ESTIMATION OF BIVARIATE REGRESSION DATA VIA THEIL’S ALGORITHM

Ekezie Dan Dan and Opara, Jude
Department of Statistics, Imo State University
PMB 2000, Owerri, Nigeria.

Corresponding Author: Ekezie Dan Dan

Abstract
This paper is on the estimation of bivariate regression data using Theil’s algorithm. This method was adopted since all errors in the y-direction are not normally distributed (i.e., they do not follow a Gaussian distribution) for the commonly used least squares regression method for fitting an equation into a set of (x, y)-data points using the Kolmogorov Smirnov test. The algorithms for Theils were stated in this paper. The data used for this research were collected from selected primary schools in Owerri Municipal, Imo State Nigeria. The data were on weights and shoulder heights of 100 randomly selected pupils in primary four, five and six. The use of a programming language software known as “R Development” was used to write an appropriate program in this paper. From the analysis, the result revealed that there exists a significant relationship between weights and shoulder heights of primary school pupils, and the estimated fitted theil’s is \( \hat{y}_i = 42.5833 + 0.1177 \hat{x}_i \) and it was observed that both the intercept and slope are significant.

Keywords: theil’s regression, weighted theil’s regression, weighted median, OLS, pairwise slope

INTRODUCTION

ORDINARY LEAST SQUARE ESTIMATION
One of the most popular methods to model the function relationship between variables is the Ordinary Least Squares (OLS) estimation procedure which is very simple and straightforward to apply. However, for OLS estimators to be ideal some conditions are needed. One of these conditions is that the error terms \( \varepsilon_i \) are assumed to be independently, identically distributed (iid) random variables with mean zero and a constant variance \( \sigma^2 \).

The simple linear regression model is the traditional equation representing the relationship between two variables; the dependent and the independent variables. What is of interest is the estimation of the parameters of the model? The most popular method is the Least Square Method (LSM). But when the data fail to fulfill the assumptions such as the normality, then this parametric method of estimation of the parameters fails to give a valid estimate. In the alternative, a non parametric method becomes very effective. Non-parametric (or distribution-free) statistical methods are those, which make no assumptions about the population distribution from which the data are taken.

Suppose that the distribution of the errors is not normal. If the errors are coming from a population that has a mean of zero, then the OLS estimates may not be optimal, but they at least have the property of being unbiased. If we further assume that the variance of the error population is finite, then the OLS estimates have the property of being consistent and asymptotically normal. However, under these conditions, the OLS estimates and tests may lose much of their efficiency and they can result in poor performance (Mutan; 2004). To deal with these situations, two approaches can be applied. One is to try to correct non-normality, if non-normality is determined and the other is to use alternative regression methods, which do not depend on the assumption of the normality (Birkes and Dodge; 1993). In this paper, the researchers were motivated to adopt this algorithm because all errors in the y-direction of the parametric equivalent are not normally distributed; hence we employed the non-parametric Theil’s regression method.

In a simple linear model, Theil (1950) proposed the median of pairwise slopes as an estimator of the slope parameter. Sen (1968) extended this estimator to handle ties. The Theil-Sen estimator (TSE) is robust with a high breakdown point 29.3%, has a bounded influence function, and possesses a high asymptotic efficiency. Thus it is very competitive to other slope estimators (e.g., the least squares estimators), see Sen (1968), Dietz (1989) and Wilcox (1998). The TSE has been acknowledged in several popular textbooks on nonparametric and robust statistics, e.g., Sprent (1993), and Hollander, and Rousseuw and Leroy (1986).

The proposed estimators contain an integer variable which controls the amount of robustness and efficiency. The maximal possible robustness (in terms of break-down point) is attained when the integer variable is chosen to be the number of the parameters to be estimated; while the maximal efficiency is achieved when the variable assumes the sample size; any value of the variable taking in between results in an estimator which gives a compromise between robustness and efficiency.

In straight-line regression, the least squares estimator of the slope is sensitive to outliers and the associated confidence interval is affected by non-normality of the dependent variable. A simple and robust alternative to least squares regression is Theil regression, first proposed by Theil (1950). Theil’s method actually yields an estimate of the slope of the regression line. Several approaches exist for obtaining a nonparametric estimate of the intercept. In this paper, we shall write a suitable program for estimating the parameters using the R
Software. This paper shall be of paramount significant to future researchers who may wish to carry out a similar research, knowing when and how to use the parametric and non-parametric methods.

REVIEW OF RELATED LITERATURES

Many research works have been investigated, which in one way or the other relates to this research investigation. Hence, there is need to review some past researches.

Mutans (2004) carried out a research on comparison of regression techniques via Monte Carlo Simulation. In the study, the simple linear regression model was investigated for conditions in which the distribution of the error term was Generalised Logistic. Some robust and nonparametric methods such as Modified Maximum Likelihood (MML), Least Absolute Deviations (LAD), Winsorized Least Squares, Least Trimmed Squares (LTS), Theil and Weighted Theil were compared via computer simulation. In order to evaluate the estimator, performance, the researcher computed the mean, variance, bias, Mean Square Error (MSE) and Relative Mean Square Error (RMSE).

Hekimoğlu et al. (2009) carried out a research on outlier detection by means of robust regression estimators for use in engineering science. The study compared the ability of different robust regression estimators to detect and classify outliers. Well-known estimators with high breakdown points were compared using simulated data. Mean Success Rates (MSR) were computed and used as comparison criteria. The results showed that the Least Median of Squares (LMS) and Least Trimmed Square (LTS) were the most successful methods for data that included leverage point, masking and swamping effects or critical and concentrated outliers.

Nevitt and Tam (1998) in their study titled “A Comparison of Robust and Nonparametric Estimators under the Simple Linear Regression Model” investigated parameter estimation under the simple linear regression model for situations in which the underlying assumptions of ordinary least squares (OLS) estimation are untenable. Classical nonparametric estimation methods were directly compared against some robust estimation methods for conditions in which varying degrees of outliers were present in the observed data. Additionally, estimator performance was considered under conditions in which the normality assumption regarding error distributions was violated. The study addresses the problem via computer simulation methods. The study design includes three sample sizes (n = 10, 30, 50) crossed with five types of error distributions (unit normal, 10% contaminated normal, 30% contaminated normal, lognormal, t-5(df)). Variance, bias, mean square error, and relative mean square error are used to evaluate estimator performance. Recommendations to applied researchers and direction for further study were considered.

Dang et al. (2002) researched on Theil-Sen estimators in a multiple linear regression model. In their article, they proposed the Theil-Sen estimators of parameters in a multiple linear regression model based on a multivariate median, generalizing the Theil-Sen estimator in a simple linear regression model. The proposed estimator was shown to be robust, consistent and asymptotically normal under mild conditions, and super-efficient when the error distribution is discontinuous. According to their result, it can be chosen to satisfy the pre-specified possible robustness and efficiency. Simulations were conducted to compare robustness and efficiency with least squares estimators and to validate super-efficiency.

Castillo et. al. (2001) investigates the influence of observations on LAD estimates and they recommend methods for regression diagnostics which are illustrated by numeric and real life examples. Mathew and Nordstrom (1993) propose to estimate the regression coefficients by minimizing the maximum of a weighted sum of squared deviations, or the sum of absolute deviations. Their aim is to establish robustness property of the LAD criterion. In another study, performed by Lawrence and Shier (1981), the methods of OLS and LAD are compared for estimating the Weibull parameters. Brown (1980) concentrates on the case of simple linear regression for LAD estimates and shows how this method works in small or moderate samples.

Theil’s regression method is a nonparametric procedure which is expected to perform well without regard to the distribution of the error terms. This procedure is based on ranks and uses the median as robust measures rather than using the mean as in OLS. To calculate the slope of a line that fits the data points, the slopes of all pairs of data points are computed and the median of all these pairwise slopes is expressed as the Theil’s slope estimator, $\hat{\theta}_{(THL)}$. The median of the $y_i - \hat{\theta}_{(THL)} x_i$ terms is the estimator for the y-intercept of the regression line passing through all n observations (Birkes and Dodge, 1993; Nevitt and Tam, 1998).

Weighted Theil’s regression method is a modified version of the Theil’s original method to calculate the slope of the regression line. Nevitt and Tam (1998) say that in this method, each of the pairwise slopes is weighted using a weighting scheme, $w_{ij}$, where $w_{ij}$ is equal to $x_j - x_i$ or $x_i - x_j$. The median of these weighted pairwise slopes is then called as the weighted Theil’s slope estimator. Also, the y-intercept estimator is calculated in a similar fashion as it is done in Theil’s original method, but using $\hat{\theta}_{(wtd.THL)}$ instead of $\hat{\theta}_{(THL)}$. In this paper, we shall examine the estimate of bivariate regression date using the Theil’s algorithm.

THEIL’S AND WEIGHTED THEIL’S REGRESSION

Theil’s regression is a nonparametric method which is used as an alternative to robust methods for data sets with outliers. Although the nonparametric procedures perform reasonably well for almost any possible distribution of errors and they lead to robust regression lines, they require a lot of computation. This method is suggested by
Theil (1950), and it is proved to be useful when outliers are suspected, but when there are more than few variables, the application becomes difficult.

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

Hussain and Sprent (1983) say that no generality is lost if we take \(1 \leq i < j \leq n\) assuming that the \(x_i\) are arranged in ascending order. Note that \(b_{ij} = b_{ji}\). According to these results the Theil’s slope estimator is\[
\hat{\theta}_{\text{THL}} = \text{med}\{b_{ij} | x_j \neq x_i\},
\]
where \(x_1 \leq x_2 \leq \ldots \leq x_n\).

It is known that median estimators are less affected compared to the mean estimators. Therefore, these estimators are resistant to outliers in the sample data.

Nevitt and Tam (1998) state that there are several methods for computing the \(y\)-intercept. One of these methods is to calculate
\[
a_{ij} = \frac{x_j y_i - x_i y_j}{x_j - x_i}, \quad i < j, x_i \neq x_j
\]
(2)
and taking the median of these \(a_{ij}\) values will give us the \(y\)-intercept.

Nevitt and Tam (1998) also investigate a different approach which is proposed by Theil (1950). For this approach, these \(a_{ij}\) values need not to be explicitly calculated. For a \(\hat{\theta}_{\text{THL}}\) slope estimator, the \(y_i - \hat{\theta}_{\text{THL}} x_i\) is computed for each observation, and the median of these terms are expressed as the \(y\)-intercept of the regression line.

To reduce the effect of outlying observations, some modifications are applied to Theil’s method and each of the pairwise slopes, \(b_{ij}\)’s, are weighted by some weighting procedures. The weighted Theil slope estimator for the \(n\) observations in the sample data is the weighted median of these \(b_{ij}\)’s. Note that
\[
\hat{\theta}_{\text{THL}} = \text{med} w_{ij} b_{ij}
\]
where
\[
w_{ij} = \frac{(x_i - x_j)^2}{\sum (x_i - x_j)^2}
\]
(3)
\(\Sigma\) represents \(n(n-1)/2\) pairs of integers \(i\) and \(j\) with \(1 \leq i < j \leq n\).

Birkes and Dodge (1993) explain how a weighted median can be calculated as follows:

First \(x_i\)’s are ordered in an increasing sequence, so that \(x_1 \leq x_2 \leq \ldots \leq x_n\). Note that the weights, \(w_i\)’s, are nonnegative and add up to 1. Obtaining the index \(k\) where
\[
w_1 + w_2 + \ldots + w_{k-1} < 0.5
\]
\[
w_1 + w_2 + \ldots + w_{k-1} + w_k > 0.5
\]
(4)
x_i is the weighted median. Note that if \(w_1 + w_2 + \ldots + w_{k-1} = 0.5\), then the weighted median is \((x_{k-1} + x_k)/2\). If the weights are exactly equal to each other (i.e. \(w_i = 1/n\)), the weighted median will be the ordinary median.

The weighted Theil slope estimator of \(\theta_1\) is the pairwise slopes \(b_{ij} = (y_i - y_j)/(x_i - x_j)\), with weights
\[
w_{ij} = |x_i - x_j|/\sum |x_i - x_i|.
\]
and \(\hat{\theta}_{\text{THL}}\) is the ordinary median of
\[
y_i - \hat{\theta}_{\text{THL}} x_i.
\]

The parameters of Theil regression can also be estimated based on Graybill and Iyer (1994). Let us consider \(\alpha\) and \(\beta\) to indicate the intercept and slope of the true regression line.

**ESTIMATE OF THE SLOPE**

Let \((x_i, y_i), i = 1, \ldots, n\), denote the data values. Without loss of generality, assume that \(x_1 \leq x_2 \leq \ldots \leq x_n\) and \(x_1 < x_n\) (i.e., there are at least two distinct \(x\) values in the dataset).

Then let
\[
N = \sum_{1 \leq i < j \leq n} \text{sign}(x_j - x_i),
\]
where
\[
\text{sign}(x) = \begin{cases} 
+1 & \text{if } x > 0 \\
0 & \text{if } x = 0 \\
-1 & \text{if } x < 0 
\end{cases}
\]
So \(N\) is the number of positive differences \(x_j - x_i\). Now consider the set \(S\) of \(N\) distinct pairs \((i, j)\) for which \(x_j > x_i\) and define
\[
S_{ij} = \frac{(y_j - y_i)}{(x_j - x_i)}, (i, j) \in S.
\]
(5)

Thus, the \(S_{ij}\)’s are the slopes of the line segments connecting pairs of points \((x_i, y_i)\) and \((x_j, y_j)\) where \(x_i \neq x_j\). Arrange the \(N\) quantities in equation (5) in ascending order and denote the \(r\)th smallest value of \(S_{ij}\) by \(S(r)\) for \(r = 1, 2, 3, \ldots, N\). We write \(2M\) if \(N\) is even or \(N = 2M + 1\) if \(N\) is odd. Then the estimate of the slope is given by
In other words, the estimate of the slope is given by the median of the pairwise slopes.

**P-VALUE FOR THE SLOPE**

If \( n > 10 \), then the p-value for the slope can be computed by normal approximation or simulation. An approximate P-value for the slope can be obtained by using the normal approximation. If

\[
\beta = \frac{1}{2} \left( S_{(M)} + S_{(M+1)} \right),
\]

\( N = 2M \quad (6) \)

Then the p-value for the slope can be approximated by

\[
Z = \frac{U'(0)}{\sqrt{\text{Var}(U'(0))}},
\]

\( \text{where} \)

\[
U'(0) = \left( \begin{array}{c} n \\ 2 \end{array} \right)^{1/2} \sum_{i<j<n} \text{sign}(x_j-x_i)\text{sign}(y_j-y_i)
\]

Then the p-value for the slope can be approximated by

\[
2\left(1 - \text{prob}(Z \leq Z)\right),
\]

where \( Z \) is a Normal random variable with mean 0 and variance 1. The estimate will only be used if \( n > 10 \).

**ESTIMATE OF THE INTERCEPT**

The estimate of the intercept \( \hat{\alpha} \) is computed using the method described by Graybill and Iyer (1994). The procedure is as follows.

If several y values are available for a given x value, we let \( y_i \) be their mean so we can assume that the \( x_i \)'s are distinct. Arrange the \( n \) (distinct) observation in ascending order of \( x \). If \( n \) is an odd number, say \( n=2m+1 \), then discard the middle observation so there are \( 2m \) observations. Of course if \( n \) is even, no observation is discarded and \( n = 2m \). The observations are arranged in Table 1. From Table 1, compute the quantities \( z,w,u,v \) and \( t \) which are shown in Table 2.

### Table 1

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td>( x_1 )</td>
<td>( y_{m+1} )</td>
<td>( x_{m+1} )</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>( x_2 )</td>
<td>( y_{m+2} )</td>
<td>( x_{m+2} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( y_m )</td>
<td>( x_m )</td>
<td>( y_{2m} )</td>
<td>( x_{2m} )</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>( z )</th>
<th>( w )</th>
<th>( u )</th>
<th>( v )</th>
<th>( t )</th>
</tr>
</thead>
</table>
| \( z_1 = y_{m+1} \) | \( w_1 = x_{m+1} \) | \( u_1 = y_1 \) | \( v_1 = y_m \) | \( t_1 = u_1 - 
| \( z_2 = y_{m+2} \) | \( w_2 = x_{m+2} \) | \( u_2 = y_2 \) | \( v_1 = y_m \) | \( t_2 = u_2 - 

Compute \( q_i^* \), where \( q_i^* = \frac{t_i}{w_i} \) for \( i = 1, 2, \ldots, m \).

### Table 3

| \( Z_m = y_{2m} \) | \( w_m = x_{2m} \) | \( u_m = y_4 \) | \( v_m = y_2 \) | \( t_m = u_m \) |

### DATA PRESENTATION

The data used for this research were gotten from selected primary schools in Owerri Municipal Council of Imo State Nigeria, they are on weights and shoulder heights of 100 randomly selected pupils in primary four, five and six, and it is presented in Table 3. The work is limited to bivariate data using only the non-parametric Theil’s regression method. It covers the weights (the response variable) measured in kg and the shoulder heights (response variable) measured in centimeter of 100 randomly selected primary school pupils.
Table 3: Data on weights ($y_i$) in kg and shoulder heights ($x_i$) in centimeter of 100 primary school pupils, Imo State Nigeria

<table>
<thead>
<tr>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
<th></th>
<th>$y_i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>41.2</td>
<td>18</td>
<td>38</td>
<td>41.8</td>
<td>35</td>
<td>30</td>
<td>50</td>
<td>52</td>
<td>35</td>
<td>46.2</td>
<td>69</td>
<td>26</td>
<td>45.9</td>
<td>86</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>43.7</td>
<td>19</td>
<td>40</td>
<td>47</td>
<td>36</td>
<td>36</td>
<td>52.1</td>
<td>53</td>
<td>41</td>
<td>46</td>
<td>70</td>
<td>30</td>
<td>46.8</td>
<td>87</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>47.8</td>
<td>20</td>
<td>46</td>
<td>48.6</td>
<td>37</td>
<td>34</td>
<td>44.6</td>
<td>54</td>
<td>38</td>
<td>48.8</td>
<td>71</td>
<td>26</td>
<td>46</td>
<td>88</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>48.9</td>
<td>21</td>
<td>43</td>
<td>52.6</td>
<td>38</td>
<td>35</td>
<td>47.8</td>
<td>55</td>
<td>43</td>
<td>51.2</td>
<td>72</td>
<td>24</td>
<td>41.6</td>
<td>89</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>46.2</td>
<td>22</td>
<td>29</td>
<td>39.3</td>
<td>39</td>
<td>38</td>
<td>44</td>
<td>56</td>
<td>51</td>
<td>46.4</td>
<td>73</td>
<td>29</td>
<td>43.6</td>
<td>90</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>35</td>
<td>46.2</td>
<td>23</td>
<td>34</td>
<td>50.2</td>
<td>40</td>
<td>40</td>
<td>47.2</td>
<td>57</td>
<td>50</td>
<td>46.8</td>
<td>74</td>
<td>31</td>
<td>41.2</td>
<td>91</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>48.8</td>
<td>24</td>
<td>37</td>
<td>49.8</td>
<td>41</td>
<td>41</td>
<td>49.3</td>
<td>58</td>
<td>38</td>
<td>52.1</td>
<td>75</td>
<td>28</td>
<td>47.6</td>
<td>92</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>38</td>
<td>51.2</td>
<td>25</td>
<td>30</td>
<td>50.2</td>
<td>42</td>
<td>39</td>
<td>44.6</td>
<td>59</td>
<td>27</td>
<td>43.6</td>
<td>76</td>
<td>24</td>
<td>38.2</td>
<td>93</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>43</td>
<td>46.4</td>
<td>26</td>
<td>36</td>
<td>52.1</td>
<td>43</td>
<td>27</td>
<td>49.2</td>
<td>60</td>
<td>26</td>
<td>45.9</td>
<td>77</td>
<td>27</td>
<td>49.2</td>
<td>94</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>46.8</td>
<td>27</td>
<td>34</td>
<td>44.6</td>
<td>44</td>
<td>26</td>
<td>41.6</td>
<td>61</td>
<td>30</td>
<td>46.3</td>
<td>78</td>
<td>26</td>
<td>41.6</td>
<td>95</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>52.1</td>
<td>28</td>
<td>35</td>
<td>47.8</td>
<td>45</td>
<td>30</td>
<td>46</td>
<td>62</td>
<td>29</td>
<td>48.6</td>
<td>79</td>
<td>30</td>
<td>46</td>
<td>96</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>38</td>
<td>46.2</td>
<td>29</td>
<td>38</td>
<td>44</td>
<td>46</td>
<td>31</td>
<td>50.2</td>
<td>63</td>
<td>31</td>
<td>45.8</td>
<td>80</td>
<td>31</td>
<td>50.2</td>
<td>97</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>35</td>
<td>46</td>
<td>30</td>
<td>33</td>
<td>39.9</td>
<td>47</td>
<td>28</td>
<td>47.6</td>
<td>64</td>
<td>38</td>
<td>45.2</td>
<td>81</td>
<td>28</td>
<td>47.6</td>
<td>98</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>41</td>
<td>48.8</td>
<td>31</td>
<td>31</td>
<td>44.0</td>
<td>48</td>
<td>26</td>
<td>42.4</td>
<td>65</td>
<td>27</td>
<td>48.3</td>
<td>82</td>
<td>26</td>
<td>42.4</td>
<td>99</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>38</td>
<td>51.2</td>
<td>32</td>
<td>29</td>
<td>39.3</td>
<td>49</td>
<td>23</td>
<td>48.4</td>
<td>66</td>
<td>33</td>
<td>45.1</td>
<td>83</td>
<td>23</td>
<td>48.4</td>
<td>100</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>29</td>
<td>47.2</td>
<td>33</td>
<td>34</td>
<td>50.2</td>
<td>50</td>
<td>33</td>
<td>59.8</td>
<td>67</td>
<td>42</td>
<td>46.8</td>
<td>84</td>
<td>27</td>
<td>39.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>32</td>
<td>37.6</td>
<td>34</td>
<td>37</td>
<td>49.8</td>
<td>51</td>
<td>38</td>
<td>48.9</td>
<td>68</td>
<td>27</td>
<td>43.6</td>
<td>85</td>
<td>29</td>
<td>36.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We shall analyzed the data using the written program stated in this paper. The plot of these sample data are given in Figures 1 to 5.

**ANALYSIS USING THE R-PROGRAM FOR THEIL**

Implementing the written computer program for Theil’s regression in this paper, the output is presented below;

Output for the R-Computer Software

```
mblm(formula = y ~ x)
```

Residuals:
```
Min       1Q  Median       3Q  Max
-14.1615  -3.0562  -0.1146   2.5547  14.0385
```

Coefficients:
```
               Estimate MAD     V value     Pr(>|V|)
(Intercept)   42.5833 7.2647  5049          < 2e-16 ***
x                0.1177 0.2303  3590          1.17e-06 ***
```

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

Residual standard error: 4.452 on 98 degrees of freedom

> anova(dan)

Analysis of Variance Table

Response: y

```
Df  Sum Sq Mean Sq F value Pr(>F)
x     1   57.83  57.831   2.9182  0.09075 .
Residuals 98 1942.12  19.818
```

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

From the computer output, the fitted line is
```
\( \hat{y}_i = 42.5833 + 0.1177 \times_i \), i = 1, 2, ..., 100
```

The ANOVA Table from the output shows that there exists a significant relationship between weights and shoulder heights of the 100 selected pupils. The MSE of the Theil’s regression is 19.818.

The results from the output also revealed that both the slope and the intercept are significant, since their p-values are less than 0.05.
From the analysis, the result revealed that there exists a significant relationship between weights and shoulder heights of primary school pupils, and the estimated fitted Theil’s is \( \hat{y}_i = 42.5833 + 0.1177 x_i \), and it was observed that both the intercept and slope are significant.

REFERENCES


