Jurnal Teknologi

Full paper

Robust Weighted Least Squares Estimation of Regression Parameter in the Presence of Outliers and Heteroscedastic Errors

Bello Abdulkadir Rasheed[a*], Robiah Adnan[a], Seyed Ehsan Saffari[b], Kafi dano Pati[a]

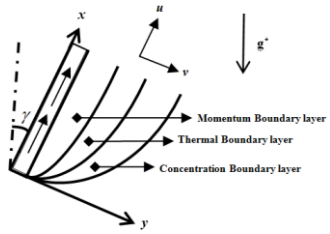
[a]*Department of mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*
[b]*Centre of Education, Sabzevar University of Medical Sciences, Sabzevar, Iran*

*Corresponding author: arasheedbello@yahoo.com

Article history

Received :2 February 2014
Received in revised form :
3 August 2014
Accepted :15 October 2014

Graphical abstract



Abstract

In a linear regression model, the ordinary least squares (OLS) method is considered the best method to estimate the regression parameters if the assumptions are met. However, if the data does not satisfy the underlying assumptions, the results will be misleading. The violation for the assumption of constant variance in the least squares regression is caused by the presence of outliers and heteroscedasticity in the data. This assumption of constant variance (homoscedasticity) is very important in linear regression in which the least squares estimators enjoy the property of minimum variance. Therefor e robust regression method is required to handle the problem of outlier in the data. However, this research will use the weighted least square techniques to estimate the parameter of regression coefficients when the assumption of error variance is violated in the data. Estimation of WLS is the same as carrying out the OLS in a transformed variables procedure. The WLS can easily be affected by outliers. To remedy this, We have suggested a strong technique for the estimation of regression parameters in the existence of heteroscedasticity and outliers. Here we apply the robust regression of M-estimation using iterative reweighted least squares (IRWLS) of Huber and Tukey Bisquare function and resistance regression estimator of least trimmed squares to estimating the model parameters of state-wide crime of united states in 1993. The outcomes from the study indicate the estimators obtained from the M-estimation techniques and the least trimmed method are more effective compared with those obtained from the OLS.

Keywords: Robust estimation; robust weighted least squares; robust least trimmed squares; heteroscedasticity; outliers

Abstrak

Dalam model regresi linear, kaedah kuasa dua terkecil (OLS) dianggap kaedah terbaik untuk menganggar parameter regresi jika andaian dipenuhi. Walau bagaimanapun, jika data tidak memenuhi andaian asas, keputusan akan terpesong. Andaian varians malar dalam regresi kuasa dua terkecil tidak dipenuhi disebabkan oleh kehadiran titik terpencil dan heteroskedastisiti dalam data. Andaian varians malar (homoskedastik) adalah sangat penting dalam regresi linear di mana penganggar kuasa dua terkecil mempunyai varians yang minimum. Oleh itu kaedah regresi teguh diperlukan untuk mengendalikan masalah titik terpencil dalam data. Walau bagaimanapun, kajian ini akan menggunakan teknik wajaran kuasa dua terkecil (WLS) untuk menganggar parameter pekali regresi apabila andaian varians ralat tidak dipenuhi dalam data. Anggaran WLS adalah sama seperti menjalankan OLS dalam prosedur transformasi pembolehubah. Pengaggar WLS dengan mudah boleh dipengaruhi oleh titik terpencil. Untuk mengatasi perkara ini, kami telah mencadangkan satu teknik yang kuat untuk anggaran parameter regresi dalam kewujudan heteroskedastik dan titik terpencil. Di sini kita menggunakan regresi teguh M- anggaran berdasarkan lelaran wajaran kuasa dua terkecil (IRWLS) daripada Huber dan Tukey Bisquare fungsi dan pengaggar rintangan r regresi kuasa dua terkurang untuk menganggar parameter model bagi data jenayah di seluruh Amerika Syarikat pada tahun 1993. Hasil daripada kajian menunjukkan penganggar yang diperolehi daripada teknik-teknik M- anggaran dan kaedah kuasa dua terkurang adalah lebih berkesan jika dibandingkan dengan yang diperoleh daripada OLS .

Kata kunci: Menunjukkan anggaran; menunjukkan wajaran kuasa dua; menunjukkan kuasa dua terkurang; heteroskedastisiti; titik terpencil

1.0 INTRODUCTION

In classical linear regression analysis the ordinary least squares (OLS) method is generally used to estimate the parameter of the regression model due to its optimal properties and straightforward computation. There are several assumptions that have to be possessed in making the OLS estimators very attractive and valid. One of the assumptions within the OLS regression model is the assumption of homoscedasticity which is rather severe. Researchers frequently encountered difficult situations where the variance from the respond variable relates to the values of a number of independent variables, leading to heteroscedasticity [1], [2].

In this type of situation, the variance of a model according to the explanatory variables can produce weights for the weighted least squares estimator [2], [3], [4]. Weighted least squares, is a special case from the generalized least squares estimator, is optimal when the structure of heteroscedasticity error variance is known. But unfortunately usually, the structure of heteroscedasticity error variance is not known in advance. For that situation, researchers may use estimated generalized least squares [3], [4].

In the existence of heteroscedasticity, the OLS estimators remained unbiased. However, probably the most harmful consequence of heteroscedasticity in regression model would be that the OLS estimator from the parameter covariance matrix (OLSCM), whose elements in the diagonal are utilized to estimate the standard errors of the regression coefficients, becomes biased and unreliable [4], [5]. As a result, the t-tests for individual coefficients are generally too liberal or too conservative, with respect to the type of heteroscedasticity. White [4] and Rana *et al.* [6] suggested a heteroscedasticity consistent covariance matrix (HCCM) to resolve the inconsistency problem from the estimator.But there is evidence that with a couple of outliers this could make all the estimation and methods meaningless [5], [6], [7]. In the existence of outliers we possess some robust approaches for the recognition of heteroscedasticity [6], [7].

However we do not have enough robust techniques obtainable in the literature for the estimation of parameters in the existence of outliers and heteroscedasticity error variance. Although heteroscedasticity does not cause any biasness problem towards the OLS estimators, the OLS may be easily affected by the presence of outliers. The weighted least squares also suffer exactly the same problems in the existence of outliers and can produce a huge interpretive issue in the estimation method [6], [7], [8]. In most cases, no estimation techniques work effectively unless of course we eliminate the influence of outliers in a heteroscedastic regression model.

This problem inspires us to build a new and better estimation technique, that provide resistant result when heteroscedasticity and outliers happened at the same time. In this study the OLS regression estimation method will be compared with the robust regression methods of M-estimate based on Huber weighted function and tukey bisquare function and the resistant regression estimator of least trimmed squares. We expect the recommended methods could be less responsive to outliers and simultaneously have the ability to remedy the problem of heteroscedasticity

2.0 METHODOLOGY OF HETEROSCEDASTIC REGRESSION MODEL

Consider the following classical linear regression model

$$y = X\beta + \varepsilon \quad (1)$$

where y is the usual $n\times1$ vector of the observed dependent values, X is the $n\times p$ matrix of the predictor variables including the intercept, β is a $p\times1$vector of regression parameters, and ε is the $n\times1$ vector of errors. The errors are assumed to be normally distributed, with mean 0 and constant variance σ^2.
The estimator of OLS regression coefficients is

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2)$$

with the variance matrix giving by

$$\mathbf{var}(\hat{\boldsymbol{\beta}})=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (3)$$

Where $E(\varepsilon\varepsilon^{T}) = \Omega,$ is a positive definite matrix. Equation (3) simplifies to the following:

$$\mathrm{var}(\hat{\beta})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (4)$$

If the errors are homoscedastic, then is $\mathrm{var}(\hat{\beta})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .but if the errors are heteroscedastic, the parameter Ω become

$\Omega=\sigma^2 V$, and equation (3) becomes

$$\mathrm{var}(\hat{\beta})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (5)$$

The mentioned problem above can be overcome by transforming our model to a new set of observations that fulfilled the underlying standard assumptions of least squares. Then the OLS is used on the transformed data. Since the covariance matrix of the errors is denoted by $\sigma^2 v$, Then υ must be non-negative and non-singular definite, then

$$\boldsymbol{\beta}_{\mathbf{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{y} \quad (6)$$

is the estimate of the generalized least squares (GLS) of β. When the errors ε have unequal variances and are uncorrelated, the covariance matrix of ε is written as

$$\sigma^2 V = diag[1/w_i] \quad , i = 1, 2, ..., n$$

Consequently, the GLS now becomes a solution to the heteroscedastic model. Assuming we defined $W = V^{-1}$, as a the diagonal matrix with diagonal elements or weights $\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n$. From equation (6), the estimator of weighted least squares is now written as

$$\mathbf{WLS} = (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{Wy} \quad (7)$$

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathbf{WLS}})=\boldsymbol{\sigma}^2_{\mathbf{WLS}}(\mathbf{X}'\mathbf{WX})^{-1}、 \quad (8)$$

Where

$$\boldsymbol{\sigma}^2\mathbf{WLS}=\frac{\sum \mathbf{W_i}^2\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}^2}{\mathbf{n-p}} \quad (9)$$

If the error structure of heteroscedastic in the regression model is known, the computation of weights W matrix is simple, and consequently the WLS regression serve a good solution of the heteroscedastic model. Unfortunately, even though in practice, the hetroscedastic error structure of the regression model is unknown.

3.0 ESTIMATION OF ROBUST WEIGHTED LEAST SQUARES REGRESSION (RWLSR)

Robust regression analysis provides an alternative choice to a least squares regression when fundamental assumptions are unfulfilled while using the character within the data [8].The qualities of efficiency, breakdown, and leverage points are broadly-knowledgeable about define robust techniques performance within the theoretical sense. One justification for robust estimators may well be a highest finite sample breakdown point $\grave{\varepsilon}^*_n$ defined by [9], [10], [11]. The breakdown point may be defined as the smallest percentage of contaminated data that can cause the estimator to take on arbitrary large aberrant Values [10]. Hence, the breakdown point is simply the initial time any record test becomes swamped due to contaminated data. Some regression estimators offer the smallest possible breakdown point of just 0/*n* or 1/*n*. basically; only one outlier will make the OLS regression equation being made useless. Other estimators offer the finest possible breakdown cause of *n*/2 or 50%. If robust estimation method includes a 50% breakdown point, then 50% of the data could contain outliers together with the coefficients that would remain useful [12], [13], [14], [15], [16]

4.0 COMPARISON OF ROBUST REGRESSION METHODS

In general, the three broad categories of robust regression models that play the most impotant role are; M-estimators (extending from M-estimates of location by considering the size of the residuals); L-estimators (based on linear combinations of order statistics), and R-estimators (based on the ranks of the residuals). Each category of the estimators contains a class of models derived under similar conditions and with comparable theoretical statistical properties. Least Trimmed Squares Estimate (LTS), M-estimate, Yohai -MM-estimate, Least Median Squares (LMS) and S-estimate are among popular techniques used in estimating the parameters of the regression line. In this study the M-estimator and least trimmed squares are used and will be briefly described in the next sections.
Suppose we define *n* sample of data points as

$$\mathbf{Z} = \{(\mathbf{x_{11}}, .., \mathbf{x_{1p}}, \mathbf{y_1}), .., (\mathbf{x_{n1}}, .., \mathbf{x_{np}}, \mathbf{y_n})\} \quad (10)$$

and let T be an estimator of regression. This indicates that by applying T to such a sample Z will produce a vector of regression coefficients.

$$T(Z) = \hat{\boldsymbol{\theta}}$$

Now let consider Z^T that are obtained by replacing any of the *m* original data points by arbitrary values. Let us denote by bias (m; T, Z) and thus the maximum bias cause as a result of such contamination

$$bias\,(m;T,Z) = \sup_{z'} \left\| T(Z') - T(Z) \right\| \quad (11)$$

where the supremum is over all possible Z'. If bias (m; T, Z) is infinite, this means that *m* outliers can have an arbitrary large effect of T which may be expressed by saying that estimator breaks down. Therefore, the (finite-sample) breakdown point of the estimator T for sample Z is described as follows,

$$\varepsilon^*_n(T, Z) = \min\left\{\frac{m}{n}, bias(m,T,Z) is\ finite\right\} \quad (12)$$

is infinite. In other words, it is the smallest fraction of contamination that can cause the estimator T to take on values arbitrarily far from T(Z). For least squares, we have seen that one outlier is sufficient to carry T over all bounds. Therefore, its breakdown point is

$$\varepsilon^*_{n(T,Z)} = \frac{1}{n} \quad (13)$$

which tends to zero for increasing size *n*, so it can be said that the LS has breakdown point of 0% This again reflects the extreme sensitivity of the LS techniques to outliers [12].

5.0 M-ESTIMATION

The performance of linear least square estimates behaves badly when the distribution of error is not normal, more especially when the errors are heavily tailed. One solution to this is to eliminate the influential observation from the least squares fit. The group of M-estimator models consists of all models that are derived from maximum likelihood models. The most frequent general method of robust regression is M-estimation, produced by Huber (1964) that is nearly as efficient as OLS [12].
Consider the following linear model

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\ &= x_i'\beta + \varepsilon_i \end{aligned} \quad (14)$$

For the i^{th} of *n* observations the fitted model is

$$\begin{aligned} y_i &= \alpha + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + e_i \\ &= x_i' b + e_i \end{aligned} \quad (15)$$

The general M-estimator minimizes the objective function rather than minimizes the sum of squared errors since the aim is to minimize the function ρ of the errors with M-estimate. The M-estimate target function is,

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i' b) \quad (16)$$

The contribution of each residual is given by the function ρ to the objective function. A suitable ρ should have the following characteristics.

$$\rho(e) \geq 0,$$

$$\rho(0) = 0,$$

$$\rho(e) = \rho(-e) \quad \text{and}$$

$$\rho(e_i) \geq \rho(e_i) \quad for \quad |e_i| \geq |e_i|$$

For example, for least squares estimation,

$$\rho(e_i) = e_i$$

The devices of normal equations to solve this minimization problem was discovered if the partial derivatives with respect to β are set to 0, produces a system of *k*+1 estimating equations for the coefficients

$$\sum_{i=1}^{n} \varphi(yi - x_i' b) x_i' = 0$$

where ψ is derivative of ρ. The preference of the ψ function is dependent on the choice of how much weight to specify outliers. A monotone ψ function does not consider weight on outliers as much as least squares (e.g. 10σ outlier would receive the same weight as a 3σ outlier). A descending ψ function increases the weight specify to an outlier until a specified distance and then reduce the weight to 0 as the outlying distance gets considerable. Newton-Raphson and Iteratively Reweighted Least Squares (IRLS) are the two methods to solve the M estimates nonlinear normal equations. But for this research, the iterative reweighted least squares robust regression is used. IRLS expresses the normal equations as,

$$\mathbf{X}^{'}\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{'}\mathbf{W}\mathbf{y} \quad (17)$$

where W is an *n* x *n* diagonal matrix of weights. The initial vector of parameter estimates, α and β are typically obtained from OLS. IRLS updates these parameter estimates with

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^{'}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{'}. \quad (18)$$

However, the weights depend upon the residuals; the residuals depend upon the estimated coefficients, and the estimated coefficients depend upon the weights. An iterative solution called iteratively re-weighted least squares, IRLS is therefore required. The following step describe the IRLS procedure:

Step 1: Select the weight function for weighting all the cases. But in this study, we will make use of the Huber and the Tukey. The weights function which is defined as follows:

Huber: $w_{2i} = \begin{cases} 1 & |u| \le 1.345 \\ \dfrac{1.345}{|u|} & |u| > 1.345 \end{cases}$

The constant 1.345 is called a turning constant and the standardized residual

$$s_i^{(0)} = \frac{median|e_i^{(0)} - median(e_i^0)|}{0.6745}$$

The corresponding bisquare method is defined as:

$$w_i^{(0)} = \left\{ \begin{array}{ll} \left[1-\left(\dfrac{u_i^{(0)}}{4.685}\right)^2\right]^2, & \left|u_i^{(0)}\right| \le 4.685 \\ 0 & \left|u_i^{(0)}\right| > 4.685 \end{array} \right\}, \quad i = 1, 2, ..., 51$$

Step 2: Obtain the starting weight for all the cases.

$$\mathbf{b^t = [x^t w^{t-1} x] x^i w^{t-i} y} \quad (19)$$

Where x is the model matrix, with x_i as its ith row, and

$\mathbf{w^{t-i=diag\{w_i^{t-i}\}}}$ is the current weight matrix.

Step 3: Use the starting weights in weighted least squares to obtain the residual e_i^{t-1} from the fitted regression function.

Step 4: Use the obtained residuals in step 3 to obtain the revised weight

$$w_i^{t-1}[e_i^{t-i}] \quad (20)$$

Step 5: Continue the iteration until convergence is obtained. The asymptotic covariance matrix of $\hat{\beta}$ is

$$v(b) = \frac{E(\Psi^2)}{(E(\Psi))}(x^t x)^{-1} \quad (21)$$

Step 6: Finally carry out a WLS regression using the final weights w_i. The coefficients of regression realized from this WLS regression are the required estimate of the heteroscedastic model. Step2 and Step3 are repeated until the estimated coefficient converges. The Procedure continues until some convergence criteria is satisfied. The estimate of scaled residuals may be updated after every initial estimate.

6.0 LEAST TRIMMED SQUARES (LTS) ESTIMATE

Another form of robust regression estimation is the least trimmed squares regression method (LTS) [9]. He develops the least trimmed squares (LTS) estimation method as a high efficiency alternative to least median squares regression (LMSR) and this technique is observed from minimizing

$\hat{\beta}_{LTS} = arq\,Min\,Q_{LTS}(\beta)$ where

$$Q_{LTS} = \sum_{i=1}^{h} e^2$$

where $e^2_{(1)} \le e^2_{(2)} \le ..... \le e^2_{(n)}$ are the ordered squared residuals from the smallest to the largest. The values of h is obtained by $h = \left[\frac{n}{2}\right] + \left[\frac{(p+1)}{2}\right]$ with *n* and *p* being the given sample size and number of parameters involved in the model respectively. This approach is similar to least squares except usually the largest *(n-h) squared* residuals are removed (trimmed sum) from the summations which allow those outlier to be removed completely, allowing the fit to avoid the outliers. Least trimmed squares (LTS) can be very efficient when exact outlying data points are trimmed. But if there is more trimming than there are outlying data points, then some good observations will be eliminated from the computation. From the breakdown point, LTS is regarded as a high break down techniques with a BP of 50% when $h = 1/n$. The main disadvantage of robust LTS is the large number of operations required to sort the squared residuals in the objective function [15]. Another challenge is deciding the best approach for determining the initial estimate. The weighted robust least trimmed squares method consists of the following procedures:

Step 1: Regressed the response variables y_i on the explanatory variables x_{ij} by least trimmed squares and compute the regression coefficients from this fitting

Step 2: The inverse of these fitted values denoted by w_{1i} will be the values of the initial weight.

Step 3: Obtain the final weight from Huber weighted functions. Which is given as

$$w_{2i} = \begin{cases} 1, & |u| \le 1.345 \\ \dfrac{1.345}{|u|}, & |u| > 1.345 \end{cases}$$

The constant 1.345 is called a turning constant and the standardized residual u. The estimate of the scaled residuals is obtained as,

$$s_i^{(0)} = \frac{median|e_i^{(0)} - median(e_i^0)|}{0.6745}$$

The standardized residual estimate is then defined as

$$u_i^{(o)} = \frac{|e_i^{(0)}|}{s^{(0)}}, i = 1, 2, ..., n$$

Step 4: Multiply the initial weight W_{1i} with the weight W_{2i} obtained from Huber function to get the final weight W_i .

Step 5: Finally perform the WLTS with regression with the final weight W_i . The regression coefficients produced from this estimates of WLTS regression are the desired regression estimates of heteroscedasticity model regression are the desired regression estimate of heteroscedasticity model.

■7.0 NUMERICAL EXAMPLE

In this section we consider some few examples to show the importance of the robust estimators in a situation when heteroscedasticity and outliers are presence. A hetrosecdastic data taken from the state-wide crime data of United states (1993) are to be used. The data contains fifty one observation of violent crime rate (per 100,000 people in population) of y with corresponding predictor variables of x_1, x_2, x_3, x_4, x_5 and x_6, is used. The analysis begins by considering one regressor variables of x, with its corresponding response variable y, to observe the effects of outliers and heteroscedasticity using diagnostic plot. We intentionally replace the value correspond to 1st and 25th observation of the original data with a higher value such as 7610 and 4340, where the original value is 761 and 434 respectively. The OLS, RWLTS, M-estimates based on Huber function and tukey bisquare were then used in the original and modified data. The results are presented in the graphs and the tables below. In Figure 1, OLS residual plot of the original data against the regression fitted values. The situation for existence of hetroscedasticity is that when variance of the error terms are not constant, and this can be identified when the residuals are not randomly distributed around the zero residual, with an indication of systematic trend on the plot. Based on this concept, the plot clearly indicates that constant variance assumption is violated, which gives evidence that the OLS fit is improper to be used, as there are clear evidence for the presence of heterogeneous error variance. In this regard we apply the m- estimator methods based on Huber and Tukey bisquare weighted function to the data for the purpose to remedy the short coming of OLS in the presence of outlier and hetroscedesticily error variance.

To introduce this technique to the data, we first need to plot the residual against the response variable with a data set that contain outlier. Figure 2, Figure 3 and Figure 4 gives the diagnostic plot of the residual against the fitted values without outliers using the M-estimate and least trimmed squares. While Figure 5 gives the linear regression models obtained from the three robust estimation techniques and OLS. From this plot we notice that there are some differences between the estimators. This is an evident that the performance of the methods was satisfactory. In order to examine the consequence of outlier in the existence of hetrosecdasticity, modification of the data is highly important. The OLS, M-estimate and resistance regression method were used to examine the presence of outliers in the modified data and the results are presented in Figure 6, Figure 7, Figure.8 and Figure 9 below. Figure 10 gives the linear regression models obtain from the three robust estimation techniques and OLS estimation when the modified data is used. The plot of linear regression models obtained from the three robust estimation techniques using the original and modified data give a clearer picturer about the real situation. The plot of Figure 1 and Figure 6 of OLS, indicate a violation of the constant variance assumption. This signifies that the OLS estimate is inappropriate to be used. On the other hand When comepared with the RWLS and RWLTS plot of Figures 2, Figures 3, Figures 4, Figures 7, Figures 8 and Figures 9 indicate that the RWLS and RWLTS can solve the heteroscedasticity and outliers problems.

Table 1 Summary of Robust techniques performance against OLS for (Original and Modified)

Method	Data Type	Estimate	SE	t- value
OLS	Original	49.03	11.83	4.15
	Modified	54.21	36.12	1.50
RWJS HUBER	Original	35.96	11.19	3.22
	Modified	49.37	12.26	4.04
Turkey Bisquare	Original	28.45	9.54	2.98
	Modified	37.57	9.118	4.12

Table 1 gives the summary results of statistics, which include the standard errors, t-values and the estimate of the regression coefficient for the original and modified crimed data. The result of Table 1 reveals the influence of outliers on the regression model, when OLS is used to estimate the regression parameter compared with the regression parameter obtained from the RWLS and RWLTS estimate. The result of RWLS based on Huber function, psiBisquare function and RWLTS estimate of regression coefficient, standard errors and t-value of the modified and original data are similar. We can also see the weights given to the estimates on dramatically lower using the Tukey bisquare weighting function than the Huber weighting function and the parameter estimates from these two different weighting methods differ.

On the other hand when considering the estimate of the result obtained from the crime data that involve six explanatory variables.

As you can see, the results from the two analyses of original and modified data are fairly different, especially with respect to the regression coefficients and the constant (intercept) While normally we are not interested in the constant, if you had centred one or all of the predictor variables, the constant would be useful. It will be noticed that some variables are not statistically significant in either analysis, whereas some are significant in both analyses it is an evidence that M- estimate and RWLTS have partially address the problem of hetroscedastic in the presence of outleirs in the data. However, the results obtained using RWLS based on Huber and RWLTS are only slightly influence by the outliers. Different functions have advantages and drawbacks. Huber weights can have difficulties with severe outliers, and Tukey Bisquare weights can have difficulties converging or may yield multiple solutions.

The summary in Table 1 provides the result of the estimated parameter using OLS and RWLS and RWLTS for the simple linear sigression of two variables X and Y and multiple regression of both original and modified crimed data. Hence they are not reliable. However, the M-estimation (IRWLS) emerges to become plainly more effective and much more reliable because it is less affected by the outliers. The outcomes of the result appear to point out that the M-estimation based on Huber estimation, Tukey bisquare and RWLTS methods provides asubstantial Improvement within the other existing techniques.

Table 2 Summary of robust techniques performance against OLS for (original and modified)

Method	Data type	plainly β_0	β_1	β_2	β_3	β_4	β_5	β_6
OLS	Orign.	-857.62	23.42	6.24	-1.27	5.36	15.09	28.632
	Mod	519.88	-17.82	-16.07	35.90	24.82	-17.40	253.912
RWLS Huber.	Orign	-662.47	24.16	5.62	-1.14	3.34	10.48	32.170
	Mod.	-139.38	15.09	3.82	-8.58	3.93	14.63	51.781
Tukey Bisquare	Orgn.	-509.74	24.25	5.28	-1.32	2.01	8.11	33.6722
	Mod.	-376.50	23.03	5.28	-1.84	0.42	12.58	31.9234
RWLTS	Orign	-662.47	24.16	5.62	-1.14	3.34	10.48	32.170
	Mod	-139.38	15.09	3.82	-8.58	3.93	14.63	51.781

Table 3 Summary of robust techniques performance against OLS for (original and modified)

Method	Data type	S.E. β_0	S.E. β_1	S.E. β_2	S.E. β_3	S.E. β_4	S.E. β_5	S.E. β_6
OLS	Original	602.8	3.94	1.18	2.554	6.98	10.34	14.73
	Modified	3816.4	24.9	7.48	16.168	44.2	65.44	93.26
RWLS Huber	Original	602.7	3.95	1.21	2.554	6.95	10.43	14.86
	Modified	616.6	4.52	1.32	3.504	7.09	10.73	18.54
Tukey Bisquare	Original	628.7	4.11	1.23	2.664	7.28	10.78	15.36
	Modified	599.2	3.92	1.18	2.539	6.95	10.28	14.64
RWLTS	Original	602.7	3.95	1.21	2.554	6.95	10.43	14.86
	Modified	616.6	4.52	1.32	3.504	7.09	10.73	18.54

Table 4 Summary of robust techniques performance against OLS for (original and modified)

Method	Data type	t-value β_0	t-value β_1	t-value β_2	t-value β_3	t-value β_4	t-value β_5	t-value β_6
OLS	Original	-1.42	5.94	5.28	-0.50	0.77	1.46	1.94
	Modified	0.14	-0.71	-2.15	-2.22	0.56	-0.27	2.72
RWLS Huber	Original	-1.10	6.12	4.66	-0.45	0.48	1.01	2.17
	Modified	-0.23	3.34	2.91	-2.45	0.56	1.36	2.79
Tukey Bisquare	Original	-0.81	5.90	4.29	-0.50	0.28	0.75	2.19
	Modified	-0.63	5.88	4.50	-0.72	0.06	1.22	2.18
RWLTS	Original	-1.10	6.12	4.66	-0.45	0.48	1.01	2.17
	Modified	-0.23	3.34	2.91	-2.45	0.56	1.36	2.79

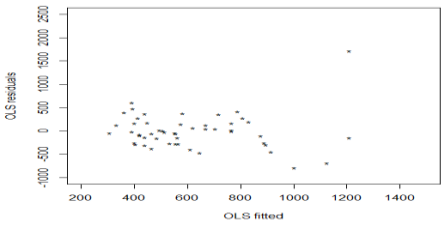


Figure 1 Plot of OLS residual versus fitted values (Original data)

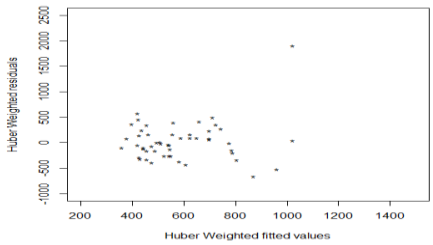


Figure 2 Plot of RWLS based on Huber residuals versus fitted values (Original data)

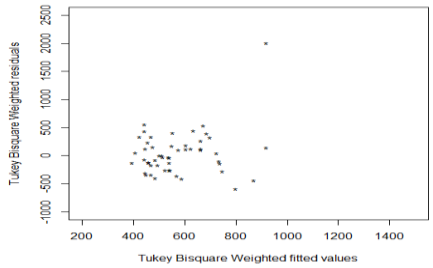


Figure 3 Plot of RWLS based on Tukey Bisquare residuals versus fitted values(Original data)

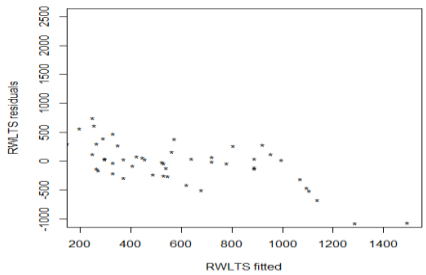


Figure 4 Plot of RWLTS residuals versus fitted values (Original data)

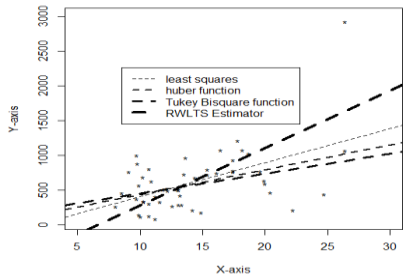


Figure 5 Plot of crime data with four estimated regression lines (Original data)

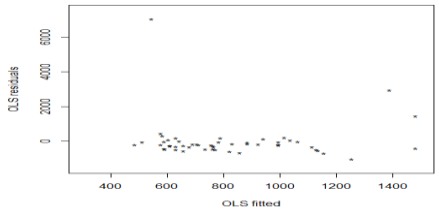


Figure 6 Plot of OLS residual versus fitted values (Modified data)

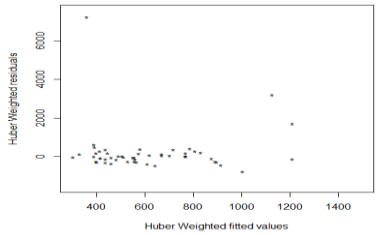


Figure 7 Plot of RWLS based on Huber residuals versus fitted values (Modified data)

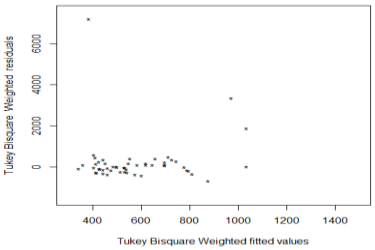


Figure 8 Plot of RWLS based on Tukey Bisquare residuals vs fittedvalues (Modified data)

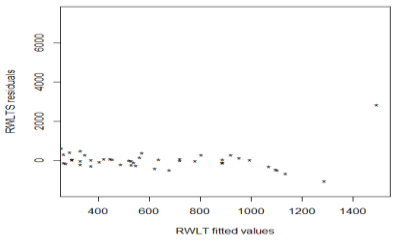


Figure 9 Plot of RWLTS residuals versus fitted values (Modified data)

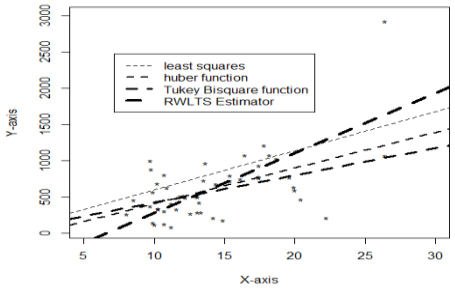


Figure 10 Plot of crime data with four estimated regression lines (MModified data)

8.0 THE BEST MODEL

The best model is by using the standard error and t-value estimated from the state-wide crime data which involve all the explanatory variables. Based on the results obtained in Table 1, Table 2, Table 3 and Table 4 it is clear that RWLS estimator using Tukey bisquare has the least standard errors with the largest t-values compared to the t-value obtain from RWLS estimator using, Huber function, RWLTS and OLS

9.0 CONCLUSION

The primary focus of this paper would be to produce reliable techniques for correcting the problem of heteroscedastic errors in the existence of outliers. The empirical study discloses the OLS estimations are often affected by the outliers.For correcting the problems of outliers and heteroscedastic errors in the data.

Acknowledgments

We would like to acknowledge the financial support from University Technologi Malaysia for the Research University Grant.

References

[1] Midi, S. Rana, and A. Rahmatullah. 2009. The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedasticity Errors. *WSEAS Transaction on Mathematics*. 351–360.

[2] S. Chatterjee, and A. S. Hadi. 2006. *Regression Analysis by Examples*. 4th ed. Wiley, New York.

[3] R. D. Cook, and S. Weisberg. Diagnostics for Heteroscedasticity in Regression. *Biometrika*. 70(983): 1–10.

[4] Halbert White, A. 1980. Heteroskedastic Consistent Covariance Matrix Estimator and a Direct Test fFor Heteroskedasticity. *Econometrica*. 48: 817–838.

[5] R. A. Maronna, R. D. Martin, and V. J. Yohai. 2006. *Robust Statistics-Theory and Methods*. Wiley, New York.

[6] M. S. Rana, H. Midi, and A. H. M. R. Imon. 2008. A Robust Modification of the Goldfeld-Quandt Test for the Detection of Heteroscedasticity in the Presence of Outliers. *Journal of mathematics and Statistics*. 4(4): 277–283.

[7] M. H. Kutner, C. J. Nachtsheim, and J. Neter. 2004. *Applied Linear Regression Models*. 4th ed., McGraw-Hill/ Irwin, New York.

[8] H. Midi and B. A. Talib. 2008. The Performance of Robust Estimator in Linear Regression Model Having both Continous and Catigorical Variables with Heteroscedasticity Errors. *Malaysia Journal of Mathematical Science*. 2(1): 25–48

[9] P. J. Rousseeuw, and A. Leroy. 1987. *Robust Regression and Outlier Detection*. Wiley, New York.

[10] A. H. Midi and R. Imon. 2009. Deletion Residuals in the Detection of Heterogencity of Variance in Linear Regression. *Journal of Applied Statistics*. 36: 347–358.

[11] R. Marona, R. Martin, V. J. Yohai. 2006. *Robust Statistics Theory and Methods*. John Wiley & Sons Ltd., England.

[12] D. L. Donoho and P. J. Huber. 1983. The Notion of Breakdown Point. In: Bickel PJ Doksum KA, Hodges JL Jr (Editors), A Festschrift for Erich L. LehmannWadsworth, Belmont. 157–184.

[13] A. Christmann. 1994. Least Median of Weighted Squares in Logistic Regression with Large Strata. *Biometrika*. 81: 413–417.

[14] P. J. Rousseeuw, A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. Wiley-Interscience, New York.

[15] P. H. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*. 35:7–101.