Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 46

5-1-2016

JMASM39: Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression (SAS)

Wan Muhamad Amir
University Science Malaysia, wmamir@usm.my

Mohamad Shafiq
University Science Malaysia, shafiqmat786@gmail.com

Hanafi A.Rahim
University Malaysia Terengganu, hanafi@umt.edu.my

Puspa Liza
University of Sultan Zainal Abidin, puspaliza@unisza.edu.my

Azlida Aleng
University Malaysia Terengganu, azlida_aleng@umt.edu.my

See next page for additional authors

DOI: 10.22237/jmasm/1462077900

JMASM39: Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression (SAS)

Erratum
This paper was originally published in JMASM Algorithms & Code without its enumeration, JMASM39.

Authors
Wan Muhamad Amir, Mohamad Shafiq, Hanafi A.Rahim, Puspa Liza, Azlida Aleng, and Zailani Abdullah

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 884-892.

ISSN 1538 − 9472

JMASM Algorithms and Code
Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression (SAS)

Wan Muhamad Amir
University of Science, Malaysia
Kelantan, Malaysia

Mohamad Shafiq
University of Science, Malaysia
Kelantan, Malaysia

Hanafi A. Rahim
University Malaysia Terengganu
Kuala Terengganu, Malaysia

Puspa Liza
Universiti Sultan Zainal Abidin
Kuala Terengganu, Malaysia

Azlida Aleng
University Malaysia Terengganu
Kuala Terengganu, Malaysia

Zailani Abdullah
University Malaysia Kelantan
Kelantan, Malaysia

The aim of bootstrapping is to approximate the sampling distribution of some estimator. An algorithm for combining method is given in SAS, along with applications and visualizations.

Keywords: Multiple linear regression, robust regression and bootstrap method

Introduction

Multiple linear regression (MLR) is an extension of simple linear regression. Table 1 displays the data for multiple linear regression.

Table 1. Data template for multiple linear regression

i	y_i	x_{i0}	x_{i1}	x_{i2}	…	x_{ip}
1	y_1	1	x_{11}	x_{12}	…	x_{1p}
2	y_2	1	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮		⋮
n	y_n	1	x_{n1}	x_{n2}	…	x_{np}

Dr. Amir bin W Ahmad is an Associate Professor of Biostatistics. Email him at: wmamir@usm.my. Mohamad Shafiq Bin Mohd Ibrahim is a postgraduate student in the School of Dental Sciences. Email him at: shafiqmat786@gmail.com.

MLR is used when there are two or more independent variables where the model using population information is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad (1)$$

where β_0 is the intercept parameter and β_0, β_1, β_2,…, β_{k-1} are the parameters associated with $k-1$ predictor variables. The dependent variable **Y** is now written as a function of k independent variables, x_1, x_2,…, x_k.

The random error term is added to make the model probabilistic rather than deterministic. The value of the coefficient β_i determines the contribution of the independent variable x_i, and β_0 is the y-intercept. (Ngo, 2012). The coefficients β_0, β_1,…, β_k are usually unknown because they represent population parameters. Below is the data presentation for multiple linear regression. General linear model in matrix form can be defined by the following vectors and matrices as below:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_{p-1} \end{bmatrix}$$

Calculation for Linear Regression using SAS

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

Approach the MM-Estimation Procedure for Robust Regression

```
/* Then we do robust regression, in this case, MM-estimation */
     proc robustreg data = temp1 method = MM;
     model y = x;
     run;
```

Procedure for Bootstrap with Case Resampling *n* = 1000

```
/* And finally we use a bootstrap with case resampling */
     ods listing close;
     proc surveyselect data = temp1 out = boot1 method = urs samprate =
     1 outhits rep=1000;
     run;
```

```
proc reg data = boot1 outest = est1(drop =_:);
model y = x;
by replicate; run;
ods listing;
```

An Illustration of a Medical Case

Case Study I: A Case Study of Triglycerides

Table 2. Description of the variables

Variables	Code	Description
Triglycerides	*Y*	Triglycerides level of patients (mg/dl)
Weight	*X*1	Weight (kg)
Total Cholesterol	*X*2	Total cholesterol of patients (mg/dl)
Proconvertin	*X*3	Proconvertin (%)
Glucose	*X*4	Glucose level of patients (mg/dl)
HDL-Cholesterol	*X*5	High density lipoprotein cholesterol (mg/dl)
Hip	*X*6	Hip circumference (cm)
Insulin	*X*7	Insulin level of patients (IU/ml)
Lipid	*X*8	Taking lipid lowering medication (0 = no, 1 = yes)

Sources: Ahmad and Ibrahim (2013), Ahmad, Ibrahim, Halim, and Aleng (2014)

Algorithm for Combining Robust and Bootstrap in Multiple Linear Model Regression

```
Title 'Alternative Modeling on 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   48.04  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
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='results_ex1.rtf';

/* This first step is to make the selection of the data that have a
significant impact with triglyceride levels. The next step is performing
the procedure of modeling linear regression model */

proc reg data= Medical;
model Y =  X1  X2 X3  X4  X5  X6  X7  X8;
run;

/* Then do robust regression, in this case MM-estimation */

proc robustreg data= Medical method=MM;
model Y =  X1  X2 X3  X4  X5  X6  X7  X8/ diagnostics leverage;
```

```
 output out=robout r=resid sr=stdres;
run;

/* Use a bootstrap with case resampling */

ods listing close;
proc surveyselect data= Medical out=boot1 method=urs samprate=1 outhits
rep = 50;
run;

/* And finally use a bootstrap with robust with case resampling */
proc robustreg data=boot1 method=MM plot=fitplot(nolimits) plots=all;
model Y =  X1  X2 X3  X4  X5  X6  X7  X8;
run;

ods rtf close;
```

Results from Original Data

Below are the results from the analysis using the original data. The residual plots do not indicate any problem with the model. A normal distribution appears to fit our sample data fairly well. The plotted points form a reasonably straight line. In our case, the residual bounce randomly around the 0 line (residual vs. predicted value). This suggest that the assumption that the relationship is linear is reasonable. A higher R-squared value of 0.62 indicated how well the data fit the model and also indicates a better model.

Table 3. Parameter estimates for original data

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	***t* Value**	**Pr > |*t*|**
Intercept	1	-86.5654	102.93662	-0.84	0.4070
*x*1	1	-1.08598	0.95288	-1.14	0.2634
*x*2	1	-0.06448	0.21973	-0.29	0.7712
*x*3	1	0.61857	0.36615	1.69	0.1015
*x*4	1	1.10882	0.33989	3.26	0.0028
*x*5	1	-0.52289	0.57119	-0.92	0.3673
*x*6	1	0.81327	1.38022	0.59	0.5601
*x*7	1	2.77339	1.25026	2.22	0.0343
*x*8	1	22.40585	14.51449	1.54	0.1331

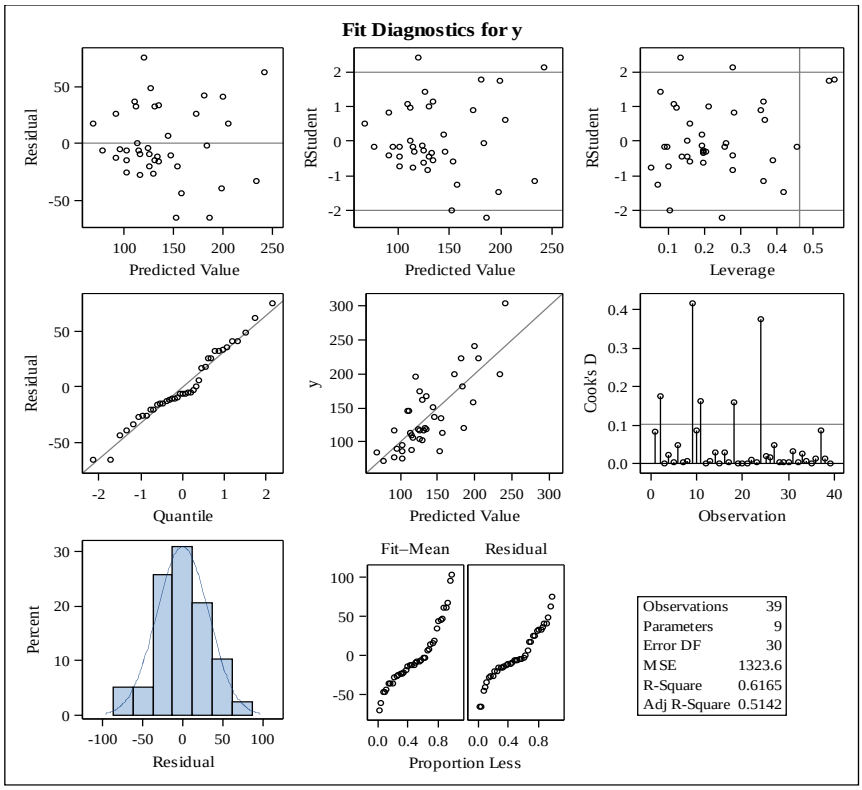

Figure 1. Fit diagnostic for *y*

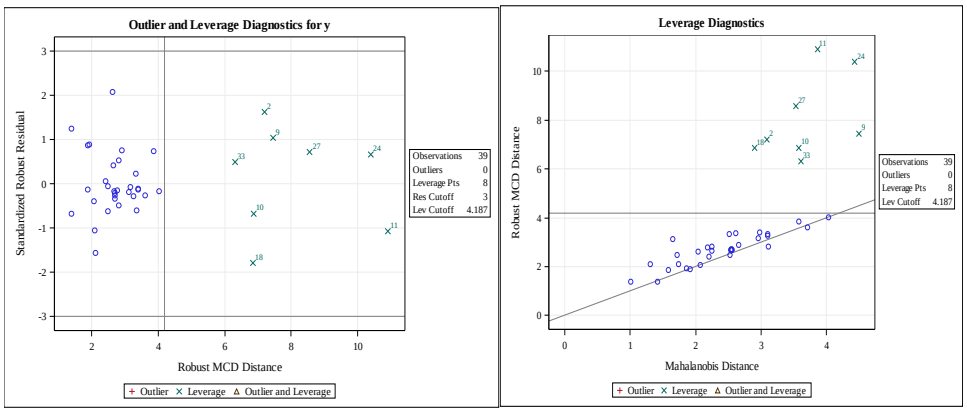

Figure 2. Outlier and Leverage Diagnostic for *y*

From Figure 2, we can see that there is no detection of outlier in observations. The leverage plots available in the SAS software are considered useful and effective in detecting multicollinearity, non-linearity, significance of the slope, and outliers (Lockwood & Mackinnon, 1998). Both of figures above indicate that this sample have no peculiarity and a data entry have no error. Figure 2 presented a regression diagnostics plot (a plot of the standardized residuals of robust regression MM versus the robust distance). Observations 2, 9, 10, 11, 18, 24, 27 and 33 are identified as leverage points. Below is the results of bootstrapping with n = 50:

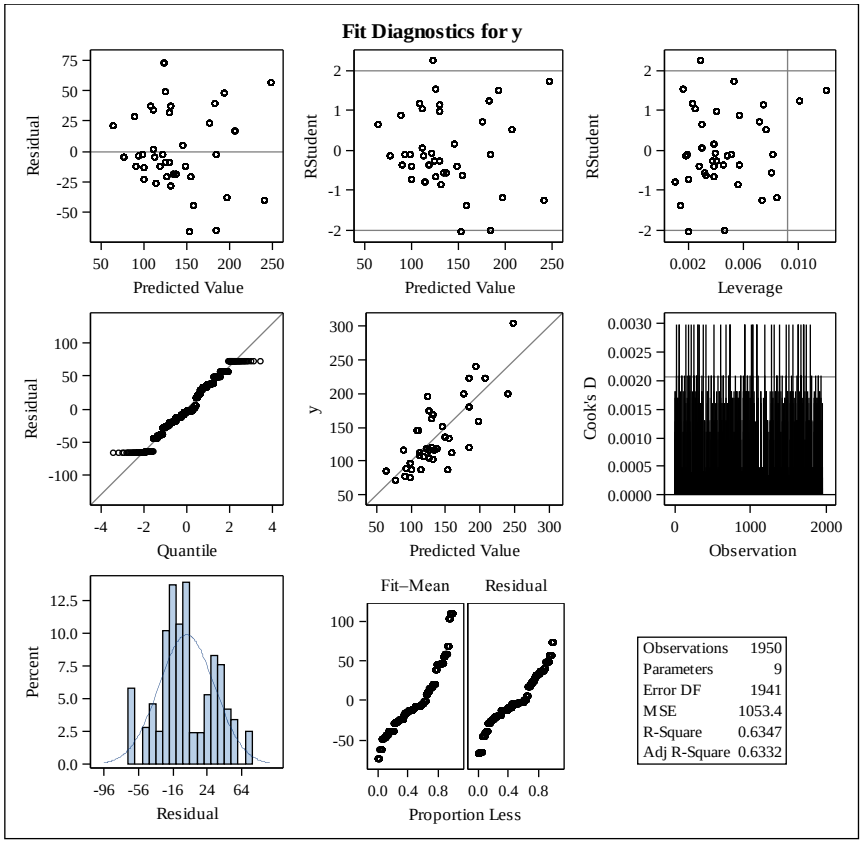

Figure 3. Fit diagnostic for *y* after bootstrapping

Table 4 shows the results by using bootstrapping method. 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). The next step is to calculate the efficiency of the

bootstrap method with the original sample data. Table 5 summarize the findings of the calculated parameter.

Table 4. Parameter estimates using bootstrapping method

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	-297.0810	9.18120	-315.0760	-279.0860	1047.02	<0.0001
*x*1	1	-1.3526	0.07910	-1.5076	-1.1977	292.69	<0.0001
*x*2	1	0.0286	0.01850	-0.0077	0.0649	2.38	0.1227
*x*3	1	0.0441	0.04360	-0.0413	0.1295	1.03	0.3112
*x*4	1	1.5405	0.03300	1.4759	1.6052	2182.31	<0.0001
x5	1	0.2976	0.04960	0.2004	0.3948	36.04	<0.0001
*x*6	1	2.6234	0.12240	2.3836	2.8632	459.66	<0.0001
*x*7	1	2.4174	0.10580	2.2100	2.6248	521.88	<0.0001
*x*8	1	24.6443	1.20480	22.2829	27.0057	418.39	<0.0001
Scale	0	27.6976					

Table 5. Comparison of parameter estimates original sample and bootstrapping method

Parameter Estimates							
Variables	**Original sample**			**Bootstrapping Method**			**Efficiency of Parameter (%)**
	Parameter Estimate	**Standard Error**	**P value**	**Estimate**	**Standard Error**	**P value**	
Intercept	-86.56544	102.93662	0.4070	-297.0810	9.1812	< 0.0001	
*x*1	-1.08598	0.95288	0.2634	-1.3526	0.0791	< 0.0001	24.55
*x*2	-0.06448	0.21973	0.7712	0.0286	0.0185	0.1227	144.35
*x*3	0.61857	0.36615	0.1015	0.0441	0.0436	0.3112	92.87
*x*4	1.10882	0.33989	0.0028	1.5405	0.0330	< 0.0001	38.93
*x*5	-0.52289	0.57119	0.3673	0.2976	0.0496	< 0.0001	156.91
*x*6	0.81327	1.38022	0.5601	2.6234	0.1224	< 0.0001	222.57
*x*7	2.77339	1.25026	0.0343	2.4174	0.1058	< 0.0001	12.83
*x*8	22.40585	14.51449	0.1331	24.6443	1.2048	< 0.0001	9.99

References

Ahmad, W. M. A. W., & Ibrahim, M. S. (2013). High density lipoprotein cholesterol predicts triglycerides level in three distinct phases of blood pressure. *International Journal of Sciences: Basic and Applied Research, 10*(1), 38-46. Retrieved from

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied&page=article&op=view&path%5B%5D=1111&path%5B%5D=1098

Ahmad, W. M. A. W., Ibrahim, M. S., Halim, N., & Aleng, N. A. (2014). A study of triglycerides level in three distinct phases of human blood pressure: A case study from previous projects. *Applied Mathematical Sciences, 8*(46), 2289-2305. doi: 10.12988/ams.2014.42145

Lockwood, C. M., & Mackinnon, D. P. (1998). Bootstrapping the standard error off the mediated effect. *Proceedings of the 23rd Annual Meeting of SAS Users Group International*. Cary, NC: SAS Institute, Inc.

Ngo, T. H. D. (2012). The steps to follow in a multiple regression analysis. *Proceedings of the SAS Global Forum 2012 Conference (paper 333-2012)*. Cary, NC: SAS Institute Inc. Retrieved from http://support.sas.com/resources/papers/proceedings12/333-2012.pdf

Yaffee, R. A. (2002). *Robust regression analysis: Some popular statistical package options*. Statistics, Social Science, and Mapping Group, New York University, NY. Retrieved from http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/RobustRegAnalysis.pdf