Dental Journal
Majalah Kedokteran Gigi
Vol. 41. No. 4 October–December 2008

Research Report

Bootstrap study to estimate linear regression parameter (Application in the study on the effect of oral hygiene on dental caries)

Ristya Widi Endah Yani
Department of Dental Public Health
Faculty of Dentistry Jember University
Jember - Indonesia

ABSTRACT

Background: *Bootstrap is a computer simulation-based method that provides estimation accuracy in estimating inferential statistical parameters.* ***Purpose:*** *This article describes a research using secondary data (n = 30) aimed to elucidate bootstrap method as the estimator of linear regression test based on the computer programs MINITAB 13, SPSS 13, and MacroMINITAB.* ***Methods:*** *Bootstrap regression methods determine* $\hat{\beta}$ *and* $\hat{Y}$ *value from OLS (ordinary least square),* $\varepsilon_i = Y_i - \hat{Y}_i$ *value, determine how many repetition for bootstrap (B), take n sample by replacement from* ε_i *to* $\varepsilon_{(i)}$*,* $Y_i = \hat{Y}_i + \varepsilon_{(i)}$ *value,* $\hat{\beta}$ *value from sample bootstrap at i vector. If the amount of repetition less than, B a recalculation should be back to take n sample by using replacement from* ε_i*. Otherwise, determine* $\hat{\beta}$ *from "bootstrap" methods as the average* $\hat{\beta}$ *value from the result of B times sample taken.* ***Result:*** *The result has similar result compared to linear regression equation with OLS method (*α *= 5%). The resulting regression equation for caries was = 1.90 + 2.02 (OHI-S), indicating that every one increase of OHI-S unit will result in caries increase of 2.02 units.* ***Conclusion:*** *This was conducted with B as many as 10,500 with 10 times iterations.*

Key words*: bootstrap, caries*

Correspondence: Ristya Widi Endah Yani, c/o: Departemen Ilmu Kedokteran Gigi Masyarakat dan Pencegahan, Fakultas Kedokteran Gigi Universitas Jember. Jl. Kalimantan No.37 Jember 68121, Indonesia. Telp. (0331) 333536. E-mail: ristya-widi@yahoo.com

INTRODUCTION

Bootstrap is a computer simulation-based method that provides estimation accuracy in estimating inferential statistical parameters. To solve problems in insufficient statistical samples (small number of samples), computer-based method such as bootstrap has rarely been employed although its use is quite simple.[1] The simplicity of application can be observed in the use of the media and more advanced method, which is an implementation of the basic concept in statistics. Being computer-based, this method does not use classical statistical method anymore, which application use a relatively complex formulation.[2]

Multiple linear regression analysis is an extension of simple linear regression analysis. Simple linear regression analysis of two variables correlation analysis is made between one dependent variable (Y) and one independent variable (X). In multiple linear regression analysis, there are one dependent variable (Y) and more than one independent variables (Xi), in which i = 1, 2, 3 ... p, with an aim to predict Y value (dependent variable) based on X values (independent variables). The correlation between one independent variable and one dependent variable is discussed in simple linear regression, and correlation between more than one independent variables in multiple linear regression analysis.[3] As an application in this study, we used data on the effect of oral hygiene on dental caries. The dependent variable was dental caries and the independent variable was oral hygiene. The problem of this study addressed the process of bootstrap method application to assess linear regression parameter. The objective of this study was to evaluate bootstrap method as an estimation of

linear regression parameter. The benefit of this study was to find the solution using bootstrap method in estimating linear regression parameter.

MATERIALS AND METHODS

The research using secondary data,[4] with independent variable (oral hygiene) and dependent variable (dental caries). Data source was secondary data entitled "Permanent Teeth Eruption and Oral Hygiene among Elementary School Children in Goiter Endemic Area, District of Jember".[5] Data analysis was undertaken using computer (MINITAB 13, SPSS 13, MacroMINITAB). The algorithm of data regression method of bootstrap result (Figure 1).

This program begins to determine $\hat{\beta}$ and $\hat{Y}$ values from OLS, $\varepsilon_i = Y_i - \hat{Y}_i$, the number of repetition bootstrap (B), then taking n sample from the return of ε_i, which is regarded as $\varepsilon_{(i)}$. Determine $Y_i = \hat{Y}_i + \varepsilon_{(i)}$ values, and then determine $\hat{\beta}$ values in i[th] sample. If the number of repetition bootstrap < B consequently taking n sample from the return of ε_i, which is regarded as $\varepsilon_{(i)}$. If the number of repetition < B consequently determining $\hat{\beta}$ from "bootstrap" method as the average of $\hat{\beta}$ of sample taking in B times.

RESULT

Thirty out of 100 secondary data were randomized. The data were tried to be firstly subjected to linear regression

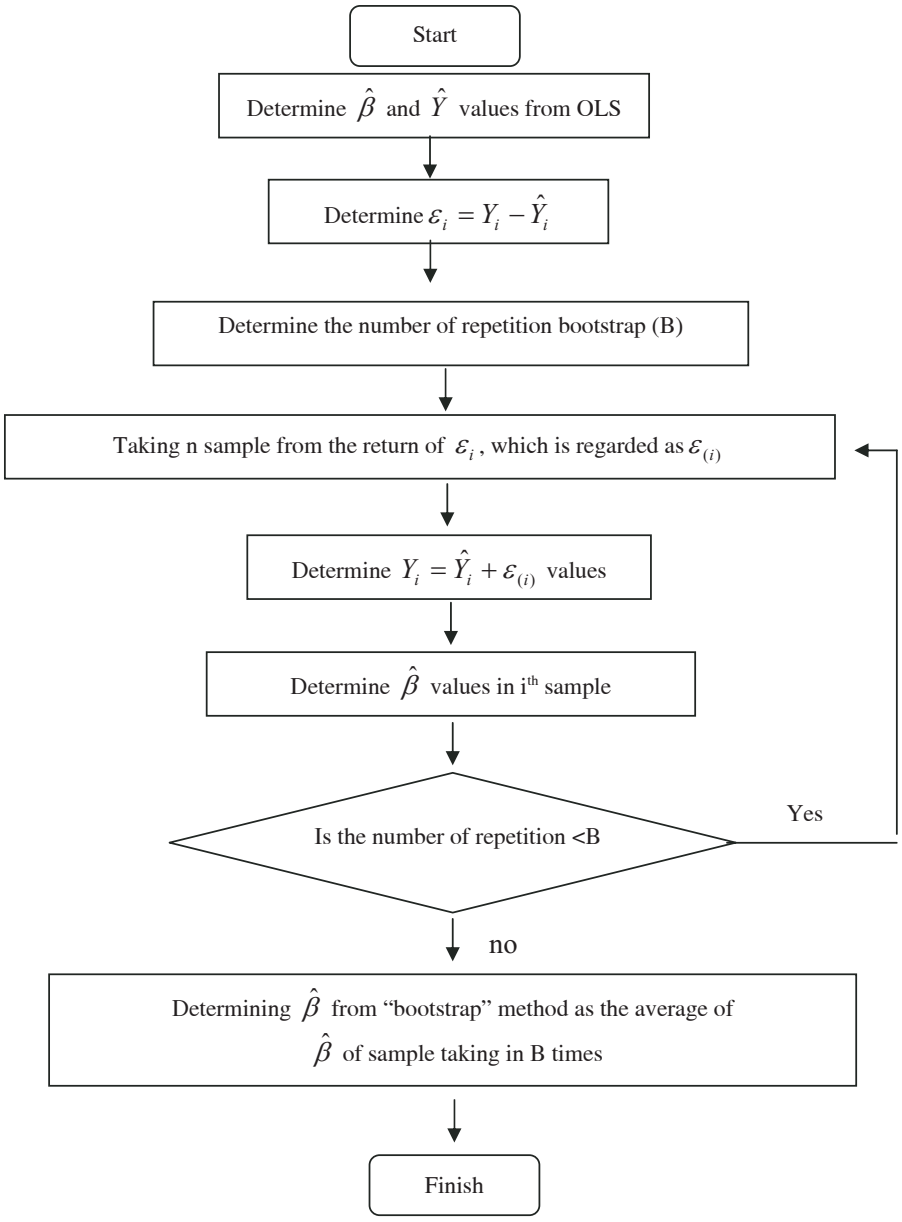

Figure 1. The algorithim of data regression method of bootstrap result.[6]

analysis, and after regression equation was obtained, they were subjected to linear regression analysis of the bootstrap result using MacroMINITAB program.

Determining regression coefficient parameter using linear regression analysis

Analysis regression: Caries vs. OHI-S
The regression equation is
Karies = 1.90 + 2.02 OHI-S

Predictor	Coef	SE Coef	T	P
Constant	1.9028	0.4557	4.18	0.000
OHI-S	2.0228	0.2660	7.61	0.000

S = 1.167 R-Sq = 67.4% R-Sq(adj) = 66.2%

Analysis of Variance

Source	DF	SS	M	F	P
Regression	1	78.818	78.818	57.85	0.000
Residual Error	28	38.149	1.362		
Total	29	116.967			

Unusual Observations

Obs	OHI-S	Karies	Fit	SE Fit	Residua	St Resid
6	1.90	10.000	5.746	0.236	4.254	3.72R

R denotes an observation with a large standardized residual
Durbin-Watson statistic = 2.00

Using enter method on the result of t test above, the p value for OHI-S variable was significant (p-value = 0.000). In variance test the significance was p = 0.000. The regression parameter of b_0 = 1.9028 and b_1 = 2.0228. The regression equation : Caries = 1.9028 + 2.0228 OHI-S or Caries = 1.90 + 2.02 OHI-S.

Several assumption tests:

ε_i was normally distributed and ε was a randomized variable with $(\Sigma\varepsilon_i) = 0$. The Kolmogorov-Smirnov test revealed "Approximate p -value" of > 0.15. Residual had normal distribution. The result of Spearman's correlation test revealed insignificant correlation between residual and OHI-S variable ("Sig" = 0.685), indicating no heteroscedacity. No correlation assumption was observed by comparing the values in Durbin Watson table to the values of Durbin Watson values from the estimation. The value of d was > du or 4-d > du, H_0 was accepted, indicating no correlation between residuals. The independent variable was only one, multicolinearity assumption test could not be performed. Plot dots were distributed around the value 0, indicating the presence of linearity.

Determining regression coefficient parameter of bootstrap result data

Bootstrap method applied was performed by resampling the residuals. The call of bootstrap command in the form of MacroMinitab with 10 iterations was %d:\bootstrap_baru.txt c1 c2 c3-c12 c13-c22 c23 c24 c25 c26 c27 c28, which was subsequently entered into Minitab program.

Description: %d:\bootstrap_baru.txt is formula bootstrap, c1 is column variable dependen, c2 is column variable independent, c3-c12 is column regression parameter b_0 . c13-c22 is column regression parameter b_1, c23 is column b_0 bootstrap, c24 is column b_1 bootstrap, c25 is column lowb_1, c26 is column up b_0, c27 is column low b_1, c28 is column up b_1.

First, we used B of 1000 as many as 1000 times iteration, and the B was augmented with the addition of 500 until reaching convergent (constant) regression parameter, with an agreement that resulted regression coefficient parameter is using two decimal places. It was found that in B = 10.500 in 10 times iteration the regression coefficient parameter value of b_0 was convergent/constant.

Table 1 shows that b_0 was 1.90, and b_1 was between 2.02 –2.03 (two decimal places). Mean of b_0 = 1.90 and mean of b_1 = 2.02. Subsequently, the variance of each bootstrap was estimated. The variance of b_0 and b_1 of B = 10.500 can be seen in table 2.

Table 2. Variance values of b_0 and b_1 in B = 10.500

B	Parameter	
	*b*0	*b*1
10.500	0.000002072	0.000002813

Table 2 shows that in B = 10.500 the variance of b_0 is 0.000002072, b_1 = 0.000002813. B = 10.500 with 10 times iteration revealed the least variance (minimum) compared to other B.

DISCUSSION

Regression equation produced using bootstrap method (with B = 10,500 and 10 times iteration) is not far different from simple linear regression equation. The resulted regression equation was Caries = 1.90 + 2.02

Table 1. Parameters of b_0 and b_1 in B = 10.500 in 10 times iteration

B= 10.500	Parameter	i = 1	i =2	i = 3	i = 4	I = 5	i = 6	i = 7	i = 8	i = 9	i = 10
	*b*0	1.8	1.9	1.9	1.9	1.9	1.8	1.8	1.8	1.9	1.9
		982	021	006	022	012	990	995	995	014	020
		5	6	5	6	1	5	5	2	4	8
	*b*1	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
		220	250	223	237	2	238	251	256	215	219
		6	6	8	3	608	5	5	2	5	8

OHI-S, indicating that every increase of one OHI-S unit will increase 2.02 unit of the caries. Linear regression analysis with bootstrap method requires a longer time, because repetition will be done until required convergent (constant) regression coefficient and minimum variance are obtained.

Prior to performing linear regression analysis using bootstrap method, it should be considered first that not all data can be bootstrapped. Bootstrap method is used only in highly necessary conditions, such as insufficient (small) number of samples, unknown data distribution, and in the measurement of parameter estimation accuracy.

From B = 10.500 the convergent (constant) regression parameter values (b_0 1.90, b_1 2.02) were obtained. The estimation of regression parameter (b) was obtained by adding the beta (b_0, b_1) in each resampling, and divided with B value. Thus, it presents as the mean of beta estimation in each resampling process.[7] There was no explanation in the literature that determines the amount of bootstrap that should be used in a study. It is apparent that bootstrap recommended in various literatures today is increasing along with the advanced capability in computerization.

In B = 100,000 in an increase of 500 in each bootstrap would quickly produce more centralized (more convergent) parameter. In this study the bootstrap was started in 1000.[8] A general guidelines, B = 1000 is the most frequently used bootstrap for the first bootstrapping. Iteration was performed 10 times to produce convergent (constant) regression coefficient parameter.[9] Iteration process is performed until obtaining convergent (constant) regression coefficient parameter.[10]

In B = 10.500 with 10 times iteration, the least (minimum) variance were produced, i.e., $b_0 = 0.000002072$ and $b_1 = 0.000002813$. The more convergent the data, the less the variance produced. However, this was not supported by Walpole and Sudjana[11] who found that the best estimator was the one with minimum variance (estimator with the least variance among all other estimators for the same parameter).[12]

OHI-S variable has effect on dental caries (p-value = 0.000). Poor dental hygiene is one cause of dental caries, either milk or permanent teeth, particularly in children who are mostly unable to brush their teeth appropriately. The better the oral hygiene, the lower the severity of the caries. In contrast, the worse the oral hygiene, the higher the severity of the caries. This confirms the assumption that oral hygiene is one of the factors that influence dental caries. The prevalence of dental caries increased in children with poor dental hygiene compared to those with good dental hygiene.[13] There was a strong correlation between poor oral hygiene, the presence of plaque, and the prevalence and severity of periodontal diseases and dental caries.[14]

Regression equation produced by using simple linear regression is not far different from bootstrap method with B = 10,500 and 10 times iteration.

Linear regression analysis with the data resulting from bootstrap should be employed in highly required conditions, such as insufficient (small) number of samples, unknown data distribution, and in the measurement of parameter estimation accuracy.

REFERENCES

1. Longini, Halloran A. Resampling-based test to detect person to person transmission of infectious disease. Journal of The Annals of Applied Statistics 2007; 1(1). Available at: http://www.litbang.depkes.go.id/download/artikel/artikel-ai/article-ai-statistics-2007.pdf. Accessed February 22, 2008.
2. Efron B, Thibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall; 1993. p. 5, 6, 10.
3. Djarwanto. Mengenal beberapa uji statistika dalam penelitian. Yogjakarta: Liberty; 1996. p. 175–6.
4. Neuman W. Lawrence, social research methods qualitative and quantitative approach. 6th ed. Boston: Pearson; 2006. p. 17.
5. Handayani ATW. Erupsi gigi permanen dan kesehatan rongga mulut pada anak sekolah dasar di daerah endemik gondok Kabupaten Jember. Tesis. Surabaya: Fakultas Kesehatan Masyarakat Universitas Airlangga; 2007. p. 48–63.
6. Atmono D. Manajemen data. Surabaya: Institut Sepuluh Nopember; 2005. p. 48.
7. Sahinler, Topuz. Bootstrap and jackknife resampling algorithms for estimation of regression parameters. Journal of Applied Quantitative Methods 2007; 2(2). Available at: http://jaqm.ro/issues/volume-2,issue-2/pdfs/sahinler_topuz.pdf. Accessed February 22, 2008.
8. Lange K. Statistics and computing, numerical analysis for statisticians. New York: Springer; 1999. p. 309–10.
9. Anonim. Bootstrapping (Statistics). Available at: http://en.wikipedia.org/wiki/Bootstrapping_(statistics). Accessed May 15, 2008.
10. Atkinson K. Elementary numerical analysis. New York: John Wiley & Sons; 1985. p. 17–9.
11. Sudjana. Metoda statistika. Bandung: Penerbit Tarsito; 1996. p. 198–9.
12. Ronald WE, Raymond MH. Ilmu peluang dan statistik untuk insinyur dan ilmuwan. Edisi keempat. Bandung: ITB; 1995. p. 363–4.
13. Kuntari S. Hubungan antara kebersihan gigi dan karies gigi pada anak usia 4–6 tahun di Kotamadya Surabaya. Majalah Kedokteran Gigi 1996; 29(1): 13–5.
14. Boedihardjo. Hubungan antara kerusakan jaringan periodontal yang disebabkan oleh plak dengan kebetuhan perawatan periodontal. Disertation. Surabaya. 1996. p. 25–32.