An alternative approach in predictive modeling using model averaging scheme for logistic regression case (case study: application in class prediction of autistic spectrum disorder data)

IOP Conf. Series: Earth and Environmental Science **299** (2019) 012039 doi:10.1088/1755-1315/299/1/012039

An alternative approach in predictive modeling using model averaging scheme for logistic regression case (case study: application in class prediction of autistic spectrum disorder data)

Septian Rahardiantoro[1]*, Anang Kurnia[1], Mulianto Raharjo[1] and Yusma Yanti[2]

[1]Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia
[2]Department of Computer Science, Pakuan University, Bogor, West Java, Indonesia

*E-mail: rahardiantoro_14@apps.ipb.ac.id

Abstract. Logistic regression has become a popular method for handling predictive modeling when the response variable has a categorical scale. The difference in category proportion in response variable could influence the prediction accuracy. This research applied the model averaging approach for logistic regression in purpose to improve the prediction accuracy in different proportion of each category. Model averaging has the idea to combine some model candidates based on the specified weight to be the final model. The model candidate in model averaging generated based on all possibilities variable selection in the model. AIC weight is chosen to apply in the combination of all possible model candidates. It is illustrated with an application to data from a classification of Autistic Spectrum Disorder data. The result of this case indicated that the logistic model averaging had better performances.

1. Introduction

Regression analysis is widely used for modeling the real phenomenon of the relationship between predictor variables with the response variable. This method also can be used for modeling the response variable which has a categorical scale, called logistic regression analysis. The main idea of logistic regression analysis is taking specified transformation, logit transformation in general used, for modeling the predictor variables with categorical response variable [1].

This research focuses on the case of prediction the categories of response variable using logistic regression analysis. However, the proportion of each category would give a significant impact on the prediction application. Because of that, this research also tried to apply the model averaging using logistic regression in the construction of model candidate [2].

The main concept of model averaging is creating some model candidate from the set of all possible models, then combining the coefficient estimators or response predictions to be the final model [3]. In this research, all possible models mean the all possibility of predictor variable to include in the model. Therefore, in this term, model candidate constructed with a different number of predictors from all possible predictors. Then, after some model candidates created, the coefficient estimators in every model candidate combined using specified weight. In this research, the weight selected is based on the AIC value of each model candidates. Higher weight belongs to the smaller AIC value of model candidate and vice versa.

The logistic model averaging method would apply in the Autistic Spectrum Disorder (ASD) Screening Data for 609 Adults with response variable is a class of ASD suffered [4]. The predictor variables selected are age, gender, born with jaundice, ASD suffered from family, used the app before, and the person who takes the test. In practice, this data will implement the logistic regression and also the logistic model averaging in the purpose of prediction the class of ASD suffered.

2. Model Averaging

This section consists of the main concept of model averaging in general. Assume the data $\boldsymbol{X}_{n\times p}$ is the p predictor variables with n observations, and $\boldsymbol{y}_{n\times 1}$ is the n observations of response to the var toiable. The first step of this method is constructing some of the model candidates $\boldsymbol{y}^l = f(\boldsymbol{X}^s) + \varepsilon; l = 1,2 \dots, m,$ which is the model from the subset of predictor variables, ${\boldsymbol{X}^s}_{n\times q}$where $q < p$ [5]. Then, the parameter estimator of the model based on the combination the parameter estimator of each model candidates as follows:

$$\widehat{\boldsymbol{\beta}}^{MA} = \sum_{l}^{m} w_l {}^{l}\widehat{\boldsymbol{\beta}} \tag{1}$$

where w_i is the weight of each model candidates, ${}^{l}\widehat{\boldsymbol{\beta}}$ is the parameter estimator of l-model candidate.

The construction of the model candidate is based on all of the possible predictor variable selection in the model. In the example, assume there are 2 predictor variable (X_1, X_2) will apply the model avaraging method. Therefore, it would be created 3 model candidates as follows:

$$y^1 = \beta_0 + \beta_1 X_1 + \varepsilon \tag{2}$$
$$y^2 = \beta_0 + \beta_2 X_2 + \varepsilon \tag{3}$$
$$y^3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{4}$$

In this research, the model averaging method would be applied in the categorical response variable case. Therefore assume the response variable is $\boldsymbol{y}_{n\times 1} = [y_i]; y_i \in \{0,1\}; i = 1,2, \dots, n$. Because of that, the model candidate that used in this research follows the logistic regression model, $logit\, p = ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} X_j$. In addition, the averaging process taken with the specify weight w_l based on the value of Akaike Information Criterion (AIC) in each model candidates. The AIC indicates a quality value of statistical model based on the given data that is better quality of model would be performed by the lowest AIC values. The formula of the weight in this research based on m model candidates follows

$$w_l = \frac{exp\left(\frac{1}{2}a_l\right)}{\sum_{l=1}^{m} exp\left(\frac{1}{2}a_l\right)} \tag{5}$$

where a_l denotes the value of AIC in the l – th model candidates, and $w_l \geq 0$; $\sum_{l=1}^{m} w_l = 1$ [6]. Based on this formula, it can be informed that better model constructed will have higher weight.

3. Data

The data used in this research is based on the research of Tabtah (2017) about Autistic Spectrum Disorder (ASD) Screening Data for Adult [4]. In this research, the data used just a part of these. The dimension of the data that used in this research is n=609 with six predictor variables: age, gender, born with jaundice, ASD suffered from family, used the app before, and the person who takes the test. The response variable is the class of ASD suffered, with "0" denotes as non ASD, and "1" denotes ASD suffered.

4. Methodology

There are two part of the method in this research, such as descriptive statistics of each of the variables in the data, and modeling using logistic regression and logistic model averaging. The descriptive statistics contains the summary of each of the variables to describe the distribution of data. Histogram and some simple statistics are used to describe the numeric predictor variable in the data, and frequency information selected to describe the categorical variables in data include the response variable.

In the modeling process, data separated to be two parts, 70% of amount of observation to be the training data, and 30% of the number of observation to be the testing data with the proportional selection of each category of the response variable. Training data selected to creating the model, the prediction evaluated in the testing data. This process would be iterated 100 times with randomly selected for observations which include in the training or testing data. In addition, the prediction performance evaluated by using the testing data with three evaluation criteria; accuracy, sensitivity, and specificity. In this case, the concept to calculate these criteria can be described below [7]

	Reference	
Predicted	1	0
1	A	B
0	C	D

where

$$Accuracy = \frac{A + D}{A + B + C + D}$$

$$Sensitifity = \frac{A}{A + C}$$

$$Specificity = \frac{D}{B + D}$$

5. Results and Discussion

There is a binary response variable used in this research. Figure 1 shows the percentage of each category of ASD, that is 29.6% ASD suffered and 70.4% non-ASD suffered. In this case, can be mention that the data is not in balance category condition in the response variable.

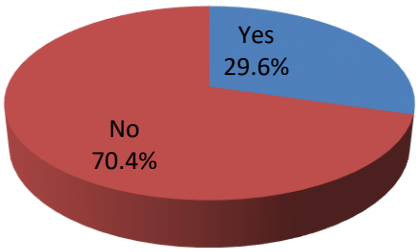
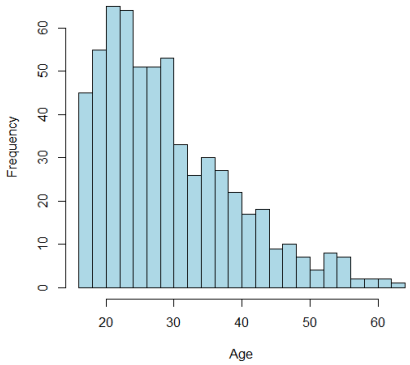


Figure 1. Pie chart of the category of ASD in the response variable.



Age	
Min.	17.00
1st Qu.	22.00
Median	27.00
Mean	29.65
3rd Qu.	35.00
Max.	64.00
Std Dev	9.69

Figure 2. Histogram and statistics summary table of Age.

There are six predictor variables in this case which one of them has a numeric scale; that is age. Figure 2 describes the distribution and statistics summary of age. The distribution of age tends to right tiled with mean 29.65 years and standard deviation of 9.69 years. The five predictor variables have the categorical variables with the description in Figure 3.

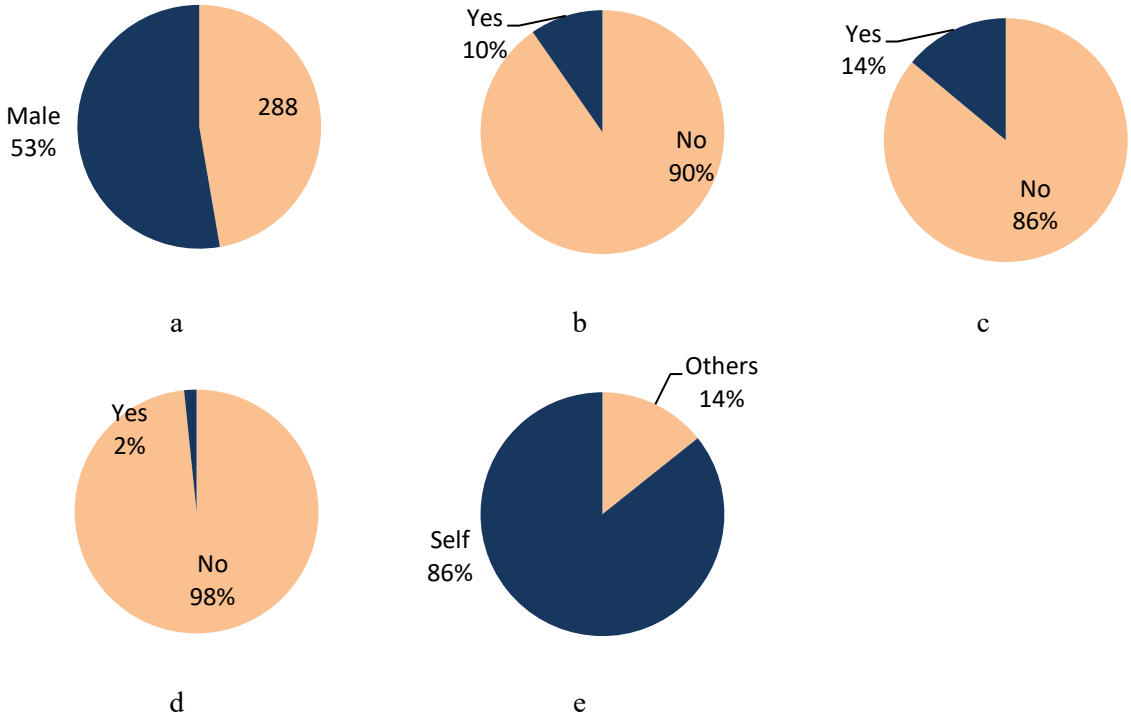
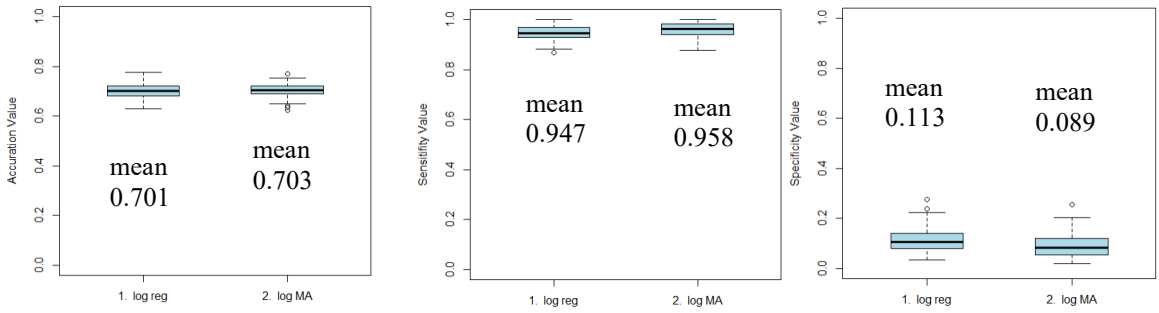


Figure 3. Pie chart of (a) gender, (b) born with jaundice, (c) ASD suffered from family, (d) used the screening app before, and (e) person who took the tests.

From the figure above, the majority of respondents who took the test are male (53%), born with no jaundice (90%), have no ASD in the family (86%), haven't used the app before (98%), and did the test by their self (86%).

The second part is the modeling of the data using logistic regression and logistic model averaging method. Figure 4 shows the distribution of 100 replications of the modeling process of prediction evaluation criteria.



(a) (b) (c)

Figure 4. Boxplots and mean values of prediction evaluation; (a) accuracy, (b) sensitivity, (c) specificity.

Based on the result, the logistic model averaging method has a higher mean of accuracy and mean sensitivity. It is a very good result in this case because, with the logistic model averaging method, class of ASD suffered can be predicted very well besides using the logistic regression. Although, mean the specificity value of the proposed method is lower than the logistic regression method. Furthermore, in this data, logistic model averaging could be a very good alternative to predict the class of ASD suffered.

6. Conclusion

Based on this case, it can be concluded that model averaging can be a good alternative in the predicting of response variable not only in the numeric scale but also in the categorical scale by implementing the logistic regression process. The ASD data has imbalance category in the response variable. The logistic model averaging method has better accuracy and sensitivity in the evaluation prediction of a class of ASD suffered.

References

[1] Myers RH. 1990. Classical and Modern Regression with Applications (Second Edition). Boston: PWS-Kent Publising Company
[2] Ghosh D, Yuan Z. 2009. An Improved Model Averaging Scheme for Logistic Regression. J Multivar Anal. 100(8): 1670–1681.
[3] Burnham KP, Anderson DR. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer-Verlag
[4] Tabtah F. 2017. Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.
[5] Perrone MP. 1993. Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization [dissertation]. Providence(US): Brown University.
[6] Claeskens G, Hjort NL. 2008. Model Selection and Model Averaging. New York (US): Cambridge University Press.
[7] Kuhn M. 2008. Building Predictive Models in R Using The caret Package. Journal of Statistical Software.