Journal of Scientific Research and Development 3 (5): 7-11, 2016
Available online at www.jsrad.org
ISSN 1115-7569

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]

[1]*School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kelantan, Malaysia*
[2]*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu*
[3]*School 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 $y_i = E(y_i \mid x_i) + e_i$, where i =1, 2, 3,...,n denoting an observation of a subject. y_i is the response variable 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 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 $n \times p$, β contains the p regression coefficients, σ^2 is the common variance of the observational and I is an $n \times n$ identity matrix. That is, $y \sim N(X\beta, \sigma^2)$. 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). 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:

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

Where, y is the observation data and λ is the model parameter. The optimal value of λ was determined and this study used, λ = 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 $Y = Z_0 + Z_1 x_1 + Z_2 x_2 + \ldots + Z_k x_k$, 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 fuzzy in nature, the same which also applies to the parameters. Our aim is to estimate these parameters. In further discussion, $Z_i's$ are assumed as symmetric fuzzy numbers which can be presented by intervals. For example, Z_i can be expressed as a fuzzy set given by $Z_i =< a_{1c}, a_{1w} >$ where a_{ic} is centre and a_{iw} 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 $Z_i =< a_{1c}, a_{1w} >$ 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 $Z_i =< a_{1c}, a_{1w} >$, 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 directly as $y_{jw} = a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}|$. y_{jw} represents radius and 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, minimised $\sum_{j=1}^{m} \left(a_{0w} + a_{1w}|x_{1j}| + \cdots + a_{nw}|x_{nj}|\right)$ 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 for this study is a sample which is composed of five variables (Table 1 and Fig. 1).

Table 1: Description of tumour data

Num.	Variables	Explanation of user variables
1.	Tumo_size	Reading of tumour size
2.	Age	Age in year
3	Ethnicity	Ethnicity 0 = Malay 1 = Non malay
4	Smoking	0 = Never 1 = Yes
5.	Lymph	Lympho vascular 0 = "-ve" 0 = "+ve"

Algorithm and Flow Chart for Modified Bayesian Linear Regression Analysis Method

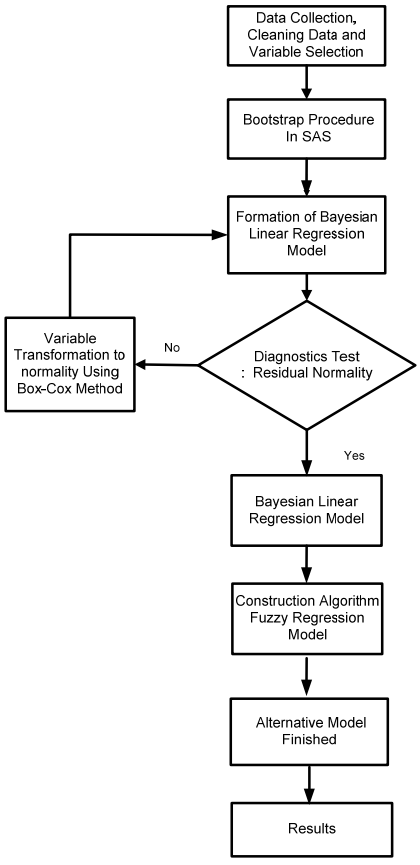

Fig. 1: Modified Bayesian linear regression analysis

```
/* BOOTSTRAPING OF THIS ANALYSIS*/
Data Tumour;
Input Age Gender Ethnicity Smoking Alcohol Lymph
TumourSize;
Cards;
66  1  1  0  1  0  2.99
```

```
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  3.01
;
ods rtf file='abc.rtf' style=journal;

/*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*/
DataTumour;
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
⋮⋮⋮⋮
49  1  1  0  1  1  3.00
49  1  1  0  1  1  3.00
49  0  1  0  0  1  3.01
;
;
odsrtffile='abc.rtf'style=journal;

odsgraphicson;
procregdata=Tooth plots=all;
modelTumourSize=  Age Ethnicity Smoking Lymph
      output   out=Residuals
p=y_hat
r=y_res;
run;
odsgraphicsoff;

odsgraphicson;
procregdata=Tooth plots=all;
modelTumourSize= Age Ethnicity Smoking Lymph/p ;
run;
odsgraphicsoff;
odsrtfclose;

/* BAYESIAN REGRESSION MODEL*/
DataTumour;
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
⋮⋮⋮⋮
49  1  1  0  1  1  3.00
49  1  1  0  1  1  3.00
49  0  1  0  0  1  3.01
;
odsrtffile='abc.rtf'style=journal;

odsgraphicson;
procgenmoddata=Tooth;

modelTumourSize= Age Gender Ethnicity Smoking
Alcohol Lymph  / dist=normal link=identity;

bayesseed=1OutPost=Post
diagnostics=allsummary=all;;
run;
odsgraphicsoff;

odsrtfclose;

/* BAYESIAN FUZZY REGRESSION*/
DataTumour;
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
⋮⋮⋮⋮
49  1  1  0  1  1  3.00
49  1  1  0  1  1  3.00
49  0  1  0  0  1  3.01
;
run;
odsrtffile='abc.rtf'style=journal;

procoptmodel;
set j= 1..115;
NumberTumour_Size_Bayes {j} Ethnicity {j}, Smoking
{j}, Lymph {j} Age {j};
readdata Tooth into [_n_] Tumour_Size_Bayes Ethnicity
Smoking Lymph Age;

/*PRINT TUMOURSIZE ETHNICITY SMOKING
LYMPH*/
PrintTumour_Size_Bayes Ethnicity Smoking Lymph
Age;

/*TOTAL OF OBSERVATIONS*/
number n init115;


/* DECISION VARIABLES BOUNDED OR NOT
BOUNDED*/
/*THESE FOUR VARIABLES ARE BOUNDED*/

var aw{1..5}>=0;

/*THESE FOUR VARIABLES ARE NOT BOUNDED*/
var ac{1..5};

/* OBJECTIVE FUNCTION*/
min z1= aw[1] * n + sum{i in j}
Ethnicity[i] * aw[2]+sum{i in j} Smoking[i]*aw[3]+
sum{i in j} Lymph[i]*aw[4]+ sum{i in j} Age[i]*aw[5];

/*LINEAR CONSTRAINTS*/
con c{i in 1..n}:
ac[1]+ Ethnicity[i]*ac[2]+
Smoking[i]*ac[3]+Lymph[i]*ac[4]+ Age [i]*ac[5]-
aw[1]- Ethnicity[i] *aw[2]-Smoking [i]* aw[3]-
Lymph[i]*aw[4]
+ Age[i] *aw[5]
<= Tumour_Size_Bayes[i];
```

```
con c1{i in 1..n}:
ac[1]+ Ethnicity[i]*ac[2]+
Smoking[i]*ac[3]+Lymph[i]*ac[4]+ Age[i]*ac[5]+
aw[1]+ Ethnicity[i]* aw[2]+Smoking[i]
*aw[3]+Lymph[i]*aw[4]+ Age[i]*aw[5]
>= Tumour_Size_Bayes [i];
expand;
/* THIS PROVIDES ALL EQUATIONS */
solve;
print ac aw;
quit;

odsrtfclose;
```

2. Results

Table 2: Results from Bayesian multiple linear regression

Analysis of Maximum Likelihood Parameter Estimates				
Parameter	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	2.1418	0.6139	0.9386	3.3449
Age	-0.0031	0.0112	-0.0251	0.0188
Ethnicity	1.3229	0.3959	0.5470	2.0987
Smoking	-1.3789	0.2161	-1.8025	-0.9553
Lymph	-0.6389	0.2248	-1.0795	-0.1984
Scale	1.0382	0.0685	0.9123	1.1814

Fitted Bayesian Multiple linear Regression is given as follows:

Tumour Size =
2.1418-0.0031 Age +1.3229 Ethnicity
Std. Error (0.6139) (0.0112) (0.2161)
-1.3789Smoking - 0.6389 Lymph (2.1)
(0.3959) (0.2248)

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

Tumour Size =
2.7557 -0.1089 Age +1.7188 Ethnicity
-1.1628Smoking –0.4141Lymph (2.2)

Lower limits

Tumour Size =
1.5279 – 0.0143 Age +0.927 Ethnicity
-1.595Smoking –0.8637Lymph (2.3)

Part II: Results from Fitted model for Fuzzy Regression

Table 3: Value of centre (AC) and radius (AW)

[1]	ac	aw
1	2.1408333	0.00305556
2	1.3220833	0.00013889
3	-1.3811111	0.00000000
4	-0.6394444	0.00138889
5	-0.0030556	0.00000000

Fitted model for fuzzy regression for

Tumour Size =
<2.1408333, 0.00305556> +
< 1.3220833, 0.00013889>Age +
<-1.3811111, 0.00000000>Ethnicity +
<-0.6394444, 0.00138889>Smoking +
<-0.0030556, 0.00000000> Lymph (2.4)

Upper or lower limits of prediction intervals are computed from the prediction equation (2) by taking the coefficient as their corresponding estimated values plus or minus standard error.

Upper limits

Tumour Size =
2.14388886 + (1.32222219 Age) –
(1.3811111 Ethnicity) –
(0.63805551Smoking) –
(0.0030556 Lymph) (2.5)

Lowers limits

Tumour Size =

2.13777774 + (1.32194441 Age)
+ (-1.3811111 Ethnicity) +(-0.64083329
Smoking) + (0.0030556 Lymph) (2.6)

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

Multiple Bayesian Linear Regression (MLR) Model Tumour Size = 2.1418 - 0.0031 Age + 1.3229 Ethnicity -1.3789Smoking - 0.6389 Lymph)			Bayesian Fuzzy Regression (FR) Model Tumour Size = <2.1408333, 0.00305556> + < 1.3220833, 0.00013889> Age + <-1.3811111, 0.00000000> Ethnicity + <-0.6394444, 0.00138889> Smoking + <-0.0030556, 0.00000000> Lymph		
Lower limit	Upper limit	Width	Lower limit	Upper limit	Width
11.6619	1.5111	4.2500	88.0294	88.0050	0.0244
8.7567	0.1449	4.2400	66.2358	66.2131	0.0228
8.7567	0.1449	4.2400	66.2358	66.2131	0.0228
8.7567	0.1449	4.2400	66.2358	66.2131	0.0228
6.8201	-0.7535	1.0000	64.9725	64.9503	0.0222
6.8201	-0.7535	1.0000	64.9725	64.9503	0.0222
10.3902	-0.0696	4.2400	86.0692	86.0422	0.0269
⋮	⋮	⋮	⋮	⋮	⋮
10.2247	.8905	9.334	65.5486	65.5350	.0136
10.2247	.8905	9.334	65.5486	65.5350	.0136
10.5514	.8476	9.704	69.5153	69.5008	.0144
Average width5.416600			Average width 0.0133		

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.

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