TWO-STEP ROBUST ESTIMATOR IN HETROSCEDASTIC REGRESSION MODEL IN THE PRESENCE OF OUTLIERS

Abstract. Although the ordinary least squares (OLS) estimates are unbiased in the presence of heteroscedasticity, these are no longer efficient. This problem becomes more complicated when the violation of constant error variances comes together with the existence of outliers. The weighted least squares (WLS) procedure is often used to estimate the regression parameters when heteroscedasticity occurs in the data. But there is evidence that the WLS estimators suffer a huge set back in the presence of outliers. Moreover, the use of the WLS requires a known form of the heteroscedastic errors structures. To rectify this problem, we proposed a new method that we call two-step robust weighted least squares (TSRWLS) method where prior information on the structure of the heteroscedastic errors is not required. In the proposed procedure, the robust technique is used twice. Firstly, the robust weights are used for solving the heteroscedastic error and secondly, the robust weighting function is used for eliminating the effect of outliers. The performance of the newly proposed estimator is investigated extensively by real data sets and Monte Carlo simulations.

Keywords: Heteroscedasticity, Weighted least squares, Two-step robust weighted least squares, Outliers, Monte Carlo simulation.

JEL Classification: C12, C22, C52, C63

1. INTRODUCTION

The linear regression model is commonly used by statistics practitioners in many different fields like engineering, physics, medicine, biology, chemistry, social science and economics. The regression parameters are often estimated by the ordinary least squares (OLS). Under the usual assumptions, the least-squares estimators possess many
desirable properties. In particular, these assumptions imply that the estimators of the parameters will be unbiased, consistent, and efficient in the class of linear unbiased estimators. A commonly used assumption is the constancy of error variances or homoskedasticity, mainly because of which the OLS estimators retain the minimum variance property. In a real life situation it is really hard to believe that the error variances will remain constant and that is why the violation of this assumption which causes the heterogeneity of error variances or heteroskedasticity is more prevalent in nature. The main problem with the violation of homoskedasticity assumption is that the usual covariance matrix estimator of the OLS becomes biased and inconsistent.

A large body of literature is now available [1-3,7-10,16,20,21,23,25,26] for correcting the problem of heteroscedasticity. The correction for heteroscedasticity is very simple by means of the weighted least squares (WLS) if the form and magnitude of heteroscedasticity are known. The WLS is equivalent to perform the OLS on the transformed variables. Unfortunately, in practice, the form of heteroscedasticity is unknown, which makes the weighting approach impractical. When heteroscedasticity is caused by an incorrect functional form, it can be corrected by making variance-stabilizing transformations of the dependent variables [4] or by transforming both sides [2]. However, the transformation procedure might be complicated when dealing with more than one explanatory variable. Montgomery et al. [20], Kutner et al. [16], and others have tried to find the appropriate weight to solve the heteroscedastic problem when the form of heteroscedasticity is unknown. White [28] proposed the heteroskedasticity-consistent covariance matrix (HCCM) estimators in this regard. Different forms of HCCM estimators such as the HC0, HC1, HC2, HC3 and HC4 have been proposed [5,6,12,13,18,28]. However, there is no general agreement among statisticians about which of the five estimators of the HCCM (HC0, HC1, HC2, HC3, HC4) should be used [5,6,17,18]. Chatterjee and Hadi [3] proposed an estimator which is weight based, but these weights depend on the known structure of the heteroscedastic data. Montgomery et al. [20] and Kutner et al. [16] proposed estimators which do not depend on the known structure of the heteroscedastic data. But the main limitation of the Montgomery et al. [20] estimator is that it cannot be applied to more than one regressor situation. The estimator proposed by Kutner et al. [16] can be applied to more than one variable and it does not depend on the known form of heteroscedasticity, but we suspect this estimator is not outlier resistant.

It is now evident that a few atypical observations (outliers) can make the entire inferential procedure meaningless [2,19,24]. The weighted least squares also suffer the same problem in the presence of outliers [19]. We also believe that the HCCM estimators should suffer from the same problem, as they are based on the OLS residuals. Generally speaking, none of the estimation techniques work well unless the effect of outliers in a heteroscedastic regression model is eliminated or reduced by robustifying the WLS or HCCM. Therefore, in this article we address the following
Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

question: which robust WLS or HCCM procedure should be used when heteroscedasticity and outliers occur at the same time? This problem motivates us to develop a new and more accurate estimation technique. However, in this article, our study is only confined to the development of the robust WLS. In the presence of outliers we have some robust techniques for the detection of heteroscedasticity [15, 22]. Unfortunately, there is not much work in the literature that deals with the estimation of the regression parameters in the presence of both heteroscedasticity and outliers when the structure of heteroscedasticity is unknown. Although Habshah et al. [9] has proposed this type of robust estimation procedure, but their procedure can be applied to only one regressor.

In this article, we propose a two-step robust weighted least squares (TSRWLS) estimator which can be applied for more than one regressor when the form of the heteroscedasticity is not known. Firstly, for solving the heteroscedastic problem we estimate the robust initial weights following the idea of Kutner et al. [16] and secondly, we estimate the parameters of the model based on Huber’s [14] weight function in order to reduce the effect of outliers. Our results show, as expected, that the existing estimators are very sensitive to outliers whereas our proposed estimator is less sensitive to outliers. The proposed TSRWLS estimator is described in section 2. Section 3 provides an illustrative example to show the better performance of the proposed method. Section 4 reports the results of a Monte Carlo simulation study which is designed to investigate the performance of the proposed method and, section 5 contains the concluding remarks.

2. TWO-STEP ROBUST WEIGHTED LEAST SQUARES (TSRWLS)

Consider the general multiple linear regression model:

\[ y = X\beta + \varepsilon \]  

(1)

where \( y = (y_1, y_2, \ldots, y_n)^T \) is an \( n \times 1 \) vector of response variable, \( X = (x_1, x_2, \ldots, x_n)^T \) is an \( n \times p \) fixed design matrix including the intercept, \( \beta \) is an \( p \times 1 \) vector of unknown linear parameters, and \( \varepsilon \) is an \( n \times 1 \) vector of errors. The traditionally used OLS estimator of \( \beta \) is \( \hat{\beta} = (X^TX)^{-1}X^Ty \). It has mean \( \beta \) (i.e., it is unbiased) and covariance matrix

\[ \text{cov}(\hat{\beta}) = (X^TX)^{-1}X^T\Omega X(X^TX)^{-1} \]  

(2)

where \( E(\varepsilon\varepsilon^T) = \Omega \), a positive definite matrix. Under homoscedasticity, we have \( \Omega = \sigma^2I_n \), and it follows that the \( \text{cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1} \), which can be estimated by \( \hat{\sigma}^2(X^TX)^{-1} \), where \( \hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon} / (n - p) \), \( \hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n) \) being the \( n \)-vector of OLS
residuals. Under heteroscedasticity, that is, \( \Omega = \sigma^2 Z \), where \( Z \) is a diagonal matrix, equation (2) becomes

\[
V(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T ZX(X^T X)^{-1}
\]  

Define \( W = Z^{-1} \), where \( W \) is a diagonal matrix with diagonal elements or weights \( w_1, w_2, \ldots, w_n \). It can be easily proved that the weighted least squares estimator is

\[
\hat{\beta}_{WLS} = (X^T WX)^{-1} X^T Wy
\]

and

\[
\text{cov}(\hat{\beta}_{WLS}) = \sigma^2_{WLS} (X^T WX)^{-1}.
\]

\( \hat{\sigma}^2_{WLS} \) also can be estimated by

\[
\hat{\sigma}^2_{WLS} (X^T WX)^{-1} \text{ where } \hat{\sigma}^2_{WLS} = \sum w_i \hat{\epsilon}_i^2 / (n - p).
\]

It is not difficult to compute the weights of the \( W \) matrix, if the heteroscedastic error structure of the regression model is known. From a standard adaptation of the Gauss-Markov theorem, one can easily prove that, if the \( W \) matrix is known, the WLS provides the best linear unbiased estimator. Moreover, under normality of the errors, it is the best unbiased estimator ever. But this situation almost never exists in real applications and the estimated weights are used instead. Although it is difficult to assess the effect of using estimated weights, but it is generally believed that small variations in the weights due to estimation do not often affect a regression analysis or its interpretation much. But the presence of outliers should have an adverse effect on the determination of weights. Likewise the OLS method, the WLS regression is also sensitive to the presence of outliers. If potential outliers are not properly addressed, they will definitely affect the parameter estimation and other aspects of a weighted least squares analysis.

In this paper, our initial goal is to find an appropriate weight matrix \( W \) in which the heteroscedastic error structure is unknown. It is worth mentioning here that the \( W \) matrix should perform well in the presence of heteroscedasticity and outliers. To find the robust weight matrix \( W \), we propose a two-step robust weighted least squares (TSRWLS) estimator. The TSRWLS is an extension of works of Habshah et al. [9] and Kutner et al. [16]. Habshah et al. [9] proposed a robust weighted least squares (RWLS) estimator to solve the heteroscedastic and outlying problem by developing robust weighting technique. Instead of fitting regression with all the data, they suggested finding several “near-neighbor” groups in the explanatory variable. The group medians represent the explanatory variable \( X \) and the groups in the response variable \( Y \) are formed in accordance with the groups formed in \( X \). The sample median absolute deviations (MAD) of each groups of \( Y \) and the median of each group of \( X \) are then computed. The square of group MADs in \( Y \) are then regressed on the corresponding group medians of \( X \) by the least trimmed of squares (LTS) [24] method and the regression coefficients from this fitting are computed. Using these coefficients and full \( X \) values, the fitted values are obtained. The inverse of these absolute fitted values then form the initial weights. The final weights are obtained after multiplying these weights by Huber’s weight [14].
Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

The main limitation of this procedure is that it cannot be applied to more than one regressor. To overcome this problem we incorporate Habshah et al. [9] and Kutner et al. [16] estimators. Hereafter we will refer to the Kutner et al. [16] estimator as KNN (Kutner, Nachtsheim and Neter) estimator. The KNN estimator starts with fitting a linear regression by the ordinary least squares and conducting some preliminary analysis of the residuals. It is obvious that the megaphone shape of absolute residuals of the OLS against the fitted values may confirm the non-constancy of error variances. Therefore, Kutner et al. [16] suggested regressing the absolute residuals against the fitted values and obtain a standard deviation regression function. To obtain the weights, the fitted values from this standard deviation regression function are computed and the inverse of the square fitted values are considered as the desirable weights. We use the LTS estimator, instead of the OLS in the KNN algorithm to get the initial robust weights. The TSRWLS consists of the following two steps. In step 1 we form the initial weight and in step 2 we obtain the final weight.

**Step1:**

(i) Find the fitted values $\hat{y}_i$ and the residuals $\hat{\varepsilon}_i$ from the regression model in equation (1), by using the least trimmed of squares (LTS) method.

(ii) Regress the absolute residuals, denoted as $s_i$ where $s_i = |\hat{\varepsilon}_i|$, on $\hat{y}_i$ also by using the LTS method.

(iii) Find the fitted values $\hat{s}_i$ from step 1(ii).

(iv) The square of the inverse fitted values would form the initial robust weights, i.e., we obtain $w_{i1} = 1/(\hat{s}_i)^2$.

**Step2:**

The robust weighting function such as the Huber function [14], the Bisquare function [27] and the Hampel function [11] can be used to obtain the final weight. However, in this study, we will use the Huber’s [14] weights function which is defined as

$$w_{2i} = \begin{cases} 
1 & |e_i| \leq 1.345 \\
\frac{1.345}{|e_i|} & |e_i| > 1.345 
\end{cases}$$

The constant 1.345 is called the tuning constant and $e_i$ is the i-th standardized residuals of the LTS obtained from step 1(i). We multiply the weight $w_{i1}$ with the weight $w_{2i}$ to get the final weight $w_i$. Finally we perform a WLS regression using the final weights $w_i$. The regression coefficients obtained from this WLS are the desired estimate of the heteroscedastic multiple regression model in the presence of outliers.
3. EXAMPLE
In this section, we consider a real data to evaluate the performance of the proposed TSRWLS method.

3.1 Education Expenditure Data

This data is taken from Chatterjee and Hadi [3] which consider the per capita income on education projected for 1975 as the response variable \((Y)\) while the three explanatory variables are \(X_1\), the per capita income in 1973; \(X_2\), the number of residents per thousand under 18 years of age in 1974, and \(X_3\), the number of residents per thousand living in urban areas in 1970 for all 30 states in USA. According to geographical regions based on the pre-assumption, the states are grouped in a sense that there exists a regional homogeneity. The four geographic regions (i) Northeast, (ii) North centre, (iii) South, and (iv) West. The LTS estimator detected that the observation 49 [Alaska (AK)] is an outlier. The residuals vs. fitted values of OLS (Standardized), KNN and TSRWLS are plotted in Fig.1. Fig.'s 1(a) - 1(c) display the residuals-fitted plots without considering Alaska. If the variances of the error terms are constant then one can expect that the residuals are randomly distributed around zero residual, without showing any systematic pattern. Fig.1 (a) clearly indicates a violation of the constant variance assumption. This signifies that the OLS fit is inappropriate here, as there is a clear indication of heterogeneous error variances. However, the KNN and TSRWLS fit, presented in:
Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

![OLS fitted values vs. residuals](image1)

![KNN fitted values vs. residuals](image2)

![TSRWLS fitted values vs. residuals](image3)

**Figure 1.** The OLS, KNN and TSRWLS fitted values vs. residuals plots without AK, (a)-(c); with AK, (d)-(f)

Fig.1(b) and Fig.1(c) respectively, do not show any symmetrical shape like the OLS fit. It shows that for this ‘clean’ data (without AK) the non-constancy of error variances is not reflected in KNN and TSRWLS. To see the effect of outliers, we include the observation Alaska and the resulting residuals and fitted values are plotted in Fig.’s 1(d)-1(f). We see that OLS residuals are affected in the presence of outliers, but the effect of AK observation is not substantial on KNN and TSRWLS estimators.

3.2 Modified Education Expenditure Data

In reality we often have to deal with multiple outliers. For this reason, we deliberately change four data points to generate big outliers. Our changed data points are cases 46,
47, 48 and 50 by taking the value from outside the well-known 3-\(\sigma\) sigma normal distance in \(Y\) direction. In fact, we replace the data points of \(Y\) for observations 46, 47, 48 and 50 by \(|y_{cont}|\) where \(y_{cont}\) are generated as \(\bar{y} \pm 9s_y\), with \(\bar{y}\) and \(s_y\) as the respective mean and standard deviation of \(Y\). In this situation, it is more likely that these points would become big outliers. With this modified data, now we have five outliers (since this data already contained one outlier, i.e., Alaska). When the LTS is employed to the data, all 5 outliers are identified.

Figure 2. The OLS, KNN and TSRWLS fitted values vs. residuals plots with 10% outliers, (a)-(c); without 10% outliers, (d)-(f)
Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

The plots of the residuals against the fitted values of the OLS, KNN and TSRWLS for the modified data are illustrated in Fig.’s 2(a)-2(f). It is observed from Fig.’s 2(a) and 2(b) that in the presence of outliers the patterns of residuals are completely destroyed. That is, the OLS and KNN are greatly affected by outliers and so they are not good estimators for the remedy of the heteroscedastic problem when outliers are present. It is interesting to note that in Fig. 2(c), the TSRWLS shows the scatter plot of the residuals except the data points which are outliers. Like as Fig.1, the residual-fitted plots without the 10% outliers for the OLS, KNN and the TSRWLS are shown in Fig.’s 2(d)-2(f). Fig. 2(d) signifies that the OLS cannot remedy the problem of heteroscedasticity but the KNN and proposed TSRWLS are successful as it is expected. It re-emphasizes our concern that the KNN might good in the absence of outliers whereas our proposed TSRWLS might be good in the presence or absence of outliers since it is keeping the scatter plot in both situations. In particular, the residuals plots of Fig.1 and Fig.2 show that the TSRWLS estimator is successful to cope with the problem of heteroscedasticity and outliers.

We know that graphical displays are always very subjective and that is why we would like to present some numerical summaries of the examples considered above. Here, we compare the performance of the proposed TSRWLS estimator with the existing estimators, such as the OLS, KNN and five versions of the HCCM estimators. Table 1 displays the summary statistics such as estimates of the parameters and their standard errors. It also considers three different situations: when there are no outliers, with only one outlier (AK), and with 5 outliers. In the absence of outliers, all estimators perform equally in terms of parameter estimates and their standard errors and the resulting values are relatively close. But things change dramatically when outliers are present in the data. All estimators except the TSRWLS are strongly affected by outlier(s). We observe that the OLS and the KNN estimators not only have more bias in comparison to the TSRWLS, but also the sign of \( \hat{\beta}_{OLS} \) and \( \hat{\beta}_{KNN} \) have been changed in some occasions. By looking at the results of standard errors it is clear that both the OLS and the KNN estimators together with the five versions of HCCM break down easily even in the presence of a single outlier. They produce much higher standard errors as compared with the TSRWLS estimator and things deteriorate when multiple outliers are present in the data. It can be concluded from Table 1 that the proposed TSRWLS is the best overall estimator as it possesses less bias and standard errors as compared to other estimators in the presence of heteroscedasticity and outliers.
Table 1: Regression estimates of the Education Expenditure Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Without outliers</th>
<th>With AK outlier</th>
<th>With multiple Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_0$</td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
</tr>
<tr>
<td>OLS</td>
<td>-277.5773</td>
<td>0.0483</td>
<td>0.8869</td>
</tr>
<tr>
<td>KNN</td>
<td>-334.4223</td>
<td>0.0550</td>
<td>0.9809</td>
</tr>
<tr>
<td>TSRWLS</td>
<td>-283.2395</td>
<td>0.0508</td>
<td>0.8827</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-556.5680</td>
<td>0.0724</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>-423.7212</td>
<td>0.0620</td>
</tr>
<tr>
<td></td>
<td>TSRWLS</td>
<td>-365.4785</td>
<td>0.0543</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>-452.0702</td>
<td>0.0821</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>-536.6901</td>
<td>0.1219</td>
</tr>
<tr>
<td></td>
<td>TSRWLS</td>
<td>-391.5358</td>
<td>0.0605</td>
</tr>
</tbody>
</table>

Standard Errors of Estimators

<table>
<thead>
<tr>
<th>Method</th>
<th>Without outliers</th>
<th>With AK outlier</th>
<th>With multiple Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>132.4229</td>
<td>0.0121</td>
<td>0.3311</td>
</tr>
<tr>
<td>KNN</td>
<td>108.2248</td>
<td>0.0111</td>
<td>0.2642</td>
</tr>
<tr>
<td>HC0</td>
<td>100.5722</td>
<td>0.0098</td>
<td>0.2590</td>
</tr>
<tr>
<td>HC1</td>
<td>109.5119</td>
<td>0.0106</td>
<td>0.2821</td>
</tr>
<tr>
<td>HC2</td>
<td>105.5744</td>
<td>0.0103</td>
<td>0.2733</td>
</tr>
<tr>
<td>HC3</td>
<td>111.0343</td>
<td>0.0108</td>
<td>0.2891</td>
</tr>
<tr>
<td>HC4</td>
<td>101.1556</td>
<td>0.0098</td>
<td>0.2609</td>
</tr>
<tr>
<td>TSRWLS</td>
<td>105.9811</td>
<td>0.0106</td>
<td>0.2732</td>
</tr>
<tr>
<td>OLS</td>
<td>123.1953</td>
<td>0.0116</td>
<td>0.3147</td>
</tr>
<tr>
<td>KNN</td>
<td>96.8830</td>
<td>0.0107</td>
<td>0.2313</td>
</tr>
<tr>
<td>HC0</td>
<td>172.6703</td>
<td>0.0157</td>
<td>0.4242</td>
</tr>
<tr>
<td>HC1</td>
<td>187.6852</td>
<td>0.0170</td>
<td>0.4611</td>
</tr>
<tr>
<td>HC2</td>
<td>222.4836</td>
<td>0.0199</td>
<td>0.5415</td>
</tr>
<tr>
<td>HC3</td>
<td>290.5865</td>
<td>0.0257</td>
<td>0.7025</td>
</tr>
<tr>
<td>HC4</td>
<td>192.3270</td>
<td>0.0173</td>
<td>0.4700</td>
</tr>
<tr>
<td>TSRWLS</td>
<td>102.6924</td>
<td>0.0105</td>
<td>0.2486</td>
</tr>
<tr>
<td>OLS</td>
<td>464.4632</td>
<td>0.0437</td>
<td>1.1864</td>
</tr>
<tr>
<td>KNN</td>
<td>182.0470</td>
<td>0.0204</td>
<td>0.4591</td>
</tr>
<tr>
<td>HC0</td>
<td>400.5560</td>
<td>0.0257</td>
<td>1.0027</td>
</tr>
<tr>
<td>HC1</td>
<td>435.3869</td>
<td>0.0279</td>
<td>1.0898</td>
</tr>
<tr>
<td>HC2</td>
<td>446.6578</td>
<td>0.0304</td>
<td>1.1106</td>
</tr>
<tr>
<td>HC3</td>
<td>513.9515</td>
<td>0.0372</td>
<td>1.2681</td>
</tr>
<tr>
<td>HC4</td>
<td>415.541</td>
<td>0.0274</td>
<td>1.0371</td>
</tr>
<tr>
<td>TSRWLS</td>
<td>161.8082</td>
<td>0.0170</td>
<td>0.3932</td>
</tr>
</tbody>
</table>
Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

4. SIMULATIONS

In this section, we report a Monte Carlo simulation study which is designed to compare the performance of the proposed TSRWLS estimator with the OLS, KNN and five versions of HCCM estimators. We re-use a design of Cribari-Neto [5]. In this simulation study the ‘good’ observations are generated according to linear regression model:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma_i \varepsilon_i, \quad i=1,2,\ldots,n. \]  

where \( \varepsilon_i \sim N(0,1) \) and \( E(\varepsilon_i \varepsilon_j) = 0 \forall i \neq j \). To generate a heteroscedastic regression model, we consider

\[ \sigma_i^2 = \sigma^2 \exp(ax_{i1} + ax_{i2}^2) \]

with \( \sigma^2 = 1 \) and \( a \) is an arbitrary constant. The covariate values are selected as random draws from the \( U(0,1) \) distribution. The level of heteroscedasticity is measured as

\[ \lambda = \max(\sigma_i^2) / \min(\sigma_i^2), \quad i=1,2,\ldots,n. \]

For each sample sizes we set \( a = .4 \) and \( a = .8 \), which yield \( \lambda \approx 2 \) and \( \lambda \approx 4 \), respectively. The values of the regression parameters used in the data generation scheme are \( \beta_0 = \beta_1 = \beta_3 = 1 \). Then we generate the contaminated model. At each step, one ‘good’ observation is substituted with an outlier. We focus on the situation where the errors are contaminated normal distribution. To generate a certain percentages of outliers, we use the regression model

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma_i \varepsilon_{i(\text{cont.})}, \quad i=1,2,\ldots,n. \]  

where \( \varepsilon_{i(\text{cont.})} \sim N(0,1) + \text{Cauchy}(0,10) \). The percentages of outliers can be varied. Since Cauchy is a longer tailed distribution, we are convinced that the contaminated normal errors would produce outliers.

The robustness measures and standard errors of the parameters of the OLS, KNN, and TSRWLS methods are investigated by considering the samples of size 50, 100 and 150. We performed 10,000 simulations using the S-Plus programming language. Summary values such as the mean estimated values

\[ \overline{\beta}_j = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}^{(k)}_j \]  

are then computed based on \( m = 10,000 \) replications. This also yields the bias \( \overline{\beta}_j - \beta_j \). The mean-squared error (MSE) is given by:
<table>
<thead>
<tr>
<th>%OT</th>
<th>Estimators</th>
<th>Coeff.</th>
<th>OLS</th>
<th>KNN</th>
<th>TSRWLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>beta0</td>
<td>-0.0059</td>
<td>-0.0013</td>
<td>-0.00198</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>0.0140</td>
<td>0.0060</td>
<td>0.006226</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>-0.0008</td>
<td>-0.0027</td>
<td>-0.00227</td>
<td>–</td>
</tr>
<tr>
<td>10%</td>
<td>beta0</td>
<td>0.9357</td>
<td>0.1849</td>
<td>0.005338</td>
<td>0.2806</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>-0.3898</td>
<td>0.1293</td>
<td>-0.01026</td>
<td>0.3632</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>-2.1139</td>
<td>-1.1011</td>
<td>-0.0035</td>
<td>0.4208</td>
</tr>
<tr>
<td>15%</td>
<td>beta0</td>
<td>-85.7342</td>
<td>-36.0054</td>
<td>0.011383</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>82.7205</td>
<td>33.1180</td>
<td>-0.0172</td>
<td>0.0079</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>32.2882</td>
<td>9.1732</td>
<td>-0.00564</td>
<td>0.0155</td>
</tr>
<tr>
<td>20%</td>
<td>beta0</td>
<td>-195.6400</td>
<td>-87.9323</td>
<td>-0.00422</td>
<td>0.0028</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>-372.1610</td>
<td>-146.0850</td>
<td>0.00136</td>
<td>0.0016</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>494.4786</td>
<td>181.1661</td>
<td>0.012381</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Size n = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>%OT</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>0%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>10%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>15%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>20%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Size n = 150</th>
</tr>
</thead>
<tbody>
<tr>
<td>%OT</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>0%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>10%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>15%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>20%</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

a. Percentages of outliers
## Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

### Table 3: Robustness measure of the parameters of the different estimators, $\lambda = 4$

<table>
<thead>
<tr>
<th>Coeff.</th>
<th>Estimators</th>
<th>OLS Bias</th>
<th>KNN Bias</th>
<th>TSRWLS Bias</th>
<th>Relative measure of RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>0%</td>
<td>-0.0105</td>
<td>0.0039</td>
<td>0.0025</td>
<td>112.4879</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0.0152</td>
<td>0.0020</td>
<td>-0.0028</td>
<td>101.3282</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0.0157</td>
<td>-0.0013</td>
<td>0.0028</td>
<td>88.8705</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>-3.5890</td>
<td>-0.7598</td>
<td>-0.0043</td>
<td>0.1139</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>6.1317</td>
<td>1.6857</td>
<td>0.0042</td>
<td>0.1631</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>-0.9029</td>
<td>-0.9975</td>
<td>0.0112</td>
<td>0.1860</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>1.3242</td>
<td>0.3470</td>
<td>0.0316</td>
<td>0.4223</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.6738</td>
<td>0.3530</td>
<td>0.0084</td>
<td>0.4953</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.0245</td>
<td>0.8270</td>
<td>-0.0787</td>
<td>0.4228</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>-5.0442</td>
<td>-1.9398</td>
<td>0.0122</td>
<td>0.2184</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>-11.4162</td>
<td>-4.4248</td>
<td>-0.0129</td>
<td>0.1864</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>15.4846</td>
<td>5.8186</td>
<td>-0.0215</td>
<td>0.1736</td>
</tr>
<tr>
<td></td>
<td>Sample Size n= 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta0</td>
<td>0%</td>
<td>-0.0035</td>
<td>-0.0048</td>
<td>-0.0041</td>
<td>111.8690</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0.0076</td>
<td>0.0185</td>
<td>0.0122</td>
<td>95.1401</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>-0.0002</td>
<td>-0.0029</td>
<td>-0.0039</td>
<td>84.9755</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>21.0349</td>
<td>5.9017</td>
<td>-0.0032</td>
<td>0.0300</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>18.5522</td>
<td>4.5442</td>
<td>0.0009</td>
<td>0.0469</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>-46.2991</td>
<td>-11.0444</td>
<td>0.0077</td>
<td>0.0252</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>-22.9743</td>
<td>-5.2984</td>
<td>0.0010</td>
<td>0.0473</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>28.5573</td>
<td>7.3703</td>
<td>-0.0056</td>
<td>0.0472</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>22.0493</td>
<td>6.4229</td>
<td>-0.0093</td>
<td>0.0701</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>6.6902</td>
<td>2.4548</td>
<td>0.0109</td>
<td>0.0737</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.7136</td>
<td>0.3268</td>
<td>-0.0034</td>
<td>0.0393</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>-11.0030</td>
<td>-3.5738</td>
<td>-0.0058</td>
<td>0.0888</td>
</tr>
<tr>
<td></td>
<td>Sample Size n= 150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta0</td>
<td>0%</td>
<td>-0.0008</td>
<td>-0.0080</td>
<td>-0.0023</td>
<td>107.5062</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0.0045</td>
<td>0.0137</td>
<td>0.0055</td>
<td>99.3414</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0.0006</td>
<td>0.0092</td>
<td>0.0006</td>
<td>101.1024</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>-6.2692</td>
<td>-2.1037</td>
<td>-0.0004</td>
<td>0.0587</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>47.2234</td>
<td>9.9197</td>
<td>-0.0022</td>
<td>0.0145</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>-52.8924</td>
<td>-12.6937</td>
<td>-0.0015</td>
<td>0.0114</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>3.1242</td>
<td>1.1391</td>
<td>-0.0052</td>
<td>0.1338</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>-1.0545</td>
<td>-0.4605</td>
<td>0.0090</td>
<td>0.2199</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>-9.7318</td>
<td>-4.4918</td>
<td>0.0017</td>
<td>0.0727</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>-5.6191</td>
<td>-0.4311</td>
<td>0.0102</td>
<td>0.0450</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>4.1717</td>
<td>1.0792</td>
<td>-0.0219</td>
<td>0.0548</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>17.2251</td>
<td>4.9047</td>
<td>0.0038</td>
<td>0.0447</td>
</tr>
<tr>
<td></td>
<td>a. Percentages of outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

[1] Percentages of outliers
<table>
<thead>
<tr>
<th>OT (%)</th>
<th>Estimators</th>
<th>Sample Size n=50</th>
<th>Sample Size n=100</th>
<th>Sample Size n=150</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OLS</td>
<td>KNN</td>
<td>HC0</td>
</tr>
<tr>
<td>0%</td>
<td>beta0</td>
<td>0.5146</td>
<td>0.3886</td>
<td>0.4282</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>0.8891</td>
<td>0.6044</td>
<td>0.6339</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>0.7555</td>
<td>0.6148</td>
<td>0.6451</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>23.3970</td>
<td>10.4365</td>
<td>17.1504</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>22.5848</td>
<td>12.0001</td>
<td>17.5826</td>
</tr>
<tr>
<td>15%</td>
<td>beta0</td>
<td>99.5604</td>
<td>52.3143</td>
<td>115.5253</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>131.3787</td>
<td>52.4478</td>
<td>116.7274</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>126.8181</td>
<td>40.8369</td>
<td>78.8850</td>
</tr>
<tr>
<td>20%</td>
<td>beta0</td>
<td>461.5606</td>
<td>119.7004</td>
<td>231.0199</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>609.0697</td>
<td>189.9633</td>
<td>487.4313</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>587.9272</td>
<td>206.0266</td>
<td>513.1987</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Standard errors of the parameters of the different estimators, $\lambda = 2$

a. Percentages of outliers
Table 5: Standard errors of the parameters of the different estimators, $\lambda = 4$

<table>
<thead>
<tr>
<th>%OT</th>
<th>Estimators</th>
<th>Standard errors of the Parameters</th>
<th>Sample Size n=50</th>
<th>Sample Size n=100</th>
<th>Sample Size n=150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coeff.</td>
<td>OLS</td>
<td>KNN</td>
<td>HC0</td>
<td>HC1</td>
<td>HC2</td>
</tr>
<tr>
<td>0%</td>
<td>beta0</td>
<td>0.7841</td>
<td>0.3701</td>
<td>0.6144</td>
<td>0.6536</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>1.9347</td>
<td>0.8175</td>
<td>0.9655</td>
<td>1.0271</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>1.4988</td>
<td>0.8770</td>
<td>1.0245</td>
<td>1.0899</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>47.0730</td>
<td>18.1229</td>
<td>34.6098</td>
<td>36.8189</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>45.4390</td>
<td>21.9545</td>
<td>36.0837</td>
<td>38.3869</td>
</tr>
<tr>
<td>15%</td>
<td>beta0</td>
<td>35.8009</td>
<td>14.8161</td>
<td>23.1191</td>
<td>24.5948</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>47.2424</td>
<td>20.1513</td>
<td>31.2127</td>
<td>33.2050</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>45.6025</td>
<td>23.8200</td>
<td>32.6295</td>
<td>34.7123</td>
</tr>
<tr>
<td>20%</td>
<td>beta0</td>
<td>62.2517</td>
<td>28.3594</td>
<td>42.9578</td>
<td>45.6997</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>82.1466</td>
<td>35.3214</td>
<td>58.6896</td>
<td>62.4358</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>79.2951</td>
<td>38.2728</td>
<td>59.2919</td>
<td>63.0765</td>
</tr>
<tr>
<td></td>
<td>beta0</td>
<td>0.6887</td>
<td>0.3983</td>
<td>0.5922</td>
<td>0.6105</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>1.0141</td>
<td>0.6564</td>
<td>0.7964</td>
<td>0.8210</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>0.9778</td>
<td>0.6713</td>
<td>0.8224</td>
<td>0.8478</td>
</tr>
<tr>
<td>10%</td>
<td>beta0</td>
<td>73.8892</td>
<td>21.7184</td>
<td>68.7456</td>
<td>70.8717</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>90.3146</td>
<td>28.0017</td>
<td>77.1267</td>
<td>79.5120</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>84.6461</td>
<td>30.8492</td>
<td>93.2082</td>
<td>96.0909</td>
</tr>
<tr>
<td>15%</td>
<td>beta0</td>
<td>71.6299</td>
<td>24.5261</td>
<td>65.2019</td>
<td>67.2185</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>87.5531</td>
<td>33.5901</td>
<td>81.1361</td>
<td>83.6454</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>82.0579</td>
<td>31.4783</td>
<td>75.9616</td>
<td>78.3110</td>
</tr>
<tr>
<td>20%</td>
<td>beta0</td>
<td>94.1637</td>
<td>28.8749</td>
<td>64.8414</td>
<td>66.8468</td>
</tr>
<tr>
<td></td>
<td>beta1</td>
<td>115.0960</td>
<td>46.7776</td>
<td>104.9376</td>
<td>108.1831</td>
</tr>
<tr>
<td></td>
<td>beta2</td>
<td>107.8721</td>
<td>38.3557</td>
<td>80.8973</td>
<td>83.3992</td>
</tr>
</tbody>
</table>

Note: Percentages of outliers
\[
MSE(\hat{\beta}_j) = (\bar{\beta}_j - \beta_j)^2 + \frac{1}{m} \sum_{k=1}^{m} (\hat{\beta}_j^{(k)} - \bar{\beta}_j)^2
\]  

Therefore, the root mean squared error (RMSE) is given by \([MSE(\hat{\beta}_j)]^{1/2}\). As a measure of robustness, we compute the ‘relative measure of RMSE’ which is the ratio of the RMSEs of the estimators of contaminated models compared with the least-squares estimators for good data. The relative bias and relative measure of RMSE of the OLS, KNN, and TSRWLS methods are presented in Tables 2 and 3. Several interesting points appear from Tables [2-3]. For ‘clean’ data, all the three estimators considered here are fairly close to one another with respect to the values of the robustness measure. By inspecting the bias and the values of robustness measures in Table 2 and 3, it is observed that the performance of both the OLS and the KNN tends to deteriorate with the increase in the percentage of outliers and they produce poor estimates at both levels \(\lambda \approx 2\) and \(\lambda \approx 4\) of heteroscedasticity. The performance of the TSRWLS is very satisfactory here. Irrespective of the percentages of outliers it maintains producing low bias and small RMSE.

Tables 4 and 5 present the standard errors of the parameter estimates of the OLS, KNN, five versions of HCCM, and TSRWLS estimators. We observe that the standard errors of the five versions of HCCM estimates also reasonably close to the KNN and TSRWLS, for the ‘clean’ data. If the form of heteroscedasticity is unknown, many authors recommend using the HCCM based estimators [5,6,12,13,17,18,28]. But these results clearly show that likewise the OLS and KNN, HCCM based estimators may breakdown even in a very small percentage of contamination and their performances also tend to deteriorate with the increase in the percentage of outliers. Nevertheless, the TSRWLS are not much affected by outliers. The biases and robustness measure of the TSRWLS are consistently small and deteriorate slightly as the percentage of outliers increases.

5. CONCLUSIONS

In this article, we propose a two-step robust weighted least squares estimator which is designed for handling the problem of heteroscedasticity and outliers in multiple regression when the form of the heteroscedasticity is unknown. We have examined the performance of the proposed TSRWLS estimator and compare its performance with other existing estimators. Although the KNN, HCCMs and TSRWLS estimators are reasonably close to one another in the presence of heteroscedasticity with clean data, but the TSRWLS is the most reliable estimator as it possesses the least bias and standard errors. However, the performance of KNN and HCCMs are much inferior to the TSRWLS when contamination occurred in the data. The empirical study reveals
Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers

that the proposed estimator is outlier(s) resistant. Larger bias in estimates and standard errors, and smaller values of robustness measures clearly prove that the OLS, KNN and the five versions of HCCM are easily get affected by outliers. To the contrary, both graphical and numerical evidences signify that the TSRWLS is capable of rectifying the problems of heteroscedasticity and outliers at the same time. Thus, the TSRWLS estimates emerge to be conspicuously more efficient and more reliable in comparison with other estimators considered in this article.

REFERENCES

Habshah Midi, Sohel Rana, A.H.M. Rahmatullah Imon