December 2017

JMASM 46: Algorithm for Comparison of Robust Regression Methods In Multiple Linear Regression By Weighting Least Square Regression (SAS)

Mohamad Shafiq  
*Universiti Sains Malaysia, Kelantan, Malaysia, shafiqmat786@gmail.com*

Wan Muhamad Amir  
*Universiti Sains Malaysia, Kelantan, Malaysia, wmamir@usm.my*

Nur Syabiha Zafakali  
*Universiti Sains Malaysia, Kelantan, Malaysia, syabiha_89@yahoo.com*

Follow this and additional works at: [http://digitalcommons.wayne.edu/jmasm](http://digitalcommons.wayne.edu/jmasm)  
Part of the [Applied Statistics Commons](http://digitalcommons.wayne.edu/jmasm), [Social and Behavioral Sciences Commons](http://digitalcommons.wayne.edu/jmasm), and the [Statistical Theory Commons](http://digitalcommons.wayne.edu/jmasm)

**Recommended Citation**  

This Algorithms and Code is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.
JMASM 46: Algorithm for Comparison of Robust Regression Methods In Multiple Linear Regression By Weighting Least Square Regression (SAS)

Mohamad Shafiq
Universiti Sains Malaysia
Kelantan, Malaysia

Wan Muhamad Amir
Universiti Sains Malaysia
Kelantan, Malaysia

Nur Syabiha Zafakali
Universiti Sains Malaysia
Kelantan, Malaysia

The aim of this study is to compare different robust regression methods in three main models of multiple linear regression and weighting multiple linear regression. An algorithm for weighting multiple linear regression by standard deviation and variance for combining different robust method is given in SAS along with an application.

Keywords: Multiple linear regression, robust regression, M, LTS, S, MM estimation

Introduction

Multiple linear regression (MLR) is a statistical technique for modeling the relationship between one continuous dependent variable from two or more independent variables. A typical data template is compiled in Table 1.

Table 1. Data template for multiple linear regression

<table>
<thead>
<tr>
<th>i</th>
<th>yi</th>
<th>x0</th>
<th>x11</th>
<th>x12</th>
<th>...</th>
<th>xip</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y1</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
<td>...</td>
<td>xip</td>
</tr>
<tr>
<td>2</td>
<td>y2</td>
<td>1</td>
<td>x21</td>
<td>x22</td>
<td>...</td>
<td>xip</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>yn</td>
<td>1</td>
<td>x01</td>
<td>x02</td>
<td>...</td>
<td>xip</td>
</tr>
</tbody>
</table>

Sources: Ahmad et al., 2016a; 2016b

Mohamad Shafiq is a postgraduate student in the School of Dental Sciences. Email him at shafiqmat786@gmail.com. Dr. Amir is an Associate Professor of Biostatistics. Email him at wqamir@usm.my.
It is used when there are two or more independent variables and a single dependent variable where the equation below shows the model population information:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + L + \beta_k x_{ik} + \varepsilon_i \]  

(1)

where

- \( \beta_0 \) is the intercept parameter, and
- \( \beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1} \) are the parameters associated with \( k - 1 \) predictor variables.

The dependent variable \( Y \) is written as a function of \( k \) independent variables, \( x_1, x_2, \ldots, x_k \). A random error term is added to equation as to make the model more probabilistic rather than deterministic. The value of the coefficient \( \beta_i \) determines the contribution of the independent variables \( x_i \), and \( \beta_0 \) is the y-intercept (Ahmad et al., 2016a; 2016b). The coefficients \( \beta_0, \beta_1, \ldots, \beta_k \) are usually unknown because they represent population parameters. Below is the data presentation for multiple linear regression. A general linear model in matrix form can be defined by the following vectors and matrices as:

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix} = 
\begin{bmatrix}
1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\
1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1}
\end{bmatrix} \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_{p-1}
\end{bmatrix} + 
\begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\]

**Robust Regression**

Robust regression is a method used when the distribution of the residual is not normally distributed and there are some outliers which affect the model (Susanti et al., 2014). It detects the outliers and provides better results (Chen, 2002). A common method of robust regression is the M estimate, introduced by Huber (1973), which is as efficient as Ordinary Least Square (OLS), and is considered the simplest approach. The Least Trimmed Squares (LTS) estimation was introduced by Rousseeuw (1984), and is a high breakdown value method. So, too, is the S estimation, another high breakdown value method with a higher statistical
efficiency than LTS estimation (Rousseeuw & Yohai, 1984). The S estimation is used to minimize the dispersion of residuals. The MM estimation, a special type of M estimation introduced by Yohai (1987), combines high breakdown value estimation and efficient estimation. The M estimation has a higher breakdown value and greater statistical efficiency than the S estimation.

**Calculation for linear Regression using SAS**

```sas
/* First do a simple linear regression */
proc reg data = temp1;
   model y = x;
   run;

/* Compute the absolute and squared residuals*/
data temp1.resid;
   set temp1.pred;
   absresid=abs(residual);
   sqresid=residual**2;

/* Run a Regression with the absolute residuals and squared residuals */
/* to get estimated standard deviation and estimated variance */
proc reg data=temp1.resid;
   model absresid=x;
   output out=temp1.s_weights p=s_hat;
   model sqresid=x;
   output out=temp1.v_weights p=v_hat;

/* Compute weight using standard deviation */
data temp1.s_weights;
   set temp1.s_weights;
   s_weight=1/(s_hat**2);
   label s_weight = "weights using absolute residuals";

/* Compute weight using variances */
data temp1.v_weights;
   set temp1.v_weights;
   v_weight=1/v_hat;
```
An Illustration of a Medical Case

A case study of triglycerides will illustrate the different methods for robust regression.

Table 1. Description of the variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triglycerides</td>
<td>Y</td>
<td>Triglycerides level of patients (mg/dl)</td>
</tr>
<tr>
<td>Weight</td>
<td>X1</td>
<td>Weight (kg)</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>X2</td>
<td>Total cholesterol of patients (mg/dl)</td>
</tr>
<tr>
<td>Proconvertin</td>
<td>X3</td>
<td>Proconvertin (%)</td>
</tr>
<tr>
<td>Glucose</td>
<td>X4</td>
<td>Glucose level of patients (mg/dl)</td>
</tr>
<tr>
<td>HDL-Cholesterol</td>
<td>X5</td>
<td>High density lipoprotein cholesterol (mg/dl)</td>
</tr>
<tr>
<td>Hip</td>
<td>X6</td>
<td>Hip circumference (cm)</td>
</tr>
<tr>
<td>Insulin</td>
<td>X7</td>
<td>Insulin level of patients (IU/ml)</td>
</tr>
<tr>
<td>Lipid</td>
<td>X8</td>
<td>Taking lipid lowering medication (0 = no, 1= yes)</td>
</tr>
</tbody>
</table>

Sources: Ahmad & Shafiq, 2013; Ahmad et al., 2014
Algorithm for Weighting Multiple Linear Model Regression by different Robust Regression Methods

Title 'Alternative Modeling on Weighting Multiple linear regression';

Data Medical;
input Y X1 X2 X3 X4 X5 X6 X7 X8;
Datalines;

168 85.77 209 110 114 37 130.0 17 0
304 58.98 228 111 153 33 105.5 28 1
72 33.56 196 79 101 69 88.5 6 0
119 49.00 281 117 95 38 104.2 10 1
116 38.55 197 99 110 37 92.0 12 0
87 44.91 184 131 100 45 100.5 18 0
136 48.09 170 96 108 37 96.0 13 1
78 69.43 163 89 111 39 103.0 8 0
223 47.63 195 177 112 39 95.0 15 0
200 55.35 218 108 131 31 104.0 33 1
159 59.66 234 112 174 55 114.0 14 0
181 68.97 262 152 108 44 114.5 20 1
134 51.49 178 127 105 51 100.0 21 0
162 39.69 248 135 92 63 93.0 9 1
96 56.58 210 122 105 56 103.4 6 0
117 63.48 252 125 99 70 104.2 10 0
106 66.70 191 103 101 32 103.3 16 0
120 74.19 238 135 142 50 113.5 14 1
119 60.12 169 98 103 33 114.0 13 0
116 36.60 221 113 88 60 94.3 11 1
109 56.40 216 128 90 49 107.1 13 0
105 35.15 157 114 88 35 95.0 12 0
88 50.13 192 120 100 54 100.0 11 0
241 56.49 206 137 148 79 113.0 14 1
175 57.39 164 108 104 42 103.0 15 0
146 43.00 209 116 93 64 97.0 13 0
199 48.84 219 104 158 44 97.0 11 0
85 41.28 171 92 86 64 95.4 5 0
90 65.79 156 80 98 54 98.5 11 1
87 56.90 247 128 95 57 106.3 9 0
103 35.15 257 121 111 69 89.5 13 0
SHAFIQ ET AL.

121  55.12  138  108  104  36  109.0  13  0
223  57.17  176  112  121  38  114.0  32  0
 76  49.45  174  121  89  47  101.0  8  0
151  44.46  213  93  116  45  99.0  10  1
145  56.94  228  112  99  44  109.0  11  0
196  44.00  193  107  95  31  96.5  12  0
113  53.54  210  125  111  45  105.5  19  0
113  35.83  157  100  92  55  95.0  13  0
;
Run;

ods rtf file='result_ex1.rtf' ;

/* This first step is to make the selection of the data that have a significant impact on triglyceride levels. The next step is to perform the procedure of modeling linear regression model and run the regression to get the residuals*/
   proc reg data= Medical;
      model Y = X1  X2  X3   X4   X5  X6  X7  X8;
      output out=work.pred r=residual;
   run;

/* Compute the Absolute and Squared Residuals*/
   data work.resid;
      set work.pred;
      absresid=abs(residual);
      sqresid=residual**2;

/* Run a Regression Compute the Absolute and Squared Residuals to Get Estimated Standard Deviation and Variances*/
   proc reg data=work.resid;
      model absresid=X1  X2  X3  X4  X5  X6  X7  X8;
      output out=work.s_weights p=s_hat;
   run;

   model sqresid=X1  X2  X3  X4  X5  X6  X7  X8;
   output out=work.v_weights p=v_hat;
   run;
/* Compute the Weight Using Estimated Standard Deviation and Variances*/
   data work.s_weights;
   set work.s_weights;
   s_weight=1/(s_hat**2);
   label s_weight = "weights using absolute residuals";

   data work.v_weights;
   set work.v_weights;
   v_weight=1/v_hat;
   label v_weight = "weights using squared residuals";

/* Do a Weighted Least Squares Using the Weight from the Estimated Standard Deviation*/
   proc reg data=work.s_weights;
   weight s_weight;
   model Y = X1 X2 X3 X4 X5 X6 X7 X8;
   run;

/* Do a Weighted Least Squares Using the Weight from the Estimated Variances*/
   proc reg data=work.v_weights;
   weight v_weight;
   model Y = X1 X2 X3 X4 X5 X6 X7 X8;
   run;

/* Do Robust Regression, a Four Estimation Method to compare which are LTS, M, MM and S-Estimation For Weighted Least Square using estimated Standard Deviation*/
   proc robustreg method=LTS data=work.s_weights;
   weight s_weight;
   model Y = X1 X2 X3 X4 X5 X6 X7 X8 / diagnostics leverage;
   run;

/* Do a Robust Regression, a Four Estimation Method compare which are LTS, M, MM and S-Estimation For Weighted Least Square using estimated Variances*/
   proc robustreg method=LTS data=work.v_weights;
   weight v_weight;
model Y = X1  X2 X3 X4 X5 X6 X7 X8 / diagnostics leverage;
run;

Results

Compiled in Table 2 are the results from the multiple regression analysis using the original data. Compiled in Table 3 are the results for the weighted least square by standard deviation and weighted least square by variance. The residual plots do not indicate any problem with the model, as can be seen in Figures 1-3. A normal distribution appears to fit the sample data fairly well. The plotted points form a reasonably straight line. In our case, the residual plots bounce randomly around the 0 line (residual vs. predicted value). This supports the reasonable assumption that the relationship is linear.

Table 2. Parameter Estimates for Original Data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-86.56544</td>
<td>102.93662</td>
<td>0.4070</td>
</tr>
<tr>
<td>x1</td>
<td>-1.08598</td>
<td>0.95288</td>
<td>0.2634</td>
</tr>
<tr>
<td>x2</td>
<td>-0.06448</td>
<td>0.21973</td>
<td>0.7712</td>
</tr>
<tr>
<td>x3</td>
<td>0.61857</td>
<td>0.36615</td>
<td>0.1015</td>
</tr>
<tr>
<td>x4</td>
<td>1.10882</td>
<td>0.33989</td>
<td>0.0028</td>
</tr>
<tr>
<td>x5</td>
<td>-0.52289</td>
<td>0.57119</td>
<td>0.3673</td>
</tr>
<tr>
<td>x6</td>
<td>0.81327</td>
<td>1.38022</td>
<td>0.5601</td>
</tr>
<tr>
<td>x7</td>
<td>2.77339</td>
<td>1.25026</td>
<td>0.0343</td>
</tr>
<tr>
<td>x8</td>
<td>22.40585</td>
<td>14.51449</td>
<td>0.1331</td>
</tr>
</tbody>
</table>

Table 3. Parameter Estimates for Weighted Multiple Linear Regression

<table>
<thead>
<tr>
<th>Variables</th>
<th>Weighted Least Square MLR (SD)</th>
<th>Weighted Least Square MLR (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimate</td>
<td>Standard Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-150.25787</td>
<td>90.05385</td>
</tr>
<tr>
<td>x1</td>
<td>-1.30694</td>
<td>0.59423</td>
</tr>
<tr>
<td>x2</td>
<td>-0.01586</td>
<td>0.17670</td>
</tr>
<tr>
<td>x3</td>
<td>0.44460</td>
<td>0.35706</td>
</tr>
<tr>
<td>x4</td>
<td>0.89106</td>
<td>0.38240</td>
</tr>
<tr>
<td>x5</td>
<td>-0.23352</td>
<td>0.44853</td>
</tr>
<tr>
<td>x6</td>
<td>1.74405</td>
<td>1.10677</td>
</tr>
<tr>
<td>x7</td>
<td>2.81731</td>
<td>1.29607</td>
</tr>
<tr>
<td>x8</td>
<td>16.87506</td>
<td>10.34963</td>
</tr>
</tbody>
</table>
Figure 1. Fit Diagnostic for $y$
Figure 2. Fit Diagnostic for $y$-weighted least square using standard deviation
ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

Figure 3. Fit Diagnostic for y-weighted least square using variances

Shown in Table 2 are the variables $x_4$ ($p = 0.0028$) and $x_7$ ($p = 0.0343$) were statistically significant for the multiple regression analysis. Shown in Table 3 are the variables $x_1$ ($p = 0.0357$), $x_4$ ($p = 0.0267$) and $x_7$ ($p = 0.0377$), which were statistically significant for weighted least square by standard deviation. The weighted least square by variance model shows the variable $x_4$ ($p = 0.0111$) and $x_7$ ($p = 0.0028$). RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data, which are to observe how close the data points are to the model predicted values. Lower value of RMSE indicated a better fit. The RMSE for weighted least square by variance (1.08) shows a lower value compared to the weighted least square standard deviation (1.31) and multiple regression (36.4). A higher R-squared value indicated how well the data fit the model and also indicates a better model. The model multiple regression analysis has R-squared of 0.62, weighted standard deviation multiple regression has R-squared of 0.67 and weighted variance multiple regression has R-squared of 0.63.
Shown in Table 4 is a comparison of the models—multiple linear regression (model 1), weighted least square by standard deviation (model 2) and weighted least square by variance (model 3)—using the four different robust methods, which are M estimation, LTS estimation, S estimation and MM estimation. The LTS estimation has high R-squared in three of the models compared to other robust methods. The S estimation also has high R-squared compared to MM and M estimation.

**Table 4. Comparison of Model by using different Robust Method**

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th></th>
<th></th>
<th></th>
<th>Model 2</th>
<th></th>
<th></th>
<th></th>
<th>Model 3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outlier</td>
<td>Leverage</td>
<td>$R^2$</td>
<td>Outlier</td>
<td>Leverage</td>
<td>$R^2$</td>
<td>Outlier</td>
<td>Leverage</td>
<td>$R^2$</td>
<td>Outlier</td>
<td>Leverage</td>
</tr>
<tr>
<td>$M$</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.4662</td>
<td>0.0769</td>
<td>0.2051</td>
<td>0.5761</td>
<td>0.0769</td>
<td>0.2051</td>
<td>0.5761</td>
<td>0.1622</td>
<td>0.1892</td>
</tr>
<tr>
<td>$LTS$</td>
<td>0.1282</td>
<td>0.2051</td>
<td>0.7289</td>
<td>0.1282</td>
<td>0.2051</td>
<td>0.7289</td>
<td>0.1282</td>
<td>0.2051</td>
<td>0.7289</td>
<td>0.1351</td>
<td>0.1892</td>
</tr>
<tr>
<td>$S$</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.5230</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.6079</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.6079</td>
<td>0.0000</td>
<td>0.1892</td>
</tr>
<tr>
<td>$MM$</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.4602</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.5843</td>
<td>0.0000</td>
<td>0.2051</td>
<td>0.5843</td>
<td>0.0000</td>
<td>0.1892</td>
</tr>
</tbody>
</table>

From Figure 4-6 there is a detection of outlier in observations. They present a regression diagnostics plot (a plot of the standardized residuals of robust regression LTS versus the robust distance). As indicated in Figure 4 and 5, observation 37 is identified as outlier. The observations of 2, 9, 24, and 27 are identified as outlier and leverage. Observations 10, 18 and 33 are identified as leverage point. In Figure 6, observation 35 is identified as outlier, observations 2, 8, 23, and 26 are identified as outlier and leverage, and observations 10, 17 and 27 are identified as leverage. The leverage plots available in SAS software are considered useful and effective in detecting multicollinearity, non-linearity, significance of the slope, and outliers (Lockwood & Mackinnon, 1998).
ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

Figure 4. Outlier and Leverage Diagnostic for Y using LTS (Model 1)

Figure 5. Outlier and Leverage Diagnostic for Y using LTS (Model 2)
Figure 6. Outlier and Leverage Diagnostic for Y using LTS (Model 3)

Conclusion
SAS code for four different methods of robust regression was considered: M estimation, LTS estimation, S estimation, and MM estimation. They provide a better understanding of the weighted multiple linear regression and different robust method underlying of relative contributions.

Acknowledgements
The authors would like to express their gratitude to Universiti Sains Malaysia for providing research funding (Grant 304/PPSG/61313187, School of Dental Sciences, Universiti Sains Malaysia).

References


