A SIMULATION STUDY OF THE ROBUSTNESS OF THE LEAST MEDIAN OF SQUARES ESTIMATOR OF SLOPE IN A REGRESSION THROUGH THE ORIGIN MODEL

by

THILANKA DILRUWANI PARANAGAMA

B.Sc., University of Colombo, Sri Lanka, 2005

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2010

Approved by:

Major Professor
Dr. Paul Nelson

Abstract

The principle of least squares applied to regression models estimates parameters by minimizing the mean of squared residuals. Least squares estimators are optimal under normality but can perform poorly in the presence of outliers. This well known lack of robustness motivated the development of alternatives, such as least median of squares estimators obtained by minimizing the median of squared residuals. This report uses simulation to examine and compare the robustness of least median of squares estimators and least squares estimators of the slope of a regression line through the origin in terms of bias and mean squared error in a variety of conditions containing outliers created by using mixtures of normal and heavy tailed distributions. It is found that least median of squares estimation is almost as good as least squares estimation under normality and can be much better in the presence of outliers.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First I would like to thank my major professor Dr. Paul Nelson, for his enormous support, encouragement and guidance throughout this project. I was fortunate to learn from his vast experience and expertise. I would also like to thank the members of my committee: Dr. John Boyer and Dr. Gary Gadbury, for their valuable inputs and feedback. I would like to express my appreciation to Dr. James Neill and the faculty of the Department of Statistics for their guidance and support and to Pam and Angie for all the help they had given me. Thank you to my colleagues and friends in the Statistics department who have provided assistance in many ways.

Finally, I thank my parents, my brother and my sister for their love and support and especially my husband, Dilan for his endless love, affection and encouragement which made this achievement possible.

Dedication

To my parents…

Chapter 1 - Introduction

This report is a simulation study of the performance of the least median of squares estimator of the slope of a regression line through the origin. Least median of squares estimation was initially proposed as hopefully being more robust with respect to outliers than the traditional least squares method of estimation. This study compares least median of squares and least squares estimation of the slope, in terms of mean squared error and bias in a variety of conditions containing outliers created by using mixtures of normal and heavy tailed distributions to model the error terms. Before explaining how least median of squares estimation is implemented, the main reason it was developed is briefly reviewed.

Theories of statistical inference are based on assumptions such as independence, normality and constant variance. However, in reality, departures from these assumptions occur, motivating the development of inference procedures that only depend weakly on assumptions. The concept of robustness signifies relative insensitivity to deviations from assumptions. Specifically, a robust estimator performs reasonably well even when the assumptions under which it is derived do not hold.

Classical linear regression models assume that the error terms are independent, normally distributed with mean zero and constant variance. The presence of outliers, a few observations vary far from the others, can be interpreted as evidence that the assumptions of normality and constant variance are invalid or that an erroneous measurement has been recorded. Least Squares estimation of regression parameters, attributed to Gauss, although widely used, can be very sensitive to the effects of outliers and inaccurate in their presence. Alternative approaches in this setting, such as the method of *least absolute deviation* developed by Edgeworth (1887) and the *M estimator* developed by Huber (1973), which is a generalization of maximum likelihood, are generally more robust with respect to outliers than least squares. More relevant to this work, considering the fact that the median is less sensitive to outliers than the mean and based on the idea of Hampel (1975), Rousseeuw (1984) introduced *Least Median of Squares* estimation, which estimates regression parameters by minimizing the *median* of the squared residuals, as described below.

The median's greater resistance to the effects of outliers than the mean can be expressed in terms of the concept of *breakdown*. The breakdown point of an estimator is the smallest percentage of contaminated data that can cause the estimator to take an arbitrarily large aberrant value, as stated in Hodges (1967). Based on this, the mean has a breakdown point of 0% since even one extreme observation has the ability to change the value of the mean drastically. On the other hand, the median, which has a breakdown point of 50% is considered to be a more robust estimator of the center of a distribution than the mean. Any breakdown point cannot exceed 50% since it is meaningless to consider more than 50% of the data as being contaminated.

1.1 Regression Through the Origin

A regression through the origin model is used when there is an explanatory variable X and the response, Y, and/or its mean is strongly believed to be zero when X is zero. This assumption is not location invariant and should therefore be used with care. The standard model for regression through the origin is as follows. Given $X_i = x_i, i = 1,.....,n,$

$$Y_i = \beta x_i + \varepsilon_i, \qquad (1.1)$$

where

Y_i - Response variable,

β - Slope parameter,

x_i - Explanatory variable (known constant),

$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$.

This model may be employed, in situations where output Y is zero when input X is zero. For example, engine speed is zero when fuel level is zero. Regression through the origin makes the origin a special point. Since in some of these cases the regression function may not change smoothly as X approaches zero, (1.1) should be used with caution if the data set contains values of X both 'close' and 'far' from zero. Note that, a standard regression model could be converted to a regression though origin by subtracting a known intercept from each response.

1.2 Least Squares (LS) Estimators

Given data $\{(x_i, y_i),\ i = 1, 2, ..., n\}$, the least squares estimate $\hat{\beta}_{LS}$ of the slope parameter in (1.1) is obtained by minimizing the mean of the squared residuals with respect to β . Specifically, $\hat{\beta}_{LS}$ minimizes $Q(b) = \sum_{i=1}^{n} (y_i - bx_i)^2 / n$ so that $Q(b) \geq Q(\hat{\beta}_{LS})$ for all possible slopes *b,* and is given by

$$\hat{\beta}_{LS} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 , \quad \text{where } \sum_{i=1}^{n} x_i^2 > 0 .$$

As described above, least squares estimation for linear models is well known to be non-robust with respect to outliers. Specifically, a small percentage of extreme observations in a data set can lead to a value of $\hat{\beta}_{LS}$ that is very different from what would be obtained if these observations were deleted, since the method of least squares tends to pull the fitted line towards these extreme values.

1.3 Least Median of Squares (LMS) Estimators

Since the median is more resistant to outliers than the mean as a measure of central tendency, in order to reduce the effect of outliers on estimators of regression parameters, Rousseeuw (1984) proposed an approach called Least Median of Squares Estimation for the slope of a linear regression model with intercept

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i , \tag{1.2}$$

which minimizes the *median* instead of the mean of the squared residuals. That is, the LMS estimators are obtained by minimizing the median of the squared residuals.

$$(\hat{\beta}_{0-LMS}, \hat{\beta}_{1-LMS}) = \arg\min\{(b_0, b_1); median\{(y_i - b_0 - b_1 x_i)^2, i = 1, 2, ...n\}\} \qquad (1.3)$$

Barreto and Maharry (2006) adapted Rousseeuw's estimator and developed an algorithm using an Excel macro written in Visual Basic for computing a least median of squares estimator of the slope of a regression line through the origin as defined in equation (1.1), given by,

$$\hat{\beta}_{LMS} = \arg\min\{b; median\{(y_i - bx_i)^2, i = 1, 2, ...n\}\}\,. \qquad (1.4)$$

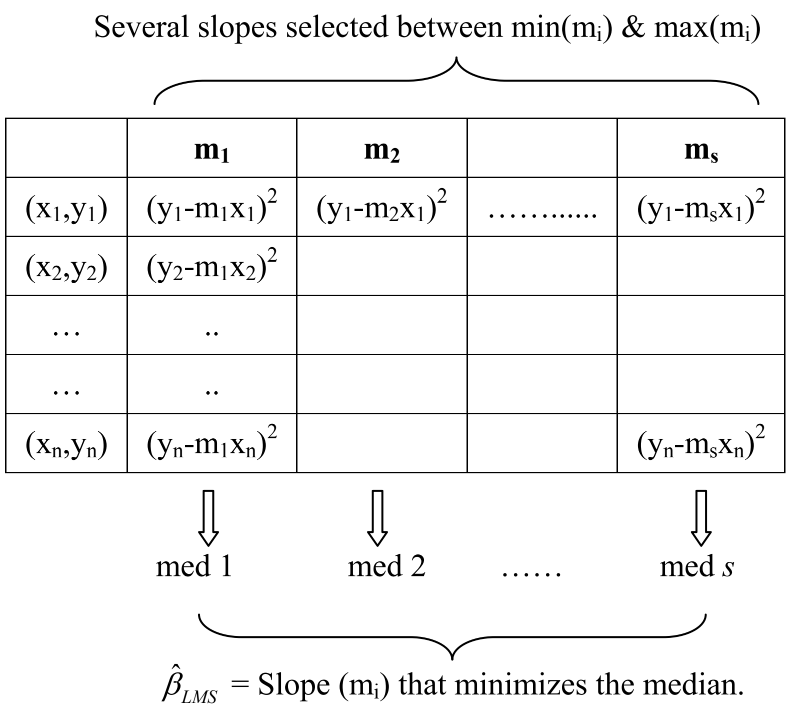
A closed form expression for least median of squares estimators is not available.

1.4 Approximate, Simplified Algorithm for Computing LMS

Based on Barreto and Maharry (2006), I developed an approximate, simplified algorithm for computing the LMS estimator of the slope in (1.1), carried out as follows:

- Calculate the slope m$_i$ = y_i / x_i for each data point (x$_i$,y$_i$) with $x_i > 0$.
- Find the maximum and minimum of { m$_i$ }
- Create several slopes between min(m$_i$) & max(m$_i$).
- Calculate the squared residuals for each data point, using all of the created slopes.
- Calculate the median of squared residuals for each slope.
- Find the slope that minimizes the median of the squared residuals.

Figure 1.1 Illustration of Finding the LMS



Several slopes selected between min(m_i) & max(m_i)

	$\mathbf{m_1}$	$\mathbf{m_2}$		$\mathbf{m_s}$
(x_1,y_1)	$(y_1-m_1x_1)^2$	$(y_1-m_2x_1)^2$	……….....	$(y_1-m_sx_1)^2$
(x_2,y_2)	$(y_2-m_1x_2)^2$			
…	..			
…	..			
(x_n,y_n)	$(y_n-m_1x_n)^2$			$(y_n-m_sx_n)^2$

med 1 med 2 …… med *s*

$\hat{\beta}_{LMS}$ = Slope (m_i) that minimizes the median.

Since there are many limitations in using an Excel macro, especially in running simulations, I wrote a program in R, given in the Appendix, to execute the above algorithm and carried out some comparisons of my program and the Excel macro given by Barreto and Maharry (2006). In making this comparison several data sets were simulated and the LMS estimates were calculated using the R (www.r-project.org) program and the Excel macro. The following results in Table 1.1 show that the estimates are very similar.

Table 1.1 Comparison of Results From Excel and the Written R Program

True Slope	LMS estimate from Excel	LMS estimate from R
2	1.912	1.997
5	5.484	5.484
-3	-2.966	-2.966
3	3.003	2.916
8	8.031	7.996
10	10.075	10.075
1	0.998	0.998
14	13.994	13.994

However, the limitation of my R program was that it took several hours to execute. Since the running time of a program is very critical in a simulation study, as an alternative to my R program, the use of the function '*lqs'* in the MASS package in R was considered. This function turned out to be more efficient than my program with respect to time and memory usage, while giving similar results as the original Excel macro. Therefore, the function '*lqs'* was used in the simulation study presented in Chapter 3.

After presenting the findings of the study in Chapter 3, an example is presented at the end to illustrate the performance of LS and LMS line for a real data set.

Chapter 2 - Simulation Outline

In carrying out the simulation study of the performance of $\hat{\beta}_{LMS}$ in comparison to $\hat{\beta}_{LS}$, data were generated using the statistical package R. To reduce the number of parameters that need to be considered, note that for $\beta \neq 0$, upon dividing (1.1) by β, it can be expressed as

$$Y = x + \varepsilon . \tag{2.1}$$

Hence, without loss of generality, I only investigated models with β = 1 or 0. Values of the independent variable 'x' were generated from a Uniform(0,1) distribution, independent of , the error terms, which were drawn from mixture densities of the form,

$$g(\varepsilon) = p\phi(\varepsilon) + (1-p)h(\varepsilon), \tag{2.2}$$

where $\phi(\cdot)$ is a standard normal density and $h(\cdot)$ is either normal, logistic or Cauchy, as described below. The mixing proportion p was taken close to one so that $h(\cdot)$ may be viewed as creating outliers at a rate of $100(1\text{-}p)\%$. Independent from *x*, an observation from $g(\varepsilon)$ could be obtained by first generating a Bernoulli random variable W~B(1, p). If W = 1, ε is sampled from $\phi(\cdot)$. Otherwise, ε is sampled from $h(\cdot)$. However, since the focus here is on studying the effect of outliers, the number of outliers was forced to be the next integer above n*(1-p).

2.1 Mixture Distributions Used for the Error Term

(i) Standard Normal + Normal

$$g(\varepsilon) = p\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} + (1-p)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{2.3}$$

where,

μ – Location parameter (mean),
σ – Scale parameter (standard deviation),
p – Proportion close to 1.

(ii) *Standard Normal + Cauchy*

$$g(\varepsilon) = p\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} + (1-p)\frac{1}{\pi\gamma\left[1+\left(\frac{x-\mu}{\gamma}\right)^2\right]} \tag{2.4}$$

where,

μ – Location parameter,
γ – Scale parameter,
p – Proportion close to 1.

(iii) *Standard Normal + Logistic*

$$g(\varepsilon) = p\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} + (1-p)\frac{e^{\frac{-(x-\mu)}{\theta}}}{s\left[1+e^{\frac{-(x-\mu)}{\theta}}\right]^2} \tag{2.5}$$

where,

μ – Location parameter,
θ – Scale parameter,
p – Proportion close to 1.

In running the simulations, 1000 data sets were simulated under each different scenario created by setting the above mentioned parameters *p*, σ, μ and the sample size n at various representative values given below. The values for the scale parameters of the Cauchy (γ) and Logistic (θ) distributions were selected so that they would have the same inter-quartile range as the corresponding normal distribution, as given in Table 2.1 below.

2.2 Relationship Between Inter Quartile Range (IQR) and Scale Parameter

Table 2.1 IQR's of Distributions

Distribution	IQR
Normal	1.349 σ
Cauchy	2 γ
Logistic	2.197 θ

In order to have the same inter quartile range as the Normal distribution, the scale parameters for Cauchy and Logistic distributions should be as follows.

$$\textit{Cauchy distribution}: \quad \gamma = \frac{1.349\sigma}{2} \tag{2.6}$$

$$\textit{Logistic distribution}: \quad \theta = \frac{1.349\sigma}{2.197} \tag{2.7}$$

2.3 Generating Data for the Simulation

The representative values chosen for the parameters are given in the following table.

Table 2.2 Values Chosen for the Parameters

Parameter	Values
n	15 , 20 , 40
p	0.9 , 0.95 , 1
μ	-20 , -15 , -10 , 10 , 15
σ	0.5 , 1 , 4

n – Sample size
p – Proportion from N(0,1)
μ – Location parameter
σ – Standard deviation of the Normal

For each combination included in Table 2.2, 1000 independent data sets were simulated from the model given in (2.2). Then, the $\hat{\beta}_{LMS}$, Least Median of Squares estimate for the slope and the $\hat{\beta}_{LS}$, Least Squares estimate for the slope were stored and calculated for each data set. And it should be noted that when the parameter ‘p’ is equal to one, all the error terms will be generated from the standard normal distribution and there will be no outliers in the data set for those cases.

2.4 Measures of Accuracy

The mean squared error of an estimator $\hat{\beta}$, denoted $MSE(\hat{\beta})$, measures how close on average in squared distance, the estimated slope is from the true slope. The bias of an estimator is the difference between an estimator's expectation and the true value of the parameter being estimated. The root mean squared error and the bias of an estimator $\hat{\beta}$ were estimated from N independent simulated values $\{\hat{\beta}_i\}$ as follows.

$$SQRT(M\hat{S}E) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta)^2}\ , \tag{2.8}$$

$$Bi\hat{a}s = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i - \beta\ . \tag{2.9}$$

Using the data from the simulations, a regression analysis was carried out considering the root mean squared errors as the responses and the other parameters sample size, number of outliers, scale parameter and the location as the explanatory variables, in order to study the effect of these variables on the accuracy of the estimates. Secondly, another regression analysis was carried out considering the bias as the response and using the same explanatory variables as in the regression for the root MSE.

Chapter 3 - Simulation Results

Based on 1000 simulations for each combination of parameters and for the 3 mixed distributions, the results are summarized below in tables and plots. The estimates of β and the root mean squared errors are tabulated by sample size, number of outliers, scale and location parameters. Note that since the slope β is set equal to one in this table, root mean squared errors are actually relative root mean squared errors. As a somewhat arbitrary but useful benchmark, I will judge an estimated root mean squared error of at least one to be unsatisfactory. As shown in equation (2.2) although 100(1-p)% was set as a target for the *proportion of outliers* generated in the simulation, in the tables below, that proportion is shown as the actual *number of outliers* which was actually obtained by multiplying the proportion (1-*p*) by the sample size *n* and the result rounded up to the next integer. My simulation results are presented separately for each combination of distributions.

3.1 Standard Normal + Normal Distribution

The means of the simulated least square estimates and the least median of squares estimates of the slope and their root mean squared errors are given in the Table 3.1, where it can be seen that root mean square errors of the least squares estimator increase as: the absolute values of the location parameters increase; the number of outliers increases; and the sample size decreases. As expected, since the true slope is positive, negative location parameters have a more harmful effect than corresponding positive ones. In no case is the least squares estimator 'satisfactory' according to my benchmark. However, the least median of squares estimator is satisfactory in all cases, with a MSE decreasing with increasing sample size and is relatively stable across all other parameters. Overall, the means of the estimates of the least median of squares estimates are clearly closer to β and more stable than the means of the least squares estimator in all cases. The scale parameter appears to have very little effect on both estimators. In particular when there are one or more outliers in the data, the LMS estimates seem to provide fairly accurate estimates of the slope with small root mean squared errors, while the LS estimates perform poorly with larger root mean squared errors. This is an indication that, in this case, the LMS estimators are more robust with respect to outliers in regressions through the origin.

Table 3.1 below contains the LS estimates and LMS estimates along with their root MSE's for different sample sizes, number of outliers, scale parameters and location parameters.

Table 3.1 Root MSE's of LS and LMS Estimates for Standard Normal + Normal (β=1)

Standard Normal + Normal			Location																			
			-20				-15				-10				10				15			
			LS		LMS		LS		LMS		LS		LMS		LS		LMS		LS		LMS	
n	outliers	σ	est	√mse	est	√ mse	est	√mse	est	√ mse	est	√mse	est	√ mse	est	√mse	est	√ mse	est	√mse	est	√ mse
15	1	0.5	-1.08	2.44	0.99	0.96	-0.54	1.84	1.00	0.89	0.01	1.24	1.04	0.95	2.02	1.26	1.05	0.91	2.54	1.85	1.00	0.92
		1	-1.02	2.39	1.02	0.91	-0.62	1.93	1.01	0.96	-0.01	1.27	1.00	0.90	2.03	1.28	0.99	0.90	2.54	1.87	1.00	0.92
		4	-1.00	2.42	1.01	0.92	-0.53	1.91	1.03	0.91	-0.03	1.35	1.05	0.90	2.05	1.42	0.98	0.88	2.57	1.91	0.97	0.92
	2	0.5	-3.05	4.39	1.01	0.88	-2.07	3.36	0.99	0.91	-1.02	2.23	1.01	0.89	3.08	2.29	1.06	0.95	4.00	3.29	1.00	0.95
		1	-2.97	4.32	1.04	0.90	-2.06	3.35	1.00	0.90	-1.07	2.28	0.96	0.91	3.03	2.25	1.01	0.89	4.05	3.34	1.03	0.97
		4	-3.18	4.59	1.01	0.92	-2.03	3.39	1.02	0.91	-1.07	2.35	0.97	0.89	3.05	2.35	1.02	0.90	4.11	3.47	1.01	0.92
20	1	0.5	-0.51	1.81	1.00	0.82	-0.11	1.37	0.98	0.84	0.28	0.92	1.03	0.82	1.74	0.94	1.04	0.81	2.14	1.39	0.99	0.86
		1	-0.53	1.81	1.00	0.79	-0.12	1.36	1.00	0.85	0.21	0.98	0.98	0.83	1.76	0.96	0.96	0.86	2.15	1.38	0.99	0.88
		4	-0.56	1.91	1.01	0.85	-0.20	1.46	1.02	0.87	0.22	1.04	0.97	0.80	1.76	1.02	1.01	0.86	2.14	1.40	0.96	0.82
	2	0.5	-2.06	3.34	1.02	0.82	-1.27	2.49	1.00	0.84	-0.52	1.69	0.98	0.83	2.53	1.70	1.04	0.87	3.25	2.46	1.00	0.83
		1	-2.01	3.29	1.02	0.86	-1.34	2.55	0.99	0.84	-0.53	1.70	0.99	0.83	2.52	1.68	1.03	0.83	3.29	2.52	1.04	0.85
		4	-2.04	3.36	0.97	0.83	-1.32	2.59	0.97	0.84	-0.49	1.74	1.00	0.85	2.51	1.76	0.98	0.84	3.29	2.55	1.01	0.81
40	2	0.5	-0.53	1.67	1.05	0.62	-0.16	1.27	1.01	0.64	0.23	0.87	0.99	0.66	1.77	0.87	0.99	0.62	2.14	1.27	1.00	0.64
		1	-0.54	1.69	1.03	0.65	-0.12	1.23	1.04	0.64	0.25	0.86	1.00	0.68	1.76	0.88	1.04	0.64	2.14	1.27	0.99	0.63
		4	-0.49	1.65	1.04	0.62	-0.14	1.29	0.97	0.63	0.24	0.91	0.97	0.64	1.77	0.91	1.00	0.64	2.13	1.27	1.02	0.62
	4	0.5	-2.00	3.12	1.01	0.64	-1.28	2.38	0.98	0.65	-0.50	1.58	1.00	0.64	2.54	1.62	1.01	0.61	3.31	2.41	1.02	0.63
		1	-2.05	3.18	0.99	0.63	-1.27	2.38	1.01	0.63	-0.52	1.60	1.01	0.64	2.52	1.60	1.00	0.64	3.27	2.37	0.98	0.64
		4	-2.02	3.16	0.99	0.63	-1.25	2.38	0.97	0.61	-0.54	1.66	0.98	0.61	2.51	1.64	0.98	0.64	3.29	2.43	0.96	0.63

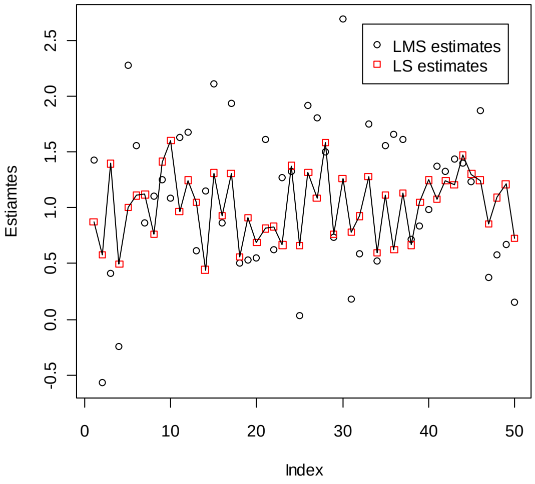
Table 3.2 summarizes what happens when there are no outliers. That is, in situations where all the error terms are been drawn from a standard normal distribution.

Table 3.2 LS and LMS Estimates with No Outliers

Standard Normal + Normal				
	LS		**LMS**	
n	**est**	**√*mse***	**est**	**√ *mse***
15	1.00	0.47	0.99	0.93
20	1.00	0.39	0.99	0.84
40	1.00	0.28	1.00	0.64

Here we see that when there are no outliers, both the LS estimate and the LMS estimates are satisfactory. However, unlike Table 3.1 above, here the root mean squared errors of the LMS estimates are somewhat larger than that of the LS estimates, which are optimal in this case. To further explore this observation, LS and LMS estimates of slope are plotted below in Figure 3.1 for 50 randomly generated data sets of sample size 25 with no outliers.

Figure 3.1 Variation in LS and LMS Estimates in the Presence of No Outliers

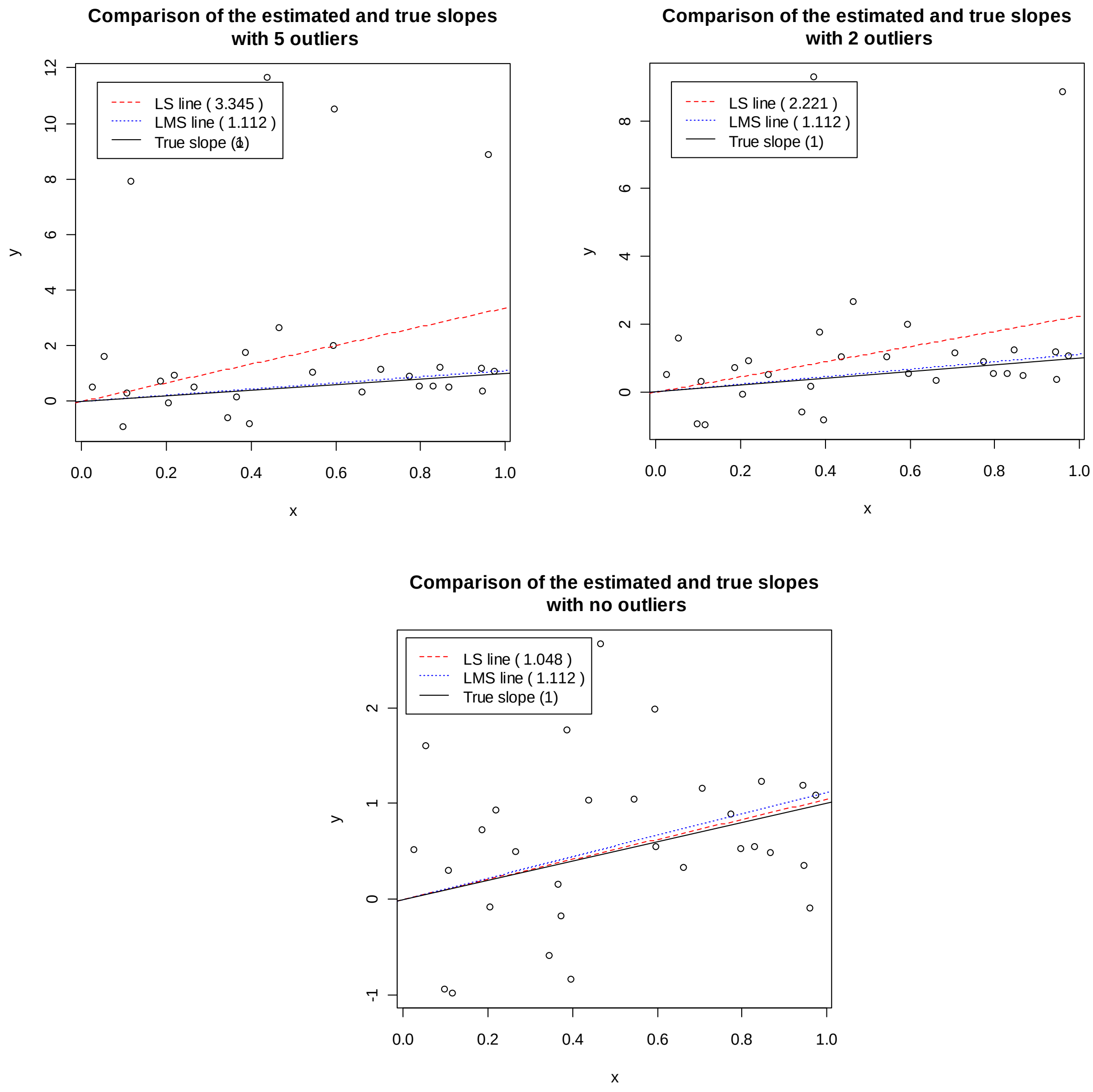


- Mean of the LMS estimates = 1.103 with MSE = 0.450
- Mean of the LS estimates = 1.020 with MSE = 0.088

Although in Figure 3.1 both the estimates are pretty close to the true slope 1, it is evident from the MSE’s and the line drawn in the plot, that the LS estimates for the slope perform marginally better when compared to LMS estimates in situations where there are no outliers.

To further compare and illustrate the performance of the two estimators, scatter plots of two data sets, each having 30 observations with five and two outliers along with the true, least squares and least median of squares lines are presented in Figure 3.2.

Figure 3.2 Comparison of Estimated and True Slopes for Simulated Data (n = 30)



Comparison of the estimated and true slopes with 5 outliers

Comparison of the estimated and true slopes with 2 outliers

Comparison of the estimated and true slopes with no outliers

Figure 3.2 shows that the LS line deviates away from the true line and leans toward the outliers. On the other hand, the outliers have very little effect on the LMS line.

Having evaluated the LS estimates and LMS estimates with respect to MSE's, the estimated bias of those estimates are presented in Table 3.3 below, computed using equation (2.9). Similar to what was seen in the analysis of MSE's, the bias of the LS estimates increased with increasing number of outliers and the shift of the location parameter away from zero. In cases where the mean estimate was negative, the bias was further inflated. However, the bias of the LMS estimates outperforms the bias of the LS estimates, being small and stable throughout the table ranging from -0.04 to 0.06. Since the conclusions drawn from the bias of the estimators were not different from the conclusions drawn from the MSE's, the bias results are not presented for the other mixture distributions.

As mentioned in chapter 2, although the main interest is in analyzing the cases when the true slope equals one, simulations were also carried out for $\beta = 0$. The results are presented in Table 3.4. Due to the similarity of the two cases, $\beta = 0$ and $\beta = 1$, zero slope results are not presented for the other mixed distributions.

Table 3.3 Bias of LS and LMS Estimates for Standard Normal + Normal (β=1)

Standard Normal +			Location																			
Normal			-20				-15				-10				10				15			
	β=1		LS		LMS		LS		LMS		LS		LMS		LS		LMS		LS		LMS	
n	**outliers**	**σ**	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***	**est**	***Bias***
15	1	0.5	-1.08	*-2.08*	0.99	*-0.01*	-0.54	*-1.54*	1.00	*0.00*	0.01	*-0.99*	1.04	*0.04*	2.02	*1.02*	1.05	*0.05*	2.54	*1.54*	1.00	*0.00*
		1	-1.02	*-2.02*	1.02	*0.02*	-0.62	*-1.62*	1.01	*0.01*	-0.01	*-1.01*	1.00	*0.00*	2.03	*1.03*	0.99	*-0.01*	2.54	*1.54*	1.00	*0.00*
		4	-1.00	*-2.00*	1.01	*0.01*	-0.53	*-1.53*	1.03	*0.03*	-0.03	*-1.03*	1.05	*0.05*	2.05	*1.05*	0.98	*-0.02*	2.57	*1.57*	0.97	*-0.03*
	2	0.5	-3.05	*-4.05*	1.01	*0.01*	-2.07	*-3.07*	0.99	*-0.01*	-1.02	*-2.02*	1.01	*0.01*	3.08	*2.08*	1.06	*0.06*	4.00	*3.00*	1.00	*0.00*
		1	-2.97	*-3.97*	1.04	*0.04*	-2.06	*-3.06*	1.00	*0.00*	-1.07	*-2.07*	0.96	*-0.04*	3.03	*2.03*	1.01	*0.01*	4.05	*3.05*	1.03	*0.03*
		4	-3.18	*-4.18*	1.01	*0.01*	-2.03	*-3.03*	1.02	*0.02*	-1.07	*-2.07*	0.97	*-0.03*	3.05	*2.05*	1.02	*0.02*	4.11	*3.11*	1.01	*0.01*
20	1	0.5	-0.51	*-1.51*	1.00	*0.00*	-0.11	*-1.11*	0.98	*-0.02*	0.28	*-0.72*	1.03	*0.03*	1.74	*0.74*	1.04	*0.04*	2.14	*1.14*	0.99	*-0.01*
		1	-0.53	*-1.53*	1.00	*0.00*	-0.12	*-1.12*	1.00	*0.00*	0.21	*-0.79*	0.98	*-0.02*	1.76	*0.76*	0.96	*-0.04*	2.15	*1.15*	0.99	*-0.01*
		4	-0.56	*-1.56*	1.01	*0.01*	-0.20	*-1.20*	1.02	*0.02*	0.22	*-0.78*	0.97	*-0.03*	1.76	*0.76*	1.01	*0.01*	2.14	*1.14*	0.96	*-0.04*
	2	0.5	-2.06	*-3.06*	1.02	*0.02*	-1.27	*-2.27*	1.00	*0.00*	-0.52	*-1.52*	0.98	*-0.02*	2.53	*1.53*	1.04	*0.04*	3.25	*2.25*	1.00	*0.00*
		1	-2.01	*-3.01*	1.02	*0.02*	-1.34	*-2.34*	0.99	*-0.01*	-0.53	*-1.53*	0.99	*-0.01*	2.52	*1.52*	1.03	*0.03*	3.29	*2.29*	1.04	*0.04*
		4	-2.04	*-3.04*	0.97	*-0.03*	-1.32	*-2.32*	0.97	*-0.03*	-0.49	*-1.49*	1.00	*0.00*	2.51	*1.51*	0.98	*-0.02*	3.29	*2.29*	1.01	*0.01*
40	2	0.5	-0.53	*-1.53*	1.05	*0.05*	-0.16	*-1.16*	1.01	*0.01*	0.23	*-0.77*	0.99	*-0.01*	1.77	*0.77*	0.99	*-0.01*	2.14	*1.14*	1.00	*0.00*
		1	-0.54	*-1.54*	1.03	*0.03*	-0.12	*-1.12*	1.04	*0.04*	0.25	*-0.75*	1.00	*0.00*	1.76	*0.76*	1.04	*0.04*	2.14	*1.14*	0.99	*-0.01*
		4	-0.49	*-1.49*	1.04	*0.04*	-0.14	*-1.14*	0.97	*-0.03*	0.24	*-0.76*	0.97	*-0.03*	1.77	*0.77*	1.00	*0.00*	2.13	*1.13*	1.02	*0.02*
	4	0.5	-2.00	*-3.00*	1.01	*0.01*	-1.28	*-2.28*	0.98	*-0.02*	-0.50	*-1.50*	1.00	*0.00*	2.54	*1.54*	1.01	*0.01*	3.31	*2.31*	1.02	*0.02*
		1	-2.05	*-3.05*	0.99	*-0.01*	-1.27	*-2.27*	1.01	*0.01*	-0.52	*-1.52*	1.01	*0.01*	2.52	*1.52*	1.00	*0.00*	3.27	*2.27*	0.98	*-0.02*
		4	-2.02	*-3.02*	0.99	*-0.01*	-1.25	*-2.25*	0.97	*-0.03*	-0.54	*-1.54*	0.98	*-0.02*	2.51	*1.51*	0.98	*-0.02*	3.29	*2.29*	0.96	*-0.04*

Table 3.4 Root MSE's of LS and LMS Estimates for Standard Normal + Normal (β=0)

Standard Normal +			Location																			
Normal			-20				-15				-10				10				15			
	β=0		LS		LMS		LS		LMS		LS		LMS		LS		LMS		LS		LMS	
n	outliers	σ	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse
15	1	0.5	-2.13	*2.51*	-0.10	*0.93*	-1.55	*1.85*	-0.04	*0.91*	-1.07	*1.29*	-0.04	*0.90*	1.02	*1.29*	0.01	*0.92*	1.60	*1.90*	0.01	*0.89*
		1	-2.08	*2.44*	-0.02	*0.95*	-1.50	*1.81*	0.00	*0.91*	-1.00	*1.25*	0.00	*0.92*	1.00	*1.27*	0.00	*0.89*	1.59	*1.92*	0.00	*0.90*
		4	-2.08	*2.48*	0.01	*0.90*	-1.52	*1.90*	0.04	*0.89*	-1.00	*1.36*	0.01	*0.93*	0.99	*1.34*	-0.02	*0.91*	1.52	*1.89*	-0.01	*0.88*
	2	0.5	-4.08	*4.41*	-0.03	*0.90*	-3.08	*3.34*	0.02	*0.90*	-2.05	*2.26*	0.03	*0.91*	2.08	*2.28*	-0.06	*0.95*	3.08	*3.34*	-0.01	*0.89*
		1	-4.10	*4.45*	-0.03	*0.86*	-3.09	*3.38*	0.05	*0.89*	-2.04	*2.25*	-0.04	*0.90*	2.04	*2.25*	0.02	*0.93*	2.99	*3.29*	-0.01	*0.88*
		4	-4.04	*4.46*	0.02	*0.96*	-3.10	*3.44*	0.02	*0.89*	-2.04	*2.33*	0.01	*0.88*	2.08	*2.37*	-0.03	*0.92*	3.08	*3.42*	-0.02	*0.89*
20	1	0.5	-1.48	*1.78*	-0.02	*0.84*	-1.19	*1.42*	0.05	*0.81*	-0.74	*0.94*	0.05	*0.85*	0.76	*0.97*	0.01	*0.81*	1.17	*1.42*	0.00	*0.82*
		1	-1.56	*1.85*	0.00	*0.86*	-1.15	*1.39*	-0.02	*0.84*	-0.74	*0.94*	0.00	*0.83*	0.77	*0.98*	0.00	*0.81*	1.14	*1.39*	0.05	*0.83*
		4	-1.48	*1.80*	0.02	*0.88*	-1.19	*1.46*	0.00	*0.83*	-0.79	*1.04*	-0.02	*0.85*	0.77	*1.03*	-0.05	*0.87*	1.19	*1.48*	0.00	*0.84*
	2	0.5	-2.99	*3.27*	-0.03	*0.83*	-2.30	*2.50*	0.01	*0.84*	-1.54	*1.70*	0.01	*0.84*	1.55	*1.71*	0.01	*0.82*	2.29	*2.51*	-0.04	*0.85*
		1	-3.03	*3.30*	-0.05	*0.83*	-2.31	*2.52*	0.04	*0.78*	-1.49	*1.67*	0.03	*0.87*	1.50	*1.67*	0.00	*0.82*	2.28	*2.50*	-0.04	*0.86*
		4	-3.00	*3.31*	0.02	*0.83*	-2.33	*2.60*	-0.03	*0.85*	-1.55	*1.79*	0.01	*0.83*	1.55	*1.78*	-0.02	*0.83*	2.33	*2.58*	0.04	*0.84*
40	2	0.5	-1.49	*1.64*	0.00	*0.65*	-1.12	*1.24*	0.01	*0.64*	-0.76	*0.86*	0.01	*0.65*	0.75	*0.85*	0.00	*0.66*	1.17	*1.28*	0.02	*0.66*
		1	-1.50	*1.65*	-0.02	*0.64*	-1.14	*1.26*	0.00	*0.63*	-0.77	*0.87*	0.01	*0.64*	0.74	*0.85*	-0.03	*0.61*	1.14	*1.26*	0.02	*0.64*
		4	-1.49	*1.65*	0.02	*0.63*	-1.12	*1.27*	-0.01	*0.61*	-0.77	*0.90*	-0.01	*0.63*	0.76	*0.89*	0.03	*0.63*	1.15	*1.30*	-0.03	*0.63*
	4	0.5	-3.01	*3.16*	0.04	*0.64*	-2.26	*2.37*	0.02	*0.62*	-1.48	*1.57*	0.01	*0.62*	1.52	*1.59*	0.00	*0.66*	2.29	*2.41*	-0.03	*0.62*
		1	-3.04	*3.18*	-0.01	*0.63*	-2.25	*2.36*	-0.01	*0.63*	-1.53	*1.62*	0.02	*0.62*	1.51	*1.60*	0.00	*0.63*	2.26	*2.37*	0.00	*0.62*
		4	-3.05	*3.22*	-0.03	*0.64*	-2.27	*2.40*	-0.01	*0.65*	-1.52	*1.63*	-0.03	*0.62*	1.52	*1.64*	0.01	*0.63*	2.28	*2.41*	-0.04	*0.64*

3.1.1 Regression Analysis

In order to investigate the effect of sample size, number of outliers etc. on the accuracy of the estimates in terms of root mean squared errors, a regression analysis was carried out with response y = Root mean squared error and independent variables x1 = sample size, x2 = number of outliers, x3 = scale parameter and x4 = location parameter.

- LMS Estimates

Figure 3.3 Regression Output for Root Mean Squared Error of LMS Estimates

```
 Coefficients:

                Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.0753233  0.0057342 187.530   <2e-16 ***

n           -0.0107811  0.0002034 -52.995   <2e-16 ***

outliers    -0.0030738  0.0017618  -1.745   0.0834 .

scale       -0.0010174  0.0013201  -0.771   0.4423

Loc          0.0002061  0.0001465   1.407   0.1619

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.02371 on 130 degrees of freedom

Multiple R-squared: 0.9626,     Adjusted R-squared: 0.9614

F-statistic: 835.7 on 4 and 130 DF,  p-value: < 2.2e-16
```

The fitted regression equation is given by

$$\sqrt{MSE} = 1.075 - 0.011(n) - 0.003(outliers) - 0.001(scale) + 0.0002(loc) . \qquad (3.1)$$

Overall, the R-Squared value of 0.96 indicates that the regression model given in equation 3.1 provides an adequate fit to the simulated root MSE’s of the LMS estimates. From the output above, all other predictors remaining fixed, we estimate that root mean square error decreases: by 0.0101 per unit increase in sample size and by 0.0031 per unit increase in the number of outliers. Both effects are small and the effect of the number of outliers is only marginally statistically significant, with a p-value of 0.0834. These conclusions are consistent with what was observed in Figures 3.1 and 3.2.

The same analysis was carried out using the root MSE’s of the least squares estimates and is presented below.

- LS Estimates

Figure 3.4 Regression Output for Root Mean Squared Error of LS Estimates

```
Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.645739   0.140099  11.747  < 2e-16 ***

n           -0.048456   0.004970  -9.749  < 2e-16 ***

outliers     0.719564   0.043045  16.716  < 2e-16 ***

scale        0.012606   0.032254   0.391  0.69656

Loc         -0.011034   0.003579  -3.083  0.00250 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.5792 on 130 degrees of freedom

Multiple R-squared: 0.7002,     Adjusted R-squared: 0.691

F-statistic: 75.92 on 4 and 130 DF,  p-value: < 2.2e-16
```

The fitted regression equation is given by

$$\sqrt{MSE} = 1.646 - 0.048(n) + 0.720(outliers) + 0.013(scale) - 0.011(loc)\,. \qquad (3.2)$$

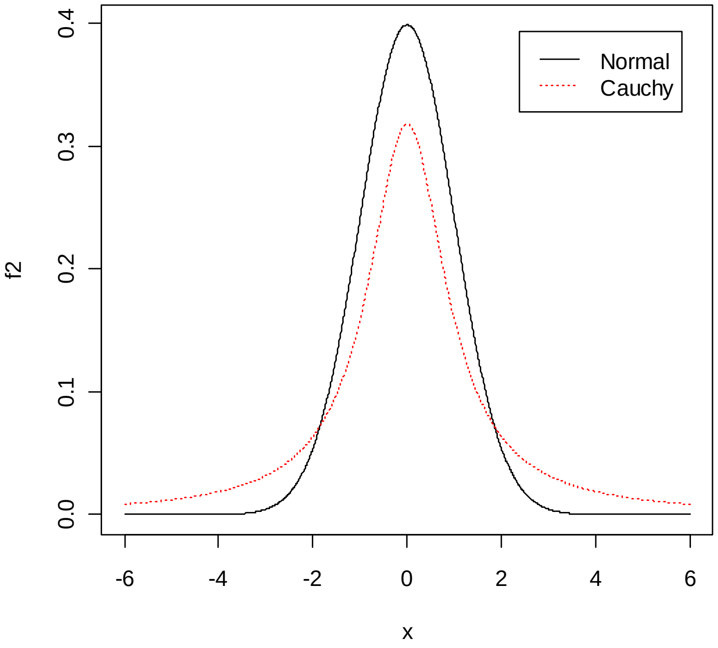
When analyzing the root MSE's of the LS estimates, the R-Squared value of 0.7 is an indication of the adequacy of the fitted regression model to the simulated data. By further examining the predictors in the model it is evident that the variables sample size, number of outliers and location have a significant effect on the root MSE's of the LS estimates. Holding all the other variables constant, we estimate that the root MSE decreases: by 0.0485 per unit increase in sample size and by 0.011 per unit increase in the location parameter. And on the other hand, the root MSE increases by 0.7196 per unit increase in the number of outliers, which is a significant reduction in the accuracy of the estimate.

As shown in the previous Figure 3.3 the model (3.1) that was fitted for the root MSE's of LMS estimates yielded small coefficients for the predictors and hence did not affect the accuracy by large amounts. However by examining the model (3.2) for the LS estimates, it is evident that some of the parameters, especially the number of outliers has a significant impact on the accuracy of the LS estimates, which is consistent with what was seen in Figure 3.2.

3.2 Standard Normal + Cauchy Distribution

The second mixture distribution analyzed in this report is the ‘Standard normal + Cauchy’. Even though the normal and Cauchy densities, pictured below in Figure 3.5, are both mound shaped and symmetric about the origin, the Cauchy has much heavier tails than the normal and does not have a mean.

Figure 3.5 Standard Normal and Standard Cauchy Distributions



The parameter settings used in this section are given in Table 2.2. Recall that the Cauchy scale parameter γ is chosen so that the Cauchy distribution has the same inter-quartile range as the normal distribution. Simulation results for this mixture model are given in Table 3.5.

Table 3.5 Root MSE's of LS and LMS Estimates for Standard Normal + Cauchy

Standard Normal +			Location																			
Cauchy			-20				-15				-10				10				15			
	β=1		LS		LMS		LS		LMS		LS		LMS		LS		LMS		LS		LMS	
n	outliers	σ	est	√mse	est	√ mse	est	√mse	est	√ mse	est	√mse	est	√ mse	est	√mse	est	√ mse	est	√mse	est	√ mse
15	1	0.5	-1.07	2.76	1.00	0.95	-0.64	5.43	0.97	0.93	-0.21	5.88	0.99	0.92	1.66	11.57	1.01	0.91	2.56	2.21	0.97	0.95
		1	-0.99	2.91	1.03	0.89	0.19	20.91	0.99	0.90	0.16	10.93	0.99	0.96	2.55	14.41	1.03	0.91	2.72	5.25	1.03	0.92
		4	-1.31	10.11	1.01	0.85	-0.48	4.24	1.01	0.90	0.53	9.50	1.01	0.92	2.11	5.87	1.04	0.90	2.70	5.39	1.01	0.86
	2	0.5	-3.13	7.84	1.01	0.97	-2.09	3.52	0.99	0.86	-1.14	3.91	1.04	0.93	3.00	3.34	1.00	0.88	2.40	47.55	0.94	0.94
		1	-3.13	5.63	0.95	0.92	-1.85	8.63	0.96	0.91	-1.23	4.17	0.93	0.87	2.73	10.80	0.96	1.00	5.08	23.56	0.95	0.90
		4	-2.22	22.19	0.99	0.89	-1.55	34.60	1.01	0.91	1.37	61.94	0.95	0.92	2.95	22.09	1.03	0.94	4.54	13.71	0.97	0.92
20	1	0.5	-0.50	1.81	1.06	0.87	-0.05	3.02	0.99	0.83	0.29	1.34	1.04	0.80	1.71	1.94	1.07	0.82	2.12	1.47	0.97	0.87
		1	-0.43	2.44	1.00	0.82	-0.10	1.68	1.02	0.84	0.27	2.48	0.98	0.85	1.73	1.23	0.95	0.84	2.07	2.21	1.04	0.85
		4	-0.69	5.62	0.97	0.84	-0.38	8.54	0.97	0.82	4.05	139.04	1.05	0.86	1.84	9.08	0.98	0.82	1.53	9.56	1.03	0.84
	2	0.5	-1.87	4.09	1.01	0.84	-1.32	2.60	0.99	0.85	-0.53	1.93	1.03	0.83	2.52	1.93	1.03	0.82	3.18	3.51	0.98	0.84
		1	-2.06	4.42	1.01	0.82	-1.05	5.92	1.01	0.85	-9.91	299.62	1.04	0.84	2.62	4.58	1.01	0.81	0.37	86.41	1.03	0.84
		4	-2.53	13.61	0.97	0.85	-1.09	11.05	1.01	0.86	-3.95	108.58	1.04	0.81	0.93	41.47	1.02	0.83	2.80	18.83	1.03	0.83
40	2	0.5	-0.49	2.45	1.01	0.64	-0.16	1.37	1.02	0.63	0.26	0.89	0.99	0.63	1.74	0.94	0.96	0.64	2.14	1.29	1.00	0.66
		1	-0.49	2.10	1.00	0.66	-0.11	4.33	0.96	0.63	0.26	1.16	1.00	0.61	1.78	2.71	1.02	0.63	2.14	1.66	0.98	0.65
		4	-0.79	13.50	0.99	0.64	0.04	4.59	0.97	0.65	0.15	4.35	0.99	0.64	2.04	5.79	1.03	0.61	1.85	8.79	0.97	0.63
	4	0.5	-2.06	3.47	0.97	0.64	-1.28	2.50	1.05	0.63	-0.47	2.66	1.01	0.62	2.48	1.74	1.04	0.64	3.11	4.66	1.03	0.64
		1	-2.32	11.99	1.01	0.62	-1.26	6.68	1.03	0.63	-0.62	2.45	0.96	0.64	2.47	2.47	0.98	0.63	3.33	3.54	0.99	0.65
		4	-2.13	14.53	1.00	0.65	-1.27	11.98	1.01	0.63	-0.83	13.98	1.01	0.65	3.18	17.26	0.99	0.62	4.19	66.86	0.99	0.59

Table 3.6 LS and LMS Estimates with No Outliers

	LS		LMS	
n	est	√mse	est	√ mse
15	1.00	0.46	0.99	0.92
20	1.00	0.39	1.00	0.85
40	1.00	0.28	1.00	0.64

As shown in Table 3.6 when there are no outliers both the LS estimate and the LMS estimate provide estimates very close to the true estimate of the slope which is 1. However, the LS estimates have smaller root MSE s when compared to that of the LMS estimates.

In the presence of outliers as shown in Table 3.5 the LS estimates tend to be far from the true slope and have relatively large root MSEs, especially in certain cases. For instance, when the sample size is 40 with 4 outliers and the location parameter has a value of 15, LS estimates of the slope average from 3.11 to 4.19 with root MSEs going from 4.66 to 66.86. In this setting, the LMS estimates are close to the true slope with root MSE's ranging from 0.64 to 0.59, a much better performance. Overall, in Table 3.5 we see that mean LMS estimates fall within the range from 0.92 to 1.06 with small root MSEs that are below 1. Further, the root MSE's of the *LMS* estimates decrease as the sample size increases. Compared to the *LS* estimates, the *LMS* estimates seem to perform significantly better in almost all the situations.

3.1.1 Regression Analysis

As a further analysis a regression analysis was carried out to understand the relationship between the parameters used and the root MSE s. Similar to the analysis that was done for the 'Standard Normal+ Normal' mixed distribution, a regression was fitted for LS estimates and LMS estimates separately.

- LMS Estimates

Figure 3.6 Regression Output for Root Mean Squared Error of LMS Estimates

```
 Coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0766356  0.0062325 172.744   <2e-16 ***
n           -0.0107569  0.0002211 -48.648   <2e-16 ***
outliers    -0.0030447  0.0019149  -1.590    0.114
scale       -0.0023571  0.0014349  -1.643    0.103
Loc          0.0002569  0.0001592   1.614    0.109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.02577 on 130 degrees of freedom
Multiple R-squared: 0.9559,     Adjusted R-squared: 0.9546
F-statistic: 704.9 on 4 and 130 DF,  p-value: < 2.2e-16
```

The fitted regression equation is given by

$$\sqrt{MSE} = 1.077 - 0.011(n) - 0.003(outliers) - 0.002(scale) + 0.0002(loc). \qquad (3.3)$$

As shown in the above regression output, the fitted regression line provides a good fit with a R-squared value of 0.95. By looking at the p-values for each explanatory variable, it can be observed that the variable, sample size, is the only significant factor in predicting the response variable root MSE of LMS estimates, adjusting for other variables. All other predictors remaining fixed, it is estimated that the root MSE decreases by 0.0108 per unit increase in sample size.

In order to analyze the behavior of root MSE of LS estimates another regression analysis was carried out using the same predictor variables.

- LS Estimates

Figure 3.7 Regression Output for Root Mean Squared Error of LS Estimates

```
Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.17631    7.39622   1.511  0.13319

n           -0.56252    0.26240  -2.144  0.03392 *

outliers     6.30553    2.27248   2.775  0.00634 **

scale        2.79698    1.70279   1.643  0.10289

Loc          0.01311    0.18895   0.069  0.94480

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 30.58 on 130 degrees of freedom

Multiple R-squared: 0.08348,    Adjusted R-squared: 0.05528

F-statistic:  2.96 on 4 and 130 DF,  p-value: 0.02224
```

The fitted regression equation is given by

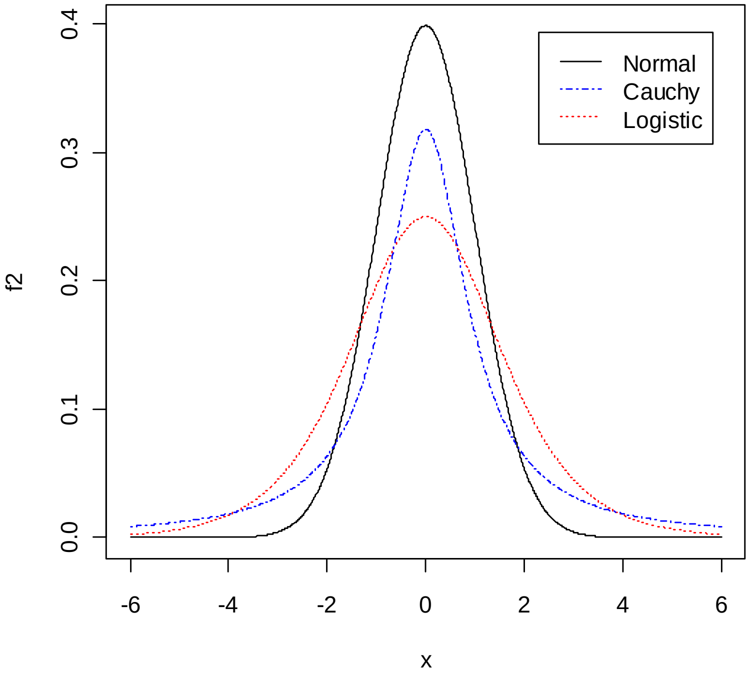
$$\sqrt{MSE} = 11.176 - 0.563(n) + 6.306(outliers) + 2.797(scale) + 0.013(loc)\,. \qquad (3.4)$$

The R-Squared value of the above regression for the LS estimates is 0.06 which indicates an inadequate fit. Hence, a residual analysis was carried out and it could be seen that there were few data points that were extreme outliers. Those identified data points were removed from the data set and another regression analysis was carried out. However, since this change did not result in any significant improvement in the goodness of the fit with respect to the R-squared value, it was decided to continue with the original analysis. To increase the goodness of the fit, higher order terms could be added to the model, but that would make the regression model more complex and harder to interpret. Therefore, based on the regression model given in (3.4), with all other predictors fixed, it is estimated that the root MSE decreases by 0.5625 per unit increase in sample size and root MSE increases by 6.3055 per unit increase in number of outliers. This indicates that the accuracy of the estimate of the slope diminishes drastically with the existence of outliers.

3.3 Standard Normal + Logistic Distribution

In this section of the report 'Standard normal + Logistic' mixed distribution is analyzed. The following Figure 3.8 illustrates the shapes of the standard Normal, standard Cauchy and standard Logistic densities.

Figure 3.8 Standard Normal, Standard Cauchy and Standard Logistic Distributions



The same parameter values given in the Table 2.2 under the simulation outline chapter are used in this mixed distribution. Note that the logistic scale parameter θ was chosen so that the logistic distribution has the same inter-quartile range as the normal distribution that was considered in the first mixed distribution. Simulation results for this mixture model are given in Table 3.7.

Table 3.7 Root MSE's of LS and LMS Estimates for Standard Normal + Logistic

Standard Normal + Logistic			Location																			
			-20				-15				-10				10				15			
	β=1		LS		LMS		LS		LMS		LS		LMS		LS		LMS		LS		LMS	
n	outlier	σ	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse	est	√mse
15	1	0.5	-1.02	*2.40*	0.99	*0.91*	-0.59	*1.89*	0.95	*0.91*	-0.01	*1.25*	1.02	*0.89*	2.01	*1.26*	0.97	*0.87*	2.57	*1.89*	1.03	*0.92*
		1	-1.04	*2.43*	1.00	*0.91*	-0.59	*1.89*	0.96	*0.88*	0.01	*1.25*	1.02	*0.96*	2.00	*1.26*	0.97	*0.92*	2.59	*1.90*	1.05	*0.91*
		4	-1.11	*2.57*	0.95	*0.90*	-0.51	*1.87*	0.95	*0.85*	-0.05	*1.39*	0.93	*0.95*	1.97	*1.31*	0.99	*0.92*	2.55	*1.97*	0.99	*0.90*
	2	0.5	-3.18	*4.52*	0.98	*0.88*	-2.04	*3.31*	0.97	*0.85*	-1.07	*2.27*	0.97	*0.88*	3.05	*2.26*	1.02	*0.90*	4.01	*3.29*	1.04	*0.91*
		1	-3.04	*4.39*	0.99	*0.89*	-2.04	*3.31*	1.03	*0.89*	-1.04	*2.27*	1.00	*0.94*	3.07	*2.28*	1.00	*0.88*	4.10	*3.38*	0.98	*0.92*
		4	-3.17	*4.59*	1.01	*0.89*	-2.08	*3.49*	1.01	*0.92*	-1.06	*2.39*	0.99	*0.90*	3.03	*2.36*	1.00	*0.94*	4.05	*3.37*	1.03	*0.88*
20	1	0.5	-0.51	*1.80*	1.05	*0.85*	-0.16	*1.41*	1.03	*0.85*	0.23	*0.97*	1.03	*0.85*	1.76	*0.96*	0.98	*0.83*	2.16	*1.41*	1.01	*0.83*
		1	-0.55	*1.85*	0.98	*0.86*	-0.13	*1.38*	0.99	*0.86*	0.25	*0.96*	1.03	*0.82*	1.78	*0.99*	1.04	*0.87*	2.15	*1.40*	0.99	*0.88*
		4	-0.51	*1.85*	1.05	*0.82*	-0.16	*1.46*	0.97	*0.85*	0.21	*1.05*	0.94	*0.87*	1.75	*1.02*	0.99	*0.82*	2.14	*1.46*	1.00	*0.82*
	2	0.5	-1.94	*3.21*	1.02	*0.83*	-1.27	*2.48*	1.02	*0.87*	-0.49	*1.66*	1.00	*0.80*	2.48	*1.64*	0.97	*0.84*	3.30	*2.50*	0.99	*0.84*
		1	-2.04	*3.31*	1.00	*0.82*	-1.24	*2.46*	0.99	*0.84*	-0.54	*1.71*	1.01	*0.83*	2.55	*1.71*	1.02	*0.77*	3.25	*2.48*	1.02	*0.83*
		4	-2.04	*3.35*	1.04	*0.86*	-1.23	*2.51*	1.02	*0.83*	-0.48	*1.74*	0.99	*0.80*	2.51	*1.74*	0.97	*0.82*	3.27	*2.56*	0.97	*0.83*
40	2	0.5	-0.54	*1.68*	0.98	*0.65*	-0.09	*1.21*	1.03	*0.64*	0.25	*0.85*	1.00	*0.62*	1.75	*0.85*	0.97	*0.65*	2.15	*1.26*	1.01	*0.64*
		1	-0.47	*1.62*	1.01	*0.62*	-0.14	*1.26*	0.98	*0.63*	0.26	*0.84*	0.97	*0.64*	1.75	*0.86*	1.02	*0.65*	2.13	*1.25*	0.97	*0.63*
		4	-0.50	*1.67*	1.02	*0.67*	-0.17	*1.32*	1.01	*0.66*	0.23	*0.91*	0.99	*0.64*	1.77	*0.91*	1.03	*0.65*	2.15	*1.30*	1.01	*0.64*
	4	0.5	-2.06	*3.19*	1.02	*0.63*	-1.26	*2.37*	1.02	*0.62*	-0.52	*1.60*	0.99	*0.65*	2.52	*1.60*	1.00	*0.61*	3.28	*2.38*	1.04	*0.65*
		1	-1.99	*3.13*	0.99	*0.64*	-1.26	*2.37*	1.03	*0.61*	-0.52	*1.61*	0.99	*0.61*	2.51	*1.60*	0.99	*0.63*	3.25	*2.36*	1.01	*0.63*
		4	-2.00	*3.15*	0.99	*0.64*	-1.31	*2.44*	1.00	*0.64*	-0.52	*1.66*	1.01	*0.65*	2.50	*1.63*	0.95	*0.61*	3.27	*2.41*	1.00	*0.61*

Table 3.8 LS and LMS Estimates with No Outliers

	LS		LMS	
n	est	√*mse*	est	√ *mse*
15	1.00	0.46	1.00	0.93
20	0.99	0.40	0.97	0.84
40	1.00	0.28	1.01	0.64

By looking at the results shown in Table 3.7 when the outliers are generated from a Logistic distribution, the LS estimates of the slope parameter in the regression through the origin seem to be quite different from the true slope which is 1. All throughout the table the average LS estimates ranges from -3.18 to 4.10 with root MSEs ranging from 0.85 to 4.59. For a given sample size, scale and location parameter value, the LS estimate tends to move further away from 1 as the number of outliers increases. Moreover, as the location parameter moves away from zero either to the positive side or the negative side, the LS estimate seems to move away from the true slope and have larger MSEs.

On the other hand, the LMS estimates presented in the same Table 3.7 perform significantly better than the LS estimates with outliers. The LMS estimates ranges only from 0.95 to 1.05 all throughout the table with MSEs ranging from 0.61 to 0.96 which is an indication of a well behaved estimator.

As discussed in the previous sections Table 3.8 shows that when there are no outliers both the LS estimate and the LMS estimate provide estimates very close to the true estimate of the slope which is 1. However, the LS estimates have smaller root MSE s when compared to that of the LMS estimates.

3.3.1 Regression Analysis

Following are the results of the regression analysis carried out for the LMS estimates and the LS estimates considering the root MSE as the response.

- LMS Estimates

Figure 3.9 Regression Output for Root Mean Squared Error of LMS Estimates

```
Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.0595325  0.0057331 184.810  < 2e-16 ***

n           -0.0104364  0.0002034 -51.310  < 2e-16 ***

outliers    -0.0049802  0.0017615  -2.827  0.00544 **

scale        0.0015700  0.0013199   1.189  0.23642

Loc         -0.0000881  0.0001465  -0.602  0.54854

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.0237 on 130 degrees of freedom

Multiple R-squared: 0.9608,     Adjusted R-squared: 0.9596

F-statistic: 797.3 on 4 and 130 DF,  p-value: < 2.2e-16
```

The fitted regression equation is given by

$$\sqrt{MSE} = 1.060 - 0.010(n) - 0.005(outliers) + 0.002(scale) - 0.0001(loc). \qquad (3.5)$$

Similar to what was learned in the other two mixed distribution, it can be seen here that the above regression model which was fitted to the LMS estimates, provides a very good fit with a R-squared value of 0.96. Furthermore, it is evident that in predicting the root MSE of the estimate, the variables sample size and the number of outliers are significant, adjusting for other variables in the model. It is estimated that remaining all other variables fixed, the root MSE decreases by 0.0104 per unit increase in sample size and by 0.0050 per unit increase in number of outliers.

- LS Estimates

Figure 3.10 Regression Output for Root Mean Squared Error of LS Estimates

```
Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.652826   0.140839  11.736  < 2e-16 ***

n           -0.048743   0.004997  -9.755  < 2e-16 ***

outliers     0.718541   0.043273  16.605  < 2e-16 ***

scale        0.013677   0.032425   0.422  0.67386

Loc         -0.011128   0.003598  -3.093  0.00243 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.5823 on 130 degrees of freedom

Multiple R-squared: 0.698,      Adjusted R-squared: 0.6887

F-statistic: 75.11 on 4 and 130 DF,  p-value: < 2.2e-16
```

The fitted regression equation is given by

$$\sqrt{MSE} = 1.653 - 0.049(n) + 0.718(outliers) + 0.014(scale) - 0.011(loc)\,. \tag{3.6}$$

With an overall R-squared value of 0.70, this regression model fits the data fairly adequately in predicting root MSE of the LS estimates with the variables; sample size, number of outliers and location parameter being significant. By looking at the estimates of the coefficients, it can be seen that adjusting for other variables, the root MSE increases by 0.7185 per unit increase in the number of outliers. In comparison to the estimate 0.005 seen in the LMS output, this is a clear indication that the number of outliers has a noteworthy effect on the LS estimate.

3.4 An Illustrative Example

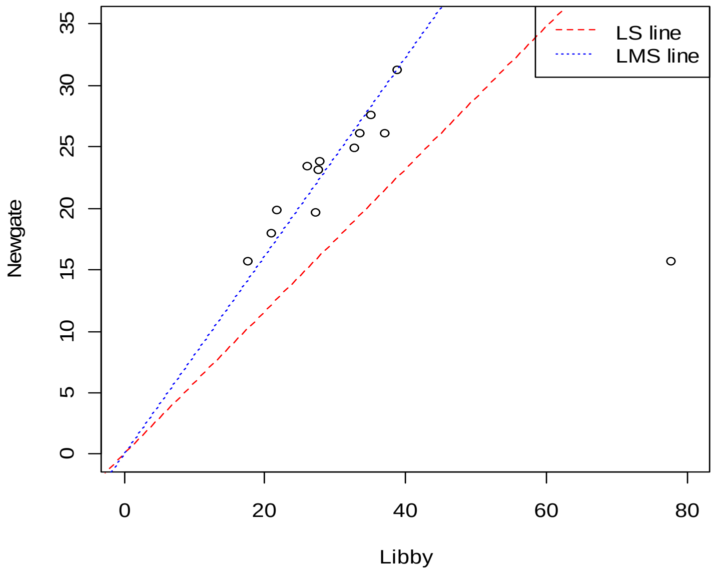
This example is adapted from 'Robust regression and outlier detection' By Peter J. Rousseeuw, Annick M. Leroy. The data consists of measurements of water flow (in 100 Cubic Feet per Second) at two different points (Libby, Montana and Newgate, British Columbia) on the Kootenay river in January for the years 1931-1943. The original data came from Ezekiel and Fox (1959, pp. 57-58) and Hampel et al.(1986) changed the Newgate measurement for the year 1934 from 44.9 to 15.7 for illustrative purposes.

And for the data presented in Table 3.9, the LS line and the LMS was drawn as shown in Figure 3.11. Validating the findings of the simulations it can be seen that the LMS line outperforms the LS line by providing a good fit to the data.

Table 3.9 Example Data

Year	Libby	Newgate
31	27.1	19.7
32	20.9	18
33	33.4	26.1
34	77.6	15.7
35	37	26.1
36	21.6	19.9
37	17.6	15.7
38	35.1	27.6
39	32.6	24.9
40	26	23.4
41	27.6	23.1
42	38.7	31.3
43	27.8	23.8

Figure 3.11 LS and LMS line for the Example Data



Chapter 4 - Conclusion

The objective of this report was to assess the robustness of the Least Median of Squares estimator of the slope in a regression through the origin, in comparison to the Least Squares estimator in the presence of outliers. The performance of the estimators was evaluated mainly, with respect to their Mean Squared Errors.

In simulating data three different mixed distributions, namely, 'Standard Normal + Normal', 'Standard Normal + Cauchy' and 'Standard Normal + Logistic' were considered. As mentioned above the smaller portions of the mixed distribution were generated from normal, Cauchy or logistic distributions, which created outliers in the data sets. These data were generated with a known slope and the fitted values of this slope using LS and LMS estimators were compared.

Having done numerous simulations creating a variety of outliers, it was discovered that the LMS estimators of the slope were very close to the true slope and were not much affected by the outliers in the data sets, which provides evidence of robustness against outliers. On the other hand, the LS estimators performed rather poorly and deviated away from the true slope, in the presence of outliers in the data set, which pulled the LS line away from most of the data.

However, when there are no outliers both the LS estimator and the LMS estimators gave fairly accurate estimates, with the LS estimator performing better than the LMS estimator with smaller MSE's.

Overall, the LMS estimator can be considered as being more robust with respect to outliers as compared to the LS estimators. This conclusion is mainly based on point estimation. In practice, the use of LMS is limited by the absence of formulas for standard errors. Therefore, as a suggestion for a future study, this issue could possibly be addressed by using bootstrap methods.

References

Å. Björck, Numerical Methods for Least Squares Problems, SIAM, 1996.

Barreto, H. and Maharry, D. (2005) 'Least Median of Squares and Regression Through the Origin', *Computational Statistics and Data Analysis*, **50**, 1391-1397.

Frank R. Hampel 'A General Qualitative Definition of Robustness', Ann. Math. Statist. Volume 42, Number 6 (1971), 1887-1896.

Hodges, J.L., Jr. (1967), 'Efficiency in normal samples and tolerance of extreme values for some estimates of location', Proc. 5th Berkeley Symp. 1

Peter J. Huber, Elvezio Ronchetti. 'Robust statistics' Page 196 2nd ed 2009

Rousseeuw Peter J, Annick M. Leroy, 'Robust regression and outlier detection', John Wiley and Sons 1987, 63-64

Rousseeuw, P.J. (1984) 'Least Median of Squares Regression', *JASA*, **79**, 871-880.

Seigel, A.F (1982) 'Robust regression using repeated medians', *Biomrtrica*, **69**, 242-244.

Appendix – R Code

```
## norm(0,1)+norm(loc,sigma) ##

rm(list=ls())
library(MASS)

m=1
n=c(15,20,40)
p=c(.9,.95,1)
sigma=c(.5,1,4)
loc=c(-20,-15,-10,10,15)

outout=NULL
out=NULL
for (i in n){
 for (j in p){
  for (k in sigma){
   for (l in loc){
   out=NULL
    for (N in 1:1000){
          n1=floor(i*j)
          n2=i-n1
          x=runif(i, min=0, max=1)
          e1=rnorm(n1,mean=0,sd=1)
          e2=rnorm(n2,mean=l,sd=k)
          e=c(e1,e2)
          y=m*x+e
          fit=lm(y~-1+x)
          lms1.est=lqs(x,y,intercept=F,method="lms")
          lms.est=as.numeric(lms1.est$coeff)
          ls.est=as.numeric(fit$coeff)
          out=rbind(out,c(LS=ls.est,LMS=lms.est))
          }

     MSE.ls=mean((out[,1]-m)^2)# mse of LS
     MSE.lms=mean((out[,2]-m)^2)# mse of LMS

outout=rbind(outout,c(n=(n1+n2),outliers=n2,Sigma=k,Loc=l,
apply(out, 2,mean), LSmse=MSE.ls,LMSmse=MSE.lms))
}}}}

outout
write.csv(outout,file="data1.csv")
```

```
## norm(0,1)+cauch(loc,scale) ##

rm(list=ls())
library(MASS)

m=1
n=c(15,20,40)
p=c(.9,.95,1)
sigma=c(.5,1,4)
loc=c(-20,-15,-10,10,15)

outout=NULL
out=NULL
for (i in n){
 for (j in p){
  for (k in sigma){
   for (l in loc){
   out=NULL
    for (N in 1:1000){
          n1=floor(i*j)
          n2=i-n1
          x=runif(i, min=0, max=1)
          e1=rnorm(n1,mean=0,sd=1)
          e2=rcauchy(n2,loc=l,scale=k*1.349/2)
          e=c(e1,e2)
          y=m*x+e
          fit=lm(y~-1+x)
          lms1.est=lqs(x,y,intercept=F,method="lms")
          lms.est=as.numeric(lms1.est$coeff)
          ls.est=as.numeric(fit$coeff)
          out=rbind(out,c(LS=ls.est,LMS=lms.est))
          }

     MSE.ls=mean((out[,1]-m)^2)# mse of LS
     MSE.lms=mean((out[,2]-m)^2)# mse of LMS

outout=rbind(outout,c(n=(n1+n2),outliers=n2,Sigma=k,Loc=l,
          apply(out, 2, mean),LSmse=MSE.ls,LMSmse=MSE.lms))
}}}}
outout
write.csv(outout,file="data1.csv")
```

```
## norm(0,1)+logis(loc,scale) ##

rm(list=ls())
library(MASS)

m=1
n=c(15,20,40)
p=c(.9,.95,1)
sigma=c(.5,1,4)
loc=c(-20,-15,-10,10,15)

outout=NULL
out=NULL
for (i in n){
 for (j in p){
  for (k in sigma){
   for (l in loc){
   out=NULL
    for (N in 1:1000){
          n1=floor(i*j)
          n2=i-n1
          x=runif(i, min=0, max=1)
          e1=rnorm(n1,mean=0,sd=1)
          e2=rlogis(n2,location=l,scale=k*1.349/2.197)
          e=c(e1,e2)
          y=m*x+e
          fit=lm(y~-1+x)
          lms1.est=lqs(x,y,intercept=F,method="lms")
          lms.est=as.numeric(lms1.est$coeff)
          ls.est=as.numeric(fit$coeff)
          out=rbind(out,c(LS=ls.est,LMS=lms.est))
          }

     MSE.ls=mean((out[,1]-m)^2)# mse of LS
     MSE.lms=mean((out[,2]-m)^2)# mse of LMS

outout=rbind(outout,c(n=(n1+n2),outliers=n2,Sigma=k,Loc=l,
          apply(out, 2, mean),LSmse=MSE.ls,LMSmse=MSE.lms))
}}}}
outout
write.csv(outout,file="data1.csv")
```

```
   ## Regression Analysis ##

   rm(list=ls())
data1=read.table("C:\\Thil\\Research\\LMS5\\N+N.txt",header=T)
   attach(data1)

   ## Reg for N+N for LS root mse ##

   reg1.1=lm(sqrt(LSmse)~n+outliers+scale+Loc)
   summary(reg1.1)

   ## Reg for N+N for LMS root mse ##

   reg1.2=lm(sqrt(LMSmse)~n+outliers+scale+Loc)
   summary(reg1.2)

   ####################################################

Data2=read.table("C:\\Thil\\Research\\LMS5\\N+C.txt",header=T)
   attach(data2)

   ## Reg for N+C for LS rootmse ##

   reg2.1=lm(sqrt(LSmse)~n+outliers+scale+Loc)
   summary(reg2.1)

   ## Reg for N+C for LMS root mse ##

   reg2.2=lm(sqrt(LMSmse)~n+outliers+scale+Loc)
   summary(reg2.2)

   ####################################################
data3=read.table("C:\\Thil\\Research\\LMS5\\N+L.txt",header=T)
   attach(data3)

   ## Reg for N+L for LS root mse ##

   reg3.1=lm(sqrt(LSmse)~n+outliers+scale+Loc)
   summary(reg3.1)

   ## Reg for N+L for LMS root mse ##

   reg3.2=lm(sqrt(LMSmse)~n+outliers+scale+Loc)
   summary(reg3.2)
```

Computing LMS without using the MASS package

```
rm(list=ls())
par(mfrow=c(2,2))
########## CREATING OUTLIERS ######
n=100
m=5
pi=.95
norm.mean=0
sigma=2
scale1=1
loc=5

set.seed(185)
x=runif(n, min=0, max=1)
set.seed(170)
e=pi*rnorm(n,mean=norm.mean,sd=sigma)+(1-pi)
        *rcauchy(n,location=loc,scale=scale1)
y=m*x+e
fit=lm(y~-1+x)
plot(x,y)
abline(fit)
hist(e)

#########   Computing LMS   #######

m=y/x
m.max=max(m)
m.min=min(m)
m.max
m.min
slope=seq(m.min,m.max,by=0.01)
mx=x%*%t(slope)
sqdev=(y-mx)^2
meds=apply(X=sqdev,MARGIN=2,FUN=median)
slope[order(meds)][1]
```