Malaysian Journal of Mathematical Sciences 2(1): 25-48 (2008)

The Performance of Robust Estimator on Linear Regression Model Having both Continuous and Categorical Variables with Heteroscedastic Errors

Habshah Midi and Bashar A. Talib
Laboratory of Applied and Computational Statistics
Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia
E-mail: habshahmidi@hotmail.com

ABSTRACT

The ordinary least squares (OLS) technique is often used in practice to estimate the parameters of a multiple linear regression model with both continuous and categorical variables. It has been the most popular technique due to its optimal properties and ease of computation. Nevertheless, in the presence of outliers, the OLS can result in very poor estimates. Outliers which arise from bad data points may have unduly effect on the OLS estimates. The problem is further complicated when both outliers and heteroscedasticity or non-constant error variances are present in the data. The influence of outliers and heteroscedasticity cannot be removed or reduced by simply transforming the data using known transformation such as logarithmic transformation. In this paper, we proposed a robust technique to deal with these two problems simultaneously. A robust estimate of scales for each level of categorical variables are first estimated by using robust distance S and M (RDSM) estimates. Then we determine the weighting scheme for each level of the categorical variables and transform the model. The reweighted least squares based on RDSM (RLSRDSM) is then applied to the transformed model. The empirical evidence shows that the proposed method has reduced the heteroscedastic effect to a greater extent.

Keywords: outliers, heteroscedasticity, robust Distance, RDLI, S/M estimates, RDSM.

INTRODUCTION

The classical multiple linear regression model is given by:

$$y_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

where $\varepsilon_i \sim N\left(0, \sigma^2\right)$ i = 1, 2, ….., n

and the explanatory variables x_{ij} are often quantitative. A qualitative variable may also be added in the multiple linear regression resulting with both continuous and categorical variables in the model.

This situation often occurs in the social and economical sciences, where the explanatory variables may include gender, ethnic background, professional occupation, marital status and so on.

Conventionally we encode such categorical regressors by binary dummy variables. If we have m categorical variables with c_1, c_2, ……., c_m levels, we can write:

$$y_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} + \sum_{l=1}^{q} \gamma_l \, I_{il} + \varepsilon_i \qquad (2)$$

where $q = \sum_{k=1}^{m} (C_k - 1)$ and I_{il} is either 0 or 1.

The ordinary least squares method (OLS) are often used in practice to estimate the parameters of the model. Nontheless, the OLS method is very sensitive to the presence of outliers. Outliers which arise from bad data points may have drastic effects on the OLS estimates. In order to rectify this problem, a robust method which is not sensitive to outliers is put forward. Hubert and Rousseeuw (1997) introduced the robust distance least absolute value (RDL_1) method to overcome this problem. According to Cizek (2002) and Maronna and Yohai (1999), RDL_1 suffers from several problems, such as producing non-singular degenerate solutions. To overcome this problems, Talib and Midi (2008) proposed a reweighted least squares based on RDSM. The proposed method which we called RLSRDSM is better than the RLSRDL$_1$and does not produce any singular matrices or degenerate solutions.

There are situations where the violation of constant variance comes together with the existence of outliers. This will make the analysis more complicated. The RLSRDSM cannot handle both problems simulataneously. Hubert and Rousseeuw (1997) proposed a method to deal with these problems. They estimate the coefficients and the error scale by applying the robust distance L_1 (RDL_1) procedure. The dispersion of residuals is then modeled as a function of a two-way structure, and the parameters are estimated robustly by the median polish. This was done using the logarithmic transformation on the data. Finally, they used the median of the absolute deviations from the median to determine the weight for each observation. The final estimates are obtained by using weighted

least squares (WLS) with the respective weights determined earlier. Unfortunately, this techniques may result in some problems, such as producing singular matrices or degenerate solutions as given by Maronna and Yohai (2000) and Cizek (2002). On the other hand, Chatterjee and Hadi (2006) follow a different approach to deal with the problem of heteroscedasticity and outliers, by using the OLS residuals to form the weights, that will be used to calculate the WLS estimates. They proposed a two-stages estimation procedure, where in the first stage, they calculate the regression using the raw data before transformation, and then used the empirical residuals grouped according to the categorical variables levels to compute an estimate of residual variance for that levels. In the second stage, an estimate of scale per categorical variables level was calculated and the weight is then determined. The weakness of using the second method for estimation, is that the OLS which is non robust is used to calculate the weights. In addition to that, the OLS cannot detect the existence of outliers which usually appear with the heteroscedasticity problem. The consequence of using this approach may result in an inflated values of sub variances calculated per factor variables level in the second stage of the procedure. In this paper, we propose a method that we call the weighted RLSRDSM (WRLSRDSM) to estimate the parameters of model (2) when the problems of heteroscedasticity and outliers occur together.

THE ROBUST RDL_1 ESTIMATOR

Hubert and Rousseeuw (1997) computed the RDL_1 in three stages:

i. Identify leverage points by computing the robust distance via minimum volume ellipsoid estimator (MVE).
ii. Compute the weighted L_1 weights based on the robust distance.
iii. Calculate the estimate of the scale of the residuals

iv. Identify leverage points by computing the robust distance via minimum volume ellipsoid estimator (MVE).
v. Compute the weighted L_1 weights based on the robust distance.
vi. Calculate the estimate of the scale of the residuals

Minimum Volume Ellipsoid (MVE) and the Robust Distance (RD)

Let $X = \{X_1, X_2, ..., X_n\}$ be a data set in *p*-dimensions. The robust location estimator T(X) are found by finding the center of the smallest ellipsoid containing half of X, as well as scatter matrix C(X) given by the shape of the ellipsoid. Hubert and Rousseeuw (1997) defined the robust distance as follows:

$$RD(x_i) = \sqrt{(x_i - T(X))C(X)^{-1}(x_i - T(X))'} \qquad (3)$$

where :-

x_i : (x_{i1}, x_{i2},…, x_{ip}) are the continuous variables.

X : is a data set of explanatory variables with *p*-dimensions.

T(X) : is the center of the smallest ellipsoid covering half of X.

C(X) : is the shape of the smallest ellipsoid covering half of X.

T(X) and C(X) are consistent for the underlying parameters as verified by Davis (1997). The square of the robust distance $(RD(X_i))^2$ is approximated by χ^2_P distribution as n get large if the x_i are observational (rather than designed) with a multivariate Gaussian distribution. Hence, observations for which $RD(X_i)$ is usually large relative to that distribution can be considered as leverage point.

Based on the robust distance $RD(X_i)$, the positive weights ω_i, are given by:

$$\omega_i = \min\left(1, \frac{p}{RD^2(x_i)}\right) \qquad , \qquad i= 1, 2, ..., n \qquad (4)$$

where:-

RD as given in (3) and *p* is the expected value of chi-square distribution already mentioned (it is approximately the number of independent variables). The weighted L_1 estimators (β_j, γ_1) of model (2) are found by minimizing the sum of the weighted absolute values of the residuals.

$$\min \sum_{i=1}^{n} \omega_i \left| r_i(\beta_j, \gamma_l) \right| \tag{5}$$

The solution $(\hat{\beta}, \hat{\gamma})$ can be computed by using the algorithm of the Barrodale and Roberts (1973) and Armstrong and Frome (1977) which treats the continuous and discrete (categorical) variables separately.

Reweighted Least Squares based on RDSM

Talib and Midi (2008) proposed a reweighted least squares based on RDSM. The RDSM is computed in three stages similar to that of Hubert and Rousseeuw (1997). A slight modification of the RDL_1 is proposed on the second stage, the parameter estimates of model (2) are found by minimizing the sum of the weighted S/M of the residuals.

$$\min \sum_{i=1}^{n} \omega_i \left| r_i(\beta_j, \gamma_l) \right| \tag{7}$$

The S/M is a combination of S-estimate for the continuous variables and a Huber type M-estimate with least absolute deviation (LAD) start for the factor variables. Finally, on the third stage, the scale of the RDSM is estimated by using:

$$\hat{\sigma} = 1.4826 \underset{i}{med} \left| r_i \right| \tag{8}$$

where r_j is based on the residuals of the RDSM. The choice of constant 1.4826 is to make the estimator consistent at gaussian error.

Since the estimate is a weighted L_1, by a well known property make σ underestimates the error variability and in some situation, one would even encounter $S \equiv 0$!.

As an alternative, Maronna and Yohai (1999) proposed using :

$$\hat{\sigma} = s / 0.675 \tag{9}$$

where s is the median of the nonnull residuals, s = med $(|r_1|, |r_2|, \ldots\ldots, |r_{n1}|)$ for $r_i \neq 0$.

The entire three-stages procedure is called RDSM because the procedure combined the robust distances and SM estimator.
Outliers can be detected by flagging the observations whose absolute standardized residuals $\left|\frac{r_i}{\hat{\sigma}}\right|$ are greater than 2.5 . (10)

The new parameters is then calculated by applying reweighted least squares to the data set of model (2) with weights based on $\left|\frac{r_i}{\hat{\sigma}}\right|$ to increase the finite-sample efficiency of the estimators. We refer this estimator as RLSRDSM. In so doing we will be able to employ approximate statistical inferences.

A WEIGHTED ROBUST RDSM

In this section, we propose a weighted robust RLSRDSM to estimate the parameters of model (2) when the problems of heteroscedasticity and outliers occurs together. The weights are calculated according to how much this part of the data has higher dispersion than the other. So the weighting scheme will be different for different parts of the data (or alternatively different levels of categorical variables). To achieve that, a residual scale estimate should be calculated for each level of the categorical variable. Then each observation in the data for the dependent and independent variables are divided by a suitable weight proportional to scale estimate of this part of the data, resulting in a model having constant error dispersion. First we assume that there is a unique residual variance associated with each of the levels of the categorical variable, denoted as $(c_1\sigma)^2, (c_2\sigma)^2, ...and (c_m\sigma)^2$. Following the idea of the weighted least squares, the regression coefficients is estimated by minimizing $\hat{\sigma}_{Total} = \hat{\sigma}_{Level_1} + \hat{\sigma}_{Level_2} + ...\hat{\sigma}_{Level_m}$, where,

$$\hat{\sigma}_j = \sum_{i=1}^{n_j} \frac{1}{c^2{}_j}\left(e^2{}_i\right), \qquad j = 1,2,....,l_i. \qquad (11)$$

where j refer to the number of data per categorical variable subgroup, and the sum is taken over only those observations that are in the subgroup. The

factor $\frac{1}{c^2{}_j}$ as mentioned by Chatterjee and Hadi (2006), is the weights that determine how much each observation have influence on the coefficients estimation, so the observations with large error variance should have less weight (i.e little influence in determining the coefficients).

Chatterjee and Hadi (2006) proposed two-stage estimation procedure . In the first stage, the OLS procedure for the original data before transformation is performed. The resulting empirical residuals grouped by categorical variables levels is used for the second stage to replace c_j^2 in (11) by the estimated $\hat{\sigma}_j^2$ (i.e $c_j^2 = \hat{\sigma}_j^2$; where $\hat{\sigma}_j^2$ is the usual MSE for the subgroup data per level). For the first subgroup of the data according to the categorical variable coding, the mean square residual (i.e $\hat{\sigma}_i{}^2 = \sum_{i=1}^{n_1} e^2{}_i / n_i$) depends again on the OLS is used, and so on for other subgroups.

In the current study, a robust estimate of scale for each subgroup is first calculated depending on the RDSM residuals for each subgroup as follows:

$$\hat{\sigma}_i^2 = 1.4826 * median\left(\left|e_1(RDSM)\right|, \left|e_2(RDSM)\right|, ..., \left|e_{n_1}(RDSM)\right|\right) \quad (12)$$

where i=1, 2,….,n is the subgroup index, and n is the number of observations per subgroup. Following Chatterjee and Hadi (2006), 1/ $\hat{\sigma}_i^2$ values is used as weights for the data toward homoscedasticity. Then each observation in the overall data will have a specified weight. After transforming the data in this way, the influence of outliers (as mentioned before) may still exist and the need to estimate the linear model robustly is still a serious matter. Since the model is a combined one, the need to use an alternative to OLS exist. The RDL_1 as mentioned before suffers from some calculation problems, so using the method of least squares is non robust. The need to use more suitable method for such cases arise and the RLSRDSM is still a recommended method for the transformed data that will give more efficient estimates than other alternatives.

This weighting scheme is intuitively justified by arguing that observations with the most erratic (having large error variance) should have less influence in determining the coefficients.

This way, the transformed model with the constant variance will be:

$$y_{ijk} / c_j = \beta_0 / c_j + \sum_{i=1}^{n} \beta_i (x_i / c_j) + \sum_{j=1}^{m} \alpha_j (\gamma_j / c_j) + e_{ijk} / c_j$$

or,

$$\omega_j y_{ijk} = \omega_j \beta_0 + \sum_{i=1}^{n} \beta_i \omega_j x_i + \sum_{j=1}^{m} \alpha_j \omega_j \gamma_j + \omega_j e_{ijk} \qquad (13)$$

where n and m here are the number of continuous and categorical variables respectively, also k is the observation index. i.e. y_{135} is the fifth observation of the third continuous variable and the first categorical variable.

The resulting residuals $\omega_j e_{ijk}$ will have a common variance σ^2, and the estimated coefficients have all the standard Least Squares properties.

NUMERICAL EXAMPLE

In this section, a numerical example is presented to assess the performance of the weighted RDSM estimator. The education expenditure data (Chatterjee and Hadi (2006)) is used in order to compare the robustness of RLSRDSM, RLSRDL$_1$and the OLS estimators. This data is known to have problems of outliers and heteroscedasticity. We will also examine the pattern of their residual plots visually. The data consists of the per capita expenditure on public education in a state as dependent variable (y) in 50 states of the US, along the period between 1965-1975, and three continuous variables, the per capita personal income (x_1), the number of residents per thousand under 18 years of age (x_2), and the number of residents per thousand residing in urban areas (x_3). The regression model includes in addition to the three continuous variables, two categorical variables, introduced as follows: the data are first grouped in four regions : North East (NE), North Central (NC), South (S), and West (W). Data are available for three years: 1965, 1970, and 1975. This way, it contains also two

categorical variables, regions (NE, NC, S, and W) and years (1965,1970, and 1975).

To model this data, we introduce an indicator variables for the categorical regressors. The four regions can be coded with three indicator variables (reg_1, reg_2, and reg_3) defined by: $reg_i = 1$ if the observation belongs to the reg_i category and 0 otherwise. The time periods of three years are coded analogously by two indicator variables $year_1$ and $year_2$.

With this notation, the following linear model can be formulated:

$$y_{ijk} = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{j=1}^{3} \alpha_j reg_j + \sum_{k=1}^{2} \gamma_k year_k + e_{ijk} \qquad (14)$$

For $i = 1,2,3;\ \ j = 1,2,3,$ and $k = 1,2$, and we assume as done by Hubert and Rousseeuw (1997) that the error e_{ijk} to have a unique variance (homoscedastic error variance).

For each cell there is only one data point available, so the total number of observations equals 150. The OLS, RLSRDL1, and RLSRDSM procedures were then applied to this data.

In order to estimate the parameters of model (14), robust methods of regression estimation is used in a way that can withstand a positive percentage of outliers (y-outliers), x-leverage points, or both. As mentioned before, many of the available robust procedures cannot be applied for such a mixed (combined) variables model. Then, we tried to focus especially on the re-weighted least squares procedures, since it will produce coefficients with better finite-sample efficiency than RDL_1, S/M, and RDSM methods.

Table 1 shows the estimated coefficients, and the scale estimate of each method applied to model (14) for the education expenditure data. Following Midi(1999), the value of the percentage variances accounted for, which are denoted as scale, $100\overline{R}^2$ = 100[1-(residual mean square/total mean square)] is also presented.

The RLSRDSM seems to give the optimal estimates and provide the best fit to the data, since it has the smallest scale estimate and the values of $100\overline{R}^2$ closed to 100%.

TABLE 1 : The Values of the Parameter Estimates for the Education Expenditure Data

	LS			**RLSRDL$_1$**			**RLSRDSM**		
	Value	**Std.E**	**Pr(>|t|)**	**Value**	**Std.E**	**Pr(>|t|)**	**Value**	**Std.E**	**Pr(>|t|)**
β_0	-64.44	37.80	0.09	-58.93	42.26	0.17	-52.98	37.89	0.16
β_1	0.06	0.01	**0.00**	0.04	0.01	**0.00**	0.05	0.01	**0.00**
β_2	0.23	0.08	**0.00**	0.27	0.10	0.01	0.24	0.09	**0.00**
β_3	-0.07	0.02	**0.00**	-0.02	0.012	0.34	-0.01	0.02	0.64
(region1) α_1	2.10	4.00	0.60	1.30	3.10	0.68	-0.26	2.85	0.93
(region2) α_2	-0.51	2.31	0.83	-0.61	1.73	0.72	0.91	1.55	0.56
(region2) α_3	8.51	1.52	**0.00**	6.82	1.24	**0.00**	7.15	1.11	**0.00**
(year1) γ_1	21.97	4.99	**0.00**	28.41	3.80	**0.00**	27.28	3.43	**0.00**
(year2) γ_2	-64.44	4.90	0.11	17.23	3.85	**0.00**	16.62	3.55	**0.00**
S(e)	**30.55**			**22.5**			**19.89**		
$100\overline{R}^2$	**89.39**			**92.9185**			**94.41**		

We also performed diagnostics to the data to detect if there is any deviation from the model assumptions. The Normal Q-Q plot of Fig. 1 reveals a non-normality of the error terms for the three methods.

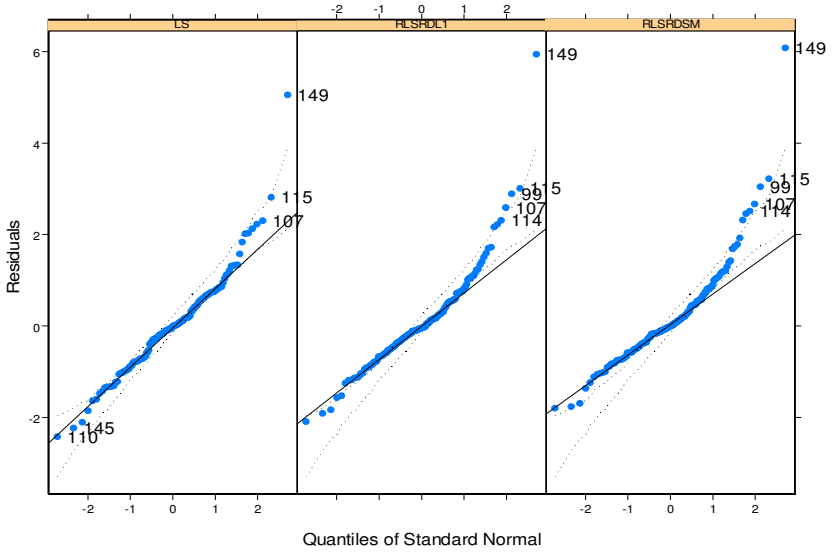

Figure 1: Normal QQ-Plot of Residuals

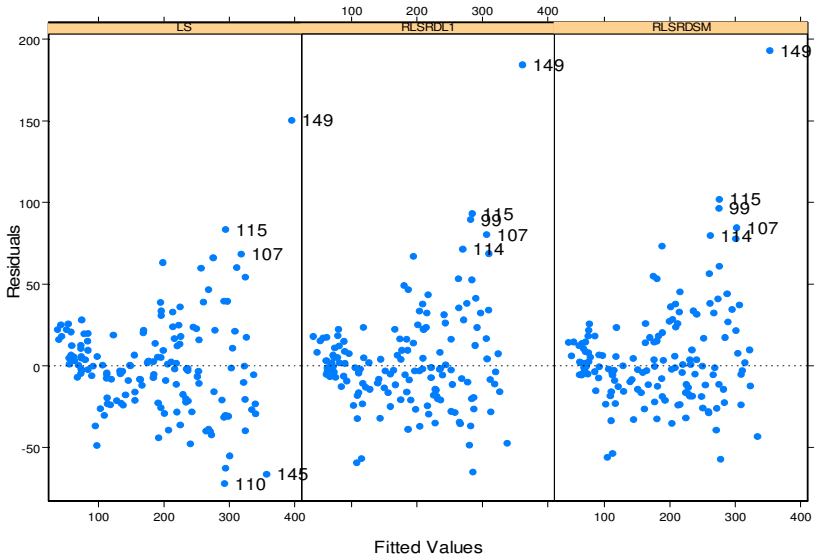

Figure 2: Residuals vs. Fitted Values

Heteroscedasticity problem as expected seems to be serious as shown in Fig. 2. The funnel shape in Fig. 2 seems more clear, revealing that the data suffer from the problem of non-constant error variance.

Figure 3 shows a scatter plot of standardized residuals versus each of the three continuous predictor variables. A serious relationship appear with the values of per capita personal income x_1, where the residual variance increase with the values of x_1. This non-constancy as we mentioned earlier can be reduced by a suitable transformation of the continuous variables. Smaller relation detected for the residuals with the number of residents per thousand residing in urban areas X_3, the relation with X_2 seems to be decreasing, where the residual variance decrease with the values of the number of residents per thousand under 18 years of age X_2. Similar results can be found in Chatterjee and Hadi (2006), but the analysis is for year 1975 only, so the relation between the standardized residuals with X_2 and X_3 have not been detected clearly, only with X_1.

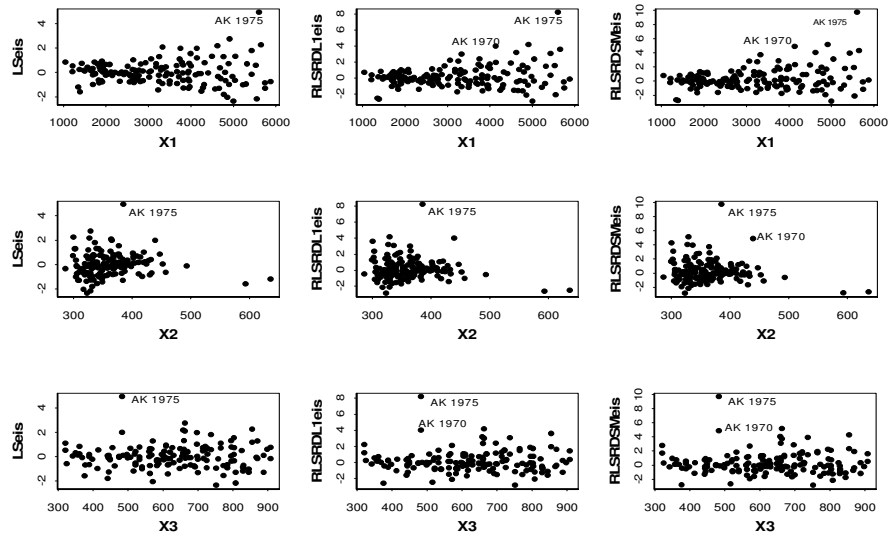

Figure 3: Plot of standardized Residuals Vs Each of the Row Variables

The state Alaska (AK) which appears in Fig. 3 separated from the bulk of the data, while less separation appear for states Alabama (AL) and Kentucky (KY) for year 1965. By looking at Figure 1-3, it is very obvious

that this data suffer from non-normality, heteroscedasticity and having outliers.

WRLSRDSM TRANSFORMATION TO HOMOGENIZE THE ERROR DISPERSION

After detecting that the residuals dispersion is non homogeneous, this problem should be treated carefully and it's influence should be lessen. This can be done by using a suitable weighting scheme as was discussed earlier. The weights can reduce the non constant dispersion of the residuals with the fitted values and can be used to calculate a weighted least squares. That will produce a results with constant error variance. Parameters of model (2) should be estimated using a weighted version of the model with weights calculated based on how much each level of the data has higher dispersion than the other, so the weighting scheme will be different for different levels of the data (different levels of the categorical variables). To achieve this, we calculate a residual scale estimate per year and per region, then divide each observation in the data (for the dependent and independent variables) by a suitable weights proportional to the scale estimate of each part of data resulting in a model having constant error dispersion. First it should assume that there is a unique residual variance associated with each of the four regions, denoted as $(c_1\sigma)^2, (c_2\sigma)^2, (c_3\sigma)^2 \; and \; (c_4\sigma)^2$. By using the concept of weighted least squares, the regression coefficients is estimated by minimizing $\hat{\sigma}_{region} = \hat{\sigma}_{NE} + \hat{\sigma}_{NC} + \hat{\sigma}_{S} + \hat{\sigma}_{W}$, where,

$$\hat{\sigma}_j = \sum_{i=1}^{n_j} \frac{1}{c_j^2}\left(e_i^2\right), \qquad j = 1,2,3,4. \tag{15}$$

In this way, the transformed model with the constant variance will be:

$$y_{ijk}/c_j = \beta_0/c_j + \sum_{i=1}^{3}\beta_i(x_i/c_j) + \sum_{j=1}^{3}\alpha_j(reg_j/c_j) + \sum_{k=1}^{2}\gamma_k(year_k/c_j) + e_{ijk}/c_j$$

or,

$$\omega_j y_{ijk} = \omega_j\beta_0 + \sum_{i=1}^{3}\beta_i\omega_j x_i + \sum_{j=1}^{3}\alpha_j\omega_j reg_j + \sum_{k=1}^{2}\gamma_k\omega_j year_k + \omega_j e_{ijk} \tag{16}$$

The resulting residuals $\omega_j e_{ijk}$ will have a common variance σ^2, and the estimated coefficients have all the standard least squares properties.

Using (12), the robust estimate of scale for the four regions depending on the RDSM residuals is given as follows:

$$\hat{\sigma}^2_{NE1965} = 1.4826 * median(|e_1(RDSM)|, |e_2(RDSM)|, ...|e_9(RDSM)|) \quad (17)$$

similarly for other three regions along the three years (i.e 1965, 1970, and 1975). The estimated region and year scale is given in Table 2. The parameters and scale estimates of the transformed data are presented in Table 3. The WRLSRDSM again gives the smallest scale estimate, since the influence of outliers is reduced by using the weighting scheme, which result in estimates that represent the bulk of data.

TABLE 2: The residual scale per region and per year

	$\hat{\sigma}_{ij}$ **(using RDSM residuals)**		
	1965	1970	1975
NE	9.973647	36.15582	44.83644
NC	12.19702	30.08506	52.14932
S	9.696966	17.85598	12.4906
W	26.50927	33.24502	27.30018

TABLE 3: The Parameter Estimates for Education Expenditure after Transformation

	WLS			**WRLSRDL$_1$**			**WRLSRDSM**		
	Value	**Std.E**	**Pr(>|t|)**	**Value**	**Std.E**	**Pr(>|t|)**	**Value**	**Std.E**	**Pr(>|t|)**
β_0	0.23	0.35	0.52	-0.18	0.32	0.57	0.22	0.25	0.38
β_1	0.06	0.00	**0.00**	0.05	0.00	**0.00**	0.05	0.00	**0.00**
β_2	0.11	0.03	**0.00**	0.25	0.04	**0.00**	0.20	0.03	**0.00**
β_3	-0.04	0.01	0.01	-0.03	0.01	0.02	-0.03	0.01	**0.00**
(region1) α_1	-40.96	7.46	**0.00**	-36.57	6.20	**0.00**	-28.60	5.04	**0.00**

(region2) α_2	-25.77	6.98	**0.00**	-25.19	5.90	**0.00**	-20.74	4.82	**0.00**
(region2) α_3	-27.58	7.03	**0.00**	-34.39	6.61	**0.00**	-24.27	5.10	**0.00**
(year1) γ_1	-33.57	9.39	**0.00**	-71.10	10.09	**0.00**	-58.60	7.97	**0.00**
(year2) γ_2	-1.90	7.91	0.81	-17.69	7.00	0.01	-16.16	6.04	0.01
S(e)	**1.188**			**0.9755**			**0.7431**		
$100\overline{R}^2$	**0.9351**			**0.9559**			**0.9660**		

Comparing Table 1 and 3, we can see that by doing the transformation to the data, the scale and the $100\overline{R}^2$ has decreased and increased respectively. Again the WRLSRDSM provide the best fit and the most efficient method. Fig. 4 through 6 show the residuals and data dispersion after transformation. It can be seen from Figure 4 and 5 that the spread of residuals has decreased and the residuals dispersion reasonably look constants.

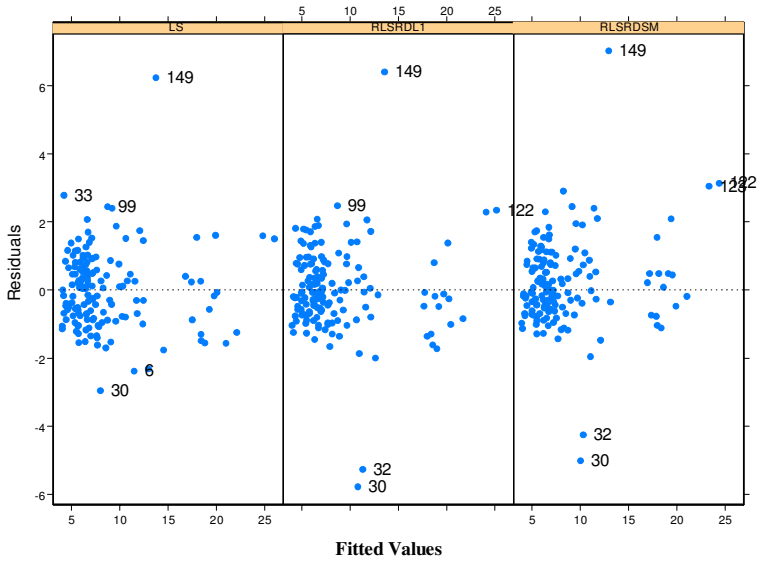

Figure 4: Residuals Vs. Fitted Values for the Transformed Data

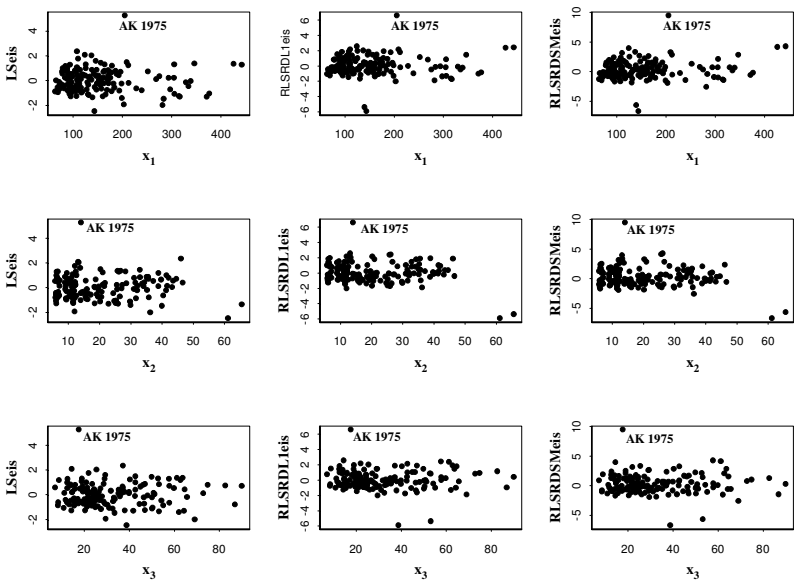

Figure 5: Plot of e_{is} versus Each of the Transformed Predictor Variable

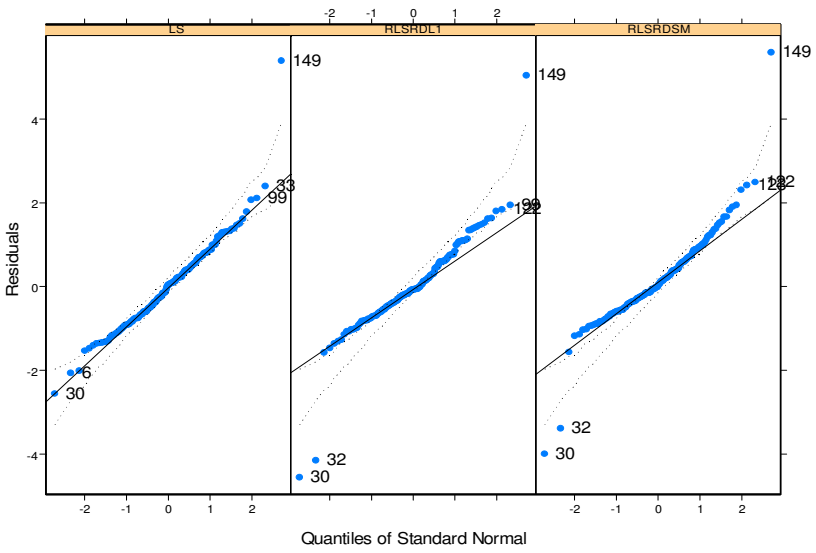

Figure 6: Normal QQ-Plot of residuals

Figure 6 of Normal Q-Q plot, again reveals the existence of outliers, and the plot reasonably more straight than Figure 4 of the original data.

Clearly from the example of education data and a few more examples not included here due to space constraint, suggest that by performing transformation, will reasonably stabilize the error variances. (All computer codes and results can be requested from the authors). We have not pursued the analysis of the example to a final conclusion, but a reasonable interpretation up to this point is that the WRLSRDSM is the least effected estimator by the combined problem of outliers and heteroscedastic followed by $WRLSRDL_1$ and WOLS. The RLSRDSM, $RLSRDL_1$ and OLS cannot cope with these problems.

SIMULATION STUDY

In order to assess the robustness of the six estimators, that is the OLS, RLSRDL1, RLSRDSM, WOLS, WRLSRDL1 and WRLSRDSM, a simulation study has been performed for two models with different dimensions, one with two continuous and two categorical variables and another model with three continuous and four ategorical variables. In fact we have performed many simulation scenarios and due to space constraints, we only report the results of the two models. The results of other models dimensions are consistent and the computer codes can be requested from the authors. Following Rousseeuw and Leroy (2003) simulation study, each of the continuous variable is generated such that $X_i \sim N(0,100)$. The models with 2 and 3 continuous variables respectively are as follows;

$$y_i = \beta_{0+} \sum_{j=1}^{2} \beta_j x_{ij} + \sum_{l=1}^{2} \gamma_l \, \mathrm{I}_{il+} \varepsilon_i$$

$$y_i = \beta_{0+} \sum_{j=1}^{3} \beta_j x_{ij} + \sum_{l=1}^{4} \gamma_l \, \mathrm{I}_{il+} \varepsilon_i$$

where β_0 is the intercept, β_j with j = 1,2,…,p are the coefficients of the linear model, i = 1,2,…,n is the index, and ε_i is the error term.

To explain the way of generating heteroscedastic errors, we first assume that there is no specific known patterns of the common heteroscedasticity (i.e when $\sigma_{e_i}^2 = \sigma_e^2 x_i^2, \sigma_e^2 x_i, \sigma_e^2 (Ey_i)$). According to that, an increasing error with respect to increasing y values is generated, by giving different dispersion (increasing or decreasing) along the sample size n. For the case of sample size equal to 100 (n=100) this can be done by writing the following code in S-PLUS:

```
S1 <- rnorm(20,0,1)
S2 <- rnorm(20,0,3)
S3 <- rnorm(20,0,5)
S4 <- rnorm(20,0,7)
S5 <- rnorm(20,0,9)
ERROR <- c(S1,S2,S3,S4,S5)
Y <- sort(Y)
```

As shown in the code above the standard deviation of the error is increasing per 20 values (1,3,5,7, then 9), and the y values is cited in ascending way (sort command in S-PLUS means to arrange the variable values ascendingly), and then having increasing error variance with respect to increased y, which is the general form of the heteroscedasticity problem.

The 4 categorical variables have been generated as factor variable with five levels resulting with four binary dummy variables. The true parameter values of the above model are such that:

$\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 1$ as suggested by Rousseeuw and Leroy (2003). We consider sample of size n= 24 and n=100 for the first and second model, respectively. For simplicity, the sample with n=24 is contaminated with 1,2, and 4 outliers which means about 4.2%, 8.3% and 16.7% outliers. The second set of data with n= 100 is contaminated with 5%, 10% and 20% outliers.

We consider three cases of generated data. The first case which is just described is the data without outliers but have heteroscedastic errors and referred as XYNORMAL. This case is just described. In the second and the third case, we generated outliers in the y (YOUTLIER) and x directions (XLEVERAGE). We deleted certain percentage of good observations as in the first case and replaced with y ouliers and x outliers for the second and

the third case respectively. The y outliers were generated by $Y_i \sim N(10,1)$. The leverage points were generated as $X_i \sim N(100,100)$. The OLS, RLSRDL1, and RLSRDSM (no transformation) and the WOLS, WRLSRDL1, and WRLSRDSM (transformation) were then applied to the simulated data. Several performance and summary measures over the two-hundred iterations (m=200) were computed:

The Mean Estimated Value :

$$\text{MEV}= \overline{\beta}_j = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_j^{(k)}$$

Variance of $\hat{\beta}_j$:

$$\text{Var}(\hat{\beta}_j)= \frac{1}{m} \sum_{k=1}^{m} (\hat{\beta}_j^{(k)} - \overline{\beta}_j)^2$$

The Bias resulting from using $\hat{\beta}_j$ to estimate β_j :

$$(\overline{\beta}_j - \beta_j)$$

The Mean Square Error (MSE (β_j))

$$\text{MSE}(\hat{\beta}_j)= (\overline{\beta}_j - \beta_j)^2 + \frac{1}{m} \sum_{k=1}^{m} (\hat{\beta}_j^{(k)} - \overline{\beta}_j)^2$$

The Root Mean Square Error (RMSE) is given by the square root of the MSE, i.e.

$$\text{RMSE}=[\text{MSE}(\hat{\beta}_j)]^{1/2}$$

Table 4-5 present only the RMSE of the parameter estimates of the models. The RMSE of the weighted estimates are in parenthesis. Other measures are not included because of space limitations. By inspecting Tables 4-5, the following conclusions can be made regarding the point estimation. We can see that by using the transformation technique has

improved the accuracy of the estimates. This is supported by looking at the RMSEs of the weighted estimator (WOLS, WRLSRDL1, WRLSRDSM) which are in most cases lower than the unweighted estimator (OLS, RLSRDL1, RLSRDSM). In the well behaved case, i.e where there are no outliers but heteroscedastic errors, the three weighted methods are closed to each other.

It is interesting to note that the presence of outliers either in the y or x directions have an effect on the parameter estimates of the categorical variables. The RMSE of the WRLSRDSM and WRLSRDL1 are reasonably smaller than the WOLS as the percentages of outliers increases. The RMSEs of the WRLSRDSM estimator is consistently slightly lower than the WRLSRDL1, although in a few cases the difference ia relatively small. Although the WRLSRDL1 appeared to be performing quite close to the WRLSRDSM, its problem of generating singular matrices and degenerate solutions limits its applicability. WOLS took the least computation time, however, it can become highly unstable and sometimes no convergence is obtained when there are outliers in the data. Naturally, the WRLSRDSM is preferred over the WRLSRDL1 and WOLS because its RMSEs are relatively slightly smaller than the other two estimates and numerically stable.

TABLE 4: RMSE values for model with 2Continuous, 2Categorical, and heteroscedastic errors; n=24, and m=200*

			XYNORMAL		
	β_0	β_1	β_2	β_3	β_4
OLS	0.3219505	1.0365071	0.9824853	1.0379773	0.8775837
WOLS	(0.2261254)	(1.0100297)	(1.0136051)	(0.9685937)	(0.5580328)
RLSRDL$_1$	0.2853840	1.0030671	0.9993191	0.7482006	0.6157523
WRLSRDL$_1$	(0.2213842)	(1.0049717)	(0.9924568)	(0.7275580)	(0.3952149)
RLSRDSM	0.2907211	0.9758210	1.0227607	0.6687230	0.5078957
WRLSRDSM	(0.1444898)	(0.9827447)	(1.0253213)	(0.4672907)	(0.1809026)
			4.2% YOUTLIER		
OLS	0.09450523	1.05435718	0.97526165	1.06016358	1.16920885
WOLS	(0.1224686)	(1.0045435)	(1.0064211)	(1.1097743)	(0.1874173)
RLSRDL$_1$	0.04589779	0.94282642	1.02901930	0.80152964	0.74463686
WRLSRDL$_1$	(0.06391192)	(0.96656835)	(0.99548353)	(0.95843551)	(0.05262141)
RLSRDSM	0.09854846	0.96907764	1.03990054	0.72502684	0.60945517
WRLSRDSM	(0.20010879)	(0.96692874)	(1.03024972)	(0.46890136)	(0.01598349)
			4.2% XLEVERAGE		
OLS	0.4251494	0.9878074	1.0007966	1.1834390	0.9269275
WOLS	(0.5318573)	(1.0135423)	(1.0221592)	(0.6609516)	(0.4569843)

$RLSRDL_1$	0.3958114	1.0186832	0.9901230	0.9497745	0.7060410
$WRLSRDL_1$	(0.4243471)	(1.0096969)	(0.9846131)	(0.5132785)	(0.3169971)
RLSRDSM	0.4045084	0.9849534	1.0216577	0.8445789	0.5716125
WRLSRDSM	(0.003537901)	(0.995618697)	(1.030125798)	(0.259302463)	(0.114590851)
			8.3% YOUTLIER		
OLS	0.5071115	1.0520870	0.9763156	1.1369068	1.4735758
WOLS	(0.01271080)	(1.00087868)	(1.00891104)	(1.17059493)	(0.07782464)
$RLSRDL_1$	0.3132786	1.0092642	1.0260427	0.9326691	1.0526465
$WRLSRDL_1$	(0.01051742)	(0.99009534)	(0.99989836)	(0.99033546)	(0.04693896)
RLSRDSM	0.2534361	1.0316115	1.0616681	0.8390891	0.8960345
WRLSRDSM	(0.20616608)	(0.99095943)	(1.00381509)	(0.43418586)	(0.07721086)
			8.3% XLEVERAGE		
OLS	0.4251494	0.9878074	1.0007966	1.1834390	0.9269275
WOLS	(0.6534483)	(1.0188099)	(1.0421668)	(0.7754247)	(0.5231300)
$RLSRDL_1$	0.3958114	1.0186832	0.9901230	0.9497745	0.7060410
$WRLSRDL_1$	(0.5200938)	(0.9952145)	(1.0145323)	(0.5845536)	(0.3649581)
RLSRDSM	0.2907211	0.9758210	1.0227607	0.6687230	0.5078957
WRLSRDSM	(0.05720748)	(0.95839936)	(1.06936744)	(0.29090268)	(0.09499025)
			16.7% YOUTLIER		
OLS	1.3302444	1.0604430	0.9833051	1.3335684	2.0216485
WOLS	(0.00680235)	(1.02146480)	(0.97617022)	(1.34229448)	(0.02373777)
$RLSRDL_1$	1.493143	1.057654	1.008476	1.141872	1.995603
$WRLSRDL_1$	(0.2002472)	(1.0104134)	(1.0230489)	(1.1759683)	(0.1092345)
RLSRDSM	1.577481	1.050906	1.010583	1.030340	1.981174
WRLSRDSM	(0.017206524)	(1.017388736)	(0.998205008)	(0.455401035)	(0.007011435)
			16.7% XLEVERAGE		
OLS	0.2895261	1.0020038	1.0130098	1.0035275	0.8591769
WOLS	(0.4594075)	(0.9989241)	(1.0317129)	(0.2935740)	(0.4997330)
$RLSRDL_1$	0.3254998	0.9927551	1.0232803	0.9029485	0.7328438
$WRLSRDL_1$	(0.4239862)	(1.0375798)	(0.8805711)	(0.1125395)	(0.3549119)
RLSRDSM	0.4580334	0.9725833	1.0016797	0.8611002	0.5794663
WRLSRDSM	(0.0669163)	(0.9512277)	(0.9865050)	(0.1365878)	(0.1917950)

TABLE 5: RMSE values for model with 3Continuous,4Categorical, and heteroscedastic errors; n=100, and m=200

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
XYNORMAL								
OLS	0.2141440	1.0096185	1.0095895	1.0067505	1.9833493	0.7943038	0.4676668	0.9940670
WOLS	(0.08614949)	(0.98912473)	(0.99586898	(0.99769769)	(0.50171628	(0.2327756	(0.20568125)	(0.36817810
RLSRDL$_1$	0.2457314	0.9830858	0.9932937	0.9983169	1.5081125	0.6109553	0.3627980	0.7116964
WRLSRDL$_1$	(0.07971148)	(1.00416061)	(1.00049142	(0.98024817)	(0.68473185	(0.1288603	(0.12435075)	(0.48572436
RLSRDSM	0.2267764	0.9951286	0.9870214	0.9975454	1.0502966	0.4847137	0.3122431	0.5256017
WRLSRDSM	(0.19403426)	(0.99458410)	(1.00182025	(0.99384200)	(0.40784234	(0.1467616	(0.08798237)	(0.34338733
5% YOUTLIER								
OLS	0.2384250	0.9844007	0.9887775	0.9898786	1.9251761	0.7575258	0.5039839	1.0722328
WOLS	(0.1307639)	(0.9888480)	(1.0010005)	(0.9877058)	(0.3862255)	(0.3163713)	(0.1362681)	(0.3347374
RLSRDL$_1$	0.3446135	0.9870919	0.9940144	0.9969406	1.3503974	0.5549136	0.3739748	0.8931730
WRLSRDL$_1$	(0.16373882)	(0.99695022)	(1.00450898	(0.98956512)	(0.59826756	(0.21104501	(0.04933537	(0.4157777
RLSRDSM	0.3962422	1.0020649	0.9901455	0.9990181	0.9688213	0.4438252	0.3476502	0.7545313
WRLSRDSM	(0.02804424)	(0.99304827)	(1.00765283	(0.98710577)	(0.35476146	(0.19289969	(0.01641317	(0.3174565
5% XLEVERAGE								
OLS	0.2680368	0.9966962	0.9877413	0.9917072	2.0363501	0.8120916	0.4754146	0.9778015
WOLS	(0.2765129)	(0.9872064)	(0.9997386)	(0.9952693)	(0.4949738)	(0.2336494)	(0.2042150)	(0.5484580)
RLSRDL$_1$	0.2133112	0.9713200	1.0002737	0.9984413	1.5087571	0.6080243	0.3623398	0.7428572
WRLSRDL$_1$	(0.2509744)	(0.9995488)	(0.9924119)	(0.9841617)	(0.6808158)	(0.1401733)	(0.1218643)	(0.6378666)
RLSRDSM	0.2116992	0.9938584	0.9850635	0.9961756	1.0478086	0.4818860	0.3109402	0.5388270
WRLSRDSM	(0.32455395)	(0.99192502)	(1.00062812	(0.99785241)	(0.42198567	(0.13901502	(0.09288285	(0.50501703)
10% YOUTLIER								
OLS	0.7273686	0.9828681	0.9870513	0.9956904	1.9220805	0.7666989	0.5962815	1.2097823
WOLS	(0.41822160)	(0.99763255)	(0.99903695	(0.99117712)	(0.26801334	(0.35862937	(0.02353971	(0.18549406)
RLSRDL$_1$	0.8722677	0.9930408	0.9943947	1.0012479	1.3406568	0.5561043	0.4553933	1.0650064
WRLSRDL$_1$	(0.46495494)	(0.98745497)	(0.98323945	(1.00053198)	(0.47759484	(0.25535697	(0.05615224	(0.25626898)
RLSRDSM	0.9751119	0.9956395	0.9867671	0.9895575	0.9224134	0.4374482	0.4250247	0.9670837
WRLSRDSM	(0.34329753)	(0.99734153)	(0.99887757	(0.98552708)	(0.29447102	(0.22735659	(0.07437805	(0.18193830)
10% XLEVERAGE								
OLS	0.2833100	0.9975232	0.9877696	0.9992115	2.0371396	0.8039088	0.4682698	0.9781895
WOLS	(0.40523364)	(0.98774360	(0.99697321)	(0.99465943)	(0.50608647	(0.18718262	(0.08928638	(0.51055000)
RLSRDL$_1$	0.2478631	0.9699690	0.9961517	1.0013570	1.4954739	0.5705492	0.3517814	0.7530187
WRLSRDL$_1$	(0.382911606	)(0.9976246	(1.008068278	)(0.98883257	)(0.7062737	(0.09331587	)(0.0076151	)(0.60214884
RLSRDSM	0.2466741	0.9919301	0.9900075	1.0013925	1.0585435	0.4625009	0.3042554	0.5334607
WRLSRDSM	(0.424669980	)(0.9931032	)(0.99935883)	(0.994705044	)(0.4583859	)(0.0909581	(0.00638466	(0.494191772
20% YOUTLIER								
OLS	1.7283910	0.9806078	0.9863155	0.9869153	2.0609726	0.9615479	1.0612274	1.3246827
WOLS	(1.26811492))	(0.99968122	(0.98738579)	(0.99554006)	(0.07510197	(0.38361189	(0.06534709	(0.40777977)

RLSRDL$_1$	1.8751124	0.9852952	0.9856361	1.0022024	1.4922844	0.7325244	0.9679926	1.1690873
WRLSRDL$_1$	(1.2965338)	(0.9863553)	(0.9927975)	(0.9901475)	(0.2745598)	(0.2756073)	(0.1117486)	(0.3121012)
RLSRDSM	2.0009517	0.9918789	0.9843991	0.9879310	1.0887448	0.6403657	0.8825399	1.1320573
WRLSRDSM	(1.0765790)	(0.9927046)	(0.9964525)	(0.9861749)	(0.1561043)	(0.2365982)	(0.1248226)	(0.2799626)
20% XLEVERAGE								
OLS	0.2381965	1.0009539	0.9912117	0.9970098	2.0324146	0.8055382	0.4739848	0.9878226
WOLS	(0.26609544)	(0.98844792	(0.99775128)	(0.99494832)	(0.56879079	(0.09235665	(0.03113616	(0.54419863)
RLSRDL$_1$	0.1989392	0.9897655	0.9964881	0.9865932	1.3201804	0.5344882	0.3326646	0.7382679
WRLSRDL$_1$	(0.25018521)	(0.98789663	(1.01197487)	(0.97953362)	(0.73974697	(0.02241450	(0.03686682	(0.63062556)
RLSRDSM	0.1508558	0.9991801	0.9904307	0.9992685	1.0394173	0.4521225	0.2887934	0.6211943
WRLSRDSM	(0.30714751)	(0.98484006	(0.99013302)	(0.98714460)	(0.65167013	(0.02453643	(0.03332308	(0.54438447)

CONCLUSION

The unweighted estimates are not efficient in the situation of the heteroscedastic errors. The empirical study shows that the weighting scheme has improved the accuracy of the three estimates. It appears that the performances of the three methods are equally good in a well behaved data, data with heteroscedastic errors without outliers. The WOLS method is not robust where outliers are present in the data. The WRLSRDSM is slightly better than WRLSRDL$_1$ and sometimes their performances are indistinguishable. Nevertheless the WRLSRDL$_1$ posed certain computational problems such as singular matrix and degenerate solution through it's many null errors produced. The WRLSRDSM does not face any computational problem. The result of this preliminary studies suggest that the WRLSRDSM is the best choice for handling problems of heteroscedasticity and outliers in the data set.

REFERENCES

Bashar, A.T. and Midi, H. 2008. Robust Efficient Estimator to deal with Regression Models Having both Continuous and Categorical Regressors. Sent to *European Journal of Scientific Research*.

Chatterjee S., and Hadi A. 2006. *Regression Analysis by Example*. 4th Ed. New York: John Wiley.

Cizek, P. 2002. Robust estimation with discrete explanatory variables, Working Paper, Institute für Statistik und Ökonometrie, CASE, Humboldt-Universität zu Berlin.

Hubert, M., and Rousseeuw, P.J. 1997. A regression analysis with categorical covariables, two-way heteroscedasticity, and hidden outliers, in *The Practice of Data Analysis: Essays in Honor of J.W. Tukey*, (edited by D.R. Brillinger, L.T. Fernholz and S. Morgenthaler). Princeton, New Jersey, Princeton University Press, 193-202.

Hubert, M. and Rousseeuw P.J. 1997. Robust regression with both continuous and binary regressors. *Journal of The Statistical Planning and Inference,* 57: 153-163.

Lin, C. 1998. A weighted least squares approach to robustify least squares estimate, Ph.D. thesis, University of Minnesota, June 1998, U.S.A.

Maronna, R. and Yohai V. 2000. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference,* 89 : 197-214.

Midi, H. 1999. Preliminary estimators for robust non-linear regression estimation. *Journal of Applied Statistics*. Vol. 26 (5), 591-600.

Rousseeuw P.J. and Leroy A.M. 2003. *Robust Regression and Outlier Detection*. New York: John Wiley.

Rousseeuw, P. J., and van Zomeren, B. C. 1990. Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85, 633-639.

Rousseeuw P.J. and Yohai V.J. 1984. Robust regression by means of S-estimators. in Robust and Nonlinear Time Series Analysis, eds. J. Franke. W. Härdle and R.D. Martin. *Lecture Notes in Statistics* 26, New York: Springer Verlag.