International Journal of Engineering & Technology, 7 (4.30) (2018) 224-228

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Model-Building of Multiple Binary Logit using Model Averaging

Siti Aisyah Mohd Padzil[1*], Khuneswari Gopal Pillay[2]*, Rohayu Mohd Salleh[3]

[1,2,3]Department of Science and Mathematics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Kampus, KM 1 Panchor Road, 84000, Muar Johor Malaysia
**Corresponding author E-mail:aisyahpadzil@gmail.com*

Abstract

Many researchers had been carried out on the study of statistical modelling, making it easier for new researchers in many sectors (social sciences, economics, medical, and etc.) to obtain knowledge in order to ease their research study. Nevertheless, there is still no agreed guidelines in obtaining the best model for multiple binary logit (MBL) using model averaging (MA). This research will demonstrate the proper guidelines to obtain best MBL model by using MA. Upper Gastrointestinal Bleed (UGIB) data were studied to illustrate the process of model-building using the proposed guidelines. This study will pinpoint the factors with high possibility leading to mortality of UGIB patients using obtained best model. Corrected Akaike Information Criteria (AIC_c) and Bayesian Information Criteria (BIC) were used to compute the weights in model averaging method. The performance of the models was computed by using Root mean square error (RMSE) and mean absolute error (MAE). Model obtained by using BIC weights showed a better performance since the RMSE and MAE values are lower compared to model obtained using AIC_c weights. The factors that affects the survivability of UGIB patients are shock score, comorbidity and rebleed. In conclusion, model-building of multiple binary logit using model averaging showed a better performance when using BIC.

Keywords: *AICc; BIC; Model Averaging; Model-building; Multiple Binary Logit; UGIB.*

1. Introduction

Model-building approach is needed when the researchers aims are to make prediction and to decide which variables should be included in making prediction. For analysis with binary dependent variables taking on value 1 (yes) and 0 (no), MBL or often known as Logistic Regression (LR) is one of the suitable modelling approach. According to [19], MBL is the extension of logit model and known as qualitative choice model. Maximum likelihood (ML) is used for parameter estimation of multiple binary logit model.

This research implemented MA method in the modelling analysis as it was proposed as an alternative to model selection (MS) which intended to overcome the underestimation of standard errors [6]. Statistical modelling using MS eliminates insignificant variables to form best fitted model with only contributing variables. According to Burham and Anderson [3], the properties of standard parameter estimates obtained from the selected model do not reflect the stochastic nature of the model selection process. As an alternative, model averaging (MA) has been proposed to overcome the underestimation of standard errors that is a consequence of model selection.

The main idea of MA is to average the weight of each possible models being studied by using model selection criteria, I_m to obtain the coefficient estimates on the weaker term so that the best model will yield a better prediction. Two commonly used I_m are Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) also known as SCHWARZ criteria. Previously, [4] had suggested Akaike Information Criteria AIC [1] to compute the weights of all possible models. There has been an issue when using AIC on small sample size as it will lead to high degree of negative bias. According to [8] as the number of parameters increases in comparison to sample size, AIC becomes a strongly negatively-biased estimator. This negative bias in AIC limits its effectiveness as a model selection criterion and can lead to over-fitting models. [8] proposed that the corrected Akaike information criterion, AIC_c to solve the problem with small samples.

This study illustrates the model-building approach by using UGIB patient's data and aim to highlight the most significant factors of mortality for UGIB patients. Two I_m (AIC_c and BIC) are compared to determine which I_m works best in MA method. Root Mean Square Error (RMSE) and Mean Absolute Error will be calculated to compare which model (using AIC_c or BIC) yield a better accuracy. The whole procedures of obtaining the best MBL model using MA were summarized and explained step by step to provide a clear guideline of model-building by using MBL model.

2. Methodology

2.1. Multiple binary logit

MBL model is a form of regression with binary dependent variable. In this research, the outcome of the study is the patient's survivability which takes on value 1(survive) and 0 (not survive).

The general MBL model [10] is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q + u, \quad (1)$$

and

$$Y_i = \ln\left[\frac{P_i}{1-P_i}\right], \quad (2)$$

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

where the binary dependent variable is denoted by Y, X_j is the jth independent variable where $j = 1,2 \dots q$, the constant term of the model is denoted by β_0, β_j is the jth coefficient of jth independent variable where $j = 1,2 \dots q$, u is the random error of the model and P_i is the probability of event occurs.

$$P_i = \frac{exp^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_q X_q}}{1 + exp^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_q X_q}}. \quad (3)$$

2.2. Model selection criteria

Model selection criteria is used to compute the weight of each possible model. AIC_c is an adjusted version of AIC which was proposed to solve the problem of small sample size. Bayesian Information Criterion which was proposed by [17] was developed from Bayesian idea where model with largest posterior probability should be chosen as the best model [7]. According to [14], choosing model with minimum BIC is the same as choosing model with maximum posterior probability. The general form of AIC_c and BIC are

$$AICc = -2logL(M) - 2\mathrm{p}\frac{\mathrm{n}}{\mathrm{n-p-1}}, \quad (4)$$

$$BIC = -2logL(M) - p[\,log(n)], \quad (5)$$

where $L(M)$ is the minimum value for likelihood function of model *M*, *n* is the number of observation *p* is the number of parameter and $I_m$is type of model selection criteria. Different I_m were tested to identify which I_m produce model with minimum error. Model with minimum error indicate a best model.

2.3. Modelling using model averaging

Model averaging (MA) is an alternative to model selection in order to overcome the underestimation of standard errors that is a consequence of model selection [6]. Instead of picking one best model with the lowest selection criteria value, model average estimator weighs all of the possible models and MA will shrink the estimates of a weaker variables [16].

Modelling of MBL model start by listing all the possible models formed from all possible combination of variables. All possible models with no interaction variables can be calculated as

$$\mathbf{N} = \sum_{j=1}^{q} ({}^{q}\boldsymbol{C}_{j}), \quad (6)$$

where, *N* represent the number of all possible models and q is number of single independent variables $j = 1,2,3 \dots, q$. The next step is to obtain weight, W_m for each possible model. [2] presented the weights W_m for a model as

$$W_m = \frac{\exp(\frac{I_m}{2})}{\sum_{m=1}^{M} \exp\frac{I_m}{2}}, \quad (7)$$

where m represent the possible models, $m = 1, 2, 3 \dots, M$ and I_m is the model selection criterion (AIC_c and BIC) for model *M*.

The third step in MBL modelling using MA is to estimate the parameters, $\hat{\beta}_p$ for each variable. The formula is as follows

$$\hat{\beta}_p = \sum_{m=1}^{M} w_m \hat{\beta}_{(p,m)}, \quad (8)$$

where $\hat{\beta}_{(p,\,m)}$ is the estimate of β_p under model for m = 1,2,…, *M*. This are the stage where the coefficient for each variable been studied were calculated based on its weight or importance in the model. Once all the coefficient had been calculated, the best model of MBL using MA using AIC_c and BIC are obtained. Figure 1 below is the flowchart of MA to obtain best model.

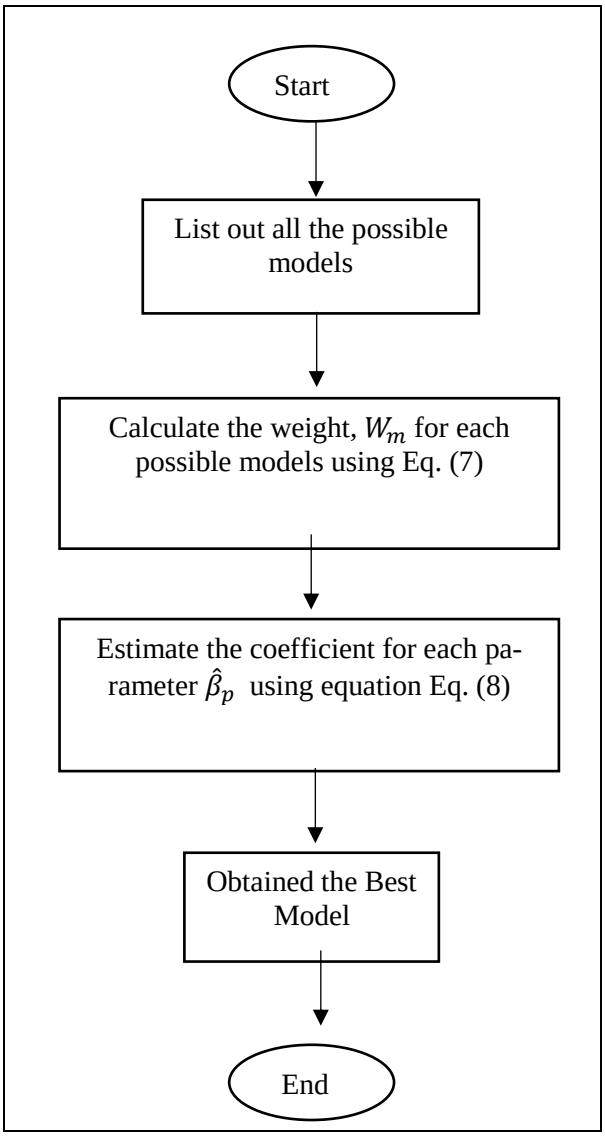

Fig.1: Flowchart of MA

2.4. Model accuracy measure

To compare the performance of model formed using AIC_c and BIC, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are computed as suggested by [11]. The formula for these accuracy measure are as following: [4]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}-y)^2}{N}}, \quad (9)$$

$$MAE = \frac{\sum_{i=1}^{N}|\hat{Y}-Y|}{N}, \quad (10)$$

where, *N* is the total number of sample, *Y* is the actual value of dependent variables and $\hat{Y}$ is the estimated value of *Y*. The smallest value of accuracy measure indicates a better performance.

3. Case Study: Upper Gastrointestinal Bleeed patients (UGIB)

Patients with UGIB suffer from bleeding in their gastrointestinal tract (oesophagus, stomach or duodenum). The patients will experience UGIB is said to be one of the common medical emergencies with 250 000 to 300 000 hospitalization every year [15]. This

research is interested in identifying the factor of UGIB patient's mortality. Rockall [15] had come out with scoring system to determine the patient's risk of death. The scoring system classes the patients scores to a Rockall group which determine whether the patients have low, medium or high risk of death. Table 1 explains Rockall scoring system

Table 1: Rockall Scoring System

Rockall Score	Rockall Group	Risk Level
0-2	1	Low
3-7	2	Medium
8-11	3	High

The calculation of Rockall score were explains deeper in [15]. The five variables involved in the calculation of Rockall scores were studied in this research to pinpoint the factors with high possibility contributes to death of UGIB patients. Table 2 explains the variables of UGIB patient's data used this study.

Table 2: Variable Descriptions

Variable	Description
Y	**Survival of Patients** 1 if the patient survives 0 if the patient not survives
X_1	**Age Score** 0: if age <60 1: if age 60-79 2: if age ≥80
X_2	**Shock Score** 0: No shock 1: Tachycardia 2: Hypotension
X_3	**Comorbidity** 0: Nil major 1: Cardiac failure, IHD, others 2: Renal failure, liver failure, disseminated malignancy
X_4	**Diagnosis Score** 0:Mallory-Weiss tear, no lesion 1: All other diagnosis 2: Malignancy of UGI
X_5	**Major Score** 0: None or Dark Spots 1: Blood in Upper GIT, adherent clot, visible spurting/ vessel
X_6	**Rebleed** 1: Yes 2: No
X_7	**Rockall Group** 1: Low Risk 2: Medium Risk 3: High Risk

4. Data Analysis

The process of obtaining the best model starts by listing all possible models. The number of all possible models were calculated using (4)

$$N = 1({}^7C_1) + 1({}^7C_2) + 1({}^7C_3) + 1({}^7C_4) + 1({}^7C_5) + 1({}^7C_6) + 1({}^7C_7) = 127\ possible\ models$$

Table 3: All possible models for Rockall score data

Number of variables	Number of Models	Model
1	7 models	M1-M7
2	21 models	M8-M28
3	35 models	M29-M63
4	35 models	M64-M98
5	21 models	M99-M119
6	7 models	M120-M126
7	1 model	M127
Total	127 models	

Table 3 shows all possible models of UGIB which are 127 possible models in total. All possible model consists of all possible combination of variable in a model. The list of all possible model were attach on the appendix.

The weight for each model were then computed based on the AIC_c and BIC values. Table 4 shows the example of weight for some models.

Table 4: Weight of each possible model

Model	**AICc**	**BIC**	W_{AIC_c}	W_{BIC}
M1	-290.63	-278.64	0.00	0.00
M2	-297.72	-285.73	0.00	0.00
M3	-305.98	-294.00	0.00	0.00
. . .	. . .	. . .	. . .	. . .
M127	-313.64	-277.95	0.01	0.01

The weight obtained from each model were used to estimate the coefficient $\hat{\beta}_p$ for each variable. As an example, the estimated β_0 can be obtained as following using (8):

$$\hat{\beta}_0 = \frac{\beta_{(0,1)}W_1 + \beta_{(0,2)}W_2 + \beta_{(0,3)}W_3 + \cdots + \beta_{(0,127)}W_{127}}{W_1 + W_2 + W_3 + \cdots + W_{127}}$$

Since,

$$\sum_{m=1}^{M} W_{I_m} = 1$$

Hence,

$$\hat{\beta}_0 = \beta_{(0,1)}W_1 + \beta_{(0,2)}W_2 + \beta_{(0,3)}W_3 + \cdots + \beta_{(0,127)}W_{127}$$

In model-building using MA, no variable is eliminated and therefore each variable will be in the final model with its corresponding weight. The weight represents the importance of the variables in the model. The higher the weight, the higher the needs for the variable to be in the model to produce good model. Table 5 shows the best model using AIC_c and BIC.

Table 5: Best Model of MA using AIC_c and BIC

Model	RMSE	MAE
Using AIC_c weights, $\hat{Y} = 0.7652 + 0.0089\,x_1 - 0.0271x_2 - 0.0288x_3 - 0.0067x_4 - 0.0073x_5 + 0.1299x_6 - 0.0263x_7$	0.0270	0.0791
Using BIC weights, $\hat{Y} = 0.7580 + 0.0074\,x_1 - 0.0289x_2 - 0.0307x_3 - 0.0085x_4 - 0.0096x_5 + 0.1332x_6 - 0.0354x_7$	0.0261	0.0579

Based on Table 5, model obtained using BIC weights showed a better performance as it has the least value for both RMSE and MAE.

Table 6: Coefficients and p-values of the best model

Variable	Coefficient	p-value
Constant	0.7580	$2E$-16
X_1	0.0074	0.6305
X_2	-0.0289	0.0273
X_3	-0.0307	0.0003
X_4	-0.0085	0.6795
X_5	-0.0096	0.3382

X_6	0.1332	0.0006
X_7	-0.0354	0.0834

Table 6 shows the coefficients and p-values the best model. p-value is often used to identify significant and insignificant variables. Variables with p-value less than 0.05 indicate an insignificant variable [10]. From the results, only three variables are found to be significant which are rebleed (X_6), comorbidity (X_3) and shock score (X_2).

$$\hat{Y} = 0.7580 - 0.0307\ Comorbidity + 0.1332\ Rebleed - 0.0289\ Shock\ Score - 0.0354\ Rock.Group - 0.0096\ MajorScore - 0.0085\ Diag.Score + 0.0074\ Age.Score$$

The best model showed that when the values for all the variables in the model above are 0, the probability of UGIB patient's survivability is

$$P_i = \frac{exp^{0.7580}}{1 + exp^{0.7580}} = 0.6809 \approx 0.68$$

If there is one unit increase in shock score (X_2), the $\hat{Y}$ will decrease by 0.0289. Therefore, the probability of UGIB patient's survivability will decrease also. Similarly, the probability of UGIB patient's survivability also will decrease if there is a comorbidity (X_3). The probability of UGIB patient's survivability will increase by 0.1332 if there is a rebleed (X_6). Whereas, the probability of UGIB patient's survivability will increase by $2 \times 0.1332 = 0.2664$ if there is no rebleed.

5. Discussion and Conclusion

Model-building of MBL can be applied on data with binary dependent variable and continuous or categorical independent variables. The modelling of MBL model using MA requires four step to obtained the best model.

Different I_m is used to compute weight of each possible model results in different model performance. [2] suggested that the weights based on AIC value but in this research AIC_c and BIC were tested and the performance were measured using RMSE and MAE. It is a good idea to compare several I_m with different performance measure before choosing the final model as it helps to be more accurate in making final decision to choose the best model. From the analysis, the results indicate that BIC produce model with lower RMSE and MAE which indicate a better performance when compared with model produce using AIC_c.

The factors comorbidity, rebleed and shock score are found to have significant contribution. From this information, the three variables were concluded as the most contributing factor that leads to mortality for UGIB patients.

Comorbidity is a term use for patients with multiple chronic diseases. [18] had stated that this multiple disease leads to a more complex clinical treatment and hence increased health care costs. Family doctors face with challenges in making decision about the treatment for patients with comorbidity because most of the clinical studies do not consider patients with multiple disease [13]. Research conducted by [5] concluded that multiple comorbidity is an important factor to predict mortality. Similarly, this research also concluded that the presence of comorbid disease is found to be an important risk factor of mortality.

Rebleed or hemorrhage is described as the most influential factor of mortality for UGIB patients according to [15]. Similarly, [18] also point out that rebleeding is an important factor of mortality and occurs in 10–30% of successfully treated patients. Research conducted by [5] and [12] also concluded that rebleeding is an influential factor of mortality. Even though there are advance treatments available in treating patients with UGIB, rebleed remains the life-threatening factor of UGIB patients [9].

Shock score is a scoring that tells if there is an abnormality in blood circulation which result in hypertension and tachycardia. Research conducted by [5,12] also conclude that Rockall scores shows medical shock to be one of the risk factor of mortality.

6. Recommendation for Future Work

This study did not involve interaction variables in building the best model. Interaction variable is very important in model-building because it will influence the dependent variable differently. It is unfair to consider a general model without including interaction variables [19] as researcher must consider all possible combination of variables. Effects of interaction variables should be explored since there are less researchers are carried out on MBL model with interaction variables.

[10] stated that the cross-product between single independent variables is called as an interaction variable. This make forming an interaction variable harder when the independent variables are in categorical. Furthermore, most statistical software does not consider interaction variable as an option in finding best model. Therefore, R-package should be developed to obtain the best multiple binary logit models with interaction variables.

The problems of under-fit and over-fit may occur when the model includes too many or too few parameter [7]. If the model includes too few parameters, it may lead to biased. Contrary, if the model includes too many parameters it will lead to poor precision. As a solution, a data screening method such as backward elimination can be implied. The backward elimination removes insignificant variables from the model based on the P-value.

References

[1] Akaike H (1978), A Bayesian Analysis of the Minimum AIC Procedure. *Annal of the Statistical Mathematics* 30(A), 9-14.

[2] Buckland TS, Burnham KP & Austin NH (1997), Model Selection: An integral part of inference. *Biometrics* 53(2). 603-618.

[3] Burnham KP & Anderson DR, *Model Selection and Multimodel Inference: A practical Information-theoretic Approach*. 2nd Ed. NewYork: Springer-Verlag, (2002).

[4] Chai T & Drexler RR (2014), Root Mean Square Error or Mean Absolute Error? Arguments against avoiding RMSE in the literatures. *Geoscientific Model Development* 7, 1247-1250.

[5] Chiu PWY, Ng EKW, Cheung FKY, Chan FKL, Leung WK, Wu JCY, Wong VWS, Yung MY, Tsoi K, Lau JYW, Sung JJY & Chung SSC (2009). Predicting Mortality in Patients with Bleeding Peptic Ulcers after Therapeutic Endoscopy. *Clinical Gastroenterology and Hepatology* 7, 311–316.

[6] Claeskens G & Hjort NL, *Model Selection and Model Averaging*. United Kingdom: University Press, Cambridge, (2008).

[7] Forster MR (2001), The new science of simplicity. *Simplicity, inference and modelling* 76(2), 83-117.

[8] Giombini G & Szroeter J (2007), Quasi Akaike and Quasi Schwarz criteria for Model Selection: A Suprising consistency result. *Economic Letters* 95, 259-266.

[9] Hurvich CM & Tsai CL (1989), Regression and time series model selection in small samples. *Biometrika* 76, 297-307.

[10] Jairath V, Rehal S, Logan R, Kahan B, Hearnshaw S, Stanworth S & Travis S (1993), Acute variceal haemorrhage in the United Kingdom: Patient characteristics, management and outcomes in a nationwide audit. *Digestive and Liver Disease* 46, 419–426.

[11] Kutner MH, Nachtsheim CJ & Neter J, *Applied Linear Regression Models*. 4th edition. Singapore: McGraw-Hill Inc, (2008).

[12] Mehdiyev N, Enke D, Fettke P & Loos P (2016), Evaluating Forecasting Methods by Considering Different Accuracy Measures. *Procedia Computer Science* 95, 264-271.

[13] Noraini A, Zainodin HJ & Rick LB (2013), Risk factor determination on UGIB patients in Kota Kinabalu, Sabah, Malaysia. *Medical Sciences* 13(7), 526-536.

[14] Osmun WE, Kim GP & Harrison ER (2015), Patients with multiple comorbidities: Simple teaching strategy. *Can. Fam. Physician*. 61(4), 378–379

[15] Posada D & Buckley TR (2004), Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology* 53(5), 793-808.
[16] Rockall TA, Logan RFA, Devlin HB, Northfield TC & the steering committee and members of the National Audit of Acute Upper Gastrointestinal Haemorrhage (1996), Risk assessment after acute upper gastrointestinal haemorrhage. *Gut* 38, 316-321.
[17] Schomaker M, Wan ATK & Heumann C (2010), Frequentist model averaging with missing observations. *Computational Statistics and Data Analysis* 54, 336-3347.
[18] Schwarz G (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
[19] Vreeburg EM, Terwee CB, Snel P, Rauws EAJ, Bartelsman JFWM, Meulen JHP & Tytgat GNJ (1999), Validation of the Rockall risk scoring system in upper gastrointestinal bleeding. *Gut* 44, 331-335.
[20] Zainodin HJ & Khuneswari GP (2007), Model-Building Approach in Multiple Binary Logit Model using Coronary Heart Disease. *Malaysian Journal of Mathematical Sciences* 4(1), 107-133.