Malaysian Journal of Mathematical Sciences 3(2): 161-181 (2009)

Robust Estimator to Deal with Regression Models Having both Continuous and Categorical Regressors: A Simulation Study

[1]Bashar A. Talib and [1,2]Habshah Midi
[1,2]Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
[2]Laboratory of Applied and Computational Statistics
Institute for Mathematical Research (INSPEM),Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia
E-mail: bashar@math.upm.edu.my,habshah@putra.upm.edu.my

ABSTRACT

The Ordinary Least Squares (OLS) method has been the most popular technique for estimating the parameters of the multiple linear regression. However, in the presence of outliers and when the model includes both continuous and categorical (factor) variables, the OLS can result in poor estimates. In this paper we try to introduce an alternative robust method for such a model that is much less influenced by the presence of outliers. A numerical example is presented to compare the performance of the OLS, the Re-weighted Least Squares based on the Robust Distance Least Absolute Value ($RLSRDL_1$), and the Re-weighted Least Squares based on the Robust Distance S/M estimator (RLSRDSM). The latter is the modification of the RDL_1. The empirical evidence shows that the performance of the RLSRDSM is fairly close to the $RLSRDL_1$ up to 20% outliers. As the percentage of outliers increases to more than 20%, the RLSRDSM is slightly better than the $RLSRDL_1$. However, the Robust Distance Least Absolute Value (RDL_1) estimator posed certain computational problems such as degenerate non-unique solutions while the RLSRDSM do not have such problem.

Keywords: Outliers, Leverage points, Robust Distance, S/M-estimates, $RLSRDL_1$, RLSRDSM

INTRODUCTION

Consider the general multiple linear regression model with additive error term:

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

where $\varepsilon_i \sim N(0, \sigma^2)$ $i = 1, 2, \ldots, n$

Additional linear model parameters may be added to model (1) when some of the independent variables are qualitative. In this situation, dummy predictor variables which only take the values 0 and 1 are incorporated in model (1). The model is then extended to cater both continuous and categorical variables.

If we have m categorical variables with $c_1, c_2,..., c_m$ levels, then (1) becomes:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{l=1}^{q} \gamma_l I_{i1} + \varepsilon_i \qquad (2)$$

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. Nevertheless, the OLS method is very sensitive to the presence of outliers. In order to rectify this problem, a robust method which is not sensitive to outliers is put forward.

M-estimation which was proposed by Huber,(1973) is frequently used method of robust regression. Armstrong and Frome,(1977) introduced the Least Absolute Value (L_1) method which is robust against outliers in the y-direction, but does not protect against leverage points, i.e points of which $(x_{i1}, x_{i2}, \ldots, x_{ip})$ are outlying. Similar to L_1, the M-estimates are still not robust to the leverage points.

Rousseeuw,(1984) and Rousseeuw and Yohai,(1984) have introduced the Least Median of Squares (LMS), the Least Trimmed of Squares (LTS) and the class of S-estimator that can withstand a positive percentage of contaminations including leverage points. However, these estimators cannot simply be applied to model (2) by treating the dummy variables (I_{il}) in the same way as the continuous regressors, since this would lead to a problem of singular matrices.

Hubert and Rousseeuw,(1997) introduced the robust distance Least Absolute Value (RDL_1) method to overcome this problem. Nevertheless, according to Cizek,(2002) and Maronna and Yohai,(1999), RDL_1 suffers from several problems, such as producing non-singular degenerate solutions and underestimating the error variances.

In this paper, we propose a Re-weighted LS based on a weighted combination between the S-estimates and the M-estimates (RDSM) and refer this estimates as RLSRDSM. The RLSRDSM is based on Weighted Least Squares (WLS) to use as an alternative to OLS and $RLSRDL_1$. Using WLS will increase the finite sample efficiency of the estimates. The weights are calculated by using the standardized residuals results from performing a weighted S/M-estimators (RDSM).

We expect that the performance of the proposed method will be close to the $RLSRDL_1$, not produce any singular matrices or degenerate solutions.

THE ROBUST RDL_1 ESTIMATOR

Hubert and Rousseeuw,(1997) describe 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

Identification of Leverage Points by Minimum Volume Ellipsoid (MVE)

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:

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

where :-

x_i : $(x_{i1}, x_{i2}, \ldots, x_{ip})$ 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,(1992). The square of the robust distance $(RD(X_i))^2$ is approximated

by χ^2_P distribution as n gets large if the x_i are observed (rather than designed) with a multivariate Gaussian distribution. Hence, observations for which $(RD(X_i))^2$ is larger $\chi^2_{P,\alpha}$ can be considered as leverage point.

Computation of Weighted L_1 based on Robust Distance

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

$$\omega_i = \min\left(1, \frac{p}{(RD(X_i))^2}\right) \quad \text{for} \quad i= 1, 2, ..., n \tag{4}$$

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_l) of model (2) are found by minimizing the sum of the weighted absolute values of the residuals $r_i(\beta_j, \gamma_l)$,

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

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.

Scale of the Residuals for LAV and RDL_1.

The residuals scale is estimated as proposed by Hubert and Rousseeuw, (1997) by

$$\hat{\sigma} = 1.4826 \underset{i}{med} \left| r_i \right| \tag{6}$$

Where r_i is the LAV residual.

The choice of constant 1.4826 (the tuning constant) 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:

$$\hat{\sigma} = s/0.675 = 1.4826\, s \tag{7}$$

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

Outliers can be detected by flagging the observations whose absolute standardized residual $\left|\frac{r_i}{\hat{\sigma}}\right|$ are greater than 2.5. Rousseeuw and Leroy,(2003) states that the 2.5 is arbitrary, but quite reasonable because in a Gaussian situation there will be very few residuals larger than $2.5\hat{\sigma}$.

RE-WEIGHTED LEAST SQUARES BASED ON RDL$_1$

Hubert and Rousseeuw,(1997) proposed applying Re-weighted Least Squares to the data set of model (2) with weights based on $\left|\frac{r_i}{\hat{\sigma}}\right|$ to increase the estimators finite-sample efficiency, where r_i is the RDL$_1$ residual. The weight is given by

$$\omega_i = \begin{cases} 1 & if \quad \left|\frac{r_i}{\hat{\sigma}}\right| \leq 2.5 \\ 0 & if \quad \left|\frac{r_i}{\hat{\sigma}}\right| > 2.5 \end{cases} \tag{8}$$

We refer this estimator as RLSRDL$_1$ (Re-weighted Least Squares based on RDL$_1$). In so doing we will be able to employ approximate statistical inferences.

The scale estimate or the residual errors of the OLS and RLSRDL$_1$ are:

$$\hat{\sigma}_{OLS} = \sqrt{(n-p-q-1)^{-1}\sum_{i=1}^{n} r_i^2} \tag{9}$$

$$\hat{\sigma}_{RLSRDL_1} = \sqrt{(\sum_{i=1}^{n} \omega_i - p - q - 1)^{-1} \sum_{i=1}^{n} \omega_i r_i^2} \tag{10}$$

respectively.

RE-WEIGHTED LEAST SQUARES BASED ON RDSM

Maronna and Yohai,(1999) introduced S-estimate/M-estimate (S/M-estimate) method for fitting linear models with both continuous and categorical predictor variables.

The M-estimator in the S/M-estimate was first introduced by Huber in 1964 (Lin,1998), and it is a class of estimators that minimize a function ρ of the residuals as follows:

$$\min \sum_{i=1}^{n} \rho(e_i) = \min \sum_{i=1}^{n} \rho(y_i - x_i'\beta) \tag{11}$$

where x_i' denote the i-th row of the independent variables matrix *X*.

This estimator is called the M-estimator, and it is a maximum likelihood estimators when the error distribution is chosen appropriately (i.e. choosing the appropriate objective function that is optimal with respect to the distribution of the error term).

The S-estimate was introduced for the first time by Rousseeuw and Yohai,(1984) as the method of estimation that can make a specified scale estimator to have minimum value, and can be defined as:

$$\hat{\beta}_S = \arg\min_{\beta} S(\beta) \tag{12}$$

where $S(\beta)$ is a certain type of M-estimate of the scale of the residuals $r_1(\beta),...,r_n(\beta)$ as given by Rousseeuw and Leroy,(2003), and have essentially asymptotic performance the same as M-estimators, but with high breakdown point, where:

$$\frac{1}{n-p} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \beta}{\hat{S}(\beta)} \right) = 0.5 \tag{13}$$

S/M-estimates uses S-estimate for the continuous variables and for the categorical variables uses M-estimate with a least absolute deviation (L_1) influence function, depending on the fact that there is no leverage points among the categorical predictor variables.

In this paper, we proposed a Re-weighted Least Squares based on RDSM. The RDSM is computed in three stages similar to that of Hubert and Rousseeuw,(1997). The identification technique of the leverage points is the same like the RDL_1. Once the $RD(x_i)$ is identified, the weight ω_i is determined by Equation (4) and then used the weights in Equation (5). The weighted S/M estimators (β_j, γ_l) of model (2) are found by minimizing the sum of the weighted S/M values of the residuals $r_i(\beta_j, \gamma_l)$ instead of the weighted absolute values of residuals, as given in Equation (5). The RDSM residuals are then computed to obtain the residual scale as in Equation (7). The weighting scheme proposed by Hubert and Rousseew,(1997) in Equation (8) is then computed and used for the Re-weighted Least Squares based on RDSM.

A NUMERICAL EXAMPLE

In this section, we consider two data sets for assessing the performance of the RLSRDSM.

Wagner Data

Wagner Data Set which has been analyzed by Wagner,(1994), Hubert and Rousseeuw,(1997), Maronna and Yohai,(1999) and S-PLUS 6 Robust Library User's Guide,(2002) is used. This data presents the rate of employment growth(y) corresponding to four continuous explanatory variables:

PA: percentage of people engaged in production activities
HS: higher services
GPA: growth of PA
GHS: growth of HS

The rate of employment growth (y) depends also on the geographical region and the time period, where the data consist of 21 regions around Hanover in three time periods P_1= 1979-1982, P_2 : 1983-1988, and P_3 : 1989-1992. The

final model contain four continuous explanatory variables and two factor (categorical) variables. The OLS, $RLSRDL_1$, RLSRDSM were then applied to these data. The standardized residuals for each estimator are computed. The index plot of the standardized residuals are plotted in Figure 1.

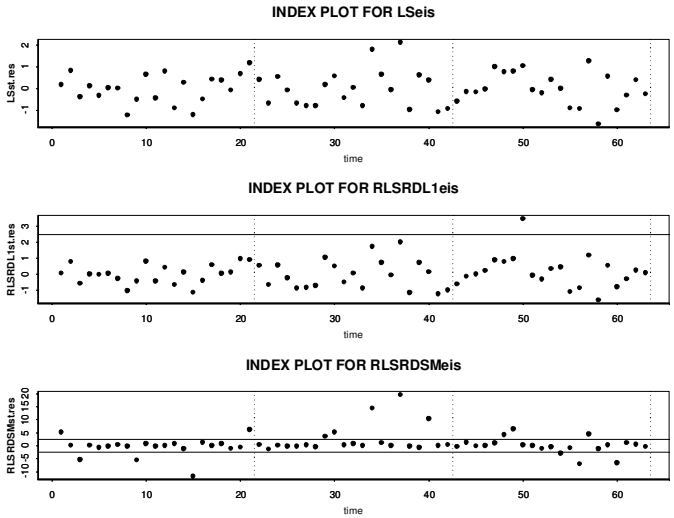

Figure1: index plots of the standardized residuals

The results of Figure 1 show that the OLS cannot detect any outlier. On the other hand, only one and Six-teen standardized residuals of $RLSRDL_1$ and RLSRDSM respectively which is greater than 2.5. The RLSRDSM detects more outliers than the $RLSRDL_1$. We then examined the scale, the standard errors and the p-values which are presented in Table 1.

TABLE 1 : The Residual errors s(e), standard errors of parameters estimates and p-value

	Method	*LS*			$RLSRDL_1$			RLSRDSM		
Coeff.	.	Value	Std.E	Pr (>|t|)	Value	Std. Error	Pr (>|t|)	Value	Std. Error	Pr (>|t|)
β_0 (intercept)		4.76	23.60	0.84	32.38	26.76	0.23	-59.3	8.12	0.00
β_1 (Period1)		1.95	1.49	0.20	2.14	1.44	0.15	1.99	0.50	0.00
β_2 (Period2)		1.09	1.10	0.33	1.31	1.06	0.23	2.83	0.36	0.00
β_3 (GHS)		6.15	1.72	0.00	4.28	1.91	0.03	6.18	0.53	0.00
β_4 β_4(HS)		5.09	2.11	0.02	1.79	2.64	0.50	5.19	0.74	0.00
β_5 (GPA)		0.07	0.56	0.90	-0.35	0.58	0.55	2.35	0.20	0.00
β_6 (PA)		-0.41	0.71	0.56	-0.98	0.74	0.19	1.60	0.24	0.00
β_7 (Region1)		-5.11	5.59	0.37	-7.81	5.56	0.17	11.13	2.02	0.00
β_8 (Region2)		2.14	3.51	0.55	-0.17	3.58	0.96	12.88	1.24	0.00

TABLE 1(continued): The Residual errors s(e), standard errors of parameters estimates and p-value

	Method	*LS*			RLSRDL$_1$			RLSRDSM		
Coeff.	.	Value	Std.E	Pr (>|t|)	Value	Std. Error	Pr (>|t|)	Value	Std. Error	Pr (>|t|)
β_9 (Region3)		1.21	1.59	0.45	0.01	1.65	1.00	4.73	0.54	0.00
β_{10} (Region4)		-0.12	1.13	0.92	-1.09	1.19	0.37	1.52	0.35	0.00
β_{11} (Region5)		-2.51	0.87	0.01	-1.52	0.98	0.13	-3.57	0.29	0.00
β_{12} (Region6)		-0.77	0.73	0.30	-0.09	0.79	0.91	-0.89	0.22	0.00
β_{13} (Region7)		-2.47	1.26	0.06	-3.17	1.27	0.02	0.04	0.42	0.93
β_{14} (Region8)		0.45	0.48	0.36	0.92	0.52	0.09	0.20	0.21	0.35
β_{15} (Region9)		-0.03	0.40	0.93	-0.04	0.39	0.92	0.50	0.12	0.00
β_{16} (Region10)		0.21	0.43	0.62	0.26	0.41	0.53	-0.36	0.12	0.01
β_{17} (Region11)		-0.10	0.32	0.77	-0.07	0.31	0.83	-0.07	0.12	0.57
β_{18} (Region12)		0.57	0.34	0.10	0.32	0.35	0.37	0.34	0.11	0.01
β_{19} (Region13)		-0.17	0.35	0.62	-0.05	0.34	0.88	-0.58	0.11	0.00
β_{20} (Region14)		-0.27	0.31	0.38	-0.53	0.33	0.11	0.08	0.13	0.53
β_{21} (Region15)		-0.31	0.33	0.36	-0.12	0.33	0.72	-1.50	0.11	0.00
β_{22} (Region16)		0.40	0.41	0.34	0.44	0.39	0.27	0.42	0.13	0.00
β_{23} (Region17)		-0.03	0.25	0.92	0.22	0.27	0.43	-0.17	0.08	0.04
β_{24} (Region18)		0.51	0.43	0.24	0.64	0.42	0.14	-0.67	0.15	0.00
β_{25} (Region19)		-0.03	0.21	0.89	-0.03	0.21	0.90	-0.16	0.06	0.02
β_{26} (Region20)		-1.37	0.45	0.00	-0.96	-1.97	0.06	-0.91	0.15	0.00
S(e)		6.229			5.999			1.66		

The results of Table 1 show that the RLSRDSM does a credible job. The RLSRDSM method outperforms the RLSRDL$_1$ and OLS by possessing the lowest residual standard errors, lowest standard errors of the parameter estimates and the most number of significant parameters.

Salary Survey Data

Our second example presents a Salary Survey Data wchich is taken from Chatterjee, Hadi, and Price,(2000). The response variable of this data is Salary (S), and the predictors are : (1) experience (X) measured in years; (2) education (E), coded as 1 for those completing a high school (H.S.) diploma, or 2 for bachelor degree (B.S.), and 3 for advanced degree; and (3) management (M), which is coded as 1 for a person with management responsibility and 0 otherwise. The Salary Survey model will then be:

$$S = \beta_0 + \beta_1 X + \gamma_1 E1 + \gamma_2 E2 + \delta_1 M + \varepsilon \quad (14)$$

Table 2 presents the summary statistics such as the standard errors of the parameter estimates, the p-values and the scale estimates, s(e).

TABLE 2: The Residual errors s(e), standard errors of parameters estimates and p-value

	LS			$RLSRDL_1$			RLSRDSM		
	Value	Std.E	Pr(>|t|)	Value	Std.E	Pr(>|t|)	Value	Std.E	Pr(>|t|)
β_0	11031.81	383.22	0.00	11199.98	33.75	0.00	9204.14	45.88	0.00
β_1	546.18	30.5192	0.00	498.32	2.65	0.00	498.14	2.30	0.00
β_2	-2996.21	411.75	0.00	-1740.58	45.63	0.00	257.02	49.34	0.00
β_3	147.82	387.66	0.70	-356.47	42.00	0.00	1640.80	37.94	0.00
β_4	6883.53	313.92	0.00	7041.07	43.97	0.00	9038.20	35.05	0.00
S(e)	1027			73.88			70.09		

We observe from Table 2 that the RLSRDSM produce the smallest scale estimates followed by the $RLSRDL_1$ and OLS. The p-values of the three estimators are fairly close. However, the standard errors of the OLS is much larger than the $RLSRDL_1$ and RLSRDSM. Although RLSRDSM produces lower scale or residual standard errors than the $RLSRDL_1$, we observe that the standard errors of the $RLSRDL_1$ estimates are reasonably close to the RLSRDSM. We have not pursued the analysis of the examples to a final conclusion, but a reasonable interpretation up to this point is that the performance of the RLSRDSM is slightly better than the $RLSRDL_1$. These two estimators outperform the OLS estimator.

SIMULATION STUDY

In this section, a simulation study will be discussed in order to compare the OLS, $RLSRDL_1$, and RLSRDSM. We have performed many simulation scenarios and due to space constraints, we include only 3 tables. The conclusions of other results are consistent and are not presented due to space limitations. All computer codes and results can be requested from the authors.

In this section, we consider three models namely the models with 1,3, and 5 continuous variables. The variables are generated according to

Rousseeuw and Leroy,(2003) simulation study. The explanatory variables are generated such that $X_i \sim N(0,100)$. The models with 3 and 5 continuous variables respectively can be constructed as:

$$y_i = \beta_0 + \sum_{j=1}^{3} \beta_j x_{ij} + \sum_{l=1}^{4} \gamma_l I_{il} + \varepsilon_i \tag{15}$$

$$y_i = \beta_0 + \sum_{j=1}^{5} \beta_j x_{ij} + \sum_{l=1}^{4} \gamma_l I_{il} + \varepsilon_i \tag{16}$$

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, where $\varepsilon_i \sim N(0,\sigma^2)$.

The 4 categorical variables have been generated as factor variables with five levels resulting in four binary dummy variables, for each categorical variable.

On the other hand, for the simple case of one continuous and one categorical variable, the continuous and categorical variables are generated by following Cizek,(2002) simulations, as Normal and Binomial variables respectively, i.e. $i = 1,...,n$; $X_i \sim N(0,10)$ instead of $X_i \sim N(0,100)$; and $I_{i1} \sim Bi(1,0.5)$ as a factor variable;

$$y_i = \beta_0 + X_i + \gamma I_{i1} + \varepsilon_i \tag{17}$$

We used a standardized version of X_i following Rousseeuw and Leroy,(2003) where,

$$z_{ij} = \frac{x_{ij} - \underset{k}{med}\, x_{kj}}{1.4826 \underset{f}{med} \left| x_{fj} - \underset{k}{med}\, x_{kj} \right|} \tag{18}$$

where $j = 1,..., p\text{-}1$, $z_{ip} \equiv x_{ip} \equiv 1$ for the intercept term, and y can be standardized in the same manner.

The error term is generated as Normal, Student's-*t*, and Exponential distribution for the simple case (one continuous and one categorical). Error term is distributed first as Normal, then as student's-t with three degrees of freedom, and finally with Exponential distribution with parameter one, i.e. a distribution with heavy tails.

The population regression models for models (15),(16), and (17) 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).

For models (15),(16), and (17), we consider three cases of generated data, X and y without outlying observations, this case is called (XYNORMAL). Then contamination of the data was commenced. At each step, 'good' observations were deleted and replaced with 'bad' observations. The contaminated data with percentage of outliers only in the y direction is refer as YOUTLIER while contaminated data with percentage of leverage points is refer as XLEVERAGE. The contaminant data points are generated from normal distribution with different mean and different variances. In the YOUTLIER case, the response was contaminated with data values distributed as N(100,10). Similarly, the contaminated explanatory variables were generated from normal distribution, that is N(100,100).

Performance Measures

Several performance and summary measures over the two-hundred iterations (m=200) were computed:

i. The Mean Estimated Value :

$$\text{MEV} = \overline{\beta}_j = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_j^{(k)} \tag{19}$$

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

$$\text{Var}\,(\hat{\beta}_j) = \frac{1}{m} \sum_{k=1}^{m} (\hat{\beta}_j^{(k)} - \overline{\beta}_j)^2 \tag{20}$$

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

$$(\overline{\beta}_j - \beta_j) \tag{21}$$

iv. The Mean Square Error (MSE (β_j))

$$\text{MSE}(\hat{\beta}_j) = (\overline{\beta}_j - \beta_j)^2 + \frac{1}{m}\sum_{k=1}^{m}(\hat{\beta}_j^{(k)} - \overline{\beta}_j)^2 \tag{22}$$

The Root Mean Square Error RMSE

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

$$\text{RMSE} = [\text{MSE}(\hat{\beta}_j)]^{1/2} \tag{23}$$

Only RMSE will be tabulated to represent the performance of each method tested. This measure sums up and summarizes all other performance measures. Due to space constraints, we only consider 20% outliers for model (18) and 10%, 20%, 30%, for models (15) and (16).

Table 3, Table 5 and Table 7 summarizes the RMSE of the coefficients and scale for the three methods tested under the given simulation conditions. In order to get a better picture, the performance of the three estimators are evaluated by using the sum ranks of their RMSE's. An estimator with lower sum ranks indicates a better performance than an estimator with higher sum ranks. The sum ranks of each estimator are presented in Table 4, Table 6 and Table 8.

TABLE 3: RMSE values under 20% contamination percentage

		$\varepsilon_i \sim N(0,1)$			$\varepsilon_i \sim t(3)$			$\varepsilon_i \sim EXP(1)$		
CASE	Coef.	**OLS**	**RLSRDL$_1$**	**RLSRDSM**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**
XYNORMAL	β_0	0.75003	0.75195	0.74954	0.73718	0.75141	0.75288	0.24407	0.01816	0.00101
	β_1	0.00430	0.00129	0.00095	0.03008	0.01739	0.01430	0.00170	0.00619	0.00499
	γ	0.74969	0.74768	0.74751	0.74892	0.741662	0.74598	0.76566	0.75582	0.75508
	σ	0.03146	0.04328	0.07960	0.71061	0.09661	0.02607	0.01433	0.38732	0.41063

TABLE 3 (continued): RMSE values under 20% contamination percentage

		$\varepsilon_i \sim N(0,1)$			$\varepsilon_i \sim t(3)$			$\varepsilon_i \sim EXP(1)$		
CASE	Coef.	**OLS**	**RLSRDL$_1$**	**RLSRDSM**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**
YOUTLIER	β_0	1.24918	0.74840	0.75032	1.26357	0.72664	0.74566	2.04588	0.16196	0.14513
	β_1	0.02610	0.00222	0.00094	0.01612	0.01369	0.01530	0.02974	0.00866	0.01078
	γ	0.71499	0.74597	0.74594	0.71718	0.74242	0.73711	0.72838	0.75289	0.75208
	σ	3.25081	0.00349	0.00921	3.42822	0.27621	0.24105	2.84881	0.18742	0.21388
XLEVERAGE	β_0	0.78779	0.76492	0.77265	0.78276	0.77025	0.78697	0.19978	0.07645	0.05109
	β_1	0.97779	0.15658	0.27025	0.97470	0.24359	0.38750	0.97463	0.08796	0.25721
	γ	0.74760	0.74952	0.74351	0.74816	0.73562	0.74123	0.76392	0.75348	0.74787
	σ	0.45379	0.04797	0.01800	0.98757	0.28918	0.14536	0.44553	0.27153	0.22984

TABLE 4: The sum ranks values of RMSE under 20% contamination percentage

		$\varepsilon_i \sim N(0,1)$			$\varepsilon_i \sim t(3)$			$\varepsilon_i \sim EXP(1)$		
CASE	**Coef.**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**	**OLS**	**RLSRDL$_1$**	**RLSRDSM**
XYNORMAL	β_0	1	3	2	1	2	3	3	2	1
	β_1	3	2	1	3	2	1	1	3	2
	γ	3	2	1	3	1	2	3	2	1
	σ	1	2	3	2	3	1	1	2	3
Sum		8	9	7	9	8	7	8	9	7
YOUTLIER	β_0	3	1	2	3	1	2	3	2	1
	β_1	3	2	1	3	1	2	3	1	2
	γ	1	3	2	1	3	2	1	3	2
	σ	3	1	2	3	2	1	3	1	2
Sum		10	7	7	10	7	7	10	7	7
XLEVERAGE	β_0	3	1	2	2	1	3	3	2	1
	β_1	3	1	2	3	1	2	3	1	2
	γ	2	3	1	3	1	2	3	2	1
	σ	3	2	1	3	2	1	3	2	1
Sum		11	7	6	11	5	8	12	7	5

Let us first focus to Tables 3 and Table 4 for model with one continuous and one categorical variable. Several interesting points emerge from these tables. For the clean data (without outlying observation), the OLS and the

RLSRDSM are reasonably close to each other. In this situation The $RLSRDL_1$ is slightly inferior than the other two estimators. However, as the percentage of outliers increases, the OLS immediately affected by outliers. The RMSEs and the sum ranks of the OLS is the largest among the three estimators. The performance of the RLSRDSM is close to the $RLSRDL_1$ at 20% outliers in the case of one continuous and one categorical variable, irrespective of their error distributions.

TABLE 5: RMSE values for model with 3Continuous and 4Categorical variables

Case	Contam. Percent. %	Coeff. \ Method	β_0	β_1	β_2	β_3	γ_1	γ_2	γ_3	γ_4
XYNORMAL	0%	**OLS**	0.20665	0.00011	0.00015	0.00020	0.99458	0.98847	1.00951	1.20459
		$RLSRDL_1$	0.20438	0.00590	0.01035	0.02486	0.99931	0.98783	1.01252	1.20841
		RLSRDSM	0.20314	0.01094	0.01470	0.01448	0.99154	0.98683	1.00706	1.20666
YOUTLIER	10%	**OLS**	0.79268	0.00002	0.00009	0.00044	0.99674	1.31321	0.92543	0.95491
		$RLSRDL_1$	0.20262	0.00009	0.00015	0.00017	0.99894	0.99206	1.00880	1.20422
		RLSRDSM	0.20828	0.00009	0.00017	0.00018	0.99790	0.98976	1.00841	1.20539
	20%	**OLS**	1.79184	0.00023	0.00014	0.00060	0.98125	1.48910	1.13824	1.07725
		$RLSRDL_1$	0.21016	0.00009	0.00014	0.00019	0.99936	0.99187	1.01094	1.20625
		RLSRDSM	0.21062	0.00009	0.00015	0.00019	0.99955	0.99178	1.01046	1.20654
	30%	**OLS**	2.79276	0.00006	0.00013	0.00047	0.73511	1.07584	1.04698	1.33081
		$RLSRDL_1$	0.20240	0.00010	0.00012	0.00015	1.00990	0.99371	1.00720	1.20623
		RLSRDSM	0.27715	0.000001	0.00009	0.00048	0.73477	0.10764	0.10463	0.13307

TABLE 5(continued): RMSE values for model with 3Continuous and 4Categorical variables

Case	Contam. Percent. %	Method \ Coeff.	β_0	β_1	β_2	β_3	γ_1	γ_2	γ_3	γ_4
XLEVERAGE	10%	**OLS**	20.82304	0.28954	0.29550	0.30326	4.97703	0.16736	2.37686	1.47411
		RLSRDL$_1$	0.20960	0.00019	0.00008	0.00016	0.99365	0.99112	1.00678	1.20550
		RLSRDSM	0.20668	0.00016	0.00015	0.00021	0.99165	0.99159	1.00684	1.20572
	20%	**OLS**	32.13253	0.46478	0.44760	0.45177	0.41127	6.85260	1.01324	3.52442
		RLSRDL$_1$	0.21389	0.00014	0.00009	0.00012	0.99857	0.98973	1.00879	1.20709
		RLSRDSM	0.21056	0.00010	0.00014	0.00018	0.99976	0.99124	1.00952	1.20638
	30%	**OLS**	37.4465	0.59365	0.57408	0.56676	3.79062	0.16589	0.56695	0.17105
		RLSRDL$_1$	4.46823	0.05837	0.05544	0.05772	1.66615	0.85646	0.89773	0.93365
		RLSRDSM	1.60140	0.01835	0.01781	0.01423	1.08096	1.32545	0.83013	1.19472

Let us now focus to the results in Table 5 and Table 6 for model with three continuous and four categorical variables. As can be expected, similar results are obtained as with model with one continuous and one categorical variable in the case of clean data. In this situation, the OLS and the RLSRDSM are equally good and their results are superior than the RLSRDL$_1$. It is interesting to point out that the OLS is the least affected by outliers up to 10% with the RLSRDL$_1$ being the next least affected estimator. However, the RLSRDL$_1$ is reasonably close to the RLSRDSM up to 20% outliers and their performances are superior than the OLS. The RLSRDSM is the most efficient estimator followed by the RLSRDL$_1$ and OLS at 30% outliers. By looking at Tables 7 and Table 8 for model with five continuous and four categorical variables, reveal that the OLS and the RLSRDSM estimates are consistently close to each other for clean data.

TABLE 6: The sum ranks values of RMSE for model with 3Continuous and 4Categorical variables

Case	Contam. Percent. %	Method \ Coeff.	β_0	β_1	β_2	β_3	γ_1	γ_2	γ_3	γ_4	Sum
XYNORMAL	0%	**OLS**	3	1	1	1	2	3	2	1	14
		RLSRDL$_1$	2	2	2	3	3	2	3	3	20
		RLSRDSM	1	3	3	3	1	1	1	2	15

TABLE 6 (continued): The sum ranks values of RMSE for model with 3Continuous and 4Categorical variables

Case	Contam. Percent. %	Method \ Coeff.	β_0	β_1	β_2	β_3	γ_1	γ_2	γ_3	γ_4	Sum
YOUTLIER	10%	**OLS**	3	1	1	3	1	3	1	1	14
		RLSRDL$_1$	1	2	2	1	3	2	3	2	16
		RLSRDSM	2	2	3	2	2	1	2	3	17
	20%	**OLS**	3	2	1	2	1	3	3	1	14
		RLSRDL$_1$	1	1	1	1	2	2	2	2	12
		RLSRDSM	2	1	2	1	3	1	1	3	14
	30%	**OLS**	3	2	3	2	2	3	3	3	21
		RLSRDL$_1$	1	3	2	1	3	2	2	2	16
		RLSRDSM	2	1	1	3	1	1	1	1	11
XLEVERAGE	10%	**OLS**	3	3	3	3	3	1	3	3	22
		RLSRDL$_1$	2	2	1	1	2	2	1	1	12
		RLSRDSM	1	1	2	2	1	3	2	2	14
	20%	**OLS**	3	3	3	3	1	3	3	3	22
		RLSRDL$_1$	2	2	1	1	2	1	1	2	12
		RLSRDSM	1	1	2	2	3	2	2	1	14
	30%	**OLS**	3	3	3	3	3	1	1	1	18
		RLSRDL$_1$	2	2	2	2	2	2	3	2	17
		RLSRDSM	1	1	1	1	1	3	2	3	13

However, as the percentage of outliers increases, the OLS is immediately affected by outliers. The presence of outliers in the data changes the situation dramatically. In this situation, the OLS has the largest RMSE and the largest sum ranks values as the percentage of outliers increases. It is interesting to note the results of Tables 7 and Table 8. The performance of the RLSRDL is reasonably close to the RLSRDSM when the percentage of outliers is up to 20% and when the outliers are in the x and y -directions. On the other hand, the RLSRDL1 estimates emerge to be conspicuously more efficient than the other two estimators at 30% outliers in the y direction. Nonetheless, the RLSRDSM is more efficient than the RLSRDL1 at 30% outliers in the x- direction evident by its smallest sum ranks values.

TABLE 7: RMSE values for model with 5Continuous and 4Categorical variables

Case	Contam. Percent. %	Method \ Coeff.	β_0	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	γ_3	γ_4
XYNORMAL	0%	**OLS**	0.20455	0.00000	0.00008	0.00005	0.00001	0.00001	1.01318	0.99834	0.99941	1.19958
		RLSRDL$_1$	0.20464	0.00003	0.00005	0.00006	0.00009	0.00002	1.01799	1.00053	1.00159	
		RLSRDM	0.20423	0.00002	0.00001	0.00002	0.00005	0.00000	1.01407	1.00264	1.00321	
YOUTLIER	10%	**OLS**	0.79273	0.00011	0.00009	0.00027	0.00012	0.00054	1.00763	1.32066	0.91947	0.94560
		RLSRDL$_1$	0.20013	0.00006	0.00005	0.00001	0.00011	0.00006	1.01618	0.99840	1.00205	1.19475
		RLSRDSM	0.20721	0.00009	0.00008	0.00002	0.00007	0.00008	1.01440	0.99646	1.00248	1.19759
	20%	**OLS**	1.78988	0.00040	0.00023	0.00043	0.00029	0.00012	1.01292	1.49207	1.13315	1.07231
		RLSRDL$_1$	0.19647	0.00008	0.00010	0.00004	0.00013	0.00006	1.01820	1.00221	1.00296	1.19735
		RLSRDSM	0.19955	0.00008	0.00011	0.00005	0.00010	0.00007	1.01779	1.00086	1.00234	1.19821
	30%	**OLS**	2.80291	0.00004	0.00026	0.00028	0.00054	0.00036	0.75206	1.07435	1.05195	1.32101
		RLSRDL$_1$	0.12154	0.00004	0.00009	0.00004	0.00006	0.00003	0.98851	1.00544	1.01045	1.20942
		RLSRDSM	2.80025	0.00005	0.00027	0.00029	0.00057	0.00037	0.75099	1.07390	1.05215	1.32302
XLEVERAGE	10%	**OLS**	31.5544	0.39623	0.36605	0.37130	0.37749	0.38051	7.06475	1.90489	4.00353	1.54586
		RLSRDL$_1$	0.20391	0.00011	0.00003	0.00011	0.00001	0.00006	1.01224	0.99834	0.99802	1.19877
		RLSRDSM	0.20402	0.00001	0.00008	0.00004	0.00005	0.00001	1.01539	0.99915	0.99946	1.19868
	20%	**OLS**	45.49237	0.56627	0.56087	0.52574	0.54875	0.57279	0.67560	10.40737	1.14265	4.32376
		RLSRDL$_1$	0.21062	0.00001	0.00003	0.00011	0.00005	0.00002	1.01750	1.00028	1.00241	1.19919
		RLSRDSM	0.49041	0.00140	0.00337	0.00022	0.00204	0.00209	1.22202	0.81376	1.01277	1.24072
	30%	**OLS**	50.8405	0.70328	0.64544	0.64992	0.65832	0.66921	4.18745	1.21691	0.83202	0.49696
		RLSRDL$_1$	31.6677	0.39952	0.35578	0.35497	0.37545	0.36902	3.99434	2.04105	0.40710	0.33036
		RLSRDSM	19.4142	0.24411	0.21038	0.22716	0.22917	0.22395	2.81532	0.53308	0.14868	0.50932

Table 8: Sum ranks values of RMSE for model with 5Continuous and 4Categorical variables

Case	Contam. Percent. %	Method \ Coeff.	β_0	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	γ_3	γ_4	**Sum**
XYNORMAL	0%	**OLS**	2	1	3	2	1	2	1	1	1	1	15
		RLSRDL$_1$	3	3	2	3	3	3	3	2	2	0	24
		RLSRDM	1	2	1	1	2	1	2	3	3	0	16

TABLE 8 (continued): Sum ranks values of RMSE for model with 5Continuous and 4Categorical variables

Case	Contam. Percent. %	Method \ Coeff.	β_0	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	γ_3	γ_4	Sum
YOUTLIER	10%	**OLS**	3	3	3	3	3	3	1	3	1	3	26
		RLSRDL$_1$	1	1	1	1	2	1	3	2	2	2	16
		RLSRDSM	2	2	2	2	1	2	2	1	3	1	18
	20%	**OLS**	3	2	3	3	3	3	1	3	3	1	25
		RLSRDL$_1$	1	1	1	1	2	1	3	2	2	2	16
		RLSRDSM	2	1	2	2	1	2	2	1	1	3	17
	30%	**OLS**	3	2	3	3	3	3	1	3	3	1	25
		RLSRDL$_1$	1	1	1	1	2	1	3	2	2	2	16
		RLSRDSM	2	1	2	2	1	2	2	1	1	3	17
XLEVERAGE	10%	**OLS**	3	3	3	3	3	3	3	3	3	3	30
		RLSRDL$_1$	1	2	1	2	1	2	1	1	1	2	14
		RLSRDSM	2	1	2	1	2	1	2	2	2	1	16
	20%	**OLS**	3	3	3	3	3	3	3	3	3	3	30
		RLSRDL$_1$	1	1	1	2	1	2	1	2	1	1	13
		RLSRDSM	2	2	2	1	2	1	2	1	2	2	17
	30%	**OLS**	3	3	3	3	3	3	3	2	3	2	28
		RLSRDL$_1$	2	2	2	2	2	2	2	3	2	1	20
		RLSRDSM	1	1	1	1	1	1	1	1	1	3	12

CONCLUSION

The main focus of this paper is to propose an alternative approach to deal with regression models having both continuous and categorical variables. We have considered the RLSRDSM in this regard. The empirical studies and simulation experiments show that the OLS is easily affected by outliers. The RLSRDSM is reasonably close to the RLSRDL1 up to 20% outliers for model having one continuous and one categorical variable and model with three continuous and four categorical variables. In the case of five continuous and four categorical variables, the RDLSRDL1 is the most efficient estimator up to 30% outliers in the y-direction. The result of this preliminary studies suggest that the RLSRDSM is slightly better than the RLSRDL1 when the percentage of outliers is at 30%. The RLSRDSM estimator has no computational problems, and do not produce any singular matrices or degenerate solutions while RLSRDL1 faces these problems.

REFERENCES

Armstrong, R. D., and Frome, E. L. 1977. A Special Purpose Linear Programming Algorithm for Obtaining Least Absolute Value Estimators in a Linear Model with Dummy Variables. *Comm. Stat*., B6,**4**: 383-398.

Barrodale, I. and Roberts, F. D. K. 1973. An Improved Algorithm for Discrete L_1 Linear Approximations. *SIAM Journal of Numerical Analysis.* **10**: 839-848.

Chatterjee, S., Hadi A., and Price, B. 2000. *Regression analysis by example*, 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.

Davies, L. 1992. The asymptotics of Rousseeuw's minimum volume ellipsoid estimator, *Annals of statistics,* **20**: 1828-1843.

Huber, P.J. 1973. Robust estimation of location parameter. *Annals of Math. Stat*., **35**: 73-101, 1973.

Hubert, M. and Rousseeuw, P.J. 1997. Robust regression with both continuous and binary regressors. *J. 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, U.S.A.

Maronna, R. and Yohai, V. 1999. Robust regression with both continuous and categorical predictors, *J. of statistical planning and inference*, **89**: 197-214

Rousseeuw, P.J. 1984. Least median of squares regression. *JASA*, **79**: 871-880.

Rousseeuw, P.J. and Leroy, A.M. 2003. *Robust regression and outlier detection*. New York: John Wiley.

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.

S-PLUS 6 Robust Library User's Guide. 2002. Version 1.0. Insightful Corporation, Seattle, Washington.

Wagner, J. 1994. Regionale Besch~iftigungsdynamik und hfherwertige Produktionsdienste: Ergebnisse f'firden Grossraum Hannover (1979-1992). *Raumforschung und Raumordnung.* **52**: 146-150.