Robustness and Efficiency of the Theil-Sen Estimator in Simple and Multiple Linear Models*

Cameron Byrum[1], Xin Dang[2], Hanxiang Peng[3], Wei Wei[4]

December 12, 2009

Abstract

In this paper, we take a look at two regression methods, namely, the Least Squares method and the Theil-Sen estimator. The Least Squares method is the most common estimator, but it is known to lack efficiency with non-normally distributed error terms and to lack robustness to outliers. The Theil-Sen estimator is based around the median and is consequently far more robust to outliers. We take a look at both estimators in simple linear regressions and in multivariate models to compare the robustness and efficiency of each.

Keywords: Breakdown point, efficiency, Least Squares estimator, robustness, spatial median, Theil-Sen estimator.

1 Introduction

Ever since it was introduced by Gauss and Legendre in the early 1800's, the Least Squares (LS) method has become the cornerstone of classical statistics. Because the LS estimator is easy to compute and the most efficient for regressing data with normally distributed errors terms, it is commonly used for both simple and multiple regression models in many applications. However, the LS estimator has two major defects. First, it is not very efficient for regressing data with error terms that follow non-normal distributions, especially discrete ones. Second, the LS estimator is very sensitive to outliers. One bad observation may destroy the whole regression line. Therefore, in recent years, many new linear regression estimators have been introduced in hopes of solving these problems. For example, Edgeworth proposed the Least Absolute Values estimator in 1887. Huber proposed the M-estimators in 1973. Rousseeuw proposed the Least Median of Squares estimator and the Least Trimmed Squares estimator in 1984. ([6])

*This material is based upon work supported by the National Science Foundation REU supplemental project under Grant No. DMS–0809018 during the semesters of Spring and Fall 2008 by Wei Wei and Cameron Byrum as undergraduate students at the University of Mississippi.

[1] Undergraduate Student, University of Mississippi Department of Mathematics

[2] Faculty Advisor, University of Mississippi Department of Mathematics

[3] Faculty Advisor, Indiana University-Purdue University Indianapolis Department of Mathematical Sciences

[4] Graduate Student, Duke University Department of Finance

Our research is focused on the Theil-Sen (TS) estimator. Theil first proposed this method in 1950 as the median of pairwise slopes in a simple linear model. Sen then extended the Theil-Sen estimator in 1968 to handle ties. Several research studies have concluded that the TS estimator is highly robust, has a bounded influence function, and possesses high asymptotic efficiency ([3], [13]). Other studies discuss the asymptotic properties and behaviors of the TS estimator ([9], [12], [7]). Peng et al also show the super-efficiency of the TS estimator when the error term distribution is discontinuous at some point ([7]). Dang et al ([2]) extend the TS estimator to a multiple linear regression model and propose the use of multivariate medians. In particular, they implement the concept of spatial depth and use the spatial median to define the multivariate median of the estimator. Since the TS estimator is gaining popularity, it is also included in several textbooks on nonparametric and robust statistics, including works by Sprent ([10]), Hollander and Wolfe ([5]), and Rousseeuw and Leroy ([6]).

Since the TS estimator offers great robustness against outliers, simplicity in computation, analytical estimates of confidence intervals, and testable assumptions regarding residuals, it is used more and more in various fields of study. For example, Akritas et al discuss the usage and advantage of the TS estimator for linear regressions with double censored data in applications in astronomy. Also, the TS estimator is useful in estimation of circular arcs and aligned ellipses ([1]). It is also suggested as a potential replacement of the ordinary LS estimator for linear regression in remote sensing applications ([4]).

The TS estimator is already known to be more robust than the Least Squares method. The purpose of our research is to empirically test and demonstrate the efficiency of the TS estimator for both simple and multiple linear regression models. We compare the variance of the TS estimator with that of the LS estimator in a simple linear model with five different symmetric error term distributions including a normal, a uniform, a binomial, a T3, and a Cauchy distribution. We compare the variance of each TS estimate with that of the corresponding LS estimate in a multiple linear model with different error terms from the same set of five distributions as well as the covariance matrices of each of the estimates.

We find that in the simple linear model, when the error terms follow a light-tailed continuous distribution, such as the normal or the uniform distribution, the TS estimator is only slightly less efficient than the LS estimator. We see the Theil-Sen estimator perform better with a T3 distribution, and we get drastically lower variance with the Cauchy distribution in comparison to the Least Squares method. Also, when the error terms follow a discrete distribution, such as the binomial distribution, the TS estimator is far more efficient than the LS estimator.

For our multiple linear model, when the sample size is large enough (at least 50), the TS estimator is consistently less efficient than the LS estimator only in the models with normally distributed and uniformly distributed error terms. When the error terms follow a T3 or Cauchy distribution, the TS estimator is again typically better. In fact, our results suggest that as the distribution of the error terms moves from center focused to more and more heavy-tailed, the TS estimator gains efficiency over the LS estimator. For the binomial distribution, the LS estimator is better for smaller sample sizes, but as the size increases, we see the Theil-Sen estimator become the more efficient method.

The rest of the paper is organized as follows. In Section 2, we first briefly discuss the advantages the TS estimator has over the LS estimator in robustness. Then we show the simulation set up for our simple linear model and the comparison of efficiency between theoretical variance and variance from our simulation for the LS estimator. Finally, we give the results of comparison between the TS and LS estimator efficiencies. In Section 3, we first introduce the measure of spatial median, and then we apply the spatial median to our multiple linear model. We discuss the simulation set up for the multiple linear model and give the results of our relative comparison between the parameters of the TS and LS estimators. We use covariance matrices formed from our multivariate estimates to calculate the relative efficiency of the two methods. We also discuss the usage and advantage of a random stochastic procedure for the multiple linear regression. Section 4 gives a summary of our results and our conclusion after investigating both estimators.

2 Simple Linear Model

In a simple linear regression model, we typically have

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, ..., n$$

where the error term e_i is taken independently from some distribution and the intercept α and the slope β are the parameters to be estimated. We use $\hat{\alpha}$ and $\hat{\beta}$ to represent the respective estimations of α and β.

The Least Squares estimation is accomplished by minimizing the sum of squared errors (SSE). So we have

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \underset{(\alpha,\beta)^T}{\text{argmin}} \sum_{i=1}^{n} r_i^2 = \underset{(\alpha,\beta)^T}{\text{argmin}} \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

where each r_i is the residual or the difference between the observed and estimated values.

The LS estimator is given by:

$$\hat{\beta}_{LS} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

The TS estimator is proposed as the median of all $\binom{n}{2}$ pairwise slopes of the n points in a set of data. Namely,

$$\hat{\beta}_{TS} = \text{med}\left\{ \frac{y_j - y_i}{x_j - x_i}, \ x_i \neq x_j, \ 1 \leq i < j \leq n \right\}$$

Both $\hat{\beta}_{LS}$ and $\hat{\beta}_{TS}$ are consistent and unbiased estimators of β, meaning that $\hat{\beta}_n \xrightarrow{p} \beta$ and $E(\hat{\beta}) = \beta$.

When comparing methods of regression, two things we are most interested in are efficiency and robustness. The efficiency measures how well the results from the regression describe the relation between the response variable and the explanatory variables while robustness measures the extent to which the regression results will be affected by outliers. A common measure of robustness for regression estimators is the breakdown point. Generally speaking, the breakdown point is the percentage of outliers needed in a set of data to have an arbitrarily large effect on an estimator.

Two common measures of efficiency are the Mean Squared Error (MSE) and $\widehat{Var}(\hat{\beta})$, which are calculated as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta)^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{\beta}_i - \bar{\beta})^2 + (\bar{\beta} - \beta)^2\right]$$

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \bar{\beta})^2$$

where N is the number of repetitions, β is the true slope of the line, $\hat{\beta}_i$ is the estimated slope each time, and $\bar{\beta}$ is the mean of all $\hat{\beta}_i$'s. However, when the sample size is large, $\bar{\beta} \approx \beta$ since $\hat{\beta}$ is unbiased, and thus $(\bar{\beta} - \beta)^2 \approx 0$. Hence, MSE $\approx$ Var$(\hat{\beta})$. Thus the relative efficiency of one estimator to the other, in this case of $\hat{\beta}_{TS}$ to $\hat{\beta}_{LS}$, is defined as either of the following:

$$\text{RE}(\hat{\beta}_{TS}, \hat{\beta}_{LS}) = \frac{\text{MSE}(\hat{\beta}_{LS})}{\text{MSE}(\hat{\beta}_{TS})} \quad \text{or} \quad \text{RE}(\hat{\beta}_{TS}, \hat{\beta}_{LS}) = \frac{\widehat{\text{Var}}(\hat{\beta}_{LS})}{\widehat{\text{Var}}(\hat{\beta}_{TS})}$$

In this research study, we choose to use the second expression to calculate our relative efficiency. Now, we will take a look at the assumptions of the two estimation methods and the effects of those assumptions.

The classical assumption for the LS estimator is that the error terms come from a normal distribution and are homoscedastic. Under those assumptions, the LS estimator has been proven to be the optimal estimator, and its efficiency cannot be surpassed. However, when the error terms are normally distributed but heteroscedastic, the efficiency of the LS estimator can be relatively poor and the usual confidence interval for the slope can have highly unsatisfactory probability coverage. When the error terms are non-normally distributed, its efficiency is even worse ([13]). Also, since $\sum_{i=1}^{n} r_i^2$ is dominated by larger r_i's, the LS estimator is subject to the influence of the outliers. It is especially vulnerable to outliers in one of the explanatory variables ([6]).

On the other hand, the TS estimator does not require strict assumptions. It works well when the error terms follow any distribution. It is especially efficient when the error terms follow a discrete distribution, and this has been called the superefficiency property. Since the TS estimator gathers information about central tendency from the median, it is naturally less affected by the occurrence of outliers in a set of data. In addition, it is also robust against outliers in explanatory variables where as some other methods, such as the Least Absolute Deviation estimator, are only robust against y-outliers.

The robustness of Theil-Sen estimators to outliers in a simple linear regression can easily be demonstrated with a small example. Consider the set of data where $y = 10 + 3x + e$ with a normal error distribution $e \sim N(0, 4)$ and where x is a sequence from 1 to 20. Both the Least Squares method and Theil-Sen perform well for estimating the true regression line as shown in the graph below in Figure 1. However, when outliers are created in the y values of the data (shown in the graph by triangles), the robustness of the Theil-Sen estimator in comparison with Least Squares is clear. The slope of the least squares estimate is drastically affected while the Theil-Sen estimate shows almost no effect and gives a close estimation of the true regression line. This is depicted in Figure 1, and the actual estimates are shown in the following equations:

True Regression Line	$y = 10 + 3x$
LSE without outliers	$y = 10.06 + 3.03x$
TSE without outliers	$y = 10.03 + 3.05x$
LSE with outliers	$y = 25.39 + 1.99x$
TSE with outliers	$y = 10.35 + 3.00x$

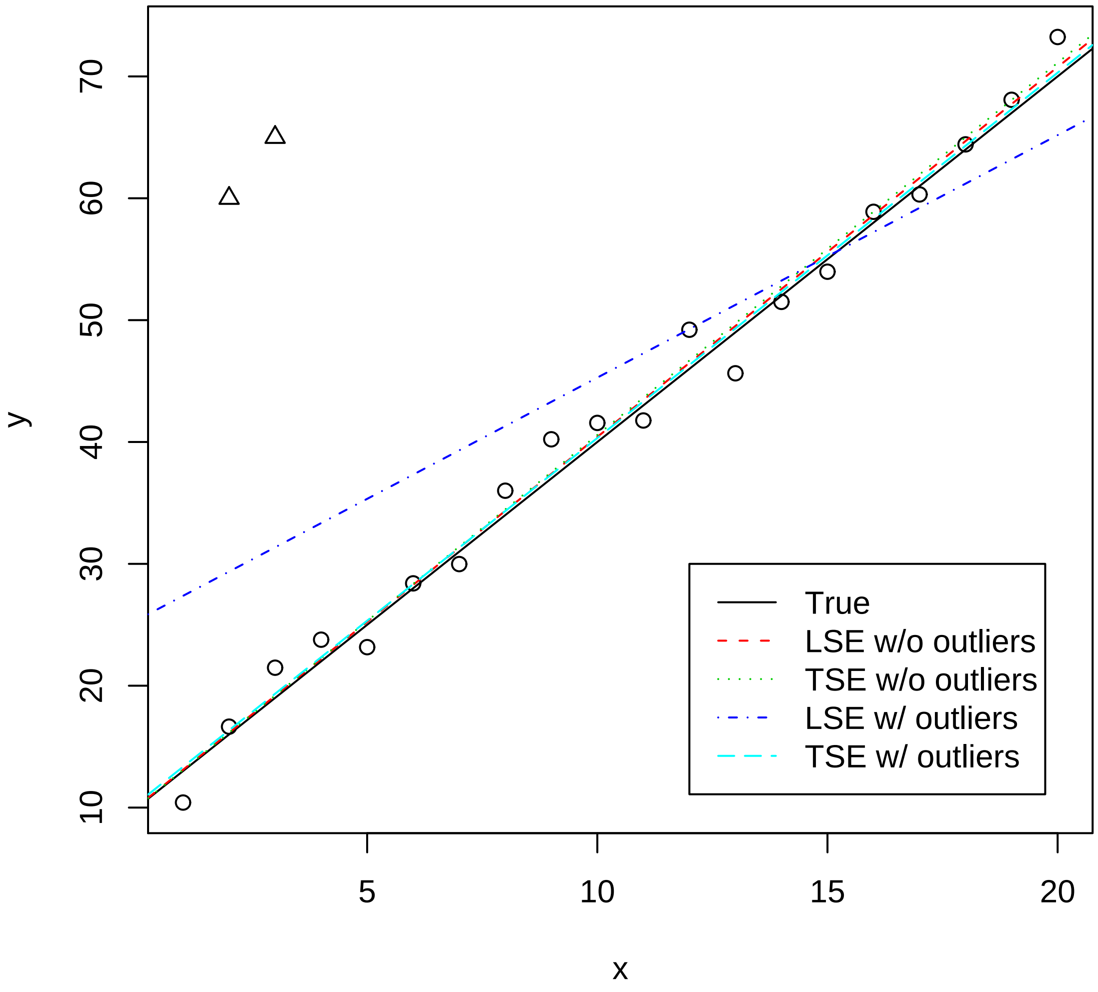

Figure 1: The Robustness of the Least Squares and Theil-Sen Estimators for Outliers in the y Values

Figure 2 illustrates an image similar to that in Figure 1, but the two added outliers are in the x values instead of the y values. Note that when the distortion is caused by outliers in explanatory variables, the LS estimate is again affected while the TS estimator still gives a close estimation of the slope. The regression lines are shown in the following equations:

True Regression Line	$y = 10 + 3x$
LSE without outliers	$y = 10.06 + 3.03x$
TSE without outliers	$y = 10.03 + 3.05x$
LSE with outliers	$y = 20.53 + 1.75x$
TSE with outliers	$y = 11.66 + 2.89x$

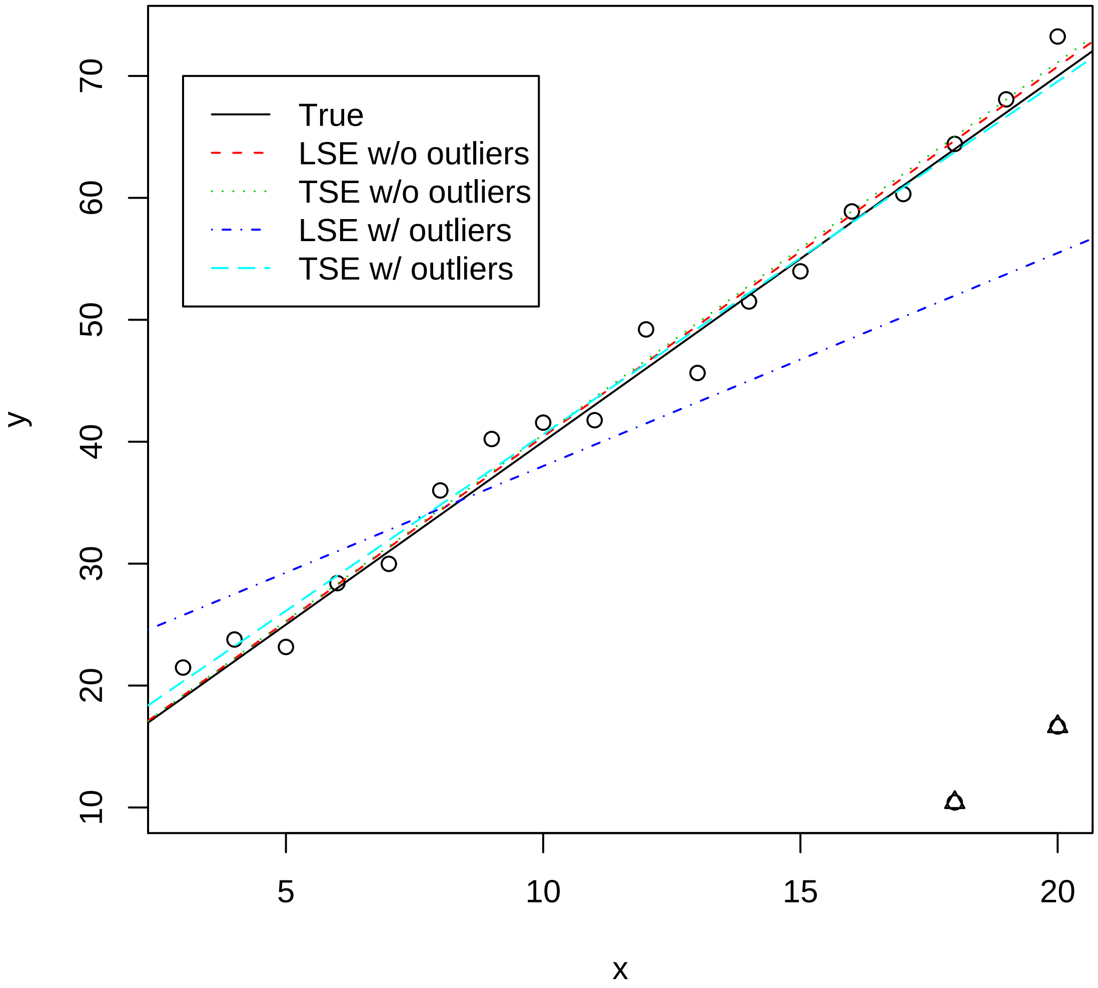

Figure 2: The Robustness of the Least Squares and Theil-Sen Estimators for Outliers in the x Values

We have already demonstrated the high sensitivity of the LS method to outliers. In Figures 1 and 2, two outliers cause the LSE to perform poorly. In fact, it only takes one extreme point to send the LS estimator to infinity. Thus, the breakdown point for LSE equals $\frac{1}{n}$ for a sample size n. This value is 0 in the limit as n approaches infinity, so the Least Squares method is said to have a breakdown point of 0%. ([6])

In Figures 1 and 2, the TS estimator was shown to be much less sensitive to outliers than the LS estimator. This is naturally the case since the TSE is created around the concept of a median, which is a much more robust measure of central tendency. For the TSE in a simple regression to remain unaffected by outliers, we need at least half of the pairwise slopes to have both points in the "good" set of data. As demonstrated by Dang et al ([2]), if we let ε represent the fraction of "bad" observations, we need $\frac{\binom{\lceil n(1-\varepsilon)\rceil}{2}}{\binom{n}{2}} > \frac{1}{2}$, where $\lceil x \rceil$ represents the smallest integer greater than or equal x. It follows that $(1-\varepsilon)^2 \geq \frac{1}{2}$, and so $\varepsilon \leq 1 - \frac{1}{2}^{\frac{1}{2}} \approx 0.293$. The breakdown point for the TSE is said to be 29.3%. ([6])

After demonstrating the advantage of TS estimators with robustness, we now conduct the following simulations using R Package software to show the efficiency of the TS estimator compared to that of the LS estimator. All of our simulations are done with the following linear model:

$$y_i = 1 + 2x + e_i, \quad i = 1, ..., n$$

So, our actual intercept, α, and actual slope, β, are 1 and 2 respectively. However, in order to keep our discussion on a one-dimension comparison, we choose to focus our investigation on the slope estimation. Also to simplify the case, we let x be a sequence from 1 to 100.

In our first simulation, each time, we randomly generate a set of 100 observations with $e \sim N(0, 3)$. We calculate $\hat{\beta}$ using both TS method and LS method. We also calculate the variance of the error terms using the residuals with the formula:

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{n-2} = \sum_{i=1}^{N} \frac{e_i^2}{n-2}$$

We repeat the above procedures for $N = 50$, 100, and 200 times. We calculate the variance of the N $\hat{\beta}_i$'s obtained from the TS method as well as the variance of the N $\hat{\beta}_i$'s obtained from the LS method. Then, we calculate the average of the N $\hat{\sigma}_i^2$'s from the LS estimation.

When we use the LS method to estimate the slope of a regression line, we can estimate the variance by:

$$\text{Var}(\hat{\beta}) = \frac{\text{Var}(e_i)}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{1}$$

Since x is a sequence from 1 to 100, we have:

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{k=1}^{100}(k - 50)^2 = 83350$$

Thus, $\text{Var}(\hat{\beta}) = \frac{\text{Var}(e_i)}{83350}$. Since $e \sim N(0, 3)$, the theoretical variance for the error terms is $\sigma^2 = 3$.

So, we can calculate the theoretical variance of the slope estimator using Equation 1. We have:

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{83350} = \frac{3}{83350} \approx 3.60e^{-5}$$

We can obtain an estimate for the variance of the slope using the average of $\hat{\sigma}_i^2$'s and Equation 1. Thus, we have:

$$\widehat{\text{Var}}(\hat{\beta})_1 = \frac{\text{Mean}(\hat{\sigma}_i^2)}{83350}$$

We can also obtain this estimate by averaging the difference between $\hat{\beta}_i$ and $\bar{\beta}$ over all the repetitions. So, we also have:

$$\widehat{\text{Var}}(\hat{\beta})_2 = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\beta}_i - \bar{\beta})^2$$

Before we make comparisons between the LSE and TSE, we first compare the theoretical variance with these two methods of estimation obtained under the LS method. The results are presented in Table 1. Note that the relative comparison is the observed value divided by the true value.

	Number of Repetitions	$\text{Var}(\hat{\beta})$	Theoretical Variance	Relative Comparison
$\widehat{\text{Var}}(\hat{\beta})_1$	50	$4.08e^{-5}$	$3.60e^{-5}$	1.133
	100	$4.04e^{-5}$		1.122
	200	$3.59e^{-5}$		0.997
$\widehat{\text{Var}}(\hat{\beta})_2$	50	$3.70e^{-5}$	$3.60e^{-5}$	1.028
	100	$3.54e^{-5}$		0.983
	200	$3.63e^{-5}$		1.008

Table 1: Comparisons with Theoretical Variance for LS Estimator where $e \sim N(0,3)$ with Sample Size $n = 100$

We see that $\widehat{\text{Var}}(\hat{\beta})_1$ and $\widehat{\text{Var}}(\hat{\beta})_2$ are fairly close to the corresponding theoretical variance. Also, both variances become smaller as the number of repetitions increases. When we repeat 200 times, both variances are within 1% difference from the theoretical variance. Therefore, we believe our simulation is valid and fair.

Finally, we compare the actual variances of the TS estimators with that of the LS estimators. We also perform the procedure described above for the cases in which $e \sim \text{Uniform}(-3,3)$, $e \sim \text{Binomial}(12, 0.5)$, $e \sim \text{T3}$, and $e \sim \text{Cauchy}$. The results for all the efficiency comparisons between the TS and LS estimates are shown in Table 2.

From the relative comparisons of LSE to TSE, we see that the TS estimator has a variance that is only slightly bigger than the variance of LS estimator when the error terms follow a light-tailed distribution such as $e_i \sim N(0,3)$ and $e_i \sim \text{Unif}(-3,3)$. Also, as the number of repetitions increases, the variance of the TS estimator becomes closer to that of the LS estimator for both error

Error Term Distribution	Number of Repetitions	Var($\hat{\beta}$)$_1$ by LSE	Var($\hat{\beta}$) by TSE	Relative Comparison
$e \sim$ Unif$(-3,3)$	50	$4.00e^{-5}$	$4.53e^{-5}$	0.883
	100	$3.03e^{-5}$	$3.24e^{-5}$	0.935
	200	$3.80e^{-5}$	$4.00e^{-5}$	0.950
$e \sim N(0,3)$	50	$4.08e^{-5}$	$4.46e^{-5}$	0.915
	100	$4.04e^{-5}$	$4.32e^{-5}$	0.935
	200	$3.59e^{-5}$	$3.77e^{-5}$	0.947
$e \sim$ T3	50	$2.86e^{-5}$	$1.54e^{-5}$	1.857
	100	$3.19e^{-5}$	$1.83e^{-5}$	1.743
	200	$5.38e^{-5}$	$2.25e^{-5}$	2.391
$e \sim$ Cauchy	50	$9.31e^{-1}$	$2.84e^{-5}$	32782
	100	1.29	$4.95e^{-5}$	26061
	200	$1.13e^{-1}$	$3.95e^{-5}$	2861
$e \sim$ Bin$(12,0.5)$	50	$4.36e^{-5}$	$4.08e^{-6}$	10.686
	100	$3.50e^{-5}$	$2.84e^{-6}$	12.324
	200	$2.96e^{-5}$	$2.48e^{-6}$	11.935

Table 2: Comparison between LSE and TSE for Simple Linear Model with Sample Size $n = 100$

distributions. When repeated 200 times, this is especially true as Var($\hat{\beta}$) of TS is within about a 5% difference of LS.

Also, when $e_i \sim$ Bin$(12, 0.5)$, a discrete distribution, the TS estimator is much more efficient than the LS estimator. In fact, the TS estimator has its smallest $Var(\hat{\beta})$ among our simulations when the error was a binomial distribution. Thus, we confirmed that the TS estimator is super efficient for data with discretely distributed error terms.

Finally, we see that the TS estimator is more efficient than the LS estimator both when $e_i \sim$ T3 and when $e_i \sim$ Cauchy. In particular, when $e_i \sim$ Cauchy, the TS estimator remains efficient whereas the LS estimate becomes nearly useless in comparison. This suggests that as the distributions of error terms change from center-focused to heavier-tailed, the TS estimator becomes relatively more and more efficient.

We repeat our five simulations for $n = 200$ error terms. The results of the fairness test when $e_i \sim N(0, 3)$ are shown in Table 3 and the results for the comparisons between the efficiencies of the LS and TS estimators are shown in Table 4. The true variance calculation is given by:

$$\text{Var}(\hat{\beta}) = \frac{\text{Var}(e_i)}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_{k=1}^{200}(k - 100)^2} = \frac{3}{666700} \approx 4.50e^{-6}$$

We find that all of our conclusions for $n = 100$ still hold when $n = 200$. Also, we note that for the discrete binomial distribution, the TS estimator provides the exact estimation for the slope and has a variance of nearly 0. This further demonstrates that TSE is super efficient for a discrete distribution of error terms. As expected, we see that when the number of error terms increases, all the variances are smaller than the values for $n = 100$.

	Number of Repetitions	$\text{Var}(\hat{\beta})$	Theoretical Variance	Relative Comparison
$\widehat{\text{Var}}(\hat{\beta})_1$	50	$4.71e^{-6}$	$4.50e^{-6}$	1.047
	100	$4.46e^{-6}$		0.991
	200	$4.31e^{-6}$		0.958
$\widehat{\text{Var}}(\hat{\beta})_2$	50	$4.54e^{-6}$	$4.50e^{-6}$	1.009
	100	$4.49e^{-6}$		0.998
	200	$4.46e^{-6}$		0.991

Table 3: Comparisons with Theoretical Variance for LS Estimator where $e \sim N(0, 3)$ with Sample Size $n = 200$

Error Term Distribution	Number of Repetitions	$\text{Var}(\hat{\beta})_1$ by LSE	$\text{Var}(\hat{\beta})$ by TSE	Relative Comparison
$e \sim \text{Unif}(-3, 3)$	50	$3.72e^{-6}$	$3.79e^{-6}$	0.981
	100	$4.00e^{-6}$	$4.00e^{-6}$	1.000
	200	$5.24e^{-6}$	$5.59e^{-6}$	0.937
$e \sim N(0, 3)$	50	$4.71e^{-6}$	$4.98e^{-6}$	0.946
	100	$4.46e^{-6}$	$4.64e^{-6}$	0.961
	200	$4.45e^{-6}$	$4.84e^{-6}$	0.919
$e \sim$ T3	50	$6.00e^{-6}$	$3.46e^{-6}$	1.734
	100	$4.12e^{-6}$	$2.00e^{-6}$	2.060
	200	$3.95e^{-6}$	$2.07e^{-6}$	1.908
$e \sim$ Cauchy	50	$2.54e^{-3}$	$5.92e^{-6}$	429
	100	$9.47e^{-3}$	$4.02e^{-6}$	2356
	200	$2.04e^{-2}$	$5.79e^{-6}$	3523
$e \sim \text{Bin}(12, 0.5)$	50	$5.07e^{-6}$	0.00	undefined
	100	$4.27e^{-6}$	0.00	undefined
	200	$4.97e^{-6}$	0.00	undefined

Table 4: Comparison between LS and TS Estimators for Simple Linear Model with Sample Size $n = 200$

3 Multiple Linear Model

In a multiple linear regression, we have many dependent variables. Thus, the model becomes:

$$y_i = \alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip} + e_i, \quad i = 1, ..., n$$

We could also write this in a matrix form as:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$$

where X represents the following $n \times (p + 1)$ design matrix, $\underset{\sim}{y}$ is the following vector of length n, and $\underset{\sim}{\beta}$ is the $(p+1)$-dimension parameter which we try to estimate:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix}, \quad \underset{\sim}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

The Least Squares estimator is accomplished by minimizing the SSE. So we have:

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \underset{(\alpha,\beta_1,...\beta_p)^T}{\text{argmin}} \sum_{i=1}^{n} r_i^2 = \underset{(\alpha,\beta_1,...\beta_p)^T}{\text{argmin}} \sum_{i=1}^{n} [y_i - (\alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip})]^2 = \underset{\underset{\sim}{\beta}}{\text{argmin}} \left\| \underset{\sim}{y} - X\underset{\sim}{\beta} \right\|^2$$

with a solution $\hat{\beta}_{LS} = (X^T X)^{-1} X^T \underset{\sim}{y}$.

However, the procedure for computing the TS estimator is more complicated in the multivariate case. Here, we employ the same method Dang et al ([2]) used in their research. First, for n observations and p dependent variables, we choose an arbitrary combination of m distinct observations, where $p+1 \leq m \leq n$. Let k_m denote the m-subset of $\{1, ..., n\}$ and let X_k denote the $m \times (p+1)$ matrix which contains the m chosen observations so that $X_k^T X_k$ is invertible. We construct a LS estimator $\hat{\beta}_k = (X_k^T X_k)^{-1} X_k^T Y_k$ based on the m observations. Then, the multivariate Theil-Sen estimator $\hat{\beta}_{TS}$ of the parameter β is the multivariate median (Mmed) of all $\binom{n}{m}$ least squares estimators:

$$\hat{\beta}_{TS} = \text{Mmed}\{\hat{\beta}_k : \forall k_m\}$$

Now, we need to define the multivariate median as there are many existing notions. As Dang et al ([2]) point out, some methods, such as the componentwise median, may perform very poorly. They use the spatial median because it represents the true "center" if the data set has one. Spatial median also has relatively small computational burden in some cases. Hence, we also decide to implement the spatial median to find $\hat{\beta}_{TS}$. In d dimensions, the spatial median of a data set $(z_1, z_2, ..., z_n)$ of size n is achieved by

$$\min_{M} \sum_{i=1}^{n} \|z_i - M\| \tag{2}$$

where z_i is a vector in d dimensions and the solution of the minimization problem (2) is the spatial median, and it can be found by solving the following equation:

$$\sum_{i=1}^{n} \frac{z_i - M}{\|z_i - M\|} = 0$$

In our TS estimator with dimension $p+1$ and size $\binom{n}{m}$, we use a modified Weiszfeld Algorithm to compute the spatial median, proposed in 2000 by Vardi and Zhang ([11]).

At this point, we will take a look at the breakdown point in the multivariate case. As previously discussed, the TS estimator is an attractive model because of its high measure of robustness. In a multivariate TS estimator for p dimensions and a sample size n where the multivariate median is taken from $\binom{n}{m}$ Least Squares estimators with $p+1 \leq m \leq n$, at least 50% of the LS estimators must contain all "good" data. If we let ε represent the portion of "bad" observations, then $\frac{\binom{\lceil n(1-\varepsilon)\rceil}{m}}{\binom{n}{m}} \geq \frac{1}{2}$. As shown in Dang et al ([2]), the TSE only breaks down when $\varepsilon \geq (\frac{1}{2})^{\frac{1}{m}}(\frac{n-m+1}{n})$. The asymptotic breakdown point has been shown to be $1 - \frac{1}{2}^{\frac{1}{m}}$.

In the case where $p = 1$ and $m = 2$, as in a simple regression, if $\varepsilon \leq 1 - (\frac{1}{2})^{\frac{1}{2}} \approx 0.293$, the TS estimator can still resist large influence of outliers. Thus the breakdown point of the TS estimator is 29.3% as we discussed in the previous section. We can see that as m increases, $1-(\frac{1}{2})^{\frac{1}{m}}$ becomes smaller and smaller. Therefore, when we choose m between $p+1$ and n for the multivariate TSE, we can select the compromise of robustness and efficiency of our model. If we select $m = p+1$, we have the highest possible measure of robustness, but our estimator will be less efficient. If we let m be the sample size n, we can have the highest measure of efficiency since this is equivalent to using the least squares estimator. However, our estimator will lose robustness. Also, either of the extremes, $p+1$ or n, provides relatively low computation burden. On the other hand, if we let $m = \lceil \frac{n}{2} \rceil$, we have a compromise of robustness and efficiency, yet a high computation burden due to performing $\binom{n}{\lceil \frac{n}{2} \rceil}$ Least Squares estimations.

Since we want to demonstrate that the TS estimator is also efficient in multiple linear regressions, we let $m = p+1$ in our simulations. Thus, the efficiency of the TS estimator obtained under our set up should be generally at its lowest. We want to show that even with $m = p+1$, the TS estimator is still efficient enough. Also, in order to keep the computation simple, we will only consider two dependent variables ($p = 2$). Thus, $m = 3$. So, the breakdown point is $1 - (\frac{1}{2})^{\frac{1}{3}} \approx 0.2063$, which is still very robust compared to the 0 breakdown point of the LS estimator.

We choose the following model for all simulations in this section:

$$y_i = 1 + 5x_1 + 10x_2 + e_i, \quad i = 1, ..., n$$

Thus, we have the parameters $\alpha = 1$, $\beta_1 = 5$, and $\beta_2 = 10$. We randomly generate n different x_1's from a standard normal distibution and n different x_2's from a uniform distribution on $(0, 1)$. As in our simple linear regression, each time we generate five sets of n observations with $e_i \sim N(0, 3)$, $e_i \sim$ Unif$(-3, 3)$, $e_i \sim$ Bin$(12, 0.5)$, $e_i \sim$ T3, and $e_i \sim$ Cauchy respectively. For each set, we first compute the LS estimator $\hat{\beta}_{LS} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2)^T$ by $(X^TX)^{-1}X^T\underset{\sim}{Y}$. To calculate the TS estimator $\hat{\beta}_{TS} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2)^T$, we choose any 3 of the n observations and solve for the LS estimator, $\hat{\beta}_k$. Then we find the spatial median of the $\binom{n}{3}$ $\hat{\beta}_k$'s.

We have two ways to evaluate multivariate estimators. One method is to consider each parameter individually and compare the relative efficiency. Our results of this comparison for $N = 20$, 50, and 100 repetitions are shown in Table 5. For a continuous light-tailed distribution of error terms such as the normal and uniform distributions, we still see that the Least Squares method again yields more efficient results. For a discrete error distribution, we see that Theil-Sen improves over the LSE as the sample size increases. For a T3 distribution, the Theil-Sen estimator is consistently more efficient. In the case of the Cauchy distribution, we see the Theil-Sen estimator return very reasonable variances whereas the LSE again gives us nonsensical values, sending the relative efficiencies to the hundreds and thousands.

Another method involves comparing the efficiency for the parameters of the LS and TS estimators by the covariance matrices constructed from our simulations. Each estimator's matrix is set up in the following way:

$$\text{Cov}(\hat{\beta}) = \begin{bmatrix} \text{Var}(\hat{\alpha}) & \text{Cov}(\hat{\alpha}, \hat{\beta}_1) & \text{Cov}(\hat{\alpha}, \hat{\beta}_2) \\ \text{Cov}(\hat{\alpha}, \hat{\beta}_1) & \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\alpha}, \hat{\beta}_2) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) \end{bmatrix}$$

We found the relative efficiency of our multivariate estimators using the method described by Serfling ([8]), which is calculated as:

$$\text{RE}(\hat{\beta}_{TS}, \hat{\beta}_{LS}) = \frac{(\det(\widehat{\text{Cov}}(\hat{\beta}_{LS})))^{\frac{1}{p+1}}}{(\det(\widehat{\text{Cov}}(\hat{\beta}_{TS})))^{\frac{1}{p+1}}}$$

This gives us a measure of comparison to indicate which estimator performed more efficiently. Values less than 1 indicate the Least Squares method was better, and values more than 1 show when the Theil-Sen estimator had the lower variance. As described by Serfling ([8]), this method also gives us a measure of the asymptotic relative efficiency and thus, gives us an estimate of the ratio of sample sizes required for our two estimators to have equivalent performance. Our results for the relative efficiency test are shown in Table 6.

As in our simple linear model, we also demonstrate that our simulations are fair by comparing the theoretical Least Squares covariance matrix with that of the LS estimator. We construct the theoretical covariance matrix as $\text{Cov}(\hat{\beta})_{LS} = (X^T X)^{-1}\sigma^2$, and the relative efficiency compared to the LS estimator is also shown in Table 6. Values close to 1 demonstrate the fairness of our simulation. This method has an advantage over our previous relative comparison since it allows us to compare the entire vector of parameters together instead of individually. For the five different error distributions, we see that this comparison method reflects the same results we have already discussed for the multivariate TSE and LSE.

To reduce the computational burden and enhance the speed of simulation, we also employed a random stochastic procedure for calculating the TS estimator. Under this procedure, instead of going through all $\binom{n}{3}$ possible k_3's, we randomly pick t of the k_3's and solve for the $\hat{\beta}_k$'s. Then we find the spatial median of the t $\hat{\beta}_k$'s. Naturally, the larger the t, the closer the results are to the actual outcome for $m = 3$. In our simulations, we set $t = 2000$ which is about 1.24% of $\binom{100}{3}$ when $n = 100$. Therefore, the procedure is much faster and allows us to compute data for larger sample sizes. Our results for the random stochastic procedure are shown in Table 7.

Our results with a random stochastic procedure prove to be very similar to our other multivariate results. We still see the Least Squares method performs best with normal and uniform error

Error Term Distribution		$N = 20$		$N = 50$		$N = 100$	
		Variance	RE	Variance	RE	Variance	RE
Unif$(-3, 3)$	α	LS 0.7079 TS 2.1321	0.3320	LS 0.2574 TS 0.6731	0.3824	LS 0.1204 TS 0.3302	0.3646
	β_1	LS 0.1991 TS 0.6557	0.3036	LS 0.0374 TS 0.0923	0.4052	LS 0.0231 TS 0.0843	0.2740
	β_2	LS 2.2780 TS 5.6973	0.3998	LS 0.6930 TS 1.5809	0.4384	LS 0.3492 TS 0.7850	0.4449
$N(0, 3)$	α	LS 0.2880 TS 0.4246	0.6783	LS 0.0587 TS 0.1281	0.4582	LS 0.0379 TS 0.0589	0.6435
	β_1	LS 0.0716 TS 0.1383	0.5177	LS 0.0245 TS 0.0390	0.6282	LS 0.0109 TS 0.0170	0.6412
	β_2	LS 0.7641 TS 1.5608	0.4896	LS 0.1975 TS 0.3226	0.6122	LS 0.0894 TS 0.1463	0.6111
T3	α	LS 0.9561 TS 0.6647	1.4384	LS 0.2565 TS 0.2391	1.0728	LS 0.1608 TS 0.0813	1.9979
	β_1	LS 0.2779 TS 0.1476	1.8828	LS 0.0484 TS 0.0375	1.2907	LS 0.0492 TS 0.0251	1.9602
	β_2	LS 2.8645 TS 2.1511	1.3316	LS 0.8928 TS 0.7708	1.1583	LS 0.6753 TS 0.2409	2.8032
Cauchy	α	LS 1956.38 TS 2.0611	949.19	LS 892.97 TS 0.2841	3143.15	LS 38.36 TS 0.0641	598.44
	β_1	LS 1318.60 TS 0.3487	3781.47	LS 407.59 TS 0.0738	5522.90	LS 50.70 TS 0.0261	1942.53
	β_2	LS 4343.27 TS 5.3516	811.58	LS 12375.81 TS 0.7653	16171.19	LS 152.97 TS 0.2447	625.13
Bin$(12, 0.5)$	α	LS 0.4768 TS 1.4418	0.3307	LS 0.2108 TS 0.3170	0.6650	LS 0.1378 TS 0.1340	1.0284
	β_1	LS 0.2064 TS 0.3434	0.6010	LS 0.0701 TS 0.0357	1.9636	LS 0.0279 TS 0.0125	2.2320
	β_2	LS 1.8691 TS 4.8340	0.3867	LS 0.5737 TS 0.5434	1.0558	LS 0.3543 TS 0.1831	1.9350

Table 5: Relative Efficiency of Each Parameter in Mutivariate TSE to LSE

Covariance	Error	N		
Matrices	Distribution	20	50	100
Theoretical vs. LS	$e \sim N(0, 3)$	1.0723	0.9614	0.9183
TS vs. LS	$e \sim \text{Unif}(-3, 3)$	0.3002	0.3717	0.2994
TS vs. LS	$e \sim N(0, 3)$	0.4606	0.5435	0.6091
TS vs. LS	$e \sim$ T3	1.2573	1.0912	2.0139
TS vs. LS	$e \sim$ Cauchy	765.33	3893.82	632.91
TS vs. LS	$e \sim \text{Bin}(12, 0.5)$	0.4545	0.9933	1.1972

Table 6: Relative Efficiency of the Covariance Matrix Comparison

distributions while the TS estimator is dominant for the Cauchy distribution and typically better for the T3 distribution. For the binomial distribution, we again see the Theil-Sen estimator surpass the LSE as the sample size increases. The results from the covariance matrix comparisons of our data calculated using the random stochastic procedure are shown in Table 8.

Error Term Distribution		$N = 50$		$N = 100$		$N = 200$	
		Variance	RE	Variance	RE	Variance	RE
Unif$(-3, 3)$	α	LS 0.2603 TS 0.7691	0.3384	LS 0.0721 TS 0.3014	0.2392	LS 0.0830 TS 0.3335	0.2489
	β_1	LS 0.0599 TS 0.1618	0.3702	LS 0.0283 TS 0.0877	0.3227	LS 0.0140 TS 0.0318	0.4403
	β_2	LS 0.8127 TS 2.0134	0.4036	LS 0.2641 TS 0.8795	0.3003	LS 0.2161 TS 0.6431	0.3360
$N(0, 3)$	α	LS 0.0851 TS 0.1691	0.5033	LS 0.0384 TS 0.1007	0.3813	LS 0.0227 TS 0.0610	0.3721
	β_1	LS 0.0262 TS 0.0554	0.4729	LS 0.0113 TS 0.0223	0.5067	LS 0.0057 TS 0.0114	0.5000
	β_2	LS 0.2470 TS 0.3669	0.6732	LS 0.1266 TS 0.2570	0.4926	LS 0.0683 TS 0.1961	0.3483
T3	α	LS 0.2693 TS 0.2532	1.0636	LS 0.1330 TS 0.1001	1.3287	LS 0.0580 TS 0.0615	0.9431
	β_1	LS 0.0488 TS 0.0648	0.7531	LS 0.0319 TS 0.0276	1.1558	LS 0.0143 TS 0.0183	0.7814
	β_2	LS 0.8380 TS 0.8547	0.9805	LS 0.3444 TS 0.2524	1.3645	LS 0.2181 TS 0.1606	1.3580
Cauchy	α	LS 112.36 TS 0.2625	428.04	LS 15.18 TS 0.1901	79.85	LS 1340.36 TS 0.0797	16817.57
	β_1	LS 55.48 TS 0.0633	876.46	LS 17.82 TS 0.0353	504.82	LS 1593.27 TS 0.0129	123509.30
	β_2	LS 561.32 TS 0.7650	733.75	LS 56.43 TS 0.6528	86.44	LS 22631.55 TS 0.2780	81408.45
Bin$(12, 0.5)$	α	LS 0.2955 TS 0.5279	0.5598	LS 0.1311 TS 0.1567	0.8366	LS 0.0594 TS 0.0612	0.9706
	β_1	LS 0.1044 TS 0.1537	0.6792	LS 0.0300 TS 0.0356	0.8427	LS 0.0132 TS 0.0050	2.6400
	β_2	LS 0.8948 TS 1.5737	0.5686	LS 0.4342 TS 0.2851	1.5230	LS 0.1430 TS 0.0072	19.8611

Table 7: Relative Efficiency of Each Parameter in Mutivariate TSE to LSE with Random Stochastic Procedure

4 Conclusion

Although the Least Squares estimator is the most commonly used method of statistical regression, it is known to have its disadvantages. As we demonstrated in section 2, the Least Squares lacks robustness to outliers. It only takes one outlier in a sample to have an arbitrarily large effect on

Covariance	Error	N		
Matrices	Distribution	50	100	200
Theoretical vs. LS	$e \sim N(0,3)$	0.9347	1.0113	1.0988
TS vs. LS	$e \sim \text{Unif}(-3,3)$	0.5438	0.3515	0.5347
TS vs. LS	$e \sim N(0,3)$	0.4185	0.4465	0.4103
TS vs. LS	$e \sim$ T3	0.9651	1.1502	1.0473
TS vs. LS	$e \sim$ Cauchy	518.03	168.90	10939.08
TS vs. LS	$e \sim \text{Bin}(12,0.5)$	0.8032	1.1833	4.1162

Table 8: Relative Efficiency of the Covariance Matrix Comparison for Random Stochastic Procedure

the estimator and to destroy the entire regression line. Hence, the breakdown point of the Least Squares is said to be 0%. The Theil-Sen estimator has been shown to be a much more robust method of regression with a breakdown point of 29.3%.

While the Least Squares estimator is the most efficient method for regressing data with a normal distribution of error terms, it is known to lack efficiency in other situations. In our study, we have compared the efficiency of the Least Squares method with that of the Theil-Sen estimator under various circumstances in both linear and multivariate regressions. As expected, we have found the Least Squares method to perform better than the Theil-Sen estimator in light-tailed error term distributions such as the uniform and normal ones. However, our results in the simple linear model show that the Theil-Sen estimator is not far behind, typically having a relative efficiency of at least 90%.

In error term distributions that are heavy-tailed, we see the Theil-Sen estimator outperform the Least Squares method in both the linear and multivariate models. For a T3 distribution, the Theil-Sen estimate gives us a variance that is about one half of the Least Squares estimate on average. In the case of the Cauchy distribution, the performance of the Least Squares estimator is poor while Theil-Sen returns reasonably small variances, and so our relative effiency is thrown into the thousands. In a binomial distribution of error terms, we see the lowest variances for the Theil-Sen estimator among all of our cases. This demonstrates the super efficiency of the Theil-Sen estimator for discrete distributions.

In our results for the multivariate model, we have designed our Theil-Sen estimator such that theoretically, the robustness is at its highest and the efficiency is at its lowest. For heavy-tailed distributions, we have demonstrated the dominance of the Theil-Sen estimator in both robustness and efficiency. For light-tailed distributions, the Least Squares method is the more efficient method of regression. However, if these data sets contain outliers, the Theil-Sen estimator is more resistent to influence, and the gain in robustness comes with a relatively small loss of efficiency in most cases but with a cost in computational burden. We have also investigated the use of a random stochastic procedure to alleviate those burdens. In conclusion, we have found the Theil-Sen estimator to be a reliable method of robust regression.

References

[1] Akritas, M., Murphy, S., & LaValley, M. (1995, March). The Theil-Sen Estimator with Doubly Censored Data and Applications to Astronomy. *Journal of the American Statistical Association*, 90, 170-177.

[2] Dang, X., Peng, H., Wang, X., & Zhang, H. (2008). The Theil-Sen Estimators in a Multiple Linear Regression Model. (submitted)

[3] Dietz, E. J. (1989). Teaching Regression in a Nonparametric Statistics Course. *The American Statistician*, 43, 35-40.

[4] Fernandes, R. & Leblanc, S. (2005). Parametric (Modified Least Squares) and Non-parametric (Theil-Sen) Linear Regressions for Predicting Biophysical Parameters in the Presence of Measurement Errors. *Remote Sensing of Environment*, 95, 303-316.

[5] Hollander, M. & Wolfe, D. (1999). Nonparametric Statistical Methods. New York: John Wiley & Sons.

[6] Leroy, A. & Rousseeuw, P. (2003). Robust Regression and Outlier Detection. New York: John Wiley & Sons.

[7] Peng, H., Wang, S., & Wang, X. (2007) Consistency and Asymptotic Distribution of the Theil-Sen Estimator. *Journal of Statistical Planning and Inference*, In press (www.sciencedirect.com).

[8] Serfling, R. (1980). Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons.

[9] Serfling, R. (1984). Generalized $L-$, $M-$, and $R-$ Statistics. *Annals of Statistics*, 1, 76-86.

[10] Sprent, P. (1993). Applied Nonparametric Statistical Methods. Chapman & Hall.

[11] Vardi, Y. and Zhang, C. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97, no. 4, 1423-1426.

[12] Wang. X. (2005). Asymptotics of the Theil-Sen Estimator in a Simple Linear Regression Model with a Random Covariate. *Journal of Nonparametric Statistics*, 17, 107-120.

[13] Wilcox, R. (1998). Simulations on the Theil-Sen Regression Estimator with Right-Censored Data. *Statistics & Probability Letters*, 39, 43-47.

[14] Wilcox, R. (2004). Some Results on Extensions and Modifications of the Theil-Sen Regression Estimator. *British Journal of Mathematical and Statistcal Psychology*, 57, 265-280.