Bootstrap Residual Ensemble Methods for Estimation of Standard Error of Parameter Logistic Regression To Hypercolesterolemia Patient Data In Health Laboratory Yogyakarta

Fransiska Grace S.W.[1] , Sri Sulistijowati H.[2], and Titin Sri Martini [3]
[1]Matematics Department FMIPA Universitas Sebelas Maret
[2,3] Statistics Department FMIPA Universitas Sebelas Maret

[1]fransiskagrace@gmail.com
[2]rr_ssh@staff.uns.ac.id
[3]titinsmartini@gmail.com

Abstract. Logistic regression is one of regression analysis to determine the relationship between response variable that have two possible values and some predictor variables. The method used to estimate logistic regression parameters is the maximum likelihood estimation (MLE) method. This method will produce a good estimate of the parameters if the estimation results have a small standard error.

In a research, the characteristics of good data must be representative of the population. If the samples taken in small size they will cause a large standard error value. Bootstrap is a resampling method that can be used to obtain a good estimate based on small data samples. Small data will be resampling so it can represent the population to obtain minimum standard error. Previous studies have discussed resampling bootstrap on residuals as much as b times. In this research we will be analyzed resampling bootstrap on the error added to the dependent variable and take the average parameter estimation ensemble logistic regression model resampling result. Next we calculate the standard value error logistic regression parameters bootstrap results.

This method is applied to the hypercholesterolemic patient status data in Health Laboratory Yogyakarta and after bootstrapping, the standard error produced is smaller than before the bootstrap resampling.

Keywords : logistic regression, standard error, bootstrap resampling, parameter estimation ensemble

1. Introduction

Logistic regression is a non linear regression where the relationship curve between the response variable and the predictor variable is not a straight line. Logistic regression is used as a method to analyze the relationship of binary response variables (0 and 1) with predictor variables. However, a problem arises when the samples taken are small in size. Whereas the characteristics of good data should be representative which means the sample data objective and describe the population so that the sample can represent the population. If the sample taken is much smaller than the size of the population then it is less representative so that the conclusions obtained produce a fairly large standard error. Therefore, a method is needed to solve the problem. Efron and Tibshirani [3] introduced a

 ISSN 2621-086X

resampling method known as the bootstrap method that can resampling small samples with the help of a computer. This method assumes that the empirically distributed sample is then considered a population and from that population resampel can be done. The size of the bootstrap resampling is better taken quite a lot in order to represent the population data so that the resulting standard error is small.

Previous studies (Sahinler and Topuz [9], Hossain and Khan [1]) discussed the bootstrap resampling algorithm for the estimation of linear regression and logistic regression parameters by resampling the residuals generated from the model. Furthermore Pardoe and Weisberg [6] discusses the conditional probability bootstrap method which is a bootstrap method if the value of a variable the response is influenced by the predictor variables.

In this study the model parameter estimation is done by the residual ensemble bootstrap method as has been done Handajani et al [10] on the spatial regression model. This method is done by calculating the residual from the estimation result of the logistic regression model and the residual value is bootstrapped k times. The bootstrap results are decomposed according to the original data and obtained by a number of groups in which each group is estimated to model parameters. Next we combine the results from the estimation of each residual group with its mean value into a logistic regression model by producing a smaller error standard.

2. Logistic Regression

Logistic regression is a regression analysis where response variables have only two possible events. Logistic regression model is started from Linear Probability Model (LPM) which is application of classical linear regression on categorical response. LPM transformed the linear regression model

$$Y = \pi(x) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \varepsilon_i \quad \text{.....................................(1)}$$

into the linear probability model as follows

$$\pi(x) = \frac{exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}+\varepsilon_i)}{1+exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}+\varepsilon_i)}.$$

The LPM in (1) above shows the right-hand side is unlimited (since the value of x is continuous) but the left-hand Y or π (x) value must be limited (0 or 1). Therefore the lefthand side of model (1) must be changed so that the left segment of value 0 and 1 can have values between -∞ to ∞ like the right-hand segment to obtain a logistic regression model (Hosmer dan Lemeshow [4]).

To estimate the parameters in nonlinear regression, especially logistic regression was used the maximum likelihood method. Basically this method gives an estimate value of β by maximizing its likelihood function (Hosmer and Lemeshow [4]).

Mathematically the probability distribution of the Y function can be expressed as follows,

$$f(y_i) = (\pi_i)_{y_i}(1-\pi_i)_{1-y_i}.$$

Each observation of y is mutually independent then the likelihood function is the multiplication of each probability distribution that is

$$L(\beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{1-y_i}$$

To find Maximum Likelihood Estimation (MLE), the log values of both are obtained so that the log likelihood function for logistic regression is

$$\log L(\beta) = \sum_{i=1}^{n} y_i \log\left(exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} + \varepsilon_i\right)\right) + \sum_{i=1}^{n}\log\left(1 + exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} + \varepsilon_i\right)\right)^{-1} .$$

Maximize the log likelihood function by partially deriving the function against parameters to be estimated and matching with zero to obtain the normal equation. To solve non linear equations on maximum likelihood method required an appropriate algorithm and this research used Fisher Scoring algorithm. Fisher Scoring is a development of the Newton-Raphson method. In the Newton-Rapshon method, the Hessian matrix uses a second derivative for the calculation at each iteration. This results in the iteration not always convergent. Based on this, the Newton-Rapshon algorithm is modified by substituting the Hessian matrix into a Fisher information matrix and herein after called the

Fisher scoring algorithm (Putri et al. [2]). The fisher scoring iteration equation is $\hat{\beta}_{(m+1)} = \hat{\beta}_{(m)} + \boldsymbol{Inf}_m^{-1}\boldsymbol{S_m}$

2.1. Bootstrap Method. The bootstrap method is a method of estimating a population distribution of small sample sizes or replacing the unknown assumption of distribution with the empirical distribution obtained from the resampling process (Efron and Tibshirani [3]; Sungkono [7]). The bootstrap approach uses a sampling method with returns. The basic idea of the bootstrap method is to build artificial samples using information from the original data. According to Teknomo [8], the bootstrap method depends on its own source or can is said to depend on the sample which is the only source owned researchers.

The bootstrap estimation for $se_F(\hat{\theta})$ is the standard error of $\hat{\theta}$ for the random number of n data taken from F. (Efron and Tibshirani [3]).

2.2. Data and Method. Secondary data taken from Health laboratory of Yogyakarta. The data used in this case is data on the cholesterol status of 20 patients. The data obtained are response variable that is patient cholesterol status and predictor variable that is LDL, HDL and triglyceride level from patient data of hypercholesterolemia. Analytical steps taken are

Step 1. Estimate the standard error of original sample logistic regression parameters
Step 2. Determine the logistic regression model so that the value of $\hat{Y}$ is obtained
Step 3. Calculates the residual value of the model obtained from the difference between Y and $\hat{Y}$

Step 4. Resampling the residue generated from the model with bootstrap
Step 5. Estimates the regression model parameters of each group of response variables that have been added with the residual value resulting from the bootstrap resampling with predictor variables

Step 6. Calculate the standard error value of the logistic regression coefficient by the bootstrap method

Step 7. Determine the ensemble regression logistic regression equation by averaging the logistic regression coefficients of k the regression equation obtained

3. Results and Discussion

The data used is secondary data obtained from Job Training Report by Prabandari [5] which comes from the internal data of the Pathology of Health laboratory in Yogyakarta. Data on the cholesterol status of 20 patients is presented in the following table.

Table 1. Data on the cholesterol status

	Y	X1	X2	X3
1	0	29	130	99
2	1	43	187	85
3	1	36	145	115
4	1	60	194	106
5	1	30	171	453
…	…	…	…	…
19	0	38	82	102
20	1	35	172	186

3.1. Estimation of Logistic Regression Parameters

To estimate logistic regression parameters with MLE on 3 independent variables using R-software and obtained the output in the following table.

Table 2. Logistic regression coefficient value and error standard

Parameter	Koefisien	Eror Standar
β_0	-360.8245972	411200.00
β_1	1.5549656	5764.00
β_2	2.3619113	2848.00
β_3	-0.1374795	1618.00

3.2. Bootstrap Replication

The bootstrap resampling is applied to the residual values generated from the logistic regression model. The amount of data from the bootstrap replication depends on

how much data there is in the observed data multiplied by the desired number of bootstrap replications. In this research, the bootstrap loop counted 100 times with each data amount of bootstrap result of 20 data.

Next we regret each group of response variable data that has been added residue value resampling bootstrap with predictor variable. After bootstrapping the residue 100 times and obtained the estimation of logistic regression parameters from the response variables that have been added with the residual result of bootstrap with predictor variable then we can calculate the standard error value of logistic regression parameters with

$$se_F(\hat{\theta}) = \left\{ \frac{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2}{B-1} \right\}^{\frac{1}{2}}$$

$$\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$$

and is presented in Table 3 below

Table 3. The standard error value of logistic regression parameters

Parameter	Standard error value
β_0	0.248727
β_1	0.001783017
β_2	0.002085892
β_3	0.000138602

After a standard error value the logical regression parameters of the bootstrap results are obtained, then compared with the standard error values in the original sample data. The result of standard error value comparison after and before bootstrapping is shown in Table 4 below.

Table 4. Comparison standard error value after and before bootstrapping

Parameter	Coefficient	Standard error	
		Original sample	Bootstrap 100 times
β_0	-1.155763	411200.00	0.248727
β_1	0.008357	5764.00	0.001783017
β_2	0.009311	2848.00	0.002085892
β_3	0.000646	1618.00	0.000138602

The results of the analysis obtained logistic regression model with 100 times recurrent bootstrap method, with the total probability model of cholesterol a patient will suffer from hypercholesterolemia is

$$\hat{Y} = \frac{exp(-1.155763 + 0.008357x_1 + 0.009311x_2 + 0.000646x_3)}{1 + exp(-1.155763 + 0.008357x_1 + 0.009311x_2 + 0.000646x_3)}$$

4. Conclusion

In the case of small amounts of data, the bootstrap method proved to provide better estimation results than using the original samples. This is evidenced by the bootstrap method is able to minimize the value of standard errors on the parameters up to 100 repetitions. The logistic regression model derived from the 100 times recurrent bootstrap method for the cholesterol total probability model of a patient will have hypercholesterolemia ie

$$\hat{Y} = \frac{exp(-1.155763 + 0.008357x_1 + 0.009311x_2 + 0.000646x_3)}{1 + exp(-1.155763 + 0.008357x_1 + 0.009311x_2 + 0.000646x_3)}$$

Acknowledgements

The authors would like to express heartfelt thanks to the Research Group of Applied Statistical and Inference for our helpful discussions. We also appreciate Sebelas Maret University for providing financial support.

References

[1] A.Hossain and H.T.A. Khan, *Nonparametric Bootstrapping for Multiple Logistic Regression Model Using R*, BRAC University Journal, Vol. I(2004), No.2, 109-113, Bangladesh.
[2] A. N.Putri, D.R.S.Saputro and P.Widyaningsih, "Informasi Fisher pada Algoritme Fisher Scoring untuk Estimasi Parameter Model Regresi Logistik Ordinal Terboboti Geografis (RLOTG)", Seminar Nasional, Jurusan Matematika, FMIPA UNS, Surakarta, 2016.
[3] B. Efron and R.J. Tibshirani, An Introduction to the Bootstrap, (Chapman & Hall, Inc, New York, 1993).
[4] D.W. Hosmer and S. Lemeshow, Applied Logistic Regression Second Edition, (John Wiley & Sons, Inc, New York, 2000).
[5] D.Prabandari, "Analisis Regresi Logistik Ganda dalam Memodelkan Faktor-Faktor Terindikasinya Penyakit Hiperkolesterol di Balai Laboratorium Kesehatan Yogyakarta", Laporan Kerja Praktek, Fakultas MIPA UGM, Yogyakarta, 2009.
[6] I. Pardoe and S.Weisberg. An Introduction to Bootstrap Methods using Arc, (University of Minnesota St.Paul, New York, 2001).
[7] J.Sungkono, *Resampling Bootstrap pada R,* FKIP UNWIDHA, Vol.XXV(2013), No.84, Klaten.
[8] K.Teknomo, Bootstrap Sampling Tutorial, Ateneo de Manila University, Manila, 2006.

[9] S. Sahinler and D.Topuz, *Bootstrap and Jackknife Resampling Algorithm for Estimation of Regression Parameters,* Journal of Applied Quantitative Method(Turkey), Vol.2(2007), No. 2, 188 – 199.
[10] S.S. Handajani, H.Pratiwi and Y. Susanti, "Penanganan Masalah Heterogenitas Eror pada Pemodelan Regresi Spasial", Laporan Penelitian, Universitas Sebelas Maret, 2017.