JMASM41: An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach (SAS)

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

Mohamad Arif Awang Nawi  
*Universiti Malaysia Terengganu, anz_wsh@yahoo.com*

Nor Azlida Aleng  
*Universiti Malaysia Terengganu, azlida_aleng@umt.edu.my*

Mohamad Shafiq  
*Universiti Sains Malaysia, Kelantan, Malaysia, shafiqmat786@gmail.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**  
W Ahmad, Wan Muhamad Amir; Awang Nawi, Mohamad Arif; Aleng, Nor Azlida; and Shafiq, Mohamad (2016) "JMASM41: An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach (SAS)," *Journal of Modern Applied Statistical Methods*: Vol. 15 : Iss. 2 , Article 44.  
DOI: 10.22237/jmasm/1478004120  
Available at: [http://digitalcommons.wayne.edu/jmasm/vol15/iss2/44](http://digitalcommons.wayne.edu/jmasm/vol15/iss2/44)
JMASM41: An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach (SAS)

Erratum
This paper was originally published in JMASM Algorithms & Code without its enumeration, JMASM41.
An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach

Wan Muhamad Amir W Ahmad
University of Science, Malaysia
Kelantan, Malaysia

Mohamad Arif Awang Nawi
University Malaysia Terengganu
Kuala Terengganu, Malaysia

Nor Azlida Aleng
University Malaysia Terengganu
Kuala Terengganu, Malaysia

Mohamad Shafiq
University of Science, Malaysia
Kelantan, Malaysia

Research on modeling is becoming popular nowadays, there are several of analyses used in research for modeling and one of them is known as applied multiple linear regressions (MLR). To obtain a bootstrap, robust and fuzzy multiple linear regressions, an experienced researchers should be aware the correct method of statistical analysis in order to get a better improved result. The main idea of bootstrapping is to approximate the entire sampling distribution of some estimator. To achieve this is by resampling from our original sample. In this paper, we emphasized on combining and modeling using bootstrapping, robust and fuzzy regression methodology. An algorithm for combining method is given by SAS language. We also provided some technical example of application of method discussed by using SAS computer software. The visualizing output of the analysis is discussed in detail.

Keywords: Multiple linear regression, robust regression, bootstrap method

Introduction

Multiple linear regression (MLR) is an extension of simple linear regression. The random error term is added to make the model 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 (Ngo & La Puente, 2012; Amir,

Dr. Wan Muhamad Amir W Ahmad is an Associate Professor. Email him at: wnamir@usm.my. Mohamad Arif Bin Awang Nawi is a postgraduate student in the Faculty Informatics and Computing. Email him at: anz_wsh@yahoo.com.
A fuzzy regression model corresponding to equation (1) can be stated as:

\[ y = A_0 + A_1 x_1 + A_2 x_2 + \ldots + A_k x_k \]  

(1)

Explanation variables \( x_i \)'s are assumed to be precise. However, response variable \( Y \) is not crisp; it is fuzzy in nature. That means the parameters are also fuzzy in nature. Hence, the objective is to estimate these parameters.

Assume \( A_i \)'s are assumed symmetric fuzzy numbers which can be presented by interval. For example, \( A_i = < a_{1c}, a_{1w} > \) where \( a_{1c} \) is center and \( a_{1w} \) is radius or has associated vagueness. The fuzzy set reflects the confidence in the regression coefficients around \( a_{1c} \) in terms of symmetric triangular memberships function. Application of this method should be given more attention when the underlying phenomenon is fuzzy which means that the response variable is fuzzy. Thus, the relationship is also considered to be fuzzy.

\[ A_i = < a_{1c}, a_{1w} > \] can be written as \( A_i = [ a_{1L}, a_{1R} ] \) with \( a_{1L} = a_{1c} - a_{1w} \) and \( a_{1R} = a_{1c} - a_{1w} \) (Kacprzyk & Fedrizzi, 1992). In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model.

\[ y_j = A_0 + A_1 x_{ij} + A_2 x_{2j} + \ldots + A_k x_{kj} \]  

(2)

Using \( A_i = < a_{1c}, a_{1w} > \) write

\[ y_j = < a_{0c}, a_{0w} > + < a_{1c}, a_{1w} > x_{1j} + \ldots + < a_{nc}, a_{nw} > x_{nj} = < a_{jc}, a_{jw} > \]  

(3)

Thus,

\[ y_{jc} = a_{0c} + a_{1c} x_{1j} + \ldots + a_{nc} x_{nj} \]  

(4)

\[ y_{jw} = a_{0w} + a_{1w} |x_{1j}| + \ldots + a_{nw} |x_{nj}| \]  

(5)

As \( y_{jw} \) represent radius and so cannot be negative, therefore on the right-hand side of equation \( y_{jw} = a_{0w} + a_{1w} |x_{1j}| + \ldots + a_{nw} |x_{nj}| \), absolute values of \( x_{ij} \) are taken. Suppose there are \( n \) data points, each comprising a \( (n + 1) \)-row vector. Then parameters \( A_i \) are estimated by minimizing the quantity, which is total vagueness.
ALTERNATIVE MULTIPLE LINEAR MODEL REGRESSION MODELING

of the model-data set combination, subject to the constraint that each data point must fall within estimated value of response variable. This can be visualized as the following linear programming problem.

Minimized \( \sum_{j=1}^{m} \left( a_{0w} + a_{iw} |x_{1j}| + \cdots + a_{w} |x_{nj}| \right) \)

Subject to

\[
\begin{align*}
\left\{ a_{0c} + \sum_{i=1}^{n} a_{ic} x_{ij} \right\} - \left\{ a_{0w} + \sum_{i=1}^{n} a_{iw} x_{ij} \right\} & \leq Y_j \\
\left\{ a_{0c} + \sum_{i=1}^{n} a_{ic} x_{ij} \right\} + \left\{ a_{0w} + \sum_{i=1}^{n} a_{iw} x_{ij} \right\} & \geq Y_j
\end{align*}
\]

and \( a_{iw} \geq 0 \). Simplex procedure is generally employed in order to solve the linear programming problem.

Calculation for linear Regression using SAS

/* First do Multiple linear regression */
proc reg data=temp1;
model y=x1 x2;
run;

Approach the MM-Estimation Procedure for Robust Regression

/* Then do robust regression, in this case, MM-estimation */
ods graphics on;
proc robustreg method = MM fwls data=biostatistics plot=fitplot(nolimits) plots=all;
model y = x1 x2 / diagnostics itprint;
output out=resids out=robout r=residual weight=weight outlier=outlier sr=stdres;
run;
ods graphics off;
Start

Data Collection

Checking Assumptions

Yes

Model Building with Robust Regression

Selection of MM or LTS estimation

No

Diagnostics Test: Checking for outliers

No

Linear Regression (Using SAS software)

Adding Bootstrap to the Linear Regression with Replication ($n = 1000$)

Data Analysis Based on Alternative Model

Bootstrap Procedure in SAS

Robust Procedure in SAS Algorithm

Results

Stop

**Figure 1.** Flow Chart of Robust, Bootstrap and Fuzzy Regression
Procedure for Bootstrap with Case Resampling \( n = 1000 \)
/* And finally, use a bootstrap with case resampling */
ods listing close;
proc surveyselect data=temp1 out=boot1 method=urs
samprate=1 outhits rep=1000;
run;

Procedure for bootstrap into fuzzy regression Model
/*Combination of Bootstrap Technique with Fuzzy Regression*/
ods listing close;
proc optmodel;
set j = 1..8;
number Fish{j}, weight{j}, height{j};
read data boot1 into [_n_] Fishweight height;

/*Print Fishweight height*/
print Fishweight height;
number n init 8; /*Total of Observations*/

/* Decision Variables bounded or not bounded*/
/*Theses three variables are bounded*/
var aw{1..3} >= 0;

/*These three variables are not bounded*/
var ac{1..3};

/* Objective Function*/
min z1 = aw[1]*n + sum{i in j} weight[i]*aw[2]+sum{i in j} height[i]*aw[3];

/*Linear Constraints*/
con c{i in 1..n}:
con c{i in 1..n}:
An Illustration of a Biostatistics Case

A Case Study of Aquaculture

Table 1. Description of the Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish</td>
<td>Y</td>
<td>Number of Fish Caught</td>
</tr>
<tr>
<td>Weight</td>
<td>X1</td>
<td>Weight in (g)</td>
</tr>
<tr>
<td>Height</td>
<td>X2</td>
<td>Height in (cm)</td>
</tr>
</tbody>
</table>

*(Talib, Jaafar, & Sirwar, 2007)*

Full Algorithm for Alternative Multiple Linear Regression Modelling

Title 'Alternative Linear programming with combining robust and bootstrap';
data Biostatistics;
input Fish weigh height;
datalines;
97.32   110.41  103.74
174.52  111.08  104.80
214.56  114.98  105.71
178.44  114.16  105.27
199.48  112.99  105.45
189.92  115.20  105.34
170.48  113.24  105.11
207.16  117.19  105.66
;
run;
ods rtf file='result_ex1.rtf' ;

/*The next step is performing the procedure of modeling linear...
/* regression model */
procreg data = biostatistics;
modelFish = weigh height;
run;

/* Then do robust regression, in this case, MM-estimation */
ods graphics on;
procrobustreg method = MM fwls data = biostatistics plot=fitplot(nolimits) plots=all;
modelFish = weigh height/ diagnostics itprint;
output out=resids out=robout r=residual weight=weight outlier=outlier sr=stdres;
run;
ods graphics off;

/* And finally use a bootstrap with case resampling */
ods listing close;
procsurveyselect data = biostatistics out = boot1 method = urs samprate = 1 outhits rep=1000;
run;

/* Combination of Bootstrap Technique with Fuzzy Regression */
ods listing close;
procoptmodel;
set j = 1..8;
numberOfFish{j}, weigh{j}, height{j};
read data boot1 into [_n_] Fish weigh height;

/* Print Fish weight height */
printFish weigh height;

/* Total of Observations */
numberOf n init 8;

/* Theses three variables are bounded */
var aw{1..3} >= 0;
/* These three variables are not bounded */
var ac{1..3};

/* Objective Function */
\min z_1 = a_1 n + \sum_{i \in j} \text{weight}_i \cdot a_2 + \sum_{i \in j} \text{height}_i \cdot a_3;

/* Linear Constraints */
con c\{i \in 1..n\}:
    a_1 + \text{weight}_i \cdot a_2 + \text{height}_i \cdot a_3 - a_1 - \text{weight}_i \cdot a_2 - \text{height}_i \cdot a_3 \leq \text{Fish}_i;

con c_1\{i \in 1..n\}:
    a_1 + \text{weight}_i \cdot a_2 + \text{height}_i \cdot a_3 + a_1 + \text{weight}_i \cdot a_2 + \text{height}_i \cdot a_3 \geq \text{Fish}_i;

expand; /* This provides all equations */
solve;
print ac aw;
quit;
ods rtf close;

Results

A higher R-squared value indicated how well the data fit the model and indicates a better model.

Table 2. Goodness-of-fit

<table>
<thead>
<tr>
<th>Goodness-of-Fit</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Square</td>
<td>0.8199</td>
</tr>
<tr>
<td>AICR</td>
<td>5.5323</td>
</tr>
<tr>
<td>BICR</td>
<td>9.4456</td>
</tr>
<tr>
<td>Deviance</td>
<td>234.4750</td>
</tr>
</tbody>
</table>

Method of Multiple linear regression (MLR), we obtained the result as shown in Table 3.

Table 4 shows the results by using bootstrapping method for fuzzy regression with \( n = 1000 \). The aim of bootstrapping procedure is to approximate the entire sampling distribution of some estimator by resampling (simple random sampling with replacement) from the original data (Yaffee, 2002). Table 4 summarizes the findings of the calculated parameter.
Table 3. Parameter Estimates for Final Weighted Least Squares Fit

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-6334.91</td>
<td>608.3789</td>
<td>-7527.31 -5142.51</td>
<td>108.43</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>x1</td>
<td>1</td>
<td>-3.0164</td>
<td>2.1608</td>
<td>-7.2516 1.2188</td>
<td>1.95</td>
<td>0.1627</td>
</tr>
<tr>
<td>x2</td>
<td>1</td>
<td>65.2183</td>
<td>7.5704</td>
<td>50.3807 80.0559</td>
<td>74.22</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>7.1356</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Method of Fuzzy Regression (FR) (OPTMODEL)

Table 4. Value of ac and aw

<table>
<thead>
<tr>
<th></th>
<th>ac</th>
<th>aw</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5764.1545</td>
<td>0.000000</td>
</tr>
<tr>
<td>2</td>
<td>-3.0958</td>
<td>0.000000</td>
</tr>
<tr>
<td>3</td>
<td>59.8722</td>
<td>0.075811</td>
</tr>
</tbody>
</table>

While using bootstrap procedure, different output for the ac and aw will be obtained:

ac1 = -5764.1545  
ac2 = -3.0958     
ac3 = 59.8722     
aw1 = 0           
aw2 = 0           
aw3 = 0.075811.

The next step is to compare the performance of multiple linear regression and fuzzy regression.

The Fitted Model for Multiple Linear Regressions

\[ Y = -6334.91 - 3.0164 \text{weight} + 62.21 \text{height} \]  

\[ \text{Standard Error (608.3789)} \quad \text{(2.1608)} \quad \text{(7.5704)} \]
The upper limits of prediction interval are computed by coefficient plus standard error

\[ Y = (-6334.91 + 608.3789) + (-3.0164 + 2.1608 \text{ weight} + (65.21 + 7.5704) \text{ height} \]
\[ Y = (-5726.53) + (-0.86 \text{ weight} + (72.78) \text{ height}) \]

The lower limits of prediction interval are computed by coefficient minus standard error

\[ Y = (-6334.91 - 608.3789) + (-3.0164 - 2.1608 \text{ weight} + (65.21 - 7.5704) \text{ height} \]
\[ Y = (-6943.29) + (-5.1772 \text{ weight} + (57.6396) \text{ height}) \]

**The Fitted Model for Fuzzy bootstrap Regression Is**

\[ Y = [-5764.1545, 0] + [-3.0958, 0] \text{ weight} + [59.8722, 0.075811] \text{ height} \quad (7) \]

The upper limits of prediction interval are computed by coefficient plus standard error

\[ Y = [-5764.1545 + 0] + [-3.0958 + 0] \text{ weight} + [59.8722 + 0.075811] \text{ height} \]
\[ Y = [-5764.15] + [-3.10] \text{ weight} + [60.00] \text{ height} \]

The lower limits of prediction interval are computed by coefficient minus standard error

\[ Y = [-5764.1545 - 0] + [-3.0958 - 0] \text{ weight} + [59.8722 - 0.075811] \text{ height} \]
\[ Y = [-5764.15] + [-3.10] \text{ weight} + [59.80] \text{ height} \]

The width of prediction intervals in respect of multiple linear regression model and fuzzy regression model corresponding to each set of observed explanatory variables is computed manually.
Table 5. Average Width for Former Multiple Linear Regression model and Fuzzy Bootstrap Regression Model

<table>
<thead>
<tr>
<th></th>
<th>Multiple Linear Regression model</th>
<th>Fuzzy Bootstrap Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower limit</td>
<td>Upper limit</td>
</tr>
<tr>
<td>Lower limit</td>
<td>1535.37</td>
<td>1728.71</td>
</tr>
<tr>
<td>-1477.74</td>
<td>1800.92</td>
<td>3278.66</td>
</tr>
<tr>
<td>-1445.48</td>
<td>1868.16</td>
<td>3313.64</td>
</tr>
<tr>
<td>-1466.60</td>
<td>1836.84</td>
<td>3303.44</td>
</tr>
<tr>
<td>-1450.17</td>
<td>1850.95</td>
<td>3301.12</td>
</tr>
<tr>
<td>-1467.93</td>
<td>1841.04</td>
<td>3308.99</td>
</tr>
<tr>
<td>-1471.06</td>
<td>1825.99</td>
<td>3297.05</td>
</tr>
<tr>
<td>-1459.81</td>
<td>1862.62</td>
<td>3322.43</td>
</tr>
<tr>
<td>Average</td>
<td>3298.68</td>
<td>Average</td>
</tr>
</tbody>
</table>

From Table 5, average width for former multiple regression was found to be 3298.68 while using fuzzy regression, the average width is 21.48. This indicates the superiority of fuzzy regression methodology. From this analysis, the most efficient method to obtain the relationship between response and explanatory variable is to apply fuzzy regression method compared to linear regression method.

Conclusion

It was explained how to combine an algorithm between robust, fuzzy regression and the bootstrap method. A small sample size (8 observations only) was used

(a) to apply a bootstrap method in order to achieve an adequate sample size.

(b) to compare the efficiency between original method and with the bootstrap method.

(c) to give a better understanding on how the algorithm works

According to biostatistics history, all the independent variables that we used in this case were significant to the number of fish caught. Without using bootstrapping, the result shows that two out of eight were significant. It is surprising that, using bootstrapping method (with $n = 1000$) the entire significant variable are included in the model as the finding from the biostatistics record. This algorithm provides us with the improved understanding of the modified method and underlying of relative contributions. For further study, it is possible to
approach response surface methodology for every each of significant variables in single algorithm.

References


