Journal of Scientific Research and Development 3 (7): 14-19, 2016
Available online at www.jsrad.org
ISSN 1115-7569

Modified bayesian regression modeling involving qualitative predictor variables: A tumor size study

Wan Muhamad Amir W Ahmad [1,*], Nor Affendy Nor Azmi [1], Nor Azlida Aleng [3], Mohamad Shafiq Bin Mohd Ibrahim [1], Ruhaya Hasan [1], Zalila Ali [2], Wan Muhammad Luqman Bin Wan Rosdi [1], Masitah Hayati Harun [1]

[1]*School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), Malaysia*
[2]*School of Mathematics Sciences, Universiti Sains Malaysia (USM), Malaysia*
[3]*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT), Malaysia*

Abstract: This paper focused on the modified algorithm of Bayesian Linear Regression (BLR) method through SAS algorithm which is involved qualitative predictor variables. This modified method can be utilized as an alternative method for data analysis (regression modeling) in biostatistics. This modified method comprises of qualitative predictor variables, normality checking of residual, bootstrapping method and improvement of Fuzzy Bayesian Linear Regression Modeling (FBLR).

Key words: Bootstrap; Bayesian and fuzzy regression

1. Introduction

Bootstrap Methods

The bootstrap methods start with an original data or sample which taken from the population; then its' 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 et al., 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 draws the samples with replacement, and it calculates statistics for each sample (it stores these statistics and creating a distribution for further analysis). After finalized the bootstrap, the data is analyzed for mean, standard deviation, confidence intervals, and any others evidence of replication (Cassel, 2010; Jung et al., 2005; Higgins, 2005). In applying the bootstrap method, the original findings from the empirical test were replicated several times to meet research requirement. In applying bootstrapped method, a sample of 23 observations was then replicate 6 times (this is equal to 115 observations). The analysis from 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 for further comprehensive study from science and non-science discipline.

Bayesian linear regression (BLR)

Bayesian Linear Regression (BLR) analysis is a Bayesian approach to linear regression in which the statistics analysis is undertaken within the context of Bayesian inference. 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. y_i 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 (Diem Ngo and La Puente, 2012). This paper provides an algorithm for Bayesian Multiple Linear Regressions (BMLR) in SAS language

Assume a BLR model where the response vector y has dimension $n \times 1$ and follow a multivariate Gaussian distribution with mean $X\beta$ and covariance matrix $\sigma^2 I$, where X the design matrix is has dimension, $n \times p$, β contains the p regression coefficients, σ^2 is the common variance of the observational and I is a $n \times n$ identity matrix. That is, $y \sim N(X\beta, \sigma^2)$. In the Bayesian approach, the data are supplemented with additional information in the form of a prior probability distribution. The prior belief about the parameters is combined with the data's likelihood function according to Bayes theorem to yield the posterior belief about the parameters β and σ (Gelman et al., 2013; Gelman and Hill, 2006).

* Corresponding Author.

Qualitative predictor variables as indicator variables in multiple linear regressions

There is a lot of quantitatively identifying the classes of a qualitative predictor variable. Usually, we use indicator variables that take on the values 0 and 1. These indicator variables are simple to use and are widely employed, but they are no means the only way to quantify a qualitative variable (Neter et al., 1996)

Qualitative predictor variables with *a* class will be represented by *a*-1 indicator variables, each taking on the values 0 and 1. Indicator variables are frequently also called as dummy variables or binary variables. If a qualitative variable has more than two classes, it requires additional indicator variables in the regression model. Let say we have (*Y*), (X_1) and (X_2),

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (1)$$

where (X_{i2}) is the variables with four classes (let's say if the class were A, B, C and D). We therefore require three indicator variables. Let us define them as follows:

$$X_2 = \begin{matrix} 0 & \text{if class A} \\ 1 & \text{Otherwise} \end{matrix}$$

$$X_3 = \begin{matrix} 0 & \text{if class B} \\ 1 & \text{Otherwise} \end{matrix}$$

$$X_4 = \begin{matrix} 0 & \text{if class C} \\ 1 & \text{Otherwise} \end{matrix}$$

A first-order regression model is given by as follows:

$$Y = \beta_0 + \beta_1 X_{i1} + +\beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

For this model, the data input for the *X* variables would be as follows:

Model	X_1	X_2	X_3	X_4
A	X_{i1}	1	0	0
B	X_{i1}	0	1	0
C	X_{i1}	0	0	1
D	X_{i1}	0	0	0

The response function for regression model is

$$E\{Y\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

To understand the meaning of the regression coefficient, we first consider what response function becomes: For the model D, with $X_2 = 0$, $X_3 = 0$ and $X_4 = 0$. So, we will obtained the equation as $E\{Y\} = \beta_0 + \beta_1 X_1$. For the model A, with $X_2 = 1$, $X_3 = 0$ and $X_4 = 0$. So, we will obtain the equation as $E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$. For model B for which $X_2 = 0$, $X_3 = 1$ and $X_4 = 0$. So, we obtained the equation as $E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1$. For model C for which $X_2 = 0$, $X_3 = 1$ and $X_4 = 0$. So, we will obtained the equation as $E\{Y\} = (\beta_0 + \beta_4) + \beta_1 X_1$

Fuzzy regression model

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 $Z_i = [a_{iL}, a_{iR}]$ with $a_{iL} = a_{ic} - a_{iw}$ and $a_{iR} = a_{ic} - a_{iw}$. 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 $\left\{\left(a_{0c} + \sum_{i=1}^{n} a_{ic} x_{ij}\right) + \left(a_{0w} + \sum_{i=1}^{n} a_{iw} x_{ij}\right)\right\} \geq Y_j$ 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 five variables.

In our case, qualitative predictor variables is variable number 4.

From the Table 1, the qualitative predictor variables are given by "TumorSite" and "Gender". Because "Gender" variable is one category, so we do not need to separate them. We have to recode back

for the "TumorSite" variables because it has more than one category.

Table 1: Description of tumor data

Num.	Variables	Explanation of user variables
1.	Sizetumor	Tumor size
2.	Age	Age in year
3	Gender	0 = Female 1 = Male
4	TumorSite	0 = Gum 1 = Tongue 2 = Lip 3 = Cheek

TumorSite { 0 = Gum; 1 = Tongue; 2 = Lip; 3 = Cheek

Qualitative predictor variables with 4 classes as above will be represented by 4-1 = 3 indicator variables. Table 3 shows the three variables that will be used in the regression model.

Table 2: Qualitative predictor variables

5.	Gump	0 = Gum 1 = Else
6.	Cheek	0 = Cheek 1 = Else
7.	Tongue	0 = Tongue 1 = Else

Table 3: Description of tumor data with qualitative predictor variables

Num.	Variables	Explanation of user variables
1.	Sizetumorbayes	Reading of Tumor
2.	Age	Age in year
3	Gender	0 = Female 1 = Male
4.	Gump	0 = Gum 1 = Else
5.	Cheek	0 = Cheek 1 = Else
6.	Tongue	0 = Tongue 1 = Else

Algorithm and Flow Chart for Modified Bayesian Linear Regression Analysis Method

Algorithm for Modified Bayesian Linear Regression Analysis using SAS software.

Data tumor;
Input Sizetumorbayes Age Gender Gump cheek tongue;
Cards;

```
4.43    80    0    1.00    0.00    0.00
4.43    80    0    1.00    0.00    0.00
4.43    80    0    1.00    0.00    0.00
2.11    66    1    0.00    0.00    1.00
1.72    48    1    0.00    0.00    1.00
1.72    48    1    0.00    0.00    1.00
⋮   ⋮   ⋮   ⋮   ⋮   ⋮
1.57    49    0    0.00    1.00    0.00
1.57    49    0    0.00    1.00    0.00
1.57    49    0    0.00    1.00    0.00
⋮   ⋮   ⋮   ⋮   ⋮   ⋮
1.50    38    1    0.00    0.00    1.00
2.03    62    1    0.00    0.00    1.00
2.03    62    1    0.00    0.00    1.00
```

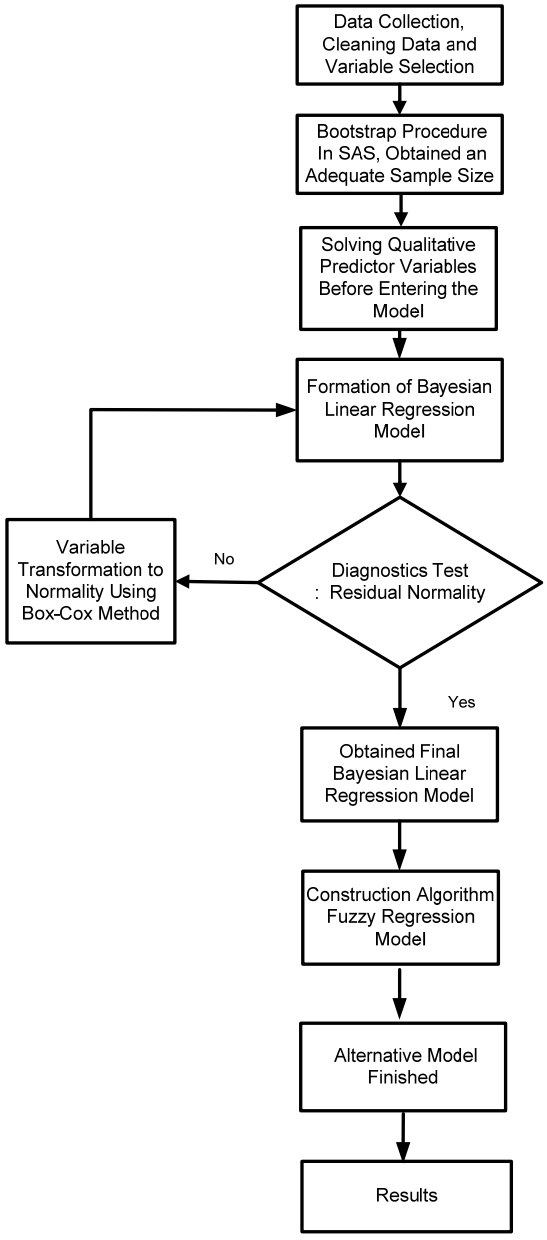

Fig. 1: Modified Bayesian linear regression analysis

```
1.86    62    0    0.00    1.00    0.00
;
ods rtf file='abc.rtf' style=journal;

/*ADDING BOOTSTRAPPING ALGORITHM TO THE METHOD */
Title "Performing bootstrap with case resampling";
Proc surveyselect data=tumor out=boot1 method=urs samprate=1 outhits rep=6;
run;

ods rtf close;

/* RESIDUAL NORMALITY CHECKING*/
Data tumor;
```

```
Input Sizetumorbayes Age Gender Gump cheek tongue;
Cards;
4.43    80    0    1.00    0.00    0.00
4.43    80    0    1.00    0.00    0.00
4.43    80    0    1.00    0.00    0.00
2.11    66    1    0.00    0.00    1.00
1.72    48    1    0.00    0.00    1.00
1.72    48    1    0.00    0.00    1.00
⋮   ⋮   ⋮   ⋮   ⋮   ⋮
1.57    49    0    0.00    1.00    0.00
1.57    49    0    0.00    1.00    0.00
1.57    49    0    0.00    1.00    0.00
⋮   ⋮   ⋮   ⋮   ⋮   ⋮
1.50    38    1    0.00    0.00    1.00
2.03    62    1    0.00    0.00    1.00
2.03    62    1    0.00    0.00    1.00
1.86    62    0    0.00    1.00    0.00
;
ods rtf file='abc.rtf' style=journal;

ods graphics on;
proc reg data=tumor plots=all;
  model Sizetumorbayes = Age Gender Gump cheek tongue
       output  out=Residuals
  p=y_hat
  r=y_res;
run;
ods graphics off;

ods graphics on;
proc reg data=tumor plots=all;
  model Sizetumorbayes = Age Gender Gump cheek tongue/p ;
run;
ods graphics off;
ods rtf close;
```

/* BAYESIAN REGRESSION MODEL*/

```
Data tumor;
Input Sizetumorbayes Age Gender Gump cheek tongue;
Cards;
4.43    80    0    1.00    0.00    0.00
4.43    80    0    1.00    0.00    0.00
4.43    80    0    1.00    0.00    0.00
2.11    66    1    0.00    0.00    1.00
1.72    48    1    0.00    0.00    1.00
1.72    48    1    0.00    0.00    1.00
⋮   ⋮   ⋮   ⋮   ⋮   ⋮
1.57    49    0    0.00    1.00    0.00
1.57    49    0    0.00    1.00    0.00
1.57    49    0    0.00    1.00    0.00
⋮   ⋮   ⋮   ⋮   ⋮   ⋮
1.50    38    1    0.00    0.00    1.00
2.03    62    1    0.00    0.00    1.00
2.03    62    1    0.00    0.00    1.00
1.86    62    0    0.00    1.00    0.00
;
run;
ods rtf file='abc.rtf' style=journal;

proc optmodel;
set j= 1..138;
Number Sizetumorbayes {j}, Age {j}, Gender {j}, Gump {j}, cheek {j}, tongue{j};
read data tumor into [_n_] Sizetumorbayes Age Gender Gump cheek tongue;


/*PRINT SIZETUMORBAYES AGE GENDER GUMP CHEEK TONGUE */
Print Sizetumorbayes Age Gender Gump cheek tongue;

/*TOTAL OF OBSERVATIONS*/
number n init 138;


/* DECISION VARIABLES BOUNDED OR NOT BOUNDED*/
/*THESE SIX VARIABLES ARE BOUNDED*/
var aw{1..6}>=0;

/*THESE SIX VARIABLES ARE NOT BOUNDED*/
var ac{1..6};

/* OBJECTIVE FUNCTION*/
min z1= aw[1]*n + sum{i in j}Age[i]*aw[2]+sum{i in j}Gender[i]*aw[3]
    +sum{i in j}Gump[i]*aw[4]+sum{i in j}cheek[i]*aw[5]+sum{i in j}tongue[i]*aw[6];

/*LINEAR CONSTRAINTS*/
con c{i in 1..n}:
   ac[1]+ Age[i]*ac[2]+ Gender[i]*ac[3]+ Gump[i]*ac[4]+ cheek[i]*ac[5]+ tongue[i]*ac[6]
   -aw[1]- Age[i]*aw[2]- Gender[i]*aw[3]- Gump[i]*aw[4]+ cheek[i]*aw[5]+tongue[i]*aw[6]
  <= Sizetumorbayes[i];

con c1{i in 1..n}:
   ac[1]+ Age[i]*ac[2]+ Gender[i]*ac[3]+ Gump[i]*ac[4]+ cheek[i]*ac[5]+ tongue[i]*ac[6]
   +aw[1]+ Age[i]*aw[2]+ Gender[i]*aw[3]+ Gump[i]*aw[4]+ cheek[i]*aw[5]+tongue[i]*aw[6]
   >= Sizetumorbayes[i];

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

2. Results and discussion

Results from Fitted model for Multiple Bayes Fuzzy Regression

Fitted model for fuzzy regression for Tumor Size is given by:

Tumor Size = (1.199107, 0.00339286)
+ (0.021786, 0.00000000) × Age
+ (-0.244107, 0.00089286) × Gender

+ (1.4850000, 0.0000000) × Gump
+ (-0.693214, 0.00000000) × cheek
+ (-0.278571, 0.00000000) × tongue .....(1)

Table 4: Value of center (AC) and radius (AW)

[1]	ac	aw
1	1.199107	0.00339286
2	0.021786	0.00000000
3	-0.244107	0.00089286
4	1.485000	0.00000000
5	-0.693214	0.00000000
6	-0.278571	0.00000000

Upper or lower limits of prediction interval are computed from the prediction equation (1) by taking the coefficient as their corresponding estimated values plus or minus standard error.

Upper limits

Tumor Size =
1.2025056 + (0.021786 × Age) +
(-0.24321414 × Gender) + (1.485000 × Gump) + (-0.693214 × cheek) +
(-0.278571 × tongue)…… (2)

Lowers limits

Tumor Size =
1.1957084 + (0.021786 × Age) +
(0.24499986 × Gender) + (1.485000 × Gump) + (-0.278571 × tongue)…… (3)

The width of prediction intervals in respect of bayesian multiple linear regression model and bayesian fuzzy regression model corresponding to each set of observed explanatory variables is computed in SPSS and the results are reported in the following Table 5.

From this table, average width was found to be 0.359823, indicating superiority of fuzzy regression methodology

Tumor Size =
(1.199107, 0.00339286)
+ (0.021786, 0.00000000) × Age
+ (-0.244107, 0.00089286) × Gender
+ (1.4850000.00000000) × Gump
+ (-0.693214, 0.00000000) × cheek
+ (-0.278571, 0.00000000) × tongue .....(1)

Bayesian Fuzzy Regression (FR) Model

Table 5: Fitting of fuzzy Bayesian regression

Lower limit	Upper limit	Width
4.4304	4.4236	0.0068
4.4304	4.4236	0.0068
4.4304	4.4236	0.0068
2.1186	2.6000	0.4814
1.7264	2.2079	0.4814
1.7264	2.2079	0.4814
2.3754	2.8568	0.4814
2.4225	2.4157	0.0068
2.1222	2.1154	0.0068
2.1215	2.8079	0.6864
⋮	⋮	⋮
1.9900	1.5086	0.4814
1.9900	1.5086	0.4814
2.5129	2.0315	0.4814
2.5129	2.0315	0.4814
2.5464	1.8600	0.6864
Average width 0.359823		

3. Summary and discussion

This paper gives an algorithm and also illustrated the procedure of modeling by using modified Bayesian linear regression with qualitative predictor variables through SAS language. Our aim for this paper is to share the algorithm and also to provide the researcher with the alternative programming, that suit for the case of small sample size. This proposed method can be applied to the small sample size data, especially where the data is very difficult to collect especially in public health.

Acknowledgements

The authors would like to express their gratitude to Universiti Sains Malaysia for providing the research funding (Grant no.304/PPSG/61313187, School of Dental Sciences, USM)

Conflict of Interest

The authors declare there is no conflict of interest regarding the publication of this paper..

References

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.

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.

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

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., andRubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, third edition.

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

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.

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

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. Applied Linear Statistical Models, 4th Edition, Richard D. Irwin, Inc., Burr Ridge, Illinois,

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