International Journal of Engineering & Technology, 7 (3.28) (2018) 343-347

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An Alternative Algorithm for Linear Regression Modeling for Efficient Decision: A New Strategy of Handling Insurance Data

Mohamad Arif Awang Nawi[1]*, Wan Muhamad Amir W Ahmad[1], Mohamad Shafiq Mohd Ibrahim[1], Mustafa Mamat[2], Rabiatul Adawiyah Abdul Rohim[1], Mohamad Afendee Mohamed[2]

[1]School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kelantan, Malaysia
[2]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UNISZA), 21030 Kuala Terengganu, Terengganu, Malaysia
**Corresponding author E-mail: mohamadarif@usm.my*

Abstract

The multiple linear regression model is an important tool for investigating relationships between several response variables and some predictor variables. The primary interest is in inference about the unknown regression coefficient matrix. In this paper, we propose to combine and compare multiple linear regression, bootstrapping and fuzzy regression methods to build alternative methods. We formalize this extension and prove its validity. A real data example and two simulated data examples, which offer some finite sample verification of our theoretical results are provided. The results, based on significant value and average width showed alternative methods produce better results than multiple linear regressions (MLR) model.

Keywords: *Multiple Linear Regression; Bootstrap method; Fuzzy Regression.*

1. Introduction

Multiple linear regression (MLR) model is an important tool for investigating relationships between several response variables and some predictor variables. MLR modeling is powerful technique in statistical and most commonly used in finance, economic, agriculture and many more. It is one of the most commonly used methods in econometric work. The relationship is described as a model for estimating the dependent variable from independent variables. This method estimates the linear relationship between a dependent (response) and independent (explanatory) variables. The multiple linear regression models are expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \ell \quad (1)$$

where Y is dependent variable, X is independently variable, β's are crisp parameters and X_n are the vector of crisp numbers. Usually, ℓ are assumed to be independent random variables with a mean of 0 and variance. The parameters are usually estimated using the method of least squares. The explanation of various aspects of multiple linear regression methodologies is given in [7]. According to [4, 6], bootstrap is actually a statistical resampling technique which is used to estimate populations' statistics, works based on random sampling with replacement. In addition, this method also provides an estimate of the statistical distribution, the coverage probability of the confidence interval, and the probability of rejecting the hypothesis test that produces accurate results. The theoretical model for bootstrap is given in the following:

$$Y^* = X\hat{\beta} + u^* \quad (2)$$

where a random term u^*, attained from the residuals $\hat{u}$ from initial regression. The model is responsible for creating a sample $\{y_i^*\}_{i=1}^n$ of size $(n, 1)$ per iteration $b(b = 1,....,B)$ basis.

By having the estimated errors larger than the OLS residue, the model's random term can be obtained from transform residuals having the same norm as that of the error terms:

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{(1-h_i)}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{u}_i}{\sqrt{(1-h_i)}}$$

Giving the theoretical bootstrap model as in (3)

$$y_i^*(b) = X_i\hat{\beta} + \tilde{u}_i^*(b), \text{ i=1...n} \quad (3)$$

where $\tilde{u}_i^*(b)$ is is resample from $\tilde{u}_i$.

Let there be another random variable z_j where $z_j = \frac{\beta_j - \beta_j}{s(\hat{\beta}_j)}$ defines the standard confidence interval derived from the assumption that z_j distribution is in accordance to a student's distribution with n-p degrees of freedom. Therefore, given a confidence level of $(1-2\alpha)$, this confidence interval can be expressed as in (4):

$$[\hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(1-\alpha),n-p}, \hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(\alpha),n-p}] \quad (4)$$

where t is the percentile valued (α) and $(1-\alpha)$ of the student's distribution with $n-p$ degrees of freedom. The percentile and

percentile-t approaches are used to construct the bootstrap confidence interval. The first method simply makes use of bootstrap estimations to obtain confidence intervals. If we have a confidence level $(1-2\alpha)$, expression (5) gives the percentile confidence interval for β_j :

$$\left[\hat{\beta}_j^*(\alpha B), \hat{\beta}_j^*\left((1-\alpha)B\right)\right] \quad (5)$$

where $\hat{\beta}_j^*(\alpha B)$ is the αB^{th} value (respectively $\hat{\beta}_j^*\left((1-\alpha)B\right)$ the $(1-\alpha)B^{th}$ value) of the ordered list of the B bootstrap replications. We choose the value for threshold such that α% of the replications produce smaller (larger) $\hat{\beta}_j^*$ than the lower (upper) bound of the percentile confidence interval.

In fuzzy method, regression plays an crucial role when dealing with imprecise data. Obviously, a simple regression equation involving a single dependent and a single independent fuzzy variable can be used for such situation. The studies in [9] showed that fuzzy regression can be a better replacement for statistical regration when the data are vaque with poor model specification. In [3], the regression problem in form of gradient-descent optimization was used to show the effectiveness of an iterative algorithm for multiple regressions with fuzzy data. Fuzzy Linear Regression (FLR) proposed for the first time by the Japanese researcher [10], provides the tools to study the problems that failing to the abovementioned assumptions. A fuzzy linear regression model corresponding to multiple linear regression equations can be stated as:

$$y = A_0 + A_1x_1 + A_2x_2 + \ldots + A_kx_k \quad (6)$$

Previously, explanatory variables x_i's are expected to be concise. However, from (6), we learnt that the response variable Y is strictly not crisp, in fact it is fuzzy in nature and hence the parameters too. Our main concerns is to get the values of these parameters estimated. Henceforth, A_i's are assumed to be symmetric fuzzy numbers, expressible in term of an interval. Consequently, A_i is expressible as fuzzy set $A_i =< a_{1c}, a_{1w} >$ where a_{ic} , a_{iw} is respectively the center and radius. Fuzzy set makes use triangular membership function to determine the confidence in the regression coefficients around a_{ic} . Whenener the nature phenomenon is fuzzy, so should the response variable and in turn its relationship. This $A_i =< a_{1c}, a_{1w} >$ can be written as $A_1 = [a_{1L}, a_{1R}]$ with $a_{1L} = a_{1c} - a_{1w}$ and $a_{1R} = a_{1c} - a_{1w}$ [8]. From the model, we estimate the parameters by minimizing the total vagueness.

$$y_j = A_0 + A_1x_{1j} + A_2x_{2j} + \ldots + A_kx_{kj} \quad (7)$$

using $A_i =< a_{1c}, a_{1w} >$ we can write

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

thus $y_{jc} = a_{0c} + a_{1c}x_{1j} + \cdots + a_{nc}x_{nj}$

$$y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}|$$

As y_{jw} represent the radius, as such it can only be non-negative, therefore from $y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}|$, we can use an absolute values of x_{ij} . Let say we have m different data points, having $a(n+1) - row$ vector each. To estimate the parameter A_i , we simply minimize the quantity, which is the total vagueness of the model data set combination, by limiting each data point belongs to the estimated value of the response variable.

2. Methodology

This section discuss the methodology used in this studies.

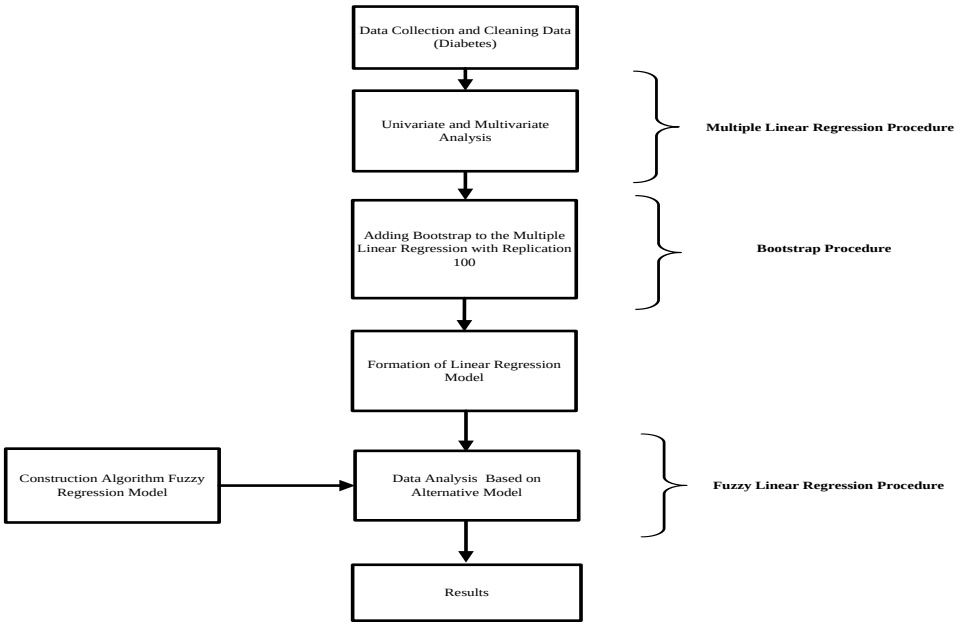

Fig. 1: Flow Chart of Alternative Linear Regression Model

2.1. This algorithm is used to develop the multiple linear regression using SAS software

```
/* First of all create multiple linear regression algorithm*/
proc reg data= for example Data 1;
model y = x1 x2;
run;
```

2.2. Next step is procedure for Bootstrap with case resampling

```
/* Next step we use a bootstrap with case resampling */
   ods listing close;
   procsurveyselect data= for example Data 1 out=boot1
   method=urs
   samprate=1 outhits rep=100 ; /* Depending on the researcher to
   do resampling data as possible */
   run;
```

2.2. Finally, procedure for bootstrap into fuzzy regression model (alternative)

```
/*Combination of Bootstrap algorithm with Fuzzy Regression*/
   ods listing close;
   proc optmodel;
   set j= 1..8;
   number y{j}, x1{j}, x2{j};
   read boot1 into [_n_] y x1 x2;

/*Print y x1 x2*/
   print y x1 x2;
   number n init 8; /*Total of Observation*/

/*Decision Variable Bounded or Not Bounded*/
   var aw{1..3}>=0; /*these three variables are bounded*/
   var ac{1..3}; /*these three variables are not bounded*/
/*Objective Function*/
   min z1= aw[1] * n + sum{i in j} x1[i] * aw[2]+sum{i in j} x2[i]
   * aw[3];

/*Linear Constraints*/
   con c{i in 1..n}:
   ac[1]+x1[i]*ac[2]+x2[i]*ac[3]-aw[1]-x1[i]*aw[2]-
   x2[i]*aw[3]<=y[i];

   con c1{i in 1..n}:
   ac[1]+x1[i]*ac[2]+x2[i]*ac[3]+aw[1]+x1[i]*aw[2]+x2[i]*aw[3]
   >=y[i];

   expand; /*This Provides All Equations*/
   solve;
   print ac aw;
   quit;
   ods rtf close;
```

2.3. Model parameter

Illustration: Data from the article of [1] is considered. They have studied the relative efficiency analysis industry of life and general insurance in Malaysia using stochastic frontier analysis (SFA). The response variable is a profitability of general and life insurance companies (Y) and the explanatory variables are net investment income (X1), total liabilities and assets (X2), management expenses (X3), annual premium (X4) and net claims paid by the company (X5). This paper explained on how to combine an algorithm between fuzzy regression and bootstrap method. The reasons why we use a small sample size are to apply a bootstrap method in order to achieve an adequate sample size and the second objective is to compare the efficiency between original method and with the bootstrapping method (with n = 100). The multiple linear regression model and fuzzy bootstrap regression model were analyzed using SAS software, version 9.3. SAS code and the results expressed as follows.

2.3.1. Multiple linear regression model

```
Data insurance;
input y x1 x2 x3 x4 x5;
Datalines;
11.5745  8.8133  11.4424  12.0755  11.5484  10.2443
12.1891  9.5173  11.9274  13.5917  12.1562  11.3366
   ⋮        ⋮        ⋮        ⋮        ⋮        ⋮
12.4004  9.8467  12.2371  13.1705  12.3107  11.6928
12.2432  9.8088  12.3660  13.1512  12.2304  11.8341
;
run;
ods rtf style=journal;
proc reg data= insurance;
model y=x1 x2 x3 x4 x5;
run;
ods rtf close;
```

2.3.2. Procedure for bootstrap for case resampling n = 100

```
/* And finally we use a bootstrap with case resampling */
   ods listing close;
   procsurveyselect data=general out=boot1 method=urs
   samprate=1outhits rep=100;
   run;
```

2.3.4. Fuzzy bootstrap linear regression (alternative)

```
Data insurance;
input y x1 x2 x3 x4 x5;
datalines;
11.5745  8.8133  11.4424    12.0755  11.5484  10.2443
12.1891  9.5173  11.9274  13.5917   12.1562    11.3366
   ⋮        ⋮        ⋮        ⋮        ⋮        ⋮
12.4004  9.8467    12.2371  13.1705  12.3107  11.6928
12.2432  9.8088    12.3660  13.1512  12.2304  11.8341
;
run;
ods rtf style=journal;
proc optmodel;
set j= 1..30;
number y{j}, x1{j}, x2{j}, x3{j}, x4{j}, x5{j};
read data insurance into [_n_] y x1 x2 x3 x4 x5;

/*Print y x1 x2 x3 x4 x5*/
   print y x1 x2 x3 x4 x5;
   number n init 30; /*total of observation*/
/*Decision Variable Bounded or Not Bounded*/
   var aw{1..6}>=0;/*these six variables are bounded*/
   var ac{1..6};      /*these six variables are not bounded*/
/*Objective Function*/
   min z1= aw[1] * n + sum{i in j} x1[i] * aw[2]+sum{i in j}
   x2[i]*aw[3]+sum{i in j} x3[i] * aw[4]+sum{i in j} x4[i]*
   aw[5]+sum{i in j} x5[i] * aw[6];
/*Linear Constraints*/
   con c{i in 1..n}:
   ac[1]+x1[i]*ac[2]+x2[i]*ac[3]+x3[i]*ac[4]+x4[i]*ac[5]+x5[i]*a
   c[6]-aw[1]-x1[i]*aw[2]-x2[i]*aw[3]-x3[i]*aw[4]-x4[i]*aw[5]-
   x5[i]*aw[6]<=y[i];

   con c1{i in 1..n}:
   ac[1]+x1[i]*ac[2]+x2[i]*ac[3]+x3[i]*ac[4]+x4[i]*ac[5]+x5[i]*a
   c[6]+aw[1]+x1[i]*aw[2]+x2[i]*aw[3]+x3[i]*aw[4]+x4[i]*aw[5]
   +x5[i]*aw[6]>=y[i];

/*This Provides All Equations*/
   expand; Solve;
   print ac aw;
   quit;
   ods rtf close;
```

3. Results and discussion

Below are the results of the analysis, which using the above algorithm. From Table 1 and Table 2 using the multiple linear regression methods, there is only one significant variable for general

insurance. While, only two significant variables for life insurance. In contrast, using the method bootstrap linear regression, all the variables contained in the study are significant for both insurances.

Table 1: Parameter estimate by multiple linear regression for general and life insurance data

	Parameter Estimates					
	Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
General	Intercept	1	0.38809	0.31025	1.25	0.2225
	*x*1	1	0.07217	0.06250	1.15	0.2591
	*x*2	1	-0.00498	0.02588	-0.19	0.8489
	*x*3	1	-0.05960	0.06480	-0.92	0.3665
	*x*4	1	0.94569	0.03511	26.94	<.0001
	*x*5	1	0.04121	0.04465	0.92	0.3648
Life	Intercept	1	1.35773	0.50536	2.69	0.0124
	*x*1	1	0.35369	0.10728	3.30	0.0028
	*x*2	1	0.25522	0.40748	0.63	0.5365
	*x*3	1	0.37671	0.38812	0.97	0.3407
	*x*4	1	-0.08891	0.02622	-3.39	0.0022
	*x*5	1	0.06315	0.04308	1.47	0.1547

Table 2: Parameter estimate by fuzzy bootstrap regression for general and life insurance data

	Parameter Estimates					
	Variable	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
General	Intercept	1	0.33154	0.02179	15.22	<.0001
	*x*1	1	0.07792	0.00438	17.81	<.0001
	*x*2	1	-0.00580	0.00185	-3.13	0.0017
	*x*3	1	-0.05765	0.00457	-12.63	<.0001
	*x*4	1	0.93892	0.00245	382.89	<.0001
	*x*5	1	0.04657	0.00310	15.03	<.0001
Life	Intercept	1	1.23973	0.03729	33.25	<.0001
	*x*1	1	0.30662	0.00795	38.55	<.0001
	*x*2	1	0.17056	0.03029	5.63	<.0001
	*x*3	1	0.50701	0.02904	17.46	<.0001
	*x*4	1	-0.10113	0.00198	-51.01	<.0001
	*x*5	1	0.07374	0.00323	22.86	<.0001

3.1. The fitted model for multiple linear regression

$$y = 0.2587 + 0.0025\,x1 + 0.003\,x2 + 0.0009\,x3 + 0.922\,x4 + 0.0579\,x5$$

Standard Error
(0.0166) (0.0038) (0.0015) (0.0037) (0.002) (0.0025)

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

$$Y = (0.2587 + 0.0166) + (0.0025 + 0.0038)\,x1 + (0.003 + 0.0015)\,x2 + (0.0009 + 0.0037)\,x3 + (0.922 + 0.002)\,x4 + (0.0579 + 0.0025)\,x5$$

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

$$Y = (0.2587 - 0.0166) + (0.0025 - 0.0038)\,x1 + (0.003 - 0.0015)\,x2 + (0.0009 - 0.0037)\,x3 + (0.922 - 0.002)\,x4 + (0.0579 - 0.0025)\,x5$$

3.2. The fitted model for fuzzy bootstrap regression an alternative method

While using the bootstrap procedure, we will obtain differ output the ac and aw of one the generated out as shown above.

$$Y = 1.307 + 0.424\,x1 - 0.0478\,x2 - 0.4355\,x3 + 0.918\,x4 + 0.1637\,x5$$

ac1=1.307	ac2=0.424	ac3=-0.048	ac4=-0.435	ac5=0.918	ac6=0.164
aw1=0	aw2=0	aw3=-0	aw4=-0	aw5=0	aw6=0.006

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

$$Y = [1.307 + 0] + [0.424 + 0]\,x1 + [-0.0478 + 0]\,x2 + [-0.4355 + 0]\,x3 + [0.918 + 0]\,x4 + [0.1637 + 0.0058]\,x5$$

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

$$Y = [1.307 - 0] + [0.424 - 0]\,x1 + [-0.0478 - 0]\,x2 + [-0.4355 - 0]\,x3 + [0.918 - 0]\,x4 + [0.1637 - 0.0058]\,x5$$

The next step is to compare the performance of multiple linear regression and fuzzy bootstrap regression (alternative) model. Manually, we computer two values, the width of prediction intervals in respect of multiple linear regression model and the other that of fuzzy regression model for each set of observed explanatory variables is computed manually. From Table 3 and Table 4, average width general and life insurance for former multiple regression was found to be 6.028 and 29.109 while using fuzzy bootstrap regression, the average width general and life insurance is 0.137, and 0.179 this indicated that the superiority of fuzzy bootstrap regression methodology. From this analysis, the most efficient method to obtained relationship between response and explanatory variable is to apply fuzzy bootstrap regression method compared to the linear regression method.

Table 3: Average width for former multiple linear regression model

Insurance	Average Lower Limit	Average Upper Limit	Width
General	8.994	15.022	6.028
Life	0.399	29.508	29.109

Table 4: Average width for fuzzy bootstrap regression model

Insurance	Average Lower Limit	Average Upper Limit	Width
General	11.964	12.101	0.137
Life	14.869	15.048	0.179

4. Conclusion

This paper explained on how to combine an algorithm between fuzzy regression and bootstrap method. This research paper contains 30 observations with six variables. A simulation study from [5] shows that the inference based on bootstrap standard error estimates may be much more accurate in small samples. According to general and life insurance data, all independent variables in this case were significantly to the profitability of general insurance companies. Without using robust and bootstrapping, the result shows that only one and two out of five variables were significant. Besides that, from the average width for general and life insurance are small compared to the multiple linear regression models. These techniques have been compared in terms of accuracy. Research demonstrates that fuzzy multiple regression models are better than linear regression equations and fuzzy models. These findings are supported by [11] an article titled Fuzzy Multiple Regression Model for Estimating Software Development Time. The goal of their research is to study models for estimating software projects. They also found that Fuzzy Multiple Regression approaches have the higher accuracy than other methods for estimation. This algorithm provides us with the improved understanding of the modified method and underlying of relative contributions. For the further study, it is possible to approach response surface methodology for every each of significant variables in the single algorithm.

Acknowledgement

The project was funded by the Universiti Sains Malaysia (USM) (Rui Grant No.1001/ppsg/8012278, School of Dental Sciences, Health Campus).

References

[1] Ahmad WMAW, Nawi MAA & Aleng NA (2013), Relative efficiency analysis industry of life and general insurance in Malaysia

using Stochastic Frontier Analysis (SFA). Applied Mathematical Sciences, 7(23), 1107-1118.
[2] Alan OS (1993), An introduction to regression analysis. Coase-Sandor Institute for Law and Economics Working Paper No. 20.
[3] Bargiela A, Pedrycz W & Nakashima T (2007), Multiple regression with fuzzy data. Fuzzy Sets and Systems, 158(19), 2169-2188.
[4] Efron B & Tibshyrani RJ (1993), An introduction to the bootstrap. Chapman and Hall.
[5] Goncalves S & White H (2005), Bootstrap standard error estimates for linear regression. J. Am. Stat. Assoc., 100, 970-979.
[6] Hall P (1992), The bootstrap and edgeworth expansion. Springer Verlag.
[7] Hoffmann JP (2010), Linear regression analysis: Applications and assumptions. Brigham Young University.
[8] Kacprzyk J & Fedrizzi M (1992), Fuzzy regression analysis. Omnitech Press.
[9] Kim KJ, Moskowitz H & Koksalan M (1996), Fuzzy versus statistical linear regression. European Journal of Operation Research, 92(2), 417-437.
[10] Tanaka H, Uejima S & Asai K (1982), Linear regression analysis with fuzzy model. IEEE Transactions on Systems, Man and Cybernetics, 12(6), 903-907.
[11] Marza V & Seyyedi MA (2009), Fuzzy multiple regression model for estimating software development time. International Journal of Engineering Business Management, 1(2), 31-34.