Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 14

5-1-2009

A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric

Oya Can Mutan
ODTU, Turkey, oya.canmutan@spk.gov.tr

Birdal Şenoğlu
Ankara University, Turkey, senoglu@science.ankara.edu.tr

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 161-172

1538 – 9472/09/$95.00

A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric

Oya Can Mutan
ODTU, Turkey

Birdal Şenoğlu
Ankara University, Turkey

The performances of the ordinary least squares (OLS), modified maximum likelihood (MML), least absolute deviations (LAD), Winsorized least squares (WIN), trimmed least squares (TLS), Theil's (Theil) and weighted Theil's (Weighted Theil) estimators are compared under the simple linear regression model in terms of their bias and efficiency when the distribution of error terms is long-tailed symmetric.

Key words: Long-tailed symmetric, ordinary least squares, modified maximum likelihood, least absolute deviations, Winsorized least squares, trimmed least squares, Theil's method, Weighted Theil's method.

Introduction

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad (1)$$

where (i = 1, 2, …, n), y_i is the response variable, x_i is a nonstochastic explanatory variable and β_0 and β_1 are the unknown parameters. Traditionally, error terms e_i $(1 \le i \le n)$ are assumed to be independently and identically distributed (iid) normal $N(0, \sigma^2)$ and the regression coefficients β_0 and β_1 are estimated by using the OLS estimators given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x}) y_i \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (2)$$

respectively.

The OLS estimators are optimal only if the error distribution is normal. However, in most real life applications, nonnormal distributions are more prevalent; see, Pearson (1932), Geary (1947), Huber (1981), Şenoğlu (2005) and Şenoğlu (2007). Additionally, the occurrence of outliers in a data set is another indication of nonnormality. Due to these weaknesses of the OLS estimators, statisticians prefer to use the alternative regression estimators which are more efficient and robust under nonnormality

However, the choice of which method to use is not defined clearly for different types of error distributions. In the literature, there exists a very limited number of researches comparing alternative regression methods, see Tam (1996) and Nevitt and Tam (1998). In this study, our main concern is to identify the most efficient method when the error distribution is long-tailed symmetric and also to see the effect of nonnormality on the efficiencies and robustness of the regression estimators.

Oya Can Mutan is a Statistician in the Capital Markets Board of Turkey. She received her B.S. (statistics), M.S. (statistics, economics) and Ph.D. (statistics) degrees from Middle East Technical University in Turkey. Email: oya.canmutan@spk.gov.tr. Birdal Şenoğlu is an Associate Professor in the Department of Statistics at Ankara University, Turkey. He received his B.S. and Ph.D. degrees (statistics) from Middle East Technical University in Turkey and his M.S. (statistics) degree from Iowa State University, USA. Email: senoglu@science.ankara.edu.tr.

Long-tailed Symmetric (LTS) Distribution

The LTS distribution has the probability density function:

$$LTS(p,\sigma):\ f(e) \propto \frac{1}{\sigma}\left\{1+\frac{e^2}{k\sigma^2}\right\}^{-p},$$
$$-\infty < e < \infty;$$

with $k = 2p-3$ and $p \geq 2$. The mean and variance of the random variable e is 0 and σ^2, respectively. See also the following table for the Pearson coefficient of kurtosis, i.e., $\beta_2 = \mu_4 / \mu_2^2$ of the $LTS(p,\sigma)$ distribution:

$p=$	2.5	3.5	5.0	10	∞
$\beta_2=$	∞	9	4.2	3.4	3.0

This reduces to the normal distribution when p is equal to ∞.

Methodology

OLS is the most popular method for estimating the parameters of the simple linear regression model. This is partly due to the relative simplicity of its computations. However, the OLS method is very sensitive to outliers and to nonnormality. To remedy these problems, alternative regression methods have been developed that are not sensitive to the violations of the assumptions of the simple linear regression model. The only disadvantage of these alternative methods is their computational difficulty. Today, however, computational difficulties are unimportant issue because of the improvements in computer technology (see Birkes & Dodge, 1993; Rousseeuw & Leroy, 1987).

The Modified Maximum Likelihood Method

The maximum likelihood (ML) estimators are the solutions of the equations

$$\partial \ln L/\partial\beta_0 = 0,$$
$$\partial \ln L/\partial\beta_1 = 0,$$

and

$$\partial \ln L/\partial\sigma = 0. \qquad (3)$$

These equations do not have explicit solutions. Tiku, et al. (2001) express likelihood equations in terms of order statistics (for a given β_1), since complete sums are invariant to ordering.

$$z_{(i)} = \frac{y_{[i]} - \beta_0 - \beta_1 x_{[i]}}{\sigma},\ (1 \leq i \leq n)$$

where $(y_{[i]}, x_{[i]})$ is that pair of observations which correspond to $z_{(i)}$ $(1 \leq i \leq n)$; $(y_{[i]}, x_{[i]})$ are called the concomitants of $z_{(i)}$. They linearize the intractable functions $g(z_{(i)}) = z_{(i)} \Big/ \left\{1+(1/k)z_{(i)}^2\right\}$ by using the first two terms of a Taylor series expansion by using the following linear approximation

$$g(z_{(i)}) \cong \alpha_i + \beta_i z_{(i)},\ 1 \leq i \leq n \qquad (4)$$

where

$$\alpha_i = \frac{(2/k)t_{(i)}^3}{\left(1+(1/k)t_{(i)}^2\right)^2}$$

and

$$\beta_i = \frac{1-(1/k)t_{(i)}^2}{\left(1+(1/k)t_{(i)}^2\right)^2}$$

$t_{(i)}$'s ($i = 1,2,.....,n$) are the expected values of the order statistics $z_{(i)}$, i. e., $t_{(i)} = E(z_{(i)})$. Incorporating (4) in (3), results in modified likelihood equations:

$$\partial \ln L^* / \partial\beta_0 = 0,$$
$$\partial \ln L^* / \partial\beta_1 = 0,$$

and

$$\partial \ln L^* / \partial\sigma = 0.$$

These equations have explicit solutions called as MML estimators:

$$\hat{\beta}_0 = \bar{y}_{[.]} - \hat{\beta}_1 \bar{x}_{[.]}$$

$$\hat{\beta}_1 = K + D\hat{\sigma}$$

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-2)}}$$

where

$$\bar{y}_{[.]} = \sum_{i=1}^{n} \beta_i y_{[i]} \Big/ m \,,\ \bar{x}_{[.]} = \sum_{i=1}^{n} \beta_i x_{[i]} \Big/ m \,,$$

$$\left(m = \sum_{i=1}^{n} \beta_i \right),$$

$$K = \sum_{i=1}^{n} \beta_i (x_{[i]} - \bar{x}_{[.]}) y_{[i]} \Big/ \sum_{i=1}^{n} \beta_i (x_{[i]} - \bar{x}_{[.]})^2 \,,$$

$$D = \sum_{i=1}^{n} \alpha_i x_{[i]} \Big/ \sum_{i=1}^{n} \beta_i (x_{[i]} - \bar{x}_{[.]})^2 \,,$$

$$B = (2p/k) \sum_{i=1}^{n} \alpha_i \left\{ y_{[i]} - \bar{y}_{[.]} - K\left(x_{[i]} - \bar{x}_{[.]}\right) \right\}$$

and

$$C = (2p/k) \sum_{i=1}^{n} \beta_i \left\{ y_{[i]} - \bar{y}_{[.]} - K\left(x_{[i]} - \bar{x}_{[.]}\right) \right\}^2 .$$

Least Absolute Deviations (LAD)

The LAD regression method was developed by Roger Joseph Boscovich in 1757, see Birkes and Dodge (1993). The LAD estimators of regression coefficients, β_0 and β_1, are found by minimizing the function:

$$F = \sum_{i=1}^{n} \left| y_i - (\beta_0 + \beta_1 x_i) \right|. \qquad (5)$$

Although the logic behind LAD is not more difficult than the concept of OLS, calculation of the LAD estimates is more troublesome. An algorithmic method is used for the calculation of the LAD estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, since there are no exact formulas.

This algorithm starts with one of the data points (x, y), say (x_1, y_1), and tries to find the best line passing through it. The line passing through (x_1, y_1) also passes through another data point denoted by (x_2, y_2). Next we find the best line passing through (x_2, y_2). As the algorithm continues, we obtain increasingly better lines and finally the most recent line obtained will be the same as the previous line. This line is the best line and it is called as LAD regression line, see Birkes and Dodge (1993) for more detailed information.

Winsorized Least Squares

The WLS which is an iterative method is another alternative to OLS method; see Yale and Forsythe (1976). Smoothing techniques based on the OLS estimation are applied to reduce the effect of the outliers in the sample. The basic idea is to replace the most extreme residual with the next closest residual in the sample in an iterative way. In the literature, the studies show that Winsorization does not worsen a good linear relationship on non-contaminated data. On the contrary, it improves the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, when the sample is contaminated with outliers.

Trimmed Least Squares

The fourth method is the TLS introduced by Rousseeuw in 1984. The TLS estimation procedure is similar to the OLS estimation, but in TLS procedure, the fit is not so much affected from the outliers, because the data points corresponding to a specified percentage of the highest residuals based on an initial OLS estimation are removed. The OLS estimates of slope and intercept for the remaining data are called TLS estimates, see Rousseeuw and Leroy (1987) and Nevitt and Tam (1998). The aim is to minimize

$$\sum_{i=1}^{h} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (6)$$

As it is seen in equation (6), rather than smoothing the data as in Winsorized regression, the outlying cases are deleted, therefore the *n-h* observations do not affect the estimators.

Theil's Method

Theil's nonparametric regression method using the median as robust measures (see Theil, 1950) is presented. In Theil's

method, the only assumption is that the error terms are identically and independently distributed (i.i.d); this is different than the robust methods.

Sprent (1993) stated that for a simple linear regression model to obtain the slope of a line that fits the data points, the set of all slopes $b_{ij} = \frac{y_j - y_i}{x_j - x_i}$ of lines joining pairs of data points (x_i, y_i), (x_j, y_j), $x_j \neq x_i$, for $1 \le i < j \le n$ should be calculated.

Hussain and Sprent (1983) say that no generality is lost if $1 \le i < j \le n$ is taken, assuming that the x_i's are arranged in ascending order (note that $b_{ij} = b_{ji}$). According to these results the Theil's slope estimator is:

$$\hat{\beta}_1 = med\left\{b_{ij} \mid x_j \neq x_i\right\}$$

where $x_1 \le x_2 \le ... \le x_n$.

It is known that median estimators are less affected from the outlying values in the data set as compared to the mean estimators, i.e., they are resistant estimators. The corresponding intercept term is defined as the median of the $y_i - \hat{\beta}_1 x_i$ terms (see Birkes & Dodge, 1993).

Weighted Theil's Method

A modified version of the Theil's method is called a Weighted Theil's Regression Method. In this method, different than the Theil's original method, each of the pairwise slopes are weighted using a weighting scheme. The weighted Theil slope estimator for the n observations in the sample data is the weighted median of these b_{ij}'s. w_{ij}, as the weighting procedure, can be taken as

$$x_j - x_i,\ j - i \text{ or } \left|x_j - x_i\right|,$$

see, for example Jaeckel (1972) and Scholz [16] and Birkes and Dodge (1993). In this study, the weights $w_{ij} = \frac{\left(x_i - x_j\right)^2}{\sum\left(x_i - x_j\right)^2}$ were used to calculate the slope estimator $\hat{\beta}_1 = \sum w_{ij} b_{ij}$. The intercept estimator is calculated in a similar fashion as in Theil's original method.

Results

The design points x_i $(1 \le i \le n)$ follow an equally spaced, sequential additive series (x_i = 1, 2, …, n) (see Hussain & Sprent, 1983) and are common to all random samples $(y_1, y_2, ..., y_n)$ for the $N = \left[100{,}000/n\right]$ (integer) Monte Carlo runs. The error terms, e_i, are generated from the long-tailed symmetric distribution given above, and β_0, β_1 and σ are taken to be 0, 1 (1 in the remainder of this article) without loss of generality. The simulated means, variances and mean square errors (MSE) of the estimators are computed for some selected values of p (2.0, 2.5, 3.0, 3.5 and 5.0) and the results are given in Table 1.

From the simulation results presented in Table 1, all of the methods of estimation produced negligible bias therefore comparisons may be made in terms of MSE for both $\hat{\beta}_0$ and $\hat{\beta}_1$. In view of MSE, the following conclusions are put forth for the intercept estimator $\hat{\beta}_0$:

- WIN20 and WIN10 outperformed other estimators at all sample sizes for $p < 3$. For moderate ($n = 20$) and large sample sizes (n=50) they had the smallest MSE when p = 3.0. For values of the shape parameter p greater than 3, WIN20 and WIN10 were the preferred estimators for large sample sizes (n=50).
- The performance of the MML is best for small sample sizes (n=10) when p =3. When p = 3.5 and 5, the highest performance was achieved by MML for small (n=10) and moderate (n=20) samples.
- LAD and TLS performed poorly at all sample sizes for all values of the shape parameter p. As expected, the performance of OLS was the worst for p = 2.5, however, it consistently increased with the value of

Table 1: Means, Variances and MSE's for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, n=10

	$\hat{\beta}_0$			$\hat{\beta}_1$		
Method	Mean	Variance	MSE	Mean	Variance	MSE
			p=2.0			
OLS	0.003516	0.442733	0.442745	0.998563	0.011514	0.011516
MML	0.004207	0.341639	0.341656	0.998604	0.009006	0.009008
LAD	-0.002318	0.362488	0.362493	0.999963	0.009799	0.009799
WIN10	0.000232	0.361181	0.361181	0.999536	0.009411	0.009411
WIN20	0.001934	0.300163	0.300167	0.999102	0.007824	0.007825
TLS	0.006592	0.329992	0.330035	0.998576	0.008706	0.008708
Theil	0.000764	0.314738	0.314738	0.999397	0.008095	0.008096
Wtd.Theil	0.001506	0.312057	0.312059	0.999328	0.008060	0.008060
			P=2.5			
OLS	-0.003356	0.461119	0.461130	1.000817	0.012041	0.012042
MML	-0.003358	0.413896	0.413908	1.000816	0.010877	0.010878
LAD	-0.001238	0.494956	0.494957	1.000694	0.013322	0.013322
WIN10	-0.003894	0.459236	0.459251	1.000988	0.012092	0.012093
WIN20	-0.001129	0.385763	0.385764	1.000446	0.010191	0.010191
TLS	-0.002565	0.445692	0.445699	1.000634	0.011855	0.011855
Theil	-0.002769	0.413026	0.413033	1.000909	0.010785	0.010786
Wtd.Theil	-0.000713	0.407067	0.407068	1.000667	0.010531	0.010531
			P=3.0			
OLS	-0.002395	0.459847	0.459853	1.000911	0.012078	0.012079
MML	-0.001450	0.410860	0.410862	1.000782	0.010912	0.010913
LAD	0.003457	0.556958	0.556970	0.999881	0.015020	0.015020
WIN10	0.002749	0.475428	0.475435	0.999938	0.012637	0.012637
WIN20	0.001543	0.415174	0.415177	1.000308	0.010967	0.010968
TLS	-0.002892	0.485833	0.485841	1.000915	0.012907	0.012908
Theil	0.000275	0.448417	0.448417	1.000647	0.011503	0.011503
Wtd.Theil	0.000458	0.438228	0.438228	1.000618	0.011210	0.011210
			P=3.5			
OLS	-0.013050	0.470511	0.470681	1.000804	0.012082	0.012082
MML	-0.010891	0.434622	0.434741	1.000796	0.011308	0.011309
LAD	-0.012993	0.594436	0.594605	1.001073	0.016032	0.016034
WIN10	-0.014704	0.510295	0.510512	1.001517	0.013519	0.013521
WIN20	-0.010134	0.446861	0.446964	1.000764	0.011649	0.011650
TLS	-0.009950	0.524629	0.524728	1.000705	0.013743	0.013743
Theil	-0.009799	0.472920	0.473016	1.000470	0.011964	0.011964
Wtd.Theil	-0.009878	0.470252	0.470350	1.000552	0.011806	0.011806
			P=5.0			
OLS	0.006726	0.473619	0.473664	0.999226	0.012242	0.012243
MML	0.006238	0.459306	0.459345	0.999366	0.011917	0.011917
LAD	0.005333	0.653941	0.653969	0.999511	0.017332	0.017332
WIN10	0.004859	0.542576	0.542600	0.999847	0.014320	0.014320
WIN20	0.006534	0.482789	0.482832	0.999342	0.012526	0.012526
TLS	0.005403	0.587715	0.587744	0.999960	0.015314	0.015314
Theil	0.005450	0.523069	0.523098	0.999733	0.013058	0.013058
Wtd.Theil	0.007827	0.507404	0.507465	0.999458	0.012595	0.012596

Table 1 (continued): Means, Variances and MSE's for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, n=20

m	$\hat{\beta}_0$			$\hat{\beta}_1$		
Method	Mean	Variance	MSE	Mean	Variance	MSE
			p=2.0			
OLS	-0.002366	0.214414	0.214420	1.000324	0.001462	0.001463
MML	-0.001976	0.141376	0.141380	1.000244	0.000964	0.000964
LAD	-0.016728	0.152118	0.152398	1.001465	0.001097	0.001099
WIN10	-0.000748	0.168078	0.168079	1.000103	0.001165	0.001165
WIN20	-0.000829	0.128047	0.128048	1.000164	0.000879	0.000879
TLS	-0.000038	0.144824	0.144824	1.000100	0.001000	0.001000
Theil	0.007148	0.132135	0.132187	0.999243	0.000896	0.000897
Wtd.Theil	-0.000949	0.130729	0.130730	1.000045	0.000880	0.000880
			P=2.5			
OLS	-0.004576	0.210169	0.210190	1.000161	0.001458	0.001458
MML	-0.005110	0.173648	0.173674	1.000131	0.001211	0.001211
LAD	-0.024790	0.207893	0.208508	1.001937	0.001483	0.001487
WIN10	0.000021	0.205904	0.205904	0.999652	0.001469	0.001469
WIN20	-0.006372	0.161651	0.161691	1.000210	0.001144	0.001144
TLS	-0.006068	0.186094	0.186131	1.000211	0.001325	0.001325
Theil	0.005634	0.173694	0.173725	0.999042	0.001185	0.001186
Wtd.Theil	-0.005509	0.171549	0.171580	1.000126	0.001166	0.001167
			P=3.0			
OLS	-0.000997	0.217897	0.217898	1.000303	0.001517	0.001518
MML	-0.001199	0.190935	0.190936	1.000301	0.001320	0.001320
LAD	-0.015553	0.236484	0.236726	1.001847	0.001681	0.001684
WIN10	-0.001128	0.227378	0.227379	1.000144	0.001614	0.001614
WIN20	0.000029	0.181401	0.181401	1.000151	0.001256	0.001256
TLS	0.002355	0.211460	0.211466	1.000057	0.001474	0.001474
Theil	0.014359	0.195260	0.195466	0.999013	0.001304	0.001305
Wtd.Theil	0.001850	0.192893	0.192896	1.000188	0.001278	0.001278
			P=3.5			
OLS	-0.005599	0.215732	0.215764	1.001062	0.001529	0.001530
MML	-0.002278	0.193426	0.193431	1.000902	0.001370	0.001371
LAD	-0.019423	0.262883	0.263260	1.002378	0.001877	0.001882
WIN10	-0.002735	0.242673	0.242680	1.000908	0.001750	0.001751
WIN20	-0.001386	0.195807	0.195809	1.000829	0.001384	0.001385
TLS	0.003151	0.232698	0.232707	1.000321	0.001637	0.001637
Theil	0.008258	0.211309	0.211377	0.999694	0.001439	0.001439
Wtd.Theil	-0.003741	0.209351	0.209365	1.000870	0.001413	0.001414
			P=5.0			
OLS	-0.001472	0.206327	0.206329	1.000286	0.001458	0.001458
MML	-0.001661	0.196991	0.196994	1.000312	0.001395	0.001395
LAD	-0.019671	0.282823	0.283210	1.002131	0.002007	0.002011
WIN10	0.002690	0.250279	0.250286	0.999782	0.001833	0.001833
WIN20	-0.002567	0.202167	0.202173	1.000406	0.001418	0.001419
TLS	-0.003974	0.243164	0.243180	1.000674	0.001704	0.001704
Theil	0.013557	0.220453	0.220637	0.999055	0.001461	0.001462
Wtd.Theil	0.001284	0.217649	0.217651	1.000161	0.001438	0.001438

Table 1 (continued): Means, Variances and MSE's for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, n=50

	$\hat{\beta}_0$			$\hat{\beta}_1$		
Method	Mean	Variance	MSE	Mean	Variance	MSE
			p=2.0			
OLS	-0.000594	0.084085	0.084085	1.000087	0.000104	0.000104
MML	0.001371	0.047691	0.047692	1.000004	0.000055	0.000055
LAD	0.000378	0.052987	0.052987	0.999985	0.000063	0.000063
WIN10	0.006258	0.065587	0.065626	0.999828	0.000079	0.000079
WIN20	0.000867	0.046586	0.046586	1.000034	0.000055	0.000055
TLS	0.001880	0.050630	0.050634	1.000005	0.000060	0.000060
Theil	-0.000014	0.047390	0.047390	1.000006	0.000054	0.000054
Wtd.Theil	-0.000386	0.047201	0.047202	1.000025	0.000053	0.000053
			P=2.5			
OLS	0.002424	0.086628	0.086634	0.999785	0.000099	0.000099
MML	-0.000641	0.066412	0.066413	0.999911	0.000076	0.000076
LAD	0.001515	0.080364	0.080366	0.999884	0.000094	0.000094
WIN10	0.004181	0.091114	0.091131	0.999756	0.000108	0.000108
WIN20	-0.000303	0.064850	0.064850	0.999878	0.000075	0.000075
TLS	-0.004784	0.076472	0.076495	1.000056	0.000087	0.000087
Theil	0.002601	0.068287	0.068294	0.999896	0.000075	0.000075
Wtd.Theil	0.002145	0.068267	0.068272	0.999915	0.000075	0.000075
			P=3.0			
OLS	-0.012133	0.085280	0.085428	1.000378	0.000100	0.000100
MML	-0.011707	0.073272	0.073409	1.000390	0.000085	0.000085
LAD	-0.012364	0.089730	0.089883	1.000416	0.000106	0.000106
WIN10	-0.007459	0.096523	0.096579	1.000233	0.000115	0.000115
WIN20	-0.009552	0.071668	0.071759	1.000327	0.000084	0.000084
TLS	-0.010372	0.078219	0.078326	1.000291	0.000094	0.000094
Theil	-0.009797	0.075452	0.075548	1.000351	0.000084	0.000084
Wtd.Theil	-0.009190	0.074861	0.074945	1.000330	0.000083	0.000083
			P=3.5			
OLS	-0.013384	0.081143	0.081322	1.000466	0.000093	0.000093
MML	-0.012900	0.070895	0.071062	1.000445	0.000082	0.000082
LAD	-0.009534	0.092041	0.092131	1.000356	0.000108	0.000108
WIN10	-0.012675	0.089156	0.089317	1.000384	0.000108	0.000108
WIN20	-0.012857	0.069653	0.069818	1.000447	0.000081	0.000081
TLS	-0.012912	0.079559	0.079725	1.000442	0.000093	0.000093
Theil	-0.012624	0.077350	0.077510	1.000469	0.000083	0.000083
Wtd.Theil	-0.012442	0.076734	0.076889	1.000476	0.000082	0.000082
			P=5.0			
OLS	0.000349	0.080924	0.080924	1.000022	0.000093	0.000093
MML	-0.002554	0.075887	0.075893	1.000110	0.000088	0.000088
LAD	-0.004909	0.110364	0.110388	1.000214	0.000129	0.000129
WIN10	0.000915	0.100494	0.100495	1.000063	0.000122	0.000122
WIN20	-0.001840	0.076396	0.076399	1.000070	0.000088	0.000088
TLS	-0.002449	0.093636	0.093642	1.000074	0.000108	0.000108
Theil	-0.003242	0.083709	0.083720	1.000146	0.000090	0.000090
Wtd.Theil	-0.002844	0.083042	0.083050	1.000120	0.000089	0.000089

the shape parameter p since OLS is the optimal method under normality and the $LTS(p,\sigma)$ distribution approaches normal as $p \to \infty$. Results were not reproduced for the sake of brevity, however.

For the slope estimator $\hat{\beta}_1$:

- For $p = 2$ and 2.5, the performances of the WIN20 and WIN10 were the best at sample sizes 10 and 20 and Wtd.Theil and Theil provide the smallest MSE for the large sample sizes $(n = 50)$.
- For $p = 3.0$, WIN20 demonstrated the strongest performance with lowest MSE at all sample sizes except for $n = 10$, in which case MML provides the smallest MSE.
- MML, WIN10 and Wtd.Theil were the preferred methods for $p = 3.5$. When $p = 5.0$, MML, WIN10 and WIN20 have the smallest MSE.
- The LAD and TLS slope estimators showed very poor performance with the largest MSE values at all sample sizes for all values of the shape parameter, p.
- The performance of the OLS slope estimator is similar to the OLS intercept estimator.

Robustness

In practice, a model is identified by Q-Q plots or goodness of fit tests. Neither of these methods, nor in fact any other method, identifies a model exactly or uniquely. In other words, the value of the shape parameter p in $LTS(p,\sigma)$ might be misspecified. Assume, for illustration, that the true distribution is the $LTS(3.5,\sigma)$. To represent a large number of plausible alternatives, consider the following sample models:

- Model (1): $LTS(2.0,\sigma)$
- Model (2): $LTS(5.0,\sigma)$
- Model (3): Outlier Model; $(n\text{-}r)$ observations from $LTS(3.5,\sigma)$ and r observations from $LTS(3.5, 4\sigma)$ where $r = [0.5+0.1n]$
- Model (4): Mixture Model; $0.90LTS(3.5,\sigma) + 0.10LTS(3.5,4\sigma)$
- Model (5): Contamination Model; $0.90LTS(3.5,\sigma) + 0.10\,Normal(0, 4)$

The simulated means, variances and MSE of the regression estimators for the alternative models are shown in Table 2. It should be noted that an estimator $\hat{\theta}$ of θ is called robust if it is fully efficient (or nearly so) for an assumed model but maintains high efficiencies for plausible alternatives to the assumed model. Based on the information in Table 2, the following conclusions are put forth for the intercept estimator $\hat{\beta}_0$:

- WIN10 and WIN20 showed the strongest performance with lowest MSE for Models (1), (3), (4) and (5) at all sample sizes except for a sample of size 50 in Models (1) and (5) in which case the Wtd. Theil provides the smallest MSE.
- MML demonstrated the strongest performance with lowest MSE as compared to other methods in Model (2).
- OLS and LAD showed very poor estimator performance at all sample sizes with largest MSE values for Models (1), (3), (4), (5) and Model (2), respectively.

For the slope estimator $\hat{\beta}_1$:

- WIN10 and WIN20 provided the smallest MSE for Models (1), (3), (4) and (5) at sample sizes 10 and 20, however, for the sample size $n = 50$, the Wtd. Theil's slope estimator had the strongest efficiency.
- The highest performance for Model (2), similar to intercept estimator $\hat{\beta}_0$, is achieved by MML.
- OLS and LAD have the highest MSE values for Models (1), (3), (4), (5) and Model (2), respectively. Therefore, they are not preferred estimators under these sample models.

Conclusion

The OLS estimation procedure provides good results when the error terms have a normal distribution. However, in real life, it is nearly impossible to find a data set that satisfies all of

Table 2: Means, Variances and MSE's for the sample models (1)-(5), n=10

	$\hat{\beta}_0$			$\hat{\beta}_1$		
Method	Mean	Variance	MSE	Mean	Variance	MSE
			Model (1)			
OLS	-0.012290	0.485174	0.485325	1.001940	0.012996	0.013000
MML	-0.010016	0.352918	0.353018	1.001675	0.009657	0.009660
LAD	-0.009981	0.357653	0.357753	1.001655	0.009603	0.009605
WIN10	-0.007079	0.365626	0.365676	1.001175	0.009806	0.009807
WIN20	-0.010102	0.317492	0.317594	1.001656	0.007995	0.007998
TLS	-0.009358	0.325208	0.325295	1.001784	0.008635	0.008638
Theil	-0.010226	0.307593	0.307697	1.001484	0.008033	0.008035
Wtd.Theil	-0.012949	0.308182	0.308350	1.001986	0.008008	0.008012
			Model (2)			
OLS	-0.006355	0.470296	0.470337	1.000850	0.012108	0.012109
MML	-0.006313	0.456893	0.456933	1.000915	0.011789	0.011790
LAD	-0.008034	0.656990	0.657055	1.001427	0.017477	0.017479
WIN10	-0.008389	0.546347	0.546417	1.001590	0.014378	0.014380
WIN20	-0.006451	0.480393	0.480435	1.001023	0.012422	0.012423
TLS	-0.005099	0.577213	0.577239	1.000909	0.015085	0.015086
Theil	-0.005327	0.520019	0.520047	1.000805	0.012966	0.012966
Wtd.Theil	-0.006527	0.507479	0.507522	1.001039	0.012562	0.012563
			Model (3)			
OLS	0.012384	1.223769	1.223923	0.998354	0.032164	0.032167
MML	0.009457	0.753716	0.753806	0.998816	0.020260	0.020262
LAD	0.016420	0.751355	0.751625	0.997478	0.019980	0.019987
WIN10	-0.010434	0.754763	0.754872	1.000511	0.019574	0.019574
WIN20	0.008255	0.630553	0.630621	0.998781	0.016398	0.016400
TLS	0.012560	0.667084	0.667242	0.998366	0.017401	0.017404
Theil	0.007114	0.660818	0.660869	0.999191	0.016871	0.016872
Wtd.Theil	0.006276	0.658813	0.658853	0.999105	0.016838	0.016839
			Model (4)			
OLS	-0.015771	1.169783	1.170031	1.003291	0.030152	0.030163
MML	-0.015086	0.776937	0.777164	1.002934	0.020509	0.020518
LAD	-0.022904	0.798735	0.799260	1.003815	0.021329	0.021343
WIN10	-0.026484	0.830451	0.831153	1.003516	0.021298	0.021310
WIN20	-0.013370	0.661862	0.662040	1.002650	0.017077	0.017084
TLS	-0.011041	0.710763	0.710885	1.002151	0.018798	0.018803
Theil	-0.016685	0.694106	0.694385	1.002835	0.017787	0.017795
Wtd.Theil	-0.015797	0.690942	0.691192	1.002849	0.017729	0.017737
			Model (5)			
OLS	-0.004107	1.179549	1.179566	1.001699	0.030212	0.030215
MML	-0.001795	0.797778	0.797782	1.001125	0.021203	0.021204
LAD	0.004313	0.797272	0.797291	0.999694	0.021461	0.021461
WIN10	-0.011044	0.839572	0.839694	1.001791	0.022213	0.022217
WIN20	0.000062	0.684177	0.684177	1.000728	0.017888	0.017889
TLS	0.002536	0.742882	0.742889	1.000399	0.019719	0.019719
Theil	-0.000683	0.701909	0.701910	1.000896	0.018376	0.018377
Wtd.Theil	-0.001727	0.715698	0.715701	1.000978	0.018841	0.018842

Table 2 (continued): Means, Variances and MSE's for the sample models (1)-(5), n=20

	$\hat{\beta}_0$			$\hat{\beta}_1$		
Method	Mean	Variance	MSE	Mean	Variance	MSE
			Model (1)			
OLS	-0.003752	0.222670	0.222685	0.999951	0.001546	0.001546
MML	-0.004145	0.138954	0.138971	1.000165	0.000966	0.000966
LAD	-0.014743	0.152306	0.152524	1.001444	0.001072	0.001074
WIN10	-0.004923	0.163155	0.163179	1.000060	0.001149	0.001149
WIN20	-0.001735	0.123830	0.123833	1.000051	0.000863	0.000863
TLS	-0.002840	0.142264	0.142272	1.000216	0.000988	0.000988
Theil	0.011090	0.129912	0.130035	0.998996	0.000873	0.000874
Wtd.Theil	0.001468	0.128770	0.128772	0.999909	0.000861	0.000861
			Model (2)			
OLS	-0.009421	0.220500	0.220589	1.000323	0.001527	0.001527
MML	-0.007871	0.208822	0.208884	1.000299	0.001453	0.001453
LAD	-0.015340	0.296896	0.297132	1.001461	0.002084	0.002086
WIN10	-0.008323	0.263561	0.263631	1.000316	0.001861	0.001861
WIN20	-0.006477	0.212390	0.212432	1.000258	0.001475	0.001475
TLS	-0.000358	0.260313	0.260313	0.999871	0.001816	0.001816
Theil	0.010322	0.231483	0.231589	0.998944	0.001514	0.001515
Wtd.Theil	-0.002871	0.228255	0.228263	1.000186	0.001491	0.001491
			Model (3)			
OLS	0.008763	0.534048	0.534125	0.998650	0.003708	0.003710
MML	0.009852	0.271805	0.271903	0.998716	0.001954	0.001955
LAD	-0.004580	0.312706	0.312727	0.999984	0.002225	0.002225
WIN10	0.007319	0.356939	0.356993	0.998841	0.002524	0.002525
WIN20	0.010996	0.254113	0.254234	0.998615	0.001813	0.001815
TLS	0.012679	0.292134	0.292295	0.998599	0.002067	0.002069
Theil	0.025097	0.270968	0.271598	0.997220	0.001856	0.001864
Wtd.Theil	0.010753	0.268535	0.268651	0.998603	0.001824	0.001826
			Model (4)			
OLS	-0.011834	0.530361	0.530501	1.000266	0.003641	0.003641
MML	-0.007635	0.285413	0.285471	1.000330	0.002019	0.002019
LAD	-0.021156	0.320834	0.321282	1.001696	0.002262	0.002265
WIN10	-0.002971	0.383144	0.383153	0.999619	0.002664	0.002664
WIN20	-0.004989	0.263167	0.263192	1.000165	0.001853	0.001853
TLS	-0.002033	0.301227	0.301231	0.999793	0.002084	0.002084
Theil	0.007851	0.274877	0.274938	0.998851	0.001875	0.001876
Wtd.Theil	-0.005549	0.272119	0.272150	1.000128	0.001839	0.001839
			Model (5)			
OLS	-0.014204	0.546967	0.547169	1.000832	0.003830	0.003830
MML	-0.007622	0.291418	0.291476	1.000401	0.002046	0.002046
LAD	-0.018763	0.323247	0.323599	1.001347	0.002247	0.002249
WIN10	-0.012408	0.388890	0.389044	1.000889	0.002683	0.002684
WIN20	-0.007292	0.271799	0.271852	1.000305	0.001893	0.001893
TLS	-0.000146	0.296508	0.296508	0.999684	0.002040	0.002040
Theil	0.006440	0.283805	0.283846	0.999057	0.001913	0.001914
Wtd.Theil	-0.007353	0.281584	0.281638	1.000388	0.001892	0.001892

Table 2 (continued): Means, Variances and MSE's for the sample models (1)-(5), n=50

	$\hat{\beta}_0$			$\hat{\beta}_1$		
Method	Mean	Variance	MSE	Mean	Variance	MSE
			Model (1)			
OLS	-0.000805	0.074541	0.074542	0.999982	0.000084	0.000084
MML	-0.000114	0.044790	0.044790	0.999976	0.000050	0.000050
LAD	0.004056	0.051563	0.051579	0.999809	0.000059	0.000059
WIN10	-0.002538	0.065713	0.065719	1.000100	0.000077	0.000077
WIN20	0.000232	0.045412	0.045412	0.999965	0.000052	0.000052
TLS	0.001835	0.048722	0.048726	0.999970	0.000055	0.000055
Theil	-0.000018	0.044688	0.044688	0.999963	0.000050	0.000050
Wtd.Theil	-0.000159	0.044719	0.044719	0.999970	0.000050	0.000050
			Model (2)			
OLS	-0.005070	0.082616	0.082642	1.000184	0.000095	0.000095
MML	-0.003343	0.078478	0.078489	1.000146	0.000091	0.000091
LAD	-0.004388	0.107137	0.107156	1.000242	0.000124	0.000124
WIN10	-0.002386	0.102003	0.102008	1.000153	0.000125	0.000125
WIN20	-0.002738	0.079004	0.079012	1.000119	0.000092	0.000092
TLS	-0.001691	0.099586	0.099589	1.000081	0.000116	0.000116
Theil	-0.001652	0.086823	0.086826	1.000123	0.000093	0.000093
Wtd.Theil	-0.001529	0.086498	0.086501	1.000126	0.000092	0.000092
			Model (3)			
OLS	0.001170	0.218986	0.218987	1.000070	0.000247	0.000247
MML	0.005980	0.138048	0.138083	0.999923	0.000149	0.000149
LAD	0.007935	0.116472	0.116535	0.999940	0.000131	0.000131
WIN10	0.006277	0.158747	0.158786	0.999961	0.000176	0.000176
WIN20	0.007735	0.102137	0.102197	0.999891	0.000112	0.000112
TLS	0.005534	0.118133	0.118164	1.000001	0.000130	0.000130
Theil	0.008831	0.099889	0.099967	0.999887	0.000108	0.000109
Wtd.Theil	0.009458	0.100008	0.100097	0.999849	0.000108	0.000108
			Model (4)			
OLS	0.007550	0.213060	0.213117	0.999828	0.000249	0.000249
MML	0.008803	0.134974	0.135051	0.999741	0.000155	0.000155
LAD	0.009354	0.118182	0.118269	0.999582	0.000142	0.000142
WIN10	0.014525	0.166463	0.166674	0.999435	0.000197	0.000197
WIN20	0.007484	0.104324	0.104380	0.999757	0.000123	0.000123
TLS	0.003825	0.115579	0.115593	0.999892	0.000138	0.000138
Theil	0.006978	0.103516	0.103565	0.999747	0.000119	0.000119
Wtd.Theil	0.007271	0.103704	0.103757	0.999727	0.000118	0.000119
			Model (5)			
OLS	0.000823	0.213641	0.213642	1.000111	0.000251	0.000251
MML	0.001214	0.139224	0.139226	1.000019	0.000158	0.000158
LAD	-0.006313	0.123031	0.123071	1.000148	0.000146	0.000146
WIN10	0.002004	0.175000	0.175004	1.000008	0.000198	0.000198
WIN20	0.000914	0.109873	0.109874	0.999948	0.000129	0.000129
TLS	0.001631	0.116706	0.116709	0.999897	0.000135	0.000135
Theil	-0.000120	0.107528	0.107528	0.999936	0.000122	0.000122
Wtd.Theil	-0.000712	0.107393	0.107394	0.999947	0.000122	0.000122

the normality assumptions, therefore, alternative regression methods are needed. In this study, efficiency and robustness properties of some prominent robust and nonparametric regression estimators have been compared via Monte Carlo simulation when the error terms come from long-tailed symmetric $LTS(p,\sigma)$ distributions.

The methods giving the smallest MSE for various shape parameters and sample models were defined clearly for different sample sizes. If the distribution of error terms is $LTS(p,\sigma)$ in a simple linear regression model, it is therefore suggested that the selection procedure for the most efficient and robust method of estimation should be accomplished according to the results given above.

References

Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression.* New York, NY: Wiley.

Geary, R. C. (1947). Testing for normality. *Biometrika*, *34*, 209-242.

Huber, P. J. (1981). *Robust statistics*. NY: John Wiley.

Hussain, S. S., & Sprent, P. (1983). Nonparametric regression. *Journal of the Royal Statistical Society*, *A146*, 182-191.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Annals of Mathematical Statistics*, *43*, 1449-1458.

Nevitt, T., & Tam, H. P. (1998). A comparison of robust and nonparametric estimators under the simple linear regression model. *Multiple Linear Regression Viewpoints*, *25*, 54-69.

Pearson, E. S. (1932). The analysis of variance in cases of nonnormal variation. *Biometrika*, *23*, 114-133.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York, NY: Wiley.

Scholz, F.W. (1978). Weighted median regression estimates. *The Annals of Statistics*, 6(*3*), 603-609.

Sprent, P. (1993). *Applied nonparametric statistical methods.* NY: Chapman and Hall.

Şenoğlu, B. (2005). Robust 2^k factorial design with Weibull error distributions. *Journal of Applied Statistics*, *32*(*10*), 1051-1066.

Şenoğlu, B. (2007). Estimating parameters in one-way analysis of covariance model with short-tailed symmetric error distributions. *Journal of Computational and Applied Mathematics*, 201, 275-283.

Tam, H. P. (1996). *A review of nonparametric regression techniques.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, *12*, 85-91.

Tiku, M. L., Islam, M. Q., & Selçuk, A. (2001). Non-normal regression II: Symmetric distributions. *Communications in Statistics Theory and Methods*, *30*, 1021-1045.

Yale, C., & Forsythe, A. B. (1976). Winsorized regression. *Technometrics*, *18*, 291-300.