail distribution.

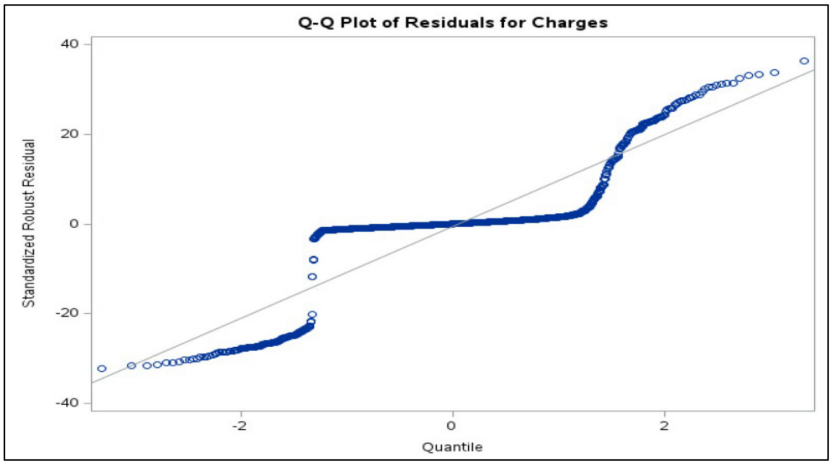

Figure 2: Plot of residuals for charges (LTS-estimator)

(b) *MM-estimator*

Table 5 below shows the parameter estimates for the MM-estimator, consisting of the estimated value of the parameters, standard error, and Pr > ChiSq value using SAS software. This table shows that age, sex, children, and smoker were the only factors that were significant since the p-value was less than 0.05, however BMI is not significant because the p-value was greater than 0.05. The parameter estimates ranged from -461.6320 to 33363.2000, except for the intercept estimate value. The standard error values for these variables were 1.5892, 44.4582, 3.7930, 18.3175, and 74.6738 respectively.

Table 5: Parameter estimates of MM-estimator

Parameter Estimates **AIC** **BIC**			
Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-3955.2700	129.8485	< 0.0001
Age	268.1284	1.5892	< 0.0001
Sex_Code	-461.6320	44.4582	< 0.0001
BMI	5.7658	3.7930	0.1285
Children	444.2160	18.3175	< 0.0001
Smoker_Code	33363.2000	74.6738	< 0.0001

The MM-estimator regression model with the R-squared value, 0.6737 is obtained as follows:

$$\hat{y} = -3955.2700 + 268.1284x_1 - 461.6320x_2 + 5.7658x_3 + 444.2160x_4 + 33363.2000x_5$$

Due to the insignificance of the BMI, it is eliminated. Table 6 shows the parameter estimates of MM-estimator following the exclusion of BMI.

Table 6: Parameter estimates of MM-estimator (BMI was removed)

Parameter Estimates AIC BIC			
Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-3790.5800	72.0238	< 0.0001
Age	268.3929	1.5716	< 0.0001
Sex_Code	-459.8940	44.2558	< 0.0001
Children	444.5126	18.2347	< 0.0001
Smoker_Code	33375.0700	73.1318	< 0.0001

By excluding BMI, the R-squared value changed to 0.6735. The robust MM-estimator regression model is obtained as follows:

$$\hat{y} = -3790.5800 + 268.3929x_1 - 459.8940x_2 + 444.5126x_4 + 33375.0700x_5$$

Figure 3 shows that 242 out of 1,338 observations were identified as outliers by MM-estimator because the standardised robust residual exceeded the cut off value. It also shows that the data contain 299 leverage points. If the value of the leverage points has standardised robust residual greater than the cut off in Table 7, this indicates that the leverage point is otherwise a bad point.

Table 7: Diagnostics summary of MM-estimator

Diagnostic Summary		
Observation Type	**Proportion**	**Cut Off**
Outlier	0.1809	3.0000
Leverage	0.2235	3.0575

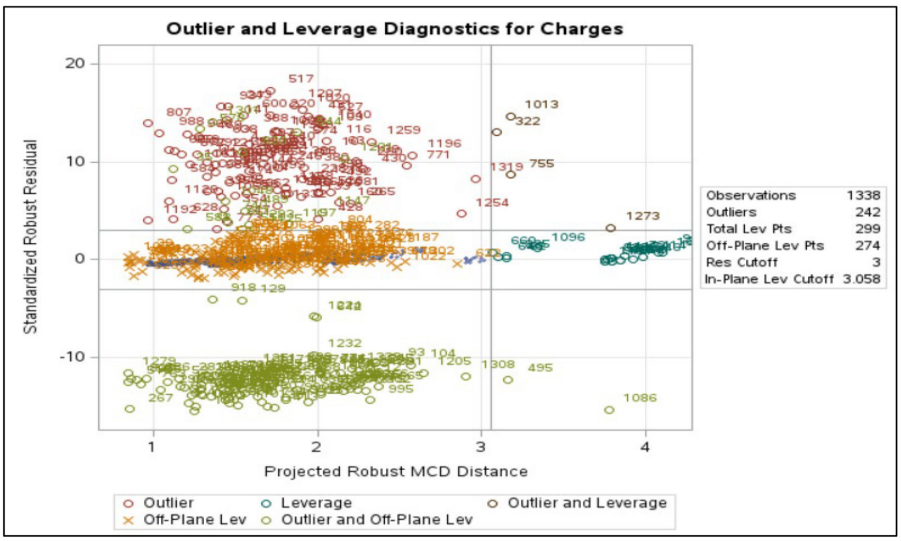

Figure 3: Outlier and leverage diagnostics for charges (MM-estimator)

The Q-Q plot presented in Figure 4 illustrates that the residual has a heavy tail distribution.

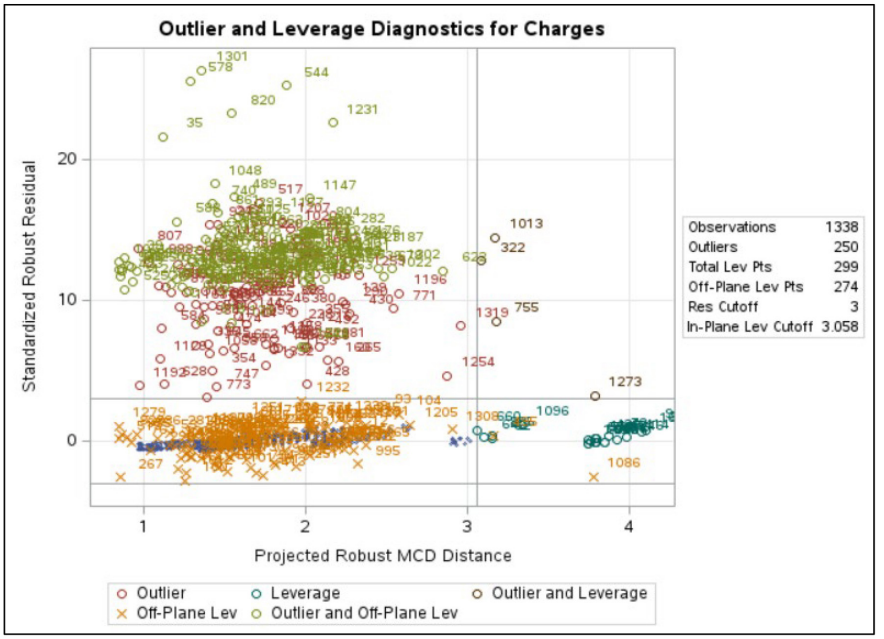

Figure 4: Plot of residuals for charges (MM-estimator)

(c) S-estimator

Table 8 below shows the parameter estimates for S-estimator, as well the standard error, and Pr > ChiSq value using SAS software. It shows that age, sex, children and smoker were the only factors that were significant since the p-value was less than 0.05, however BMI is not significant because the p-value was greater than 0.05. The parameter estimates ranged from -463.1080 to 33248.5300, except for the intercept estimate value. The standard error values for these variables were 1.5504, 43.2477, 3.6869, 17.8181, and 74.4268 respectively.

Table 8: Parameter estimates of S-estimator

Parameter Estimates **AIC** **BIC**			
Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-3848.9900	126.3861	< 0.0001
Age	267.8179	1.5504	< 0.0001
Sex_Code	-463.1080	43.2477	< 0.0001
BMI	2.8180	3.6869	0.4447
Children	439.4792	17.8181	< 0.0001
Smoker_Code	33248.5300	74.4268	< 0.0001

The robust S-estimator regression model with the R-squared value, 0.9727 is obtained as follows:

$$\hat{y} = -3848.9900 + 267.8179x_1 - 463.1080x_2 + 2.8180x_3 + 439.4792x_4 + 33248.5300x_5$$

Due to the insignificance of the BMI, it is eliminated. Table 9 shows the parameter estimates of MM-estimator following the exclusion of BMI.

Table 9: Parameter estimates of S-estimator (BMI was removed)

Parameter Estimates			
AIC			
BIC			
Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-3722.5200	70.8663	<0.0001
Age	266.8869	1.5514	<0.0001
Sex_Code	-421.9080	43.0990	<0.0001
Children	417.1979	17.7814	<0.0001
Smoker_Code	14113.3500	74.2525	<0.0001

By excluding BMI, the R-squared value changed to 0.9709 and the robust S-estimator regression model is obtained as follows:

$$\hat{y} = -3722.5200 + 266.8869x_1 - 421.9080x_2 + 417.1979x_4 + 14113.3500x_5$$

Figure 5 shows that 250 out of 1,338 observations were identified as outliers by S-estimator because the standardised robust residual exceeded the cut off value. It also shows that the data contain 299 leverage points. If the leverage points has standardized robust residual greater than the cut off in Table 10, this indicates that the leverage point is otherwise a bad point.

Table 10: Diagnostics Summary of S-estimator

Diagnostic Summary		
Observation Type	**Proportion**	**Cutoff**
Outlier	0.1868	3.0000
Leverage	0.2235	3.0575

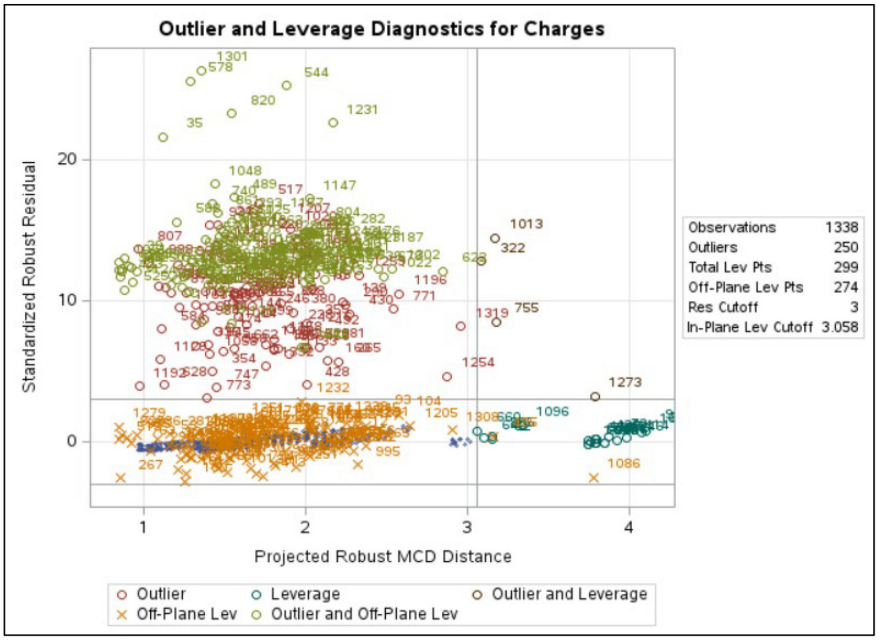

Figure 5: Outlier and leverage diagnostics for charges (S-estimator)

The Q-Q plot presented in Figure 6 illustrates that the residual has a heavy tail distribution.

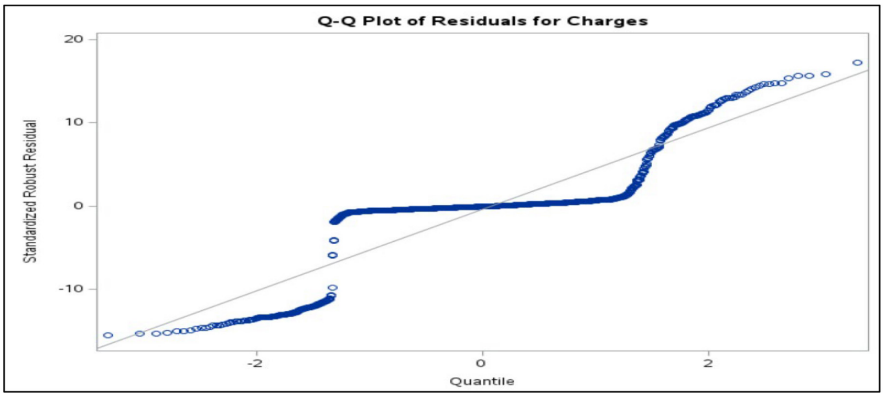

Figure 6: Plot of residuals for charges (S-estimator)

Comparison of R-squared Value Among Robust Regression Estimators

The comparison of R-squared value has been made in order to determine the best estimator among LTS-estimator, MM-estimator and S-estimator. Table 11 below illustrates the value of R-squared generated by SAS software for those estimators. It can be seen that the R-squared value of all estimators was more than 0.5000, or 50%, which means they were significant and there was a relationship between the dependent and independent variables. As shown in Table 11, the result indicates that LTS-estimator is the best estimator and fits the model the best compared with MM-estimator and S-estimator. This is because LTS-estimator shows the highest value of R-squared even after BMI was removed to achieve significant data.

Table 11: R-squared value of robust regression estimator

Estimator	R-squared
LTS-estimator	0.9814
LTS-estimator (BMI was removed)	0.9813
MM-estimator	0.6737
MM-estimator (BMI was removed)	0.6735
S-estimator	0.9727
S-estimator (BMI was removed)	0.9709

Contamination of the Data

From section 3.4, LTS-estimator is the best model among three estimators. In this section, we would like to test the robustness of LTS-estimator when the data is contaminated with different percentages of contamination in order to achieve our second objective. This study uses 10%, 20%, and 30% of contamination from the total observations. Table 12 shows the percentages of contamination of the observations.

Table 12: The number of observations contaminated with different percentages of contamination

Percentages of Contamination (%)	The Number of Contaminated Observations
10	134
20	268
30	401

(i) LTS-estimator

LTS has been applied to the contaminated dataset with different percentages of contamination. Below is Table 13 which shows the results of the parameter estimates by SAS software. All the independent variables are still significant when contaminated with different percentages of contamination since p-value is less than 0.05. The value of standard error when the data are contaminated and uncontaminated are not too far from each other, which indicates that the model is still robust despite being contaminated with the outliers.

Table 13: Parameter estimates of contaminated data of LTS-estimator

Percentages (%)	Parameter	Estimate Parameter (p-value)	Standard Error	Pr > ChiSq
10	Intercept	-3821.4800	74.9883	< 0.0001
	Age (x_1)	268.6189	1.6553	< 0.0001
	Sex_Code (x_2)	-425.9820	47.0635	< 0.0001
	Children (x_4)	458.2468	18.4184	< 0.0001
	Smoker_Code (x_5)	33394.0700	75.9781	< 0.0001
20	Intercept	-3200.3100	293.9565	< 0.0001
	Age (x_1)	280.0224	6.5607	< 0.0001
	Sex_Code (x_2)	-505.5900	186.5728	< 0.0001
	Children (x_4)	-261.1060	49.7649	< 0.0001
	Smoker_Code (x_5)	9489.665	241.0167	< 0.0001
30	Intercept	1051.7220	454.9050	< 0.0001
	Age (x_1)	180.1377	10.0655	< 0.0001
	Sex_Code (x_2)	-74.4188	291.4115	< 0.0001
	Children (x_4)	-110.2710	71.2626	< 0.0001
	Smoker_Code (x_5)	2634.6430	331.6557	< 0.0001

(ii) R-squared Value of Contamination Data

The difference in the R-squared value between the contaminated and uncontaminated data is shown in Table 14. It can be seen that these R-square values remain relatively close to uncontaminated R-squared values even when contaminated to varying percentages. When the data were contaminated with 10%, 20%, and 30% of outliers, the R-squared values of LTS-estimator were 0.9395, 0.8992, and 0.8343 respectively. Even though there are outliers, the LTS-estimator is still able to maintain its robustness.

Table 14: Comparison of R-squared value of contamination data

Percentage of Contamination (%)	R-squared Value
Uncontaminated	0.9813
10	0.9395
20	0.8992
30	0.8343

Conclusion

This study attempted to observe the medical insurance data in the U.S. using different estimators, and to test the robustness of the estimator when the data were contaminated with outliers. Three estimators were compared, LTS-estimator, MM-estimator and S-estimator. It can be seen that LTS-estimator is the best estimator compared with MM-estimator and S-estimator since it shows the highest value of R-squared and lowest value of standard error. Since LTS-estimator outperformed MM-estimator and S-estimator in modelling medical insurance U.S. data, the LTS-estimator was tested for robustness, and it was proven that the LTS-estimator maintained its robustness and showed high R-squared values.

This study has room for improvement.Additional research might include developing a hybrid model to produce a better model to the robust regression. On medical insurance data, researchers could compare other estimators like M-estimator and LMS-estimator to see which one produces a better result. In addition, other methods such as Ridge Regression, ARIMA, Quantile Regression, Multiple Linear Regression, neural network, and any suitable methods could be applied in this study. Besides, other software could be utilised on those estimators to analyse or provide for robust statistics such as Python, R programming and MATLAB. Finally, it is recommended that the robust regression approach be applied in forecasting other data where it is known that the data is contaminated.

Acknowledgement

The authors would like to thank all reviewers for their comments and suggestions for the improvement of this manuscript.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

[1] Berenguer-Rico, V., Johansen, S., & Nielsen, B. (2023). A model where the Least Trimmed Squares estimator is maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(3), 886-912.

[2] Kula, K. S., Tank, F., & Dalkilic, T. E. (2012). A study on fuzzy robust regression and its application to insurance. *Mathematical and Computational Applications*, *17*(3), 223-234.

[3] Gad, A. M., & Qura, M. E. (2016). Regression estimation in the presence of outliers: A comparative study. *International Journal of Probability and Statistics*, *5*(3), 65-72.

[4] Mahmudah, U., Chamdani, M., Tarmidzi, T., & Fatimah, S. (2020). Robust regression for estimating the impact of student's social behaviors on scientific literacy. *Jurnal Cakrawala Pendidikan*, *39*(2), 293-304.

[5] Blatna, D. (2006). Outliers in regression. *Trutnov*, *30*, 1-6.

[6] Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., & Kavsek, B. (2000). Informal identification of outliers in medical data. In *Fifth international workshop on intelligent data analysis in medicine and pharmacology* (Vol. 1, pp. 20-24).

[7] Aleng, N. A., Naing, N. N., Mohamed, N., & Mokhtar, K. (2017). Outlier detection based on robust parameter estimates. *International Journal of Applied Engineering Research*, *12*(23), 13429-13434.

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

[9] Alma, O. G. (2011). Comparison of robust regression methods in linear regression. *International Journal of Contemporary Mathematical Science*, 6(9), 409-421.

[10] Andriany, C. D., & Susanti, Y. (2021). Estimasi parameter regresi robust dengan metode estimasi Least Trimmed Squares (LTS) pada kemation ibu di Indonesia. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2021*, *20 Maret 2021* (pp. 9-14).

[11] Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In W. H. Franke, & R. D. Martin (Eds), *In robust and nonlinear time series analysis.* (pp. 256- 272). New York: Springer Verlag.

[12] Glen, S. (2021). Linear regression: Simple steps, video. find equation, coefficient, slope. *Statistic How To.* https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regressionequation

[13] Kasuya, E. (2018). On the use of r and r squared in correlation and regression. *Ecological Research*, *34*(1), 235-236.

[14] Brett, L. (2013). *Machine learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Packt Publishing.

[15] Friedman, J., Bridlington, E., Guarino, M., & Fisher, C. (2021). Unhealthy Debt: Medical costs and bankruptcies in Oregon (pp. 1-28). OSPIRG: Prontier Groop.

[16] Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1*(1), 73-79.

[17] Sakata, S., & White, H. (2001). S-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics*, *103*(1-2), 5-72.

[18] Zuo, Y., & Zuo, H. (2023). Least sum of squares of trimmed residuals regression. *Electronic Journal of Statistics*, *17*(2), 2416-2446.