International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018) pp. 1661-1663

A New Strategy for Statistical Modeling using Bootstrap and Weighted Data in Bayesian Logistic Regression

Wan Muhamad Amir W Ahmad [1], Nurfadhlina Abdul Halim[2], Nor Azlida Aleng[3], Zalila Ali[4] and Ruhaya Hasan[5]

[1, 5] *School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kelantan.*
[2, 3] *School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Malaysia.*
[4]*School of Mathematics Sciences, Universiti Sains Malaysia (USM), 11800 Minden, Pulau Pinang, Malaysia.*

Abstract

Research on statistical modeling develop very fast over the year and some of them involve with a new ideas, methods, improvement of a previous methodology and many more but, the basic things is, it must comprise of various comprehensive techniques and new ideas. To gain a better significant results, many researcher that study and emphasize on specific method been tried to build a new or improved methodology. From the origin method, many reseachers developed new idea or strategy stage by stage accordingly. This research paper, illustrate the weighted logistics regression with combination of bootstrapping method (SAS Macro) and weighted by standard and varians technique. Related algorithm with specific illustration is fully given in this paper.

Keywords: Logistic regression, weighted, standard deviation and variance

INTRODUCTION

Logistic regression analyzes the relationship between two types of variables, independent variables and dependent variables (Hosmer and Lemeshow, 2000; (Amir et al., 2010).). Suppose that we have *k* independent observations $y_1 \cdots y_k$, and that the *i*-th observation can be treat as a realization of a random variable Y_i. So, we assume that Y_i has a binomial distribution as formula follows:

$$Y_i \sim B(n_i, \pi_i) \qquad (1)$$

With binomial denominator n_i and probability π_i. The probability distribution function of Y_i. is given by

$$\Pr\{Y_i = y_i\} = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

with mean, $E(Y_i) = \mu_i = n_i \pi_i$ and variance, $\text{var}(Y_i) = \sigma_i^2 = n_i \pi_i (1-\pi_i)$ (Hosmer and Lemeshow, 2000). Consider a collection of p independent variables denoted by the vector $x' = (x_1, x_2, x_3 \ldots, x_p)$. For the momen we will assume that each of these variables is at least interval scale. Let the conditional probability that the outcome is present be denoted by $P(Y_i = 1 | x) = \pi(x)$.

The logit of the multipe logistic regression model is given by the equation

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (2)$$

in which case the logistics regression model is

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

In 1993, Efron et. al had introduced bootstrap method which is emphasize on empirical density function (EDF). Bootstrap method initiates with an original sample data which is taken from the studied population and copies the original sample in a number of times to create a pseudopopulation. In pseudopopulation process, the bootstrap draws several samples with replacement according to the needs of the study. In this method, it uses the random sampling approach with replacement, and as a result it provides different samples from the original sample. This technique stores the new set of data and creating a new distribution for further analysis. (Efron et al., 1993; Higgins, 2005; (Ahmad et al., 2017).). In our study, we used PROC GENMOD options to fits generalized linear models using Bayesian methods by considering Bayesian estimation procedures. In Bayesian estimation procedures, the posterior distribution plays an important role in statistical inferential.

Weighting technique is a very important concept in the analysis of sample data. This method also involving of an adjusting data to reflect dissimilarities in the number of population units that each respondent represents. Besides that, this weighted techniques allows us to assign different weights to the different cases in data analysis. Most of the researcher use weighting method to correct skewness of a studied sample and to obtained parametric analysis. The second obejctive of using weighted method is to make the sample more representative of a "true" population.

To obtain the weighted data, researchers must first decide what kind of weighted procedure should be used. There is a lot of selection that can be choosing as such weighted by mean, standard deviation or variances. Weights techniques make sure the sample is representative of the population of interest.

METHODOLOGY

/*ADDING BOOTSTRAPPING ALGORITHM TO THE METHOD */

```
%MACRO bootstrap (data=_last_, booted=booted, boots=2, seed=1234);
DATA &booted;

** randomly picks an integer from 1 to n;
    pickobs = INT(RANUNI(&seed)*n)+1;
** POINT tells SAS to read value pickobs
** NOBS sets n to number of obs in &Data;
** when the point option is used SAS will loop through the data step forever;
SET &data POINT = pickobs NOBS = n;

** saves number of current bootstrap;
    REPLICATE = int(i/n)+1;
    i+1;
** stop will leave data set when n*&boots obs have been created;
    IF i > n*&boots THEN STOP;
    RUN;
%MEND bootstrap;

Data boot1;
input size mobility fbody bform lhabit fhabit weight;
datalines;
6 6 0 1 4 1 0.00
6 6 0 1 4 1 0.00
4 4 2 1 1 1 0.03
7 7 0 3 2 1 0.01
7 7 0 3 2 1 0.01
3 3 1 4 4 1 0.03
3 3 1 4 4 1 0.03
3 3 3 2 5 1 0.00
3 3 0 3 3 1 0.00
3 3 0 3 2 1 0.02
3 3 0 3 2 1 0.02
3 3 0 3 2 1 0.02
3 3 0 3 2 1 0.07
3 3 2 1 1 1 0.00
6 6 1 1 1 1 0.01
6 6 1 1 1 1 0.01
6 6 2 0 4 1 0.01
5 5 0 3 4 1 0.00
3 3 0 4 3 1 0.02
5 5 0 1 2 1 0.00
4 4 2 1 1 1 0.03
4 4 2 1 2 1 0.01
3 3 3 2 5 1 0.00
;
ods rtf file='abc.rtf' style=journal;

/*GENERATE BOOTSTRAP SAMPLE*/
%bootstrap (data= boot1, boots=2);
run;

/* BOOTSTRAPING DATA PRINTING */
 proc print data=booted;
run;

/*LOGISTIC REGRESSION USING BOOSTRAPPING DATA WITH OUTPUT DATASET FOR FORECAST */
proc genmod data= booted descending;
class fhabit;
  model fhabit = size fbody bform lhabit weight / dist=binomial link=logit;
output out=data_kajian reschi=residual;
run;

/*COMPUTE AND SAVE RESIDUAL, ABSOLUTE AND SQUARED RESIDUALS*/
data bentos;
  set data_kajian;
  absresid=abs(residual);
  sqresid=residual**2;
run;
```

```
/*RUN A REGRESSION WITH THE ABSOLUTE RESIDUALS VS. INDEPENDENT VARIABLES TO GET THE ESTIMATED STANDARD DEVIATION*/
proc genmod data=bentos;
   model absresid = size fbody bform lhabit weight;
  output out= s_data_kajian p=standard_hat;
run;

/* COMPUTE THE WEIGHTS USING THE ESTIMATED STANDARD DEVIATIONS*/
data s_data_kajian;
  set s_data_kajian;
standard_deviation_weight=1/(standard_hat**2);
label standard_deviation_weight = "weights using absolute residuals";

/* DO THE WEIGHTED LEAST SQUARES USING THE WEIGHTS FROM THE ESTIMATED STANDARD DEVIATION*/
proc genmod  data= s_data_kajian;
    weight standard_deviation_weight;
     model fhabit = size fbody bform weight/dist=binomial link=logit;
run;

ods graphics on;

proc logistic descending data=s_data_kajian plots= effect plots= roc(id=prob);
    weight standard_deviation_weight;
       model fhabit(event='1') = size fbody
    bform weight/rsquare expb lackfit;
roc 'size fbody bform weight dist' size fbody bform weight ;
run;

proc genmod data= s_data_kajian descending;
  weight standard_deviation_weight;
  model fhabit = size fbody bform lhabit weight/dist=binomial link=logit;
  bayes nbi=1000 nmc=10000 thin=2 seed=1 out=posterior;
run;

ods graphics off;
ods rtf close;
```

CONCLUSION

This study provides algorithm for weighted data by standard deviation with bootstrapping bayesian logistics regression by standard and varians technique. This proposed method can be applied to small sample size data, especially when limited data is obtained. It also can be used as an alternative method of data analysis for reseacher to obtain a better precise results.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Universiti Sains Malaysia (USM) for providing the research funding (Grant no.304/PPSG/61313187, School of Dental Sciences)

REFERENCES

[1] Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall.
[2] 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
[3] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B.,Vehtari, A., andRubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, third edition.
[4] Hosmer, D.W. and S. Lemeshow, 2000. Applied logistic regression, second edition, John Wiley and Sons.
[5] W.M.A.W. Ahmad., N.A. Aleng and Z. AIi, 2010. Binary logistic regression analysis technique used in analyzing the categorical data in education sciences: A case study of Terengganu State, Malaysia. World Appl. Sci. J., 9(9): 1062-1066.
[6] W.M.A.W. Ahmad., N.A. Aleng and Z. AIi, 2017. Statistical Analysis with SAS. Pulau, Pinang. Penerbit USM.
[7] Wardle, J., Guthrie, C. A., Sanderson, S., & Rapoport, L. (2001).
[8] Development of the children's eating behaviour questionnaire. Journal of Child Psychology and Psychiatry, 42, 963–970