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 2, Mohamad Shafiq Bin Mohd Ibrahim 1, Ruhaya Hasan 1, Zalila Ali 2, Wan Muhammad Luqman Bin Wan Rosdi 1, Masitah Hayati Harun 1

1School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), Malaysia
2School of Mathematics Sciences, Universiti Sains Malaysia (USM), Malaysia
3School 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 \( y_i = E(y_i | x_i) + e_i \), where \( i = 1, 2, 3, \ldots, n \) denoting an observation of a subject. \( y \) is the response variables and \( x \) is a \( k \times 1 \) vector of independent variables. \( E(y_i | x_i) \) 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 \( n \times p \), \( \beta \) contains the \( p \) regression coefficients, \( \sigma \) is the common variance of the observational and \( I \) is a \( n \times n \) identity matrix. That is, \( y \sim N(X \beta, \sigma^2 I) \). 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 \( \beta \) and \( \sigma \) (Gelman et al., 2013; Gelman and Hill, 2006).
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 0-1 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 \( Y, (X_i) \) and \( (X_{ij}) \),

\[
Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
\]  

(1)

where \( \{X_{ij}\} \) 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:

<table>
<thead>
<tr>
<th>Model</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( X_{i1} )</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>( X_{i1} )</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>( X_{i1} )</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>( X_{i1} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 obtain the equation as

\[
E\{Y\} = \beta_0 + \beta_1 X_{i1}
\]

For the model A, with \( X_2 = 1 \) and \( X_4 = 0 \). So, we will obtain the equation as

\[
E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_{i1}
\]

For model B for which \( X_2 = 0, X_3 = 1 \) and \( X_4 = 0 \). So, we obtained the equation as

\[
E\{Y\} = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_{i1}
\]

For model C for which \( X_2 = 0, X_3 = 1 \) and \( X_4 = 0 \). So, we obtained the equation as

\[
E\{Y\} = (\beta_0 + \beta_2 + \beta_3 + \beta_4) + \beta_1 X_{i1}
\]

Fuzzy regression model

A fuzzy regression model can be written as

\[
Y = Z_1 + Z_2 X_1 + Z_3 X_2 + \ldots + Z_j X_j, \quad \text{here the explanation variables } X_1, \ldots, X_j \text{ are assumed to be precise. However, according to the equation above, response variable } Y \text{ 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_1, \ldots, Z_j \text{ are assumed symmetric fuzzy numbers which can be presented by interval. For example, } Z_1, \quad \text{can be express as fuzzy set given by } Z_1 = \langle a_{-1}, a_1 \rangle \text{ where } a_{+1} \text{ is centre and } a_{-1} \text{ is radius or vagueness associated. Fuzzy set above reflects the confidence in the regression coefficients around } a_{+1} \text{ in terms of symmetric triangular membership 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_1 = \langle a_{-1}, a_1 \rangle \text{ can be written as } Z_1 = \{a_{-1}, a_1\} \text{ with } a_{+1} = a_{+1} - a_{-1} \text{ and } a_{-1} = a_{+1} - a_{-1}. \text{ In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model. } y = Z_1 + Z_2 X_1 + Z_3 X_2 + \ldots + Z_j X_j \text{. Using } Z_1 = \langle a_{-1}, a_1 \rangle \text{ it can be written } y = \langle a_{-1}, a_1 \rangle + \langle a_{-1}, a_1 \rangle \times x_1 + \ldots + \langle a_{-1}, a_1 \rangle \times x_j. \text{ Thus this can be written as } y = a_{+1} X_1 + a_{-1} X_2 + \ldots + a_{-1} X_j \text{ and } y = a_{+1} \times x_1 + a_{-1} \times x_2 + \ldots + a_{-1} \times x_j. \text{ As } y, X_1, X_2, \ldots, X_j \text{ represent radius and so cannot be negative, therefore on the right-hand side of equation } y = a_{+1} X_1 + a_{-1} X_2 + \ldots + a_{-1} X_j, \text{ absolute values of } x_i \text{ are taken. Suppose there } m \text{ data point, each comprising } a(n + 1) \text{– row vector. Then parameters } Z_1, \ldots, Z_j \text{ 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 (a_{+1} X_1 + a_{-1} X_2 + \ldots + a_{-1} X_j) \text{ and subject to }

\[
\left\{ a_{+1} + \sum_{i=1}^n a_{+1} x_i \right\} \geq y_i \quad \text{and} \quad \left\{ a_{-1} + \sum_{i=1}^n a_{-1} x_i \right\} \leq y_i
\]

and \( a_{+1}, a_{-1} \geq 0 \). 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**

<table>
<thead>
<tr>
<th>Num.</th>
<th>Variables</th>
<th>Explanation of user variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sizetumor</td>
<td>Tumor size</td>
</tr>
<tr>
<td>2.</td>
<td>Age</td>
<td>Age in year</td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>0 = Female 1 = Male</td>
</tr>
<tr>
<td>4</td>
<td>TumorSite</td>
<td>0 = Gum 1 = Tongue 2 = Lip 3 = Cheek</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Num.</th>
<th>Variables</th>
<th>Explanation of user variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sizetumorbayes</td>
<td>Reading of Tumor</td>
</tr>
<tr>
<td>2.</td>
<td>Age</td>
<td>Age in year</td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>0 = Female 1 = Male</td>
</tr>
<tr>
<td>4.</td>
<td>Gump</td>
<td>0 = Gum 1 = Else</td>
</tr>
<tr>
<td>5.</td>
<td>Cheek</td>
<td>0 = Cheek 1 = Else</td>
</tr>
<tr>
<td>6.</td>
<td>Tongue</td>
<td>0 = Tongue 1 = Else</td>
</tr>
</tbody>
</table>

**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 0.00 1.00 2.03 62 1 0.00 0.00 1.00 0.00 2.03 62 1 0.00 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;
/*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
\ldots
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
\ldots
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/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
\ldots
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
\ldots
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}:
-aw[1]- Age[i]*aw[2]- Gender[i]*aw[3]-
<= Sizetumorbayes[i];
con c1{i in 1..n}:
>= 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) \times Age
+ (0.021786, 0.00000000) \times Gender
Tumor Size = 0.359823, indicating superiority of fuzzy regression following Table 5.

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.

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).

Acknowledgements

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

References


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.

Table 4: Value of center (c) and radius (aw)

<table>
<thead>
<tr>
<th>(i)</th>
<th>ac</th>
<th>aw</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.199107</td>
<td>0.00339286</td>
</tr>
<tr>
<td>2</td>
<td>0.021786</td>
<td>0.00000000</td>
</tr>
<tr>
<td>3</td>
<td>-0.244107</td>
<td>0.00089286</td>
</tr>
<tr>
<td>4</td>
<td>1.485000</td>
<td>0.00000000</td>
</tr>
<tr>
<td>5</td>
<td>-0.693214</td>
<td>0.00000000</td>
</tr>
<tr>
<td>6</td>
<td>-0.278571</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

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) + (-0.243214 × Gender) + (1.485000 × Gump) + (-0.693214 × cheek) + (-0.278571 × tongue) …..(1)

Table 5: Fitting of fuzzy Bayesian regression

<table>
<thead>
<tr>
<th>Limit</th>
<th>Upper limit</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>4.4304</td>
<td>0.0068</td>
</tr>
<tr>
<td>Limit</td>
<td>4.4304</td>
<td>0.0068</td>
</tr>
<tr>
<td>Limit</td>
<td>4.4304</td>
<td>0.0068</td>
</tr>
<tr>
<td>Limit</td>
<td>2.1186</td>
<td>0.4814</td>
</tr>
<tr>
<td>Limit</td>
<td>2.1764</td>
<td>0.4814</td>
</tr>
<tr>
<td>Limit</td>
<td>2.1764</td>
<td>0.4814</td>
</tr>
<tr>
<td>Limit</td>
<td>2.3754</td>
<td>0.4814</td>
</tr>
<tr>
<td>Limit</td>
<td>2.4225</td>
<td>0.0068</td>
</tr>
<tr>
<td>Limit</td>
<td>2.4225</td>
<td>0.0068</td>
</tr>
<tr>
<td>Limit</td>
<td>2.1221</td>
<td>0.6864</td>
</tr>
</tbody>
</table>

Average width 0.359823