Modeling using modified bayesian regression (Method: A tumour size study)

Wan Muhamad Amir W Ahmad 1,*, Nor Affendy Nor Azmi 1,Nor Azlida Aleng 2, Ruhaya Hasan 1, Zalila Ali 3, Masitah Hayati Harun 1

1School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kelantan, Malaysia
2School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu
3School of Mathematics Sciences, Universiti Sains Malaysia (USM), 11800 Minden, Pulau Pinang, Malaysia

Abstract: This paper focuses on the modified algorithm and the analysis of modified Bayesian Linear Regression (BLR) method through SAS algorithm. This modified method can be utilised as an alternative method of data analysis in biostatistics. This modified method comprises normality checking of residual normality, bootstrapping method and improvement of Bayesian Linear Regression Modeling (BLR). This proposed method can be applied to small sample size data, especially when limited data is obtained, for example in public health. In this paper, we illustrated the use of modified Bayesian Linear Regression (BLR) algorithm.

Key words: Bootstrap; Bayesian regression; Fuzzy regression; Tumour

1. Introduction to the models

Bayesian Linear Regression (BLR) analysis is an 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 \( y_i = \beta x_i + e_i \), where \( i = 1, 2, 3,...,n \) denoting an observation of a subject \( y \) is the response variable and \( x \) 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 Bayesian Multiple Linear Regressions (BMLR) in SAS (Diem Ngo and La Puente, 2012).

Assume a BLR model where the response vector \( y \) has dimension \( n \times 1 \) and follows a multivariate Gaussian distribution with mean \( X\beta \) and covariance matrix \( \sigma^2 I \), where \( X \) the design matrix has dimension \( n \times p \), \( \beta \) contains the \( p \) regression coefficients, \( \sigma^2 \) is the common variance of the observational and \( I \) is an \( n \times n \) 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 \( \sigma^2 \) (Gelman et al., 2013; Gelman and Hill, 2006). Data transformation tools are commonly used to improve normality of a distribution and equalizing variance to meet assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparing for statistical analyses. The traditional transformations that are commonly discussed include: adding constants, square root, converting to logarithmic scales, inverting and reflecting, and applying trigonometric transformations such as sine wave transformations (Osborne, 2010).

The study uses Box-Cox transformation. The form of Box-Cox transformation is as below:

\[
y(\lambda) = \begin{cases} 
\frac{y^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\
\log y, & \text{if } \lambda = 0 
\end{cases}
\]

Where, \( y \) is the observation data and \( \lambda \) is the model parameter. The optimal value of \( \lambda \) was determined and this study used, \( \lambda = 2 \). 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 is taken from the population and then calculated as 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 sample the same size of the original, which may include an observation several times while omitting other observations. Bootstrap method draws the samples with replacement, and calculates statistics for each sample (it stores these statistics and creates a distribution for further analysis). After finalizing the bootstrap, the data is analyzed for mean, standard deviation, confidence intervals, and

* Corresponding Author.
any other 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 the research requirement. As an example, for 1000 observations (original data), the analysis is performed by using statistical linear model. The analyses results of beta coefficients and r-squared are obtained, followed by application of bootstrapping method to the selected data. In applying the bootstrap method, a sample of 23 observations was replicated 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 that are similar to the original findings, from where it was replicated. Interestingly, the bootstrap method provides another noble opportunity for further comprehensive study of science and non-science discipline. A fuzzy regression model can be written as $Y = Z_a + Z_b + Z_c + ... + Z_k$, where 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 fuzzy in nature, the same which also applies to the parameters. Our aim is to estimate these parameters. In further discussion, $Z'$s are assumed as symmetric fuzzy numbers which can be presented by intervals. For example, $Z$ can be expressed as a fuzzy set given by $Z = [a_{ic}, a_{ic}]$ where $a_{ic}$ is centre and $a_{ic}$ is radius or vagueness associated. The fuzzy set above reflects the confidence in the regression coefficients around $a_{ic}$ in terms of symmetric triangular membership functions. Application of this method should be given more attention when the underlying phenomenon is fuzzy, which indicates that the response variable is fuzzy. $T$, the relationship is also considered to be fuzzy. This $Z = [a_{ic}, a_{ic}]$ can be written as $Z = [a_{ic}, a_{ic}]$ with $a_{ic} = a_{ic} - a_{ic}$ and $a_{ic} = a_{ic} - a_{ic}$. In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model. $y_j = Z, Z, Z, ... + Z, x_j$, then it can be written as $y_j = a_{ic} x_j + a_{ic} x_j + ... + a_{ic} x_j$. Thus this can be written as $y_j = a_{ic} + a_{ic} x_j + ... + a_{ic} x_j$. Then it can be written directly as $y_j = a_{ic} x_j + a_{ic} x_j + ... + a_{ic} x_j$. Absolute values of $x_j$ are taken. Suppose there $m$ data point, each comprising $a(n + 1)$-row vector. Then parameters $Z$ 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, minimised $\sum (a_{ic} + a_{ic} |x_j| + ... + a_{ic} |x_j|)$ and subject to $\{ (a_{ic} + \sum a_{ic} x_j) + (a_{ic} + \sum a_{ic} x_j) \} \geq y_j$ and $\{ (a_{ic} + \sum a_{ic} x_j) - (a_{ic} + \sum a_{ic} x_j) \} \leq y_j$ and $a_{ic} \geq 0$.

**Table 1: Description of tumour data**

<table>
<thead>
<tr>
<th>Num.</th>
<th>Variables</th>
<th>Explanation of user variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Tumo_size</td>
<td>Reading of tumour size</td>
</tr>
<tr>
<td>2.</td>
<td>Age</td>
<td>Age in year</td>
</tr>
<tr>
<td>3</td>
<td>Ethnicity</td>
<td>Ethnictiy</td>
</tr>
<tr>
<td>4</td>
<td>Smoking</td>
<td>Smoking</td>
</tr>
<tr>
<td>5</td>
<td>Lymph</td>
<td>Lymph vascular</td>
</tr>
</tbody>
</table>

Algorithm and Flow Chart for Modified Bayesian Linear Regression Analysis Method

**Fig. 1: Modified Bayesian linear regression analysis**

/* BOOTSTRAPING OF THIS ANALYSIS*/

Data Tumour;
Input Age Gender Ethnicity Smoking Alcohol Lymph TumourSize;
Cards;
66 1 1 0 1 0 2.99
/*ADDING BOOTSTRAPPING ALGORITHM TO THE METHOD* /
Title "Performing bootstrap with case resampling";
Procsurveyselect data=Tooth out=boot1 method=urs samprate=1 outhits rep=6;
run;
ods rtf close;

/* RESIDUAL NORMALITY CHECKING*/
Data Tumour;
Input Age Gender Ethnicity Smoking Alcohol Lymph TumourSize;
Cards;
66 1 1 0 1 0 2.99
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
49 1 1 0 1 1 3.00
49 1 1 0 1 1 3.00
49 0 1 0 0 1 3.00
;
ods rtf file='abc.rtf' style=journal;
ods rtf close;

/* BAYESIAN REGRESSION MODEL*/
Data Tumour;
Input Age Gender Ethnicity Smoking Alcohol Lymph TumourSize;
Cards;
66 1 1 0 1 0 2.99
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
50 1 1 1 1 0 1.00
49 1 1 0 1 1 3.00
49 1 1 0 1 1 3.00
49 0 1 0 0 1 3.00
;
ods rtf file='abc.rtf' style=journal;
ods rtf close;

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

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

ods rtf file='abc.rtf' style=journal;
ods rtf close;
Table 2: Results from Bayesian multiple linear regression

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.1418</td>
<td>0.6139</td>
<td>0.9386 - 3.3449</td>
</tr>
<tr>
<td>Age</td>
<td>-0.0031</td>
<td>0.0112</td>
<td>-0.0251 - 0.0188</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>1.3229</td>
<td>0.3959</td>
<td>0.5470 - 2.0987</td>
</tr>
<tr>
<td>Smoking</td>
<td>-1.3789</td>
<td>0.2161</td>
<td>-1.8025 - 0.9553</td>
</tr>
<tr>
<td>Lymph</td>
<td>-0.6389</td>
<td>0.2248</td>
<td>-1.0795 - 0.1984</td>
</tr>
<tr>
<td>Scale</td>
<td>1.0382</td>
<td>0.0685</td>
<td>0.9123 - 1.1814</td>
</tr>
</tbody>
</table>

Fitted Bayesian Multiple linear Regression is given as follows:

\[
\text{Tumour Size} = 2.1418 - 0.0031 \text{Age} + 1.3229 \text{Ethnicity} - 1.3789 \text{Smoking} - 0.6389 \text{Lymph}
\]

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

**Upper limits**

\[
\text{Tumour Size} = 2.7557 -0.1089 \text{Age} +1.7188 \text{Ethnicity} -1.1628 \text{Smoking} -0.4141 \text{Lymph}
\]

**Lower limits**

\[
\text{Tumour Size} = 2.13777774 + (1.32194441 \text{Age}) - (1.3811111 \text{Ethnicity}) +(-0.64083329 \text{Smoking}) + (0.0030556 \text{Lymph})
\]

The width of prediction intervals with respect to 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 Table 4. From this table, the average width for former was found to be 568.00, while that of the latter was only 15.93, thereby indicating the superiority of fuzzy regression methodology.

3. Summary and discussion

This paper presents an algorithm and illustrated the procedure of modeling by using modified Bayesian linear regression through SAS language. Our aim is to share the algorithm and also provide the researcher with an alternative programming that suitable for a small sample size. This proposed method can be applied to small sample size data, especially when limited data is obtained, for example in public health.
Table 4: The results of the width of prediction intervals with respect to Bayesian multiple linear regression model and Bayesian fuzzy regression model corresponding to each set of observed explanatory variables is computed in SPSS

<table>
<thead>
<tr>
<th>Lower limit</th>
<th>Upper limit</th>
<th>Width</th>
<th>Lower limit</th>
<th>Upper limit</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumour Size = 2.1418 - 0.0031 Age + 1.3229 Ethnicity -1.3789 Smoking -0.6389 Lymph)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.6619</td>
<td>1.5111</td>
<td>4.2500</td>
<td>88.0294</td>
<td>88.0050</td>
<td>0.0244</td>
</tr>
<tr>
<td>8.7567</td>
<td>0.1449</td>
<td>4.2400</td>
<td>66.2358</td>
<td>66.2131</td>
<td>0.0228</td>
</tr>
<tr>
<td>8.7567</td>
<td>0.1449</td>
<td>4.2400</td>
<td>66.2358</td>
<td>66.2131</td>
<td>0.0228</td>
</tr>
<tr>
<td>8.7567</td>
<td>-0.7535</td>
<td>1.0000</td>
<td>64.9725</td>
<td>64.9503</td>
<td>0.0222</td>
</tr>
<tr>
<td>6.8201</td>
<td>0.0696</td>
<td>4.2400</td>
<td>86.0692</td>
<td>86.0422</td>
<td>0.0269</td>
</tr>
<tr>
<td>6.8201</td>
<td>-0.7535</td>
<td>1.0000</td>
<td>64.9725</td>
<td>64.9503</td>
<td>0.0222</td>
</tr>
<tr>
<td>10.3902</td>
<td>-0.0696</td>
<td>4.2400</td>
<td>86.0692</td>
<td>86.0422</td>
<td>0.0269</td>
</tr>
</tbody>
</table>

Average width 5.416600

References


