AJAS / SPAS

ISSN 2316-0861

African Journal of Applied Statistics
Vol. 4 (1), 2017, pages 231–240.
DOI: http://dx.doi.org/10.16929/ajas/2017.231.212

ISSN 2316-0861

Robust Regression Using Multiple Repeated Median of Slope

Bamidele Mustapha Oseni[1*]

[1]Department of Statistics, Federal University of Technology, Akure Nigeria

Received June 03, 2017; Accepted December 30, 2017; Published Online February 6, 2018

Abstract. A generalization of the Repeated Median of Slope (RMS) is carried out to accommodate multiple regression models. This is obtained through the investigation of the behavior of total change in the dependent variable with respect to an independent variable. The asymptotic behavior of the estimator is investigated when certain percentage of the observations come from contamination-outlier generating unknown distribution. The performance of the estimator is compared with that of the ordinary least square (OLS) and huber estimator using bias, variance and RMSE. Expectedly, the estimator is consistent and more efficient than the OLS and huber when the observations are contaminated with outlier. Though using RMSE as evaluation criteria its performance is comparably very poor.

Key words: Outliers, Robust Statistic, Estimator, Repeated Median of Slope, Multiple Repeated Median of Slope, Contamination, Median, Regression, Monte Carlo.
AMS 2010 Mathematics Subject Classification : 62Gxx, 62Jxx

Presented by Dr Anani Lotsi, University of Ghana, Legon, Accra
Corresponding Member of the Editors Board.

---

*Bamidele Mustapha Oseni: bmoseni@futa.edu.ng

Résumé : Une généralisation de la méthode de la médiane répétée du coefficient à l'origine est mise en oueuvre pour améloirer les méthodes de regression multiples. Ceci est opérée par l'investigation du comportement de la variable dépendante par rapport à la variable indépendante lorsque la premiére subit un changement total. Le comportement asymptotique est étudiée lorsqu'un certain pourcentage des observations provient d'une distribution inconnue générant des contaminations tendant à l'aberration (outliers). L'estimateur obtenu est plus performant, dans les conditions décrites plus hait, que ceux de la méthode des moindres carrés et de la méthode Uber par le biais et la variance. Cependant, par le RMSE, elle est moins efficace.

1. Introduction

Statistical inferences are based in part upon the observations and equally in part upon the assumptions about the underlying distribution Huber (1981). These assumptions are generally formalization of the conjecture about an often blurred knowledge of the data set. However, the formalizations are simplifications of reality and their validity is at best approximate. The normality assumption is central to most classical methods. Justification of the normality assumption has been unsuccessfully attempted by several authors in the past but it is widely believe that the assumption gives approximate representation to many real data sets and at the same time is theoretically convenient Maronna *et al.* (2006). In practice, though most observations could be accounted for by underlying normal distribution, but the underlying distributions of some or part of observations are always non-normal. The non-normality of some or part of observations makes the classical method unsuitable for such observations. Such observations which are legitimate and yet comes from non-normal distribution could be said to be outlying observation if they are farther apart from the rest of the observations.

The Least Square method proposed by Legendre in 1805 is a classical method based on normality of the random term. This method was widely accepted as it was the only method of estimation that could be effectively computed before the advent of electronic computer Stigler (1986). But despite the effectiveness of this method, it has a breakdown point of zero (like most classical method). That is, the maximum fraction of contamination the estimate can tolerate before its value is completely determined by the contaminating data is zero Yohai (1987). A number of estimators with higher breakdown points have been proposed. Notably among are the Median based estimators like the Median of Pair-wise slope (MPS) by Thiel (1950) and Sen (1968), Generalized Median of Slope (GMS) by Brown and Mood (1951) and Repeated Median of Slope(RMS) by Siegel (1982). Most of these estimators have higher breakdown points than the Least Square but RMS is the most desirable due to its breakdown point of 0.5. Since the introduction of this procedure by Siegel (1982), researchers have investigated some methodology relating to the procedure while others are basically interested in the application. Among such researchers are Borowski and Fried (2011), Bernholt *et al.* (2006), Bernholt and Fried (2003) and Fried *et al.* (2003)

Unlike the least square method, RMS can only be used to estimate the slope of simple linear regression. Though other median based estimators for multiple regression such as Least Median of Slopes (LMS) by Rousseeuw and Wagner (1994) have been developed, the

generalization of the RMS into multiple repeated median of slope is still appealing due to its high breakdown point. This work attempts to generalize the RMS estimator into Multiple Repeated Median of Slope (MRMS) for the estimation of the parameters of multiple linear regression.

2. Generalization

Consider the regression through the origin

$$y_i = \beta x_i + \epsilon_i. \tag{1}$$

The slope of Siegel's Repeated Median regression line Siegel (1982) is define as

$$\hat{\beta}^{RMS} = Med_{1\leq i\leq n} Med_{j\in J_i} r(i,j). \tag{2}$$

where $J_i = \{j : (i,j) \in Integer\}$, for $1 \leq i \leq n$.

The ratio $r(i,j)$ is defined as

$$r(i,j) = \frac{y_j - y_i}{x_j - x_i}; \quad x_j \neq x_i. \tag{3}$$

where (x_i, y_i) and (x_j, y_j) are independent. Equation (3) can be written in a more compact form as

$$r = \frac{\delta y}{\delta x}; \quad \delta x \neq 0. \tag{4}$$

Suppose model (1) is a multiple regression through the origin given by

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \epsilon_i \tag{5}$$

Following a similar notation as Eq. (4),the slope of y with respect to x_1 can be written as

$$r_1 = \frac{\delta y}{\delta x_1} = \frac{\delta y}{\delta x_1}\frac{\delta x_1}{\delta x_1} + \frac{\delta y}{\delta x_2}\frac{\delta x_2}{\delta x_1} + \cdots + \frac{\delta y}{\delta x_m}\frac{\delta x_m}{\delta x_m}. \tag{6}$$

where $\delta x_1, \delta x_2, \ldots, \delta x_m \neq 0$. That is, the slope of Equation (5) with respect to x_1 can be written as

$$\begin{aligned} r_1(i,j) = \frac{\delta y}{\delta x_1} = &\left(\frac{y_j - yi}{x_{1j} - x_{1i}}\right)\left(\frac{x_{1j} - x_{1i}}{x_{1j} - x_{1i}}\right) + \left(\frac{y_j - yi}{x_{2j} - x_{2i}}\right) \\ &\left(\frac{x_{2j} - x_{2i}}{x_{1j} - x_{1i}}\right) + \cdots + \left(\frac{y_j - yi}{x_{mj} - x_{mi}}\right)\left(\frac{x_{mj} - x_{mi}}{x_{1j} - x_{1i}}\right). \end{aligned} \tag{7}$$

In general, the slope of y with respect to x_k is :

$$r_k = \frac{\delta y}{\delta x_k} = \frac{\delta y}{\delta x_1}\frac{\delta x_1}{\delta x_k} + \frac{\delta y}{\delta x_2}\frac{\delta x_2}{\delta x_k} + \cdots + \frac{\delta y}{\delta x_k}\frac{\delta x_k}{\delta x_k} + \cdots + \frac{\delta y}{\delta x_m}\frac{\delta x_m}{\delta x_k}. \tag{8}$$

for $\delta x_1, \delta x_2, \ldots, \delta x_m \neq 0$ and $1 \leq k \leq m$; $\{k \in Integer\}$, implying;

$$r_k(i,j) = \frac{\delta y}{\delta x_k} = \left(\frac{y_j - yi}{x_{1j} - x_{1i}}\right)\left(\frac{x_{1j} - x_{1i}}{x_{kj} - x_{ki}}\right) + \left(\frac{y_j - yi}{x_{2j} - x_{2i}}\right)\left(\frac{x_{2j} - x_{2i}}{x_{kj} - x_{ki}}\right) + \cdots + \left(\frac{y_j - yi}{x_{kj} - x_{ki}}\right)\left(\frac{x_{kj} - x_{ki}}{x_{kj} - x_{ki}}\right) + \cdots + \left(\frac{y_j - yi}{x_{mj} - x_{mi}}\right)\left(\frac{x_{mj} - x_{mi}}{x_{kj} - x_{ki}}\right). \quad (9)$$

Let $g_k(i,j) = \dfrac{y_j - yi}{x_{kj} - x_{ki}}$ and $g_{hk}(i,j) = \dfrac{x_{hj} - x_{hi}}{x_{kj} - x_{ki}}$

Thus $r_k(i,j)$ can be represented as

$$r_k(i,j) = g_1.g_{1k}(i,j) + g_2.g_{2k}(i,j) + \cdots + g_k.g_{kk}(i,j) + \cdots + g_m.g_{mk}(i,j). \quad (10)$$

Thus the multiple repeated median is written as[1]

$$\hat{\beta}_k^{MRMS} = \sum_{h=1}^{m} \left[(Med_{1\leq i\leq n} Med_{j\in J_i} g_h(i,j))(Med_{1\leq i\leq n} Med_{j\in J_i} g_{hk}(i,j))\right]. \quad (11)$$

3. Algorithm

Suppose the regression model (5) is written in matrix form as

$$y = X\beta + \epsilon, \quad (12)$$

where y is $n \times 1$ vector of dependent variable, X is matrix of independent variables and $\beta = (\beta_1, \beta_2, \ldots \beta_m)$ is the vector of parameters to be estimated. The algorithm goes as follows:

1. Repeat the following step for $j = 1(1)n$, where $h = 1$ and $i = 1$:
   (a) Compute
   $$g_h = \frac{y_i - y_j}{x_{hi} - x_h j}; \qquad (i,j) \in I_2, j = 1(1)n,$$
   , where $\dfrac{y_i - y_j}{x_{hi} - x_h j} = 0$, if $i = j$. This gives a row vector with n elements.
   (b) Obtain the median of the elements. Denote this median by
   $$Med_{j\in J_i} g_1(1,j)$$
   .
2. Repeat *step*1 above for $i = 2(1)n$ and $h = 1$. This gives n column vector of medians ($i = 1$ inclusive)
3. Obtain the median of the column vector in *step*2 and denote it by
   $$Med_{1\leq i\leq n} Med_{j\in J_i} g_1(i,j)$$
   .

[1] For a simple linear regression model, $m = 1$ and $h = k = 1$, thus Eq. (11) reduces to RMS.

4. For $k = 1$, carry out $steps 1 - 3$ above on

$$g_{hk} = \frac{x_{hi} - x_{hj}}{x_{ki} - x_{kj}}; \qquad (i, j) \in I_2, j = 1(1)n,$$

where $\frac{x_{hi} - x_{hj}}{x_{ki} - x_{kj}} = 0$, if $i = j$ and denote the median by

$$Med_{1 \leq i \leq n} Med_{j \in J_i} g_{11}(i, j)$$

.

5. Obtain the product

$$(Med_{1 \leq i \leq n} Med_{j \in J_i} g_1(i, j)) \,.\, (Med_{1 \leq i \leq n} Med_{j \in J_i} g_{11}(i, j))$$

.

6. Repeat $step 1 - 5$ above for $h = 2(1)m$
7. Obtain the sum;

$$\hat{\beta}_1^{MRMS} = \sum_{h=1}^{m} \left[(Med_{1 \leq i \leq n} Med_{j \in J_i} g_h(i, j))(Med_{1 \leq i \leq n} Med_{j \in J_i} g_{hk}(i, j)) \right].$$

. This is the estimate of β_1.

8. Repeat $step 1 - 6$ for $k = 2(1)m$ to obtain estimates of other parameters.

4. Monte Carlo

A Monte Carlo experiment is conducted to examine the performance and sensitivity of the estimator to different fractions of contamination. Assumed values are assigned to the parameters of the model and the variables are simulated in three different stages:

1. The regressors are simulated from uniform distribution according to Kmenta (1971). This is done with the preconception of testing the suitability of the estimator for economic data, since most economic time series data are positive numbers.
2. The random errors are simulated from $N(0, 1)$
3. Using the assumed values of the parameters, the simulated regressors and random errors, the regressand is obtained using model (5)

The φ-contamination neighborhood defined by Tukey (1960) is employed to allow for outliers in one of the regressors. This neighborhood is defined as

$$\mathfrak{F}_\varphi(F_\theta) = \{F : F = (1 - \varphi)F_\theta + \varphi F^*\}, \tag{13}$$

where φ is the fraction of contamination, F_θ is the distribution of the regressors; assumed to be a uniform$(0, 1)$ distribution; and F^* is the contamination-outlier generating unknown distribution. F^* is assumed to be uniform$(7, 10)$. This design gives control over the fraction of outliers in the simulation. The experiment is carried out for sample sizes (n) of 10, 20, 40, 60, 80 and 100 in other to examine both the small sample and large sample properties of the estimator. The contamination levels (φ) are set at 0.0, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6.

Table 1. Simulated Sample of size $n = 20$ and $\varphi = 0.2$

n	y	x_1	x_2	x_3	e
1	3.9208678	7.696101178	0.98675716	0.33107634	0.38917344
2	1.5877772	7.614867472	0.72719996	0.03299742	-1.12868708
3	0.3884104	9.776816593	0.13738814	0.76277599	-0.09546952
4	2.1795572	9.721272234	0.61011762	0.81987150	-1.03705413
5	3.5537455	0.266365227	0.65817356	0.47676172	-0.33193291
6	1.4631162	0.005320937	0.68434302	0.39929991	-0.87587793
7	2.4110814	0.708098713	0.17853648	0.20310558	0.91457350
8	1.4269872	0.477091049	0.72996698	0.02361824	0.11124994
9	4.7310033	0.334091914	0.49334642	0.92244026	-0.57565853
10	2.2506782	0.042785810	0.99334720	0.68291477	0.74083597
11	1.9876764	0.316999218	0.20024144	0.89857855	-0.14300568
12	5.0363247	0.907181498	0.91401146	0.48215578	-1.02542318
13	−0.8603620	0.028331155	0.41152147	0.06107988	1.54798131
14	3.7697408	0.334414804	0.73388429	0.84995186	1.37866657
15	1.1732457	0.045177345	0.24205835	0.83040722	-1.84259309
16	0.7929323	0.733478248	0.80711575	0.36932534	1.01564571
17	1.8999194	0.056418553	0.55033752	0.88539232	-0.45762652
18	3.6672891	0.590559146	0.56523018	0.76656049	0.75489645
19	4.7607825	0.724298270	0.84155729	0.68817198	-0.91297653
20	4.7607825	0.724298270	0.84155729	0.68817198	1.57328223

The experiment is replicated $1,000$ times.

Table 1 is a sample data from the simulation of a regression model with three regressors. The parameters of the model (β) are set at $(1.2 \quad 2 \quad 1.5)\prime$ in accordance with matrix form in Eq. (12). The sample size $n = 20$ and φ is set at 0.2. Setting $\varphi = 0.2$ allows for four sample points to come from the contamination-outlier generating distribution $U(7, 10)$ and the rest from the main distribution $U(7, 10)$. All simulations and computations programs are done using R statistical software.

This design not only allows for violation of regression distributional assumption which motivates the development of this estimator but also allows for certain number of outliers in a regressors. The performance of the estimator when no assumption is violated is investigated by assuming $\varphi = 0.0$

5. Analysis and Discussion of Results

The estimator is evaluated using asymptotic bias, variance and root mean square errors (RMSE). The results are examined both across the sample sizes and fraction of contamination. Comparison is carried out between the asymptotic properties of the estimator and that of both ordinary least square (ols) and huber estimator. For ease of evaluation, the Euclidean norms; of the values from the criteria; over the variables are used instead of the real values from the criteria. Table 2 below shows the values obtained for the bias both across the sample sizes and the fraction of contamination φ.

Table 2. Comparison of Bias of MRMS, OLS and Huber estimator

φ	Estrs	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
	mrms	0.079337	0.049610	0.017282	0.020060	0.021903
0.0	ols	0.023034	0.012189	0.013445	0.009198	0.010548
	h	0.021064	0.015159	0.011108	0.011055	0.010288
	mrms	0.661179	0.734359	0.733884	0.764685	0.763978
0.1	ols	1.169018	1.166978	1.169208	1.167135	1.168809
	h	1.170668	1.167122	1.169280	1.167258	1.1688229
	mrms	1.025203	1.050809	1.065086	1.065370	1.065714
0.2	ols	1.179500	1.178410	1.176515	1.175425	1.177344
	h	1.179191	1.178339	1.176184	1.175594	1.177205
	mrms	1.135147	1.141504	1.136983	1.137731	1.136331
0.3	ols	1.179170	1.176424	1.180804	1.179312	1.178269
	h	1.178871	1.176360	1.180225	1.179175	1.178147
	mrms	1.172789	1.170705	1.169520	1.169504	1.169137
0.4	ols	1.179737	1.179049	1.177334	1.17855	1.178793
	h	1.180013	1.179295	1.177149	1.178616	1.178790
	mrms	1.183010	1.191613	1.190554	1.189385	1.189167
0.5	ols	1.176694	1.179213	1.177333	1.177620	1.177223
	h	1.176447	1.178738	1.177159	1.177643	1.177215

If no outlier is present, though OLS is better, but MRMS competes reasonably. When the observations are contaminated, expectedly OLS breaks down while MRMS is able to absorb the shock due to outliers (especially when the fraction of contamination is small), though its asymptotic bias tends to fluctuate with increasing fraction of contamination but it tends to converge to the true value of the parameter.

To examine the general performances of the estimators irrespective of their performances with respect to each of the parameters, the Euclidean norm of the variance is also used. The variance of OLS is smaller than that of the MRMS when no outlier is present but when the observations become contaminated with outliers that of the MRMS performs better; though when fraction of contamination increases above 40%, the variance of OLS tends to be better. As expected of any good estimator, the variance of MRMS decreases consistently as sample size is increased. Thus MRMS is asymptotically consistent.
Using RMSE as evaluation criterion, it is discovered that both mrms and huber estimator performs poorly unlike the OLS. This can be attributed to better variance of OLS in comparison with the other two estimators.

6. Conclusion

A generalization of the Repeated Median of Slope (RMS) is carried out to accommodate more than one independent variable in the regression models. This is obtained through the investigation of the behavior of total change in the dependent variable with respect to an independent variable. A Monte Carlo experiment is conducted to investigate the asymptotic behavior of the estimator obtained; Multiple Repeated Median of Slope (MRMS); when certain percentage of the observations comes from contamination-outlier generating unknown

Table 3. Comparison of variance of MRMS, OLS and Huber estimator

φ	Estrs	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
	mrms	1.913983	1.325794	0.937852	0.937852	0.838022
0.0	ols	1.746626	1.221282	0.831740	0.831740	0.749025
	h	1.888921	1.305542	0.925093	0.925093	0.821508
	mrms	0.931970	0.641697	0.450296	0.450296	0.381520
0.1	ols	1.667803	1.150668	0.809852	0.809852	0.722136
	h	1.761182	1.287431	0.879185	0.879185	0.814008
	mrms	0.395660	0.261082	0.156952	0.156952	0.130952
0.2	ols	1.628530	1.143849	0.780252	0.780252	0.725138
	h	1.665506	1.248003	0.829442	0.829442	0.804302
	mrms	0.184248	0.119758	0.088998	0.088998	0.083475
0.3	ols	1.617309	1.108583	0.782661	0.782661	0.703805
	h	1.689279	1.200506	0.857375	0.857375	0.746224
	mrms	0.141642	0.101591	0.071799	0.071799	0.066578
0.4	ols	1.605376	1.096815	0.782052	0.782052	0.709216
	h	1.681168	1.215627	0.829810	0.829810	0.767835
	mrms	0.127374	0.091921	0.064898	0.064899	0.060856
0.5	ols	1.665452	1.133068	0.793207	0.793207	0.712386
	h	1.780848	1.186381	0.901578	0.901578	0.782579

Table 4. Comparison of RMSE of MRMS, OLS and Huber estimator

φ	Estrs	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
	mrms	2.736653	1.316286	0.668114	0.677499	0.541361
0.0	ols	2.167027	1.089508	0.495614	0.506909	0.418272
	h	2.617947	1.283965	0.635742	0.646337	0.522890
	mrms	2.335384	2.150211	2.436754	2.082757	2.066172
0.1	ols	2.082743	1.110979	0.575138	0.653584	0.554261
	h	2.333979	1.360319	0.675270	0.743597	0.677992
	mrms	2.309015	2.287612	2.633326	2.272637	2.274221
0.2	ols	2.006458	1.084645	0.533304	0.612771	0.569744
	h	2.125101	1.292760	0.589972	0.675903	0.646274
	mrms	2.369599	2.359320	2.721901	2.358832	2.353661
0.3	ols	1.946561	1.014107	0.530951	0.608540	0.511680
	h	2.158916	1.179065	0.633632	0.692967	0.595661
	mrms	2.413164	2.402825	2.738591	2.396628	2.396298
0.4	ols	1.935293	0.979870	0.522162	0.591837	0.509176
	h	2.123809	1.199404	0.578757	0.646806	0.603765
	mrms	2.415916	2.429854	2.763109	2.420653	2.418941
0.5	ols	2.061542	1.004630	0.511887	0.585889	0.510950
	h	2.397901	1.141492	0.676883	0.728189	0.581821

distribution. To allow for the presence of outlier in the observation the gross error model is used in simulating outlier infested dependent variables. The performance of MRMS is evaluated and compared with that of the ordinary least square (OLS) and huber estimator using bias, variance and RMSE. The bias, variance and RMSE of the MRMS decrease with

increase in sample size. As expected of median based estimators, MRMS is more consistent and efficient than the OLS and huber when the observations are contaminated with outlier, though OLS out-performs it when no outlier is present. Using RMSE, the estimator is outperformed by the other two estimators.

Acknowledgments The author wishes to address his special thanks to the presenter of the paper, for the valuable suggestions, comments and recommendations on the manuscript, which have been used to improve on the version.

References

Bernholt, T. and Fried, R.(2003). *Computing the update of repeated median regression line in linear time.* Information Processing Letters, 88(3):111–117. DOI:-10.1016/S0020-0190(03)00350-8.

Bernholt, T., Fried, R., Gather, U. and Wegener, I.(2006). *Modified repeated median filters.* Statistics and Computing, 177–192. DOI:-10.1007/s11222-006-8449-1.

Borowski, M. and Fried, R.(2011). Robust repeated median regression in moving windows with data-adaptive width selection *In: Discussion paper SFB*, 823 German Research Foundation.

Brown, G.W. and Mood, A.M.(1951). On median test for linear hypotheses. *In: Proceedings of second Berkeley Symposium on Mathematical Statistics and Probability*, University of Californian Press, Berkeley, 159–166.

Fried, R. Einbeck, J. and Gather, U.(2003). *Weighted repeated median smoothing and filtering.* Journal of American Statistical Association, 102: 1300–1308. DOI:-10.1198/016214507000001166.

Huber, P. (1981). *Robust Statistics*, Wiley, New York. DOI:-10.1002/0471725250.

Kmenta, J.(1971). *Elements of Econometrics*, MacMillian, New York.

Maronna, R., Martin, R. and Yohai, V.(2006). *Robust Statistics: Theory and Methods*, Sussex, Wiley.

Rousseeuw, P. and Wagner, J.(1994). Robust regression with a distributed intercept using least median of squares. *Computational Statistics and Data Analysis*, 17(1): 65–75. DOI:-10.1016/0167-9473(92)00063-W

Sen, P.(1968). Estimates of regression coefficient based on Kendall's tau. *Journal of American Statistical Association*, 63: 1379–1389. DOI:- 10.1080/01621459.1968.10480934.

Siegel, J.(1982). Robust regression using repeated medians. *Biometrika*, 69(1): 242–244. DOI:- 10.1093/biomet/69.1.242.

Stigler, S.(1986). *History of Statistics: The Measurement of Uncertainty before 1900*, Harvard University Press, Havard.

Theil, H.(1950). A rank-invariant method of linear and polynomial regression analysis, I, II and III. *In: Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen* 53:386–392; 521-525; 1397–1412.

Tukey, J. (1960). Survey of Sampling from Contaminated Distribution. *Contributions to Probability and Statistics I. Olkin, Ed.*, Stanford University Press, Stanford.

Yohai, V.J. (1987). High Breakdown-point and High Efficiency Robust Estimate for Regression. *The Annals of Statistic*, 15(20):642–656. DOI:- 10.1214/aos/1176350366