IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 3 Ver. III (May. - Jun. 2016), PP 29-36
www.iosrjournals.org

Modified Simple Fuzzy Nonparametric Linear Regression Method: A Technical Approach

Wan Muhamad Amir W Ahmad[1], Nurfadhlina Halim[2], Nor Azlida Aleng[3], Ruhaya Hasan[4], Zalila Ali[5], Kasypi Mokhtar[6] and Syerrina Zakaria[7]

[1, 4] *School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kelantan, Malaysia*

[5] *School of Mathematics Sciences, Universiti Sains Malaysia (USM), 11800 Minden, Pulau Pinang, Malaysia.*

[2, 3, 7] *School of Informatics and Applied Mathematics,* [6]*School of Maritime Business and Management, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu*

Abstract: *This paper focused on the algorithmic building for a simple alternative nonparametric fuzzy linear regression method through SAS algorithm. This modified method can be used as an alternative method of data analysis. This modified method comprises of normality checking, transforming data into normality, data bootstrapping and fuzzy regression modelling as improvement of regression modelling. In this paper, we provide an alternative algorithm which can be used as a tool for the option of data analyzed.*

Keywords: *Bootstrap, Bayesian and Fuzzy Regression.*

I. Introduction To The Models

Linear Regression (LR) analysis is a common used technique in data analysis. This technique can be applied to forecast the value of the response variables (dependent) when given any value of the predictor variables (independent variables). A general regression model is given by $y_i = E(y_i \mid x_i) + e_i$, where i=1, 2, 3,…, n denoting an observation of a subject is the response variables and x_i is a $k \times 1$ vector of independent variables. $E(y_i \mid x_i)$ is the expectation of y_i conditional on x_i, and e_i is the error term. This paper provides an algorithm for simple linear regressions (SLR) in SAS (Diem Ngo & La Puente, 2012).

Data transformation tools are commonly-used to improve normality of a distribution and equalizing variance to meet the assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparing for statistical analyses. The traditional transformations have commonly been discussed include: adding constants, square root, converting two logarithmic scales, inverting and reflecting and applying trigonometric transformations such as sine wave transformations (Osborne, 2010). The study uses the Box-Cox transformation. The form of Box-Cox transformation as below:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log y & if\ \lambda = 0 \end{cases}$$

where, y is the observation data and λ is the model parameter. The example for application of the method discussed by using SAS language computer software is provided (Osborne, 2010).

The bootstrap methods begin with an original data or sample that taken from the population, then calculates sample statistics. The next step is to copy the original sample several times to create a pseudo population with replacement by using the empirical density function (EDF), (Efron, Bradley and Tibshirani, 1993). The benefit of using bootstrap is its capability to develop a same size of the original sample that may include an observation several times while omitting other observations. Bootstrap methods draw the samples with replacement, and it calculates statistics for each sample (it stores these statistics and creating a distribution for further analysis). After finalizing the bootstrap, the data is analyzed for mean, standard deviation, confidence intervals, and any other evidence of replication (Cassel, 2010; Jung, Jhun, & Lee, 2005; Higgins, 2005).

In applying the bootstrap method, the original findings from the empirical test were replicated several times to meet research requirement. As an example, for 1000 observations (original data), the analysis is performed by using statistical linear model. The analysis results of beta coefficients and r-squared are obtained, then, the bootstrapping method is applied to the selected data. In applying bootstrapped method, a sample of 100 observations, then replicates 10 times (this is equal to 1000 observations). The analysis of statistical, linear model, the beta coefficients and r-squared values of bootstrap method were compared to the original results. The bootstrap method findings depict the average beta coefficients and r-squared values are similar to the original findings, from where it was replicated. Surprisingly, the bootstrap method provides another noble opportunity to further comprehensive study of science and non-science discipline.

Theil proposes a method for obtaining a point estimate of the slope coefficient β (which is known as Theil's slope estimator). We assume that the data conform to the classic regression model.

$$y_i = \alpha + \beta x_i + e_{ij} \qquad i = 1,...,n$$

Where the x_i are known constant, α and β known parameter, and y_i is an observed value of the continuous random variable *y* at x_i. For each value of x_i, we assume a subpopulation of *y* values, and the e_i are mutually independent. The x_i are all distinct (no ties), we take $x_1 < x_2 < ... < x_n$. The data consist of *n* pairs of sample observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where the *i*th pair represents measurements taken on the *i*th unit of association. To obtain Theil's estimator of β, we first form all possible sample slopes $S_{ij} = (y_j - y_i)/(x_j - x_i)$, where $i < j$. There will be $N = {}^{n}C_2$ values of S_{ij}. The estimator of β, which designate by $\hat{\beta}$, is the median of S_{ij} values.

A fuzzy regression model can be written as $Y = Z_0 + Z_1 x_1 + Z_2 x_2 + \ldots + Z_k x_k$, here the explanation variables $x_i's$ are assumed to be precise. However, according to the equation above, response variable *Y* is not crisp but is instead fuzzy in nature. That means the parameters are also fuzzy in nature. Our aim is to estimate these parameters. In further discussion, $Z_i's$ are assumes symmetric fuzzy numbers which can be presented by interval. For example, Z_i can be express as fuzzy set given by $Z_i = < a_{ic}, a_{iw} >$ where a_{ic} is centre and a_{iw} is radius or vagueness associated. Fuzzy set above reflects the confidence in the regression coefficients around a_{ic} 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. So, the relationship is also considered to be fuzzy. This $Z_i = < a_{ic}, a_{iw} >$ can be written as $Z_1 = [a_{1L}, a_{1R}]$ with $a_{1L} = a_{1c} - a_{1w}$ and $a_{1R} = a_{1c} - a_{1w}$. In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model. $y_j = Z_0 + Z_1 x_{1j} + Z_2 x_{2j} + \ldots + Z_k x_{kj}$. Using $Z_i = < a_{ic}, a_{iw} >$, it can be written $y_j = < a_{0c}, a_{0w} > + < a_{1c}, a_{1w} > x_{1j} + \ldots + < a_{nc}, a_{nw} > x_{nj}$. $= < a_{jc}, a_{jw} >$. Thus this can be written as $y_{jc} = a_{0c} + a_{1c} x_{1j} + \cdots + a_{nc} x_{nj}$ then it can be written straightly as $y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}|$. As y_{jw} represent radius and so cannot be negative, therefore on the right-hand side of equation $y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}|$, absolute values of x_{ij} are taken. Suppose there *m* data point, each comprising $a(n+1) - row$ vector. Then parameters Z_i are estimated by minimizing the quantity, which is total vagueness 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 $\sum_{j=1}^{m} \left(a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}| \right)$ and subject to

$$\left\{ \left(a_{0c} + \sum_{i=1}^{n} a_{ic} x_{ij} \right) + \left(a_{0w} + \sum_{i=1}^{n} a_{iw} x_{ij} \right) \right\} \geq Y_j \text{ and}$$

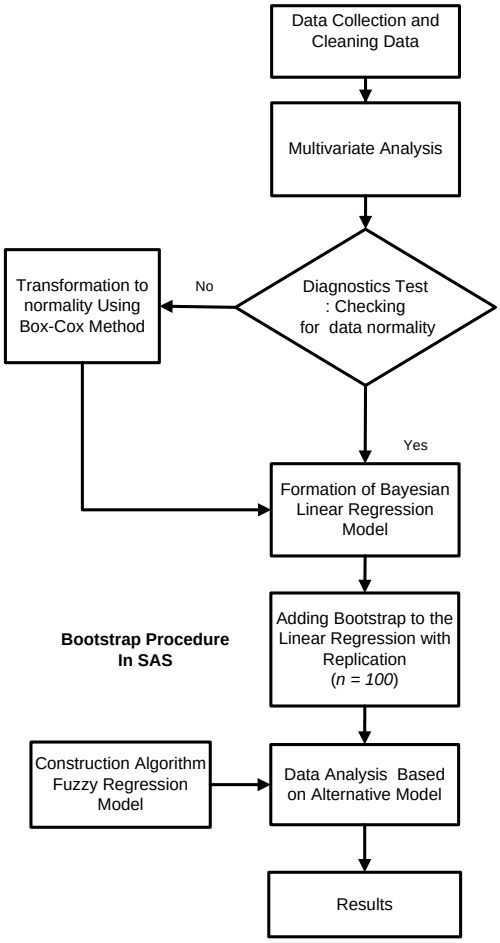
$$\left\{ \left(a_{0c} + \sum_{i=1}^{n} a_{ic} x_{ij} \right) - \left(a_{0w} + \sum_{i=1}^{n} a_{iw} x_{ij} \right) \right\} \leq Y_j .$$

and $a_{iw} \geq 0$. Simple procedure is commonly used to solve the linear programming problem. (Kacprzyk and Fedrizzi, 1992). Data of this study is a sample which composed of three variables.

Table1. Description of data among children in Bachok, Kelantan, Malaysia

Num.	Variables	Explanation of user variables
1.	Caries	Number of caries
2.	Ascore	Number of Score

1.1 Algorithm and Flow Chart for Nonparametric Fuzzy Regression



1.1 Flow Chart of Simple Nonparametric Fuzzy Regression Modeling

PART I:

```
 /* Part one of this analysis*/
/* Normality checking residual analysis, probability plot.*/

Data tooth1;
Input caries Ascore;
Cards;
15    5
15    6
11    5
 6    3
 5    4
 5    2
 5    3
 5    2
 5    3
 4    4
 4    4
 3    3
 2    3
 1    4
 1    3
 ;
run;
```

```
ods rtf file='abc.rtf' style=journal;

/*Title "Simple Linear Regression Model with no options"*/
proc reg data= tooth 1;
model caries = Ascores;
output out=temp r=residual;
proc univariate data=temp normal;

/*"Checking Residuals for Normality"*/
var residual;
histogram / normal;
qqplot residual /Normal(mu=est sigma=est color=red l=1);
run;

/*Calculation of Theil's Slope estimator. Theil proposed a method for obtaining point estimate of the slope estimation beta*/
Data tooth1;
array Ascores(8) Ascores1- Ascores8;
array caries(8) caries1-caries8;
do I = 1 to 8;
input caries(I) Ascores(I);
end;
output;
cards;
15   5
15   6
11   5
 6   3
 5   4
 5   2
 5   3
 5   2
 5   3
 4   4
 4   4
 3   3
 2   3
 1   4
 1   3
;
 run;


Data tooth2;
set tooth1;
array  Ascores(8) Ascores1-Ascores8;
array  caries(8)  caries1- caries8;
    do I=1 to 8;
do J=I+1 to 8;
Slope = (caries(J)-caries(I))/(Ascores(J)- Ascores(I));
output; end; end;
keep slope;
proc sort;
by slope;
run;

proc print;
title 'Theil's Slope estimator';
run;
```

```
proc means median;
var slope;
run;
```

PART2:

```
Data npreg3;
set npreg1;
    ARRAY Ascores(8) Ascores1- Ascores8;
    ARRAY  caries(8)  caries1- caries8;
```

/* when one assumes that the error terms are not symmetric about 0 */

```
do I = 1 to 8;
inter1 = caries(I)- 2.5000*Ascores(I);
```

/* 2.5000000 is the median of the slopes */

```
output;
end;
keep inter1;

Proc print;
var inter1;
run;

Proc means median;
var inter1;
run;
ods rtf close;
run;
```

We used the independent variables from the original data to estimate the outcome of dependent variable, then once again we estimate the dependent variable by using the estimate independent variables in order to suit the nonparametric equation. Estimation value using nonparametric regression which is given by this equation. For caries variables: $caries = -0.75 + 2.500\ Ascores$. For Ascore variables : $Ascores = \dfrac{caries + 0.75}{2.500}$.

PART 3:

```
Data tooth1;
input cariesf Ascores;
Datalines;
 11.75    6.3
 14.25    6.3
 11.75    4.7
  6.75    2.7
  9.25    2.3
  4.25    2.3
  6.75    2.3
  4.25    2.3
  6.75    2.3
  9.25    1.9
  9.25    1.9
  6.75    1.5
  6.75    1.1
  9.25    0.7
  6.75    0.7
 ;
 run;

ods rtf file='result_ex1.rtf' ;
```

```
proc optmodel;

set j= 1..15;

number cariesf{j}, Ascores{j};
read data tooth1 into [_n_]  cariesf Ascores;
/*Print cariesf Ascores */
print cariesf Ascores;

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

/* Decision Variables*/
var aw{1..2}>=0; /*Theses two variables are bounded*/
var ac{1..2};    /* These two variables are not bounded*/

/* Objective function*/
min z1= aw[1] * n + sum{i in j} Ascores[i] * aw[2] ;

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

con c1{i in 1..n}: ac[1]+Ascores[i]*ac[2]+aw[1]+Ascores[i]*aw[2] >= cariesf[i];

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

/* To compare the performance of the predicted values, let's compare the means of their residuals as follows ***/**

```
Data tooth1;
input caries Ascores;
   pred1 = -4.18+ 2.77*caries;
   resid1 = Ascores - pred1;

   np_pred2 = -0.75 + 2.500*caries;
   resid2 = Ascores - np_pred2;

   np_pred3 = 6.375 + 1.250*caries;
   resid3 = Ascores - np_pred3;

   np_pred4 = 1.375 + 1.250*caries;
   resid4 = Ascores - np_pred4;

 cards;
 15    5
 15    6
 11    5
 6     3
 5     4
 5     2
 5     3
 5     2
 5     3
 4     4
 4     4
 3     3
```

```
 2     3
 1     4
 1     3
run;

proc means;
  var resid1 resid2 resid3 resid4;
run;
```

II. Results

i. Simple Parametric Regression Model
$caries = -4.18 + 2.77\,Ascores$

ii. Simple Non-Parametric Regression Model $caries = -0.75 + 2.500\,Ascores$

iii. Simple Fuzzy Non-Parametric Regression Model
$caries = <3.875, 2.5> + <1.250, 0>\,Ascores$

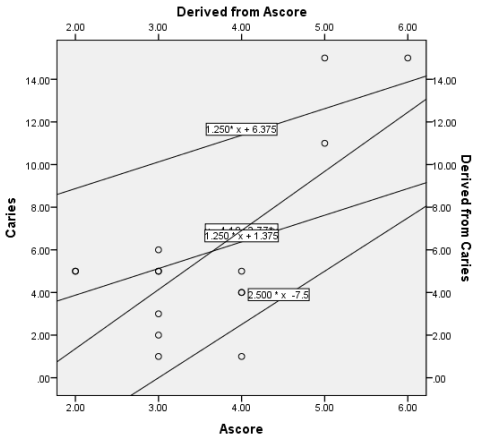
To compare the performance of the predicted values, let's compare the means of their residual as follows:

```
 pred1 = -4.18+ 2.77*caries;
  resid1 = Ascores - pred1;

  np_pred2 = -0.75 + 2.500*caries;
  resid2 = Ascores - np_pred2;

  np_pred3 = 6.375 + 1.250*caries;
  resid3 = Ascores - np_pred3;

  np_pred4 = 1.375 + 1.250*caries;
  resid4 = Ascores - np_pred4;
```

Output:

Variable	N	Mean	Std Dev	Minimum	Maximum
resid1	15	-8.28600	11.5037005	-32.37000	5.41000
resid2	15	-10.15000	10.3115746	-31.75000	2.25000
resid3	15	-10.02500	4.8133000	-20.12500	-3.62500
resid4	15	-5.02500	4.8133000	-15.12500	1.375000

In this case, seem that the nonparametric fuzzy regression equation predicts the values of caries more closely than that of the nonparametric regression.



III. Summary And Discussion

This paper gives the explanation for a modified linear regression Analysis through SAS algorithm. Our aim for this algorithm is to provide the researcher with the alternative programming, that suit for the case of small sample size and also the assumption of linear model are not met. This method can be applied to the small sample size data, especially where the data is very difficult to collect especially in dental public health.

References

[1]. Diem Ngo, T.H., La Puente, C.A. (2012). The Steps to Follow in a Multiple Regression Analysis. SAS Global Forum 2012: Statistics and Data Analysis. Paper 333-2012, Pp 1-12.

[2]. Cassel, D.L., 2010. Bootstrap Mania: Re sampling the SAS. SAS Global Forum 2010: Statistics and Data Analysis. Paper 268-279 In: Proceedings of the SAS Global Forum 2010 Conference. Cary (NC): SAS Institute Inc.

[3]. Jung, B.C., Jhun, M., Lee, J.W., 2005. Bootstrap Tests for Overdispersion in a Zero-Inflated Poisson Regression Model. Biometrics 61, pp.626-629.

[4]. Kacprzyk J. and Fedrizzi M. (1992) Fuzzy Regression Analysis, Omnitech Press, Warsaw.

[5]. Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall.

[6]. Higgins, G. E. (2005). Statistical Significance Testing: The Bootstrapping Method and an Application to Self-Control Theory. The Southwest Journal of Criminal Justice. Vol 2(1).pp 54-76

[7]. Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Pres.

[8]. Osborne, J.W., 2010. Improving your data transformations?: Applying the Box-Cox transformation. Practical Assessment, Research & Evaluation, 15(12): 1-9