Journal of Scientific Research and Development 3 (5): 51-54, 2016
Available online at www.jsrad.org
ISSN 1115-7569

An algorithm for weighted and bootstrap logistic regression modelling in benthic organism

Wan Muhammad Luqman Wan Rosdi [1,*], Wan Muhamad Amir W Ahmad [1], Ruhaya Hasan [1], Nor Azlida Aleng [2], Nurfadhlina Halim [2], Syerrina Zakaria [2], Kasypi Mokhtar [3], Zalila Ali [4]

[1]*School of Dental Sciences, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kelantan, Malaysia*
[2]*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT), Terengganu Malaysia*
[3]*School of Maritime Business and Management, Universiti Malaysia Terengganu (UMT), Terengganu Malaysia*
[4]*School of Mathematics Sciences, Universiti Sains Malaysia (USM), 11800 Minden, Pulau Pinang, Malaysia*

Abstract: This paper supplied a complete method of an alternative weighted logistic regression as a technique for analysis through SAS algorithm. This paper also aims to investigate the effect of the fish cage aquaculture activity to the benthic communities in Bidong Island. Benthic organisms have been one of the most common organisms used as indicators for assessing environmental quality in marine environments due to their diversity and known characteristics such as limited mobility. Literally, it means those benthic organisms are unable to avoid themselves from the changes in the environment. This alternative method is a manipulation technique (using bootstrap) for the small data set and gives the researcher an option to launch the analysis even there is not enough data set. This is an extension of an improvement of the current method by adding weighted and bootstrap to step of building logistic regression model.

Key words: SAS algorithm; Aquaculture; Bootstrap; Weighted and logistic regression

1. Introduction

This paper provides a road map of the practical approach of logistic regression modeling and an illustration using aquaculture dataset. The benthic impact of fish farm wastes on the benthic environment has mainly been assessed in terms of changes in macro faunal abundance and diversity (Tsutsumi, 1995). Benthic communities play important roles in the maintenance of ecosystem functioning and the links between benthic and pelagic systems (Bremner and Rogers, 2006) and these roles are determined by the biological traits species exhibit (Bremner and Rogers, 2006). Benthic infaunal communities demonstrate the ability to change in a predictable manner along gradients of natural and anthropogenic stresses (Pearson and Rosenberg, 1978). Benthic infauna have been one of the most common organismal groups used as indicators for assessing environmental quality in marine environments due to their diversity and known characteristics such as limited mobility (meaning they are unable to avoid the environmental changes as most pelagic fauna can) (Gray, 1979) and long life spans of up to several years (Nelson, 1990). Logistic regression is a type of predictive model that can be used when the target variable is a categorical variable with two categories for instance live or die, has cancer or no cancer, having coronary heart disease or not having coronary heart disease, patient survives of dies and many more (Amir et al., 2010). In logistic regression, the dependent variable is a binary or dichotomous; it only contains data coded as 1 or 0. The objective of logistic regression is to find the best fitting model to illustrate the relationship between the dichotomous characteristics of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. The parametric bootstrap method is recommended for sample size between 50 and 100 for a reliable performance (Jung et al., 2005; Cassel, 2010). A recent approach to analyse data with missing values in the covariates is weighted estimating equations and this technique appear to be highly efficient (Zhao et al., 1996). The objective of this study is to discuss the improvement between the advanced multiple logistic regression with the original version of multiple logistic regression (Hosmer and Lemeshow, 2000). We extend the advanced of multiple logistic regressions with combining bootstrap and response surface methodology. The response surface methodology (RSM) was introduced by Box and Wilson in 1951 (Lockwood and Mackinnon, 1998; Cassel, 2010). The response surface methodology (RSM) explores the relationships between several explanatory variables (X) and one or more response variables (Y). The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response through

* Corresponding Author.

linear model and second-degree polynomial. They acknowledge that this model is only an approximation, but use it because such a model is easy to estimate and apply, even when little is known about the process. According to Mead and Pike stated origin of RSM starts 1930s with use of *Response Curves* (Myers et al., 1989).

Bootstrap method is a statistical technique that falls under the broad heading of resampling. This method is very useful and can be used various especially in the estimation of nearly any statistics (Cassel, 2010). This procedure involves a relatively simple procedure, but repeated so many times depending on the need of the researcher. Bootstrap technique is heavily dependent upon computer calculation. Using the bootstrap method, we are able to determine the estimating value of a parameter that presenting the whole of a population. Without using bootstrap method, the value of the parameter of a population is impossible to measure directly. So, we use statistical sampling method and we sample a population, measure a statistic of this sample, and then use these statistics to say something about the corresponding parameter of the population (Cassel, 2010).

2. Data and methods

Data of this study is a sample which composed of six variables. Namely variables are as in Table 1.

Table 1: Description of data

Num.	Code	Explanation of user variables
1.	Y	Feeding Habit 0 = Suspension Feeder 1 = Deposit Feeder
2.	X1	Size
3.	X2	Mobility
4.	X3	Flattened Body
5	X4	Body form
6	X5	Life Habitat
7	X6	Weight
8	X7	Distribution

Multiple weighted logistic regression technique was used in the analysis of relationship between variables. The algorithm is given as follows. Indeed, Fig. 1 shows the flow chart of an alternative method of logistic regression procedure.

2.1. Case Study I: Calculation for Bootstrap Logistic Regression Using SAS

```
Data Sampling;
input size mobility fbody bform lhabit fhabit weight dist;
datalines;
4 4 0 2 3 0 0.0014 4
5 5 3 1 6 0 0.0048 4
5 5 0 1 5 0 0.0066 6
5 5 0 2 5 0 0.0152 4
6 6 0 3 5 0 0.0031 7
5 5 0 1 3 0 0.0012 13
3 3 3 2 3 0 0.0113 7
3 3 2 2 2 0 0.0009 2
3 3 2 2 6 0 0.0026 1
6 6 0 3 6 0 0.0141 19
3 3 0 2 1 0 0.0045 2
6 6 0 1 4 1 0.0031 2
4 4 0 2 4 0 0.0013 18
⋮ ⋮ ⋮ ⋮ ⋮
9 9 0 3 3 0 0.0207 9
6 6 1 0 2 1 0.0068 5
7 7 1 3 2 0 0.0012 5
3 3 0 2 2 0 0.5609 14
3 3 0 3 2 1 0.0693 1
3 3 2 1 1 1 0.0015 2
6 6 1 1 1 1 0.0066 13
3 3 0 4 3 0 0.0007 5
6 6 2 0 4 1 0.0132 3
5 5 0 3 4 1 0.0001 1
4 4 0 1 6 0 0.0019 2
3 3 0 4 3 1 0.0233 5
6 6 1 0 1 0 0.0009 6
3 3 2 1 2 1 0.0043 3

;
ods rtf file='robdunc0.rtf' style=journal;


Title "Performing bootstrap with case resampling";
Proc surveyselect data=Sampling out=boot1 method=urs samprate=1 outhits rep=60;
run;


proc print data =boot1;
run;

ods rtf close;
```

2.2. Case Study II: Calculation for Alternative Method for Logistic Regression Using SAS

```
Data boot1;
input size mobility fbody bform lhabit fhabit weight dist;
datalines;
6 6 0 1 4 1 0.00 2
6 6 0 1 4 1 0.00 2
4 4 2 1 1 1 0.03 4
7 7 0 3 2 1 0.01 51
7 7 0 3 2 1 0.01 51
3 3 1 4 4 1 0.03 4
3 3 1 4 4 1 0.03 4
3 3 3 2 5 1 0.00 3
⋮ ⋮ ⋮ ⋮ ⋮
3 3 0 3 3 1 0.00 4
3 3 0 3 2 1 0.02 18
5 5 0 2 5 0 0.02 4
6 6 0 3 5 0 0.00 7
6 6 0 3 5 0 0.00 7
5 5 0 1 3 0 0.00 13
5 5 0 1 3 0 0.00 13
```

```
6 6 0 3 6 0 0.01 19
6 6 0 3 6 0 0.01 19
6 6 0 3 6 0 0.01 19
3 3 0 2 1 0 0.00 2
3 3 0 2 1 0 0.00 2
3 3 0 2 1 0 0.00 2
4 4 0 2 4 0 0.00 18
4 4 0 2 4 0 0.00 18
4 4 0 2 4 0 0.02 8
4 4 1 1 2 0 0.11 7
;

ods rtf file='robdunc0.rtf' style=journal;


/* Run the original logistic regression + boostrap to get the residuals*/
proc genmod data=boot1 descending;
 class fhabit;
   model fhabit = size fbody bform lhabit weight dist/ dist=binomial link=logit;
 output out=work.pred  reschi=residual;
run;


/* Compute the absolute and squared residuals*/
data work.resid;
 set work.pred;
 absresid=abs(residual);
 sqresid=residual**2;
proc genmod data=work.resid;

/*Run a regression with the absolute residuals vs. independent variables to get the estimated standard deviation*/
   model absresid = size fbody bform lhabit weight dist;
   output out=work.s_weights p=s_hat;
run;

/*Compute the weights using the estimated standard deviations*/
data work.s_weights;
 set work.s_weights;
 s_weight=1/(s_hat**2);
 label s_weight = "weights using absolute residuals";


/*Do the weighted least squares using the weights from the estimated standard deviation*/
proc genmod  data=work.s_weights;
  weight s_weight;
  model fhabit = size fbody bform weight dist/ dist=binomial link=logit;
 run;


ods graphics on;
  proc logistic descending data=work.s_weights plots= effect plots= roc(id=prob);
     weight s_weight;
   model fhabit(event='1') = size fbody bform weight dist / rsquare expb lackfit;
   roc 'size  fbody bform weight dist' size  fbody bform weight dist;
  run;
ods graphics off;


ods rtf close;
```

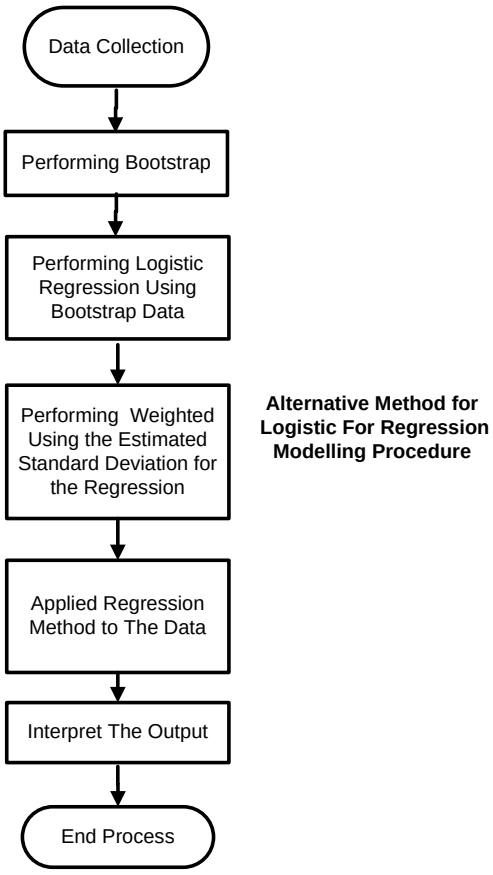

Fig. 1: Flow chart of an alternative analysis

3. Summary and conclusion

This paper explained on how an alternative programming method of bootstrap weighted multiple logistic regression procedure using SAS software. This method can be applied for the small sample size data especially where the data is very difficult to collect. By resampling (using bootstrap method), it provides the pleminary comprehensive information and also give the general overview on how the data behaviour even though the original data is not enough (small sample size). In our case, smaller standard error of the estimate parameter will tell us how accurate our estimate parameter is likely to be. It is not easier to understand the behaviour of the data in studies when it is not reaching the actual sample size needed in an analysis (Naing, 2003).

References

Amir WA, Nor Azlida Aleng and Zalila Ali. 2010. Binary Logistic Regression Analysis Technique

Used in Analyzing the Categorical Data In Education Sciences: A Case Study of Terengganu State, Malaysia. World Applied Sciences Journal 9 (9): 1062-1066,

Bremner, J., and Rogers, S. I. (2006). Ecological indicators. Methods for describing ecological functioning of marine benthic assemblages using biological traits analysis (BTA), 6, 609-622.

Bremner, J., and Rogers, S. I. (2006). journal of marine system. Matching biological traits to environmental conditions in marine benthic ecosystems, 60, 302-316.

Cassel D.L.(2010). Bootstrap Mania: Re sampling the SAS. SAS Global Forum 2010 : Statistics and Data Analysis. Paper 268-2010: Pp 1-11

Cassel D.L.(2010). Bootstrap Mania: Re sampling the SAS. SAS Global Forum 2010 : Statistics and Data Analysis. Paper 268-2010: Pp 1-11

D. W. Hosmer and S. Lemeshow. 2000. Applied logistic regression, second edition, John Wiley and Sons.

Gray, J.S. 1979. Pollution-induced changes in populations. Philosophical Transactions of the Royal Society of London Series B, 268:545-561.

Jung, B. C., Jhun, M., and Lee, J. W. (2005). Bootstrap Tests for Overdispersion in a Zero-Inflated Poisson Regression Model. *Biometrics*,*61*(2), 626-628.

Lockwood, C.M., and Mackinnon, D.P. (1998). Bootstrapping the standard error off the mediated effect. Proceedings of the 23rd Annual Meeting of SAS users Group International (pp. 997-1002). Cary, NC: SAS Institute, Inc.

Myers R.H., Khuri A.I., Carter W.H. (1989), "Response surface methodology : 1966- 1988. Technometrics, 15, 301–317.

Naing, N. N. (2003). Determination of sample size. *The Malaysian journal of medical sciences: MJMS*, *10*(2), 84-86

Nelson, W.G. 1990. Prospects for development of an index of biotic integrity for evaluating habitat degradation in coastal ecosystems. Chemistry and Ecology, 4: 197-210.

Pearson, T., Rosenberg, R. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology Annual Review 16:229-311.

Tsutsumi, H., 1995. Impact of a fish net pen culture on the benthic environment of a cove in South Japan. Estuaries 18 (1A), 108–115.

Zhao LP, Lipsitz SR, Lew D. (1996) Regression analysis with missing covariate data using estimating equations . Biometrics; 52:1165 – 1182.