Bootstrapping Residuals to Estimate the Standard Error of Simple Linear Regression Coefficients

Muhammad Hasan Sidiq Kurniawan[1)]

[1)*] Department of Statistics, Universitas Islam Indonesia. hasan.sidiq@uii.ac.id (156111303@uii.ac.id)

ABSTRACT

Regression models are the statistical methods that widely used in many fields. The models allow relatively simple analysis of complicated situations. The aim of the regression models is to analyze the relationship between the predictor and response. In order to do that, we have to estimate the regression coefficient. In case of simple linear regression, the method to estimate the regression coefficient is either least square method or maximum likelihood estimation. Also, the standard error of the regression coefficient is being estimated. In this paper, we apply the bootstrap method to estimate the standard error of the regression coefficient. We compare the result of the bootstrapping method with the least square method. From this study, we know that the standard error estimation value of regression model using the bootstrap method is close to the value if we use the least square method. So we can say that the bootstrap method can be used to estimate the standard error of another regression models coefficient which does not have the closed-form formula.

***Keywords*:** bootstrap, simple linear regression, least square, residuals, standard error.

Introduction

Regression analysis has been used in many fields to analyze the relationship between the predictor and response variables. Also, we can use the model to predict the response variable with the predictor. One of the steps that must be done in regression analysis is estimating the regression coefficients. The goodness of the estimating coefficient can be shown using the standard error. Generally, on regression analysis, we have a closed-form formula to compute the standard error. But in some cases, we do not have one. Therefore we can estimate the standard error using bootstrap method. In this paper, we focus our study in simple linear regression, which has the closed-form formula for the standard error. We choose the model in order to compare the least square and the bootstrap standard error. If the value of bootstrap standard error is close to the actual one, then we can say that the bootstrap method can be applied to another regression models which don't have the closed-form formula of the standard error of its coefficient.

The bootstrap method has been applied in many fields. It can be used to determine the response and predictor variable in regression analysis. It is also used to approach the confidence interval of the estimated statistics. The time series analysis also uses the bootstrap method to

p. ISSN: 1411-1047
e. ISSN: 2503-2364

simulate the stationary process of the data[4]. We already know that many things in this world can't be expressed by mathematical models. Also, in research, we often do not have enough sample size. So the analysis we have conducted is not as we expected. The existence of bootstrap method is like an oasis in the middle of desert. It can overcome both problems and the method is quiet simple. The researcher should not be worried anymore because we can simulate the outcome of the research using bootstrap method.

This paper contains five sections. Section two discusses about the simple linear regression. Section three discusses the bootstrap method to estimate the standard error. Section four contains the results of the study, the case study is included. And finally, section five gives the conclusion of this study.

Simple Linear Regression and its Standard Error

The linear regression analysis is one of the most popular statistical methods. The ability to predict the response variable value based on the predictor value has made the regression analysis widely used in many fields. The simplest regression model is the simple linear regression. It is called the simple linear regression because the analysis involved only one predictor variable (covariate). Let $z_1, z_2, \cdots, z_n$ is n-sized random sample where

$$z_i = (x_i, y_i); \quad i = 1,2, \cdots, n \tag{1}$$

$x_i = (1, x_{1i})$ is the covariate vector and y_i is the response variable. The simple linear regression model is expressed as

$$y_i = x_i\beta + \varepsilon_i, \quad i = 1,2, \cdots, n \tag{2}$$

The regression parameter vector $\beta = (\beta_0, \beta_1)$ is unknown so we have to estimate it. The error ε_i is assumed have a zero expectation and following normal distribution. Therefore the expected value of y_i if the value of x_i is known is

The regression model in (2) can be rewritten as:

$$\mu_i = E(y_i | x_i) = E\left((x_i\beta + \varepsilon_i) | x_i\right)$$

$$= E(x_i\beta | x_i) + E(\varepsilon_i | x_i) = x_i\beta \tag{3}$$

$$\underline{y} = \underline{x\beta} + \underline{\varepsilon} \tag{4}$$

where $\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\underline{x} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix}$, $\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$,

and $\underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

Using the least-square, the estimation for $\underline{\beta}$ is

p. ISSN: 1411-1047
e. ISSN: 2503-2364

$$\underline{\hat{\beta}} = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y} \quad (5)$$

Now we have to estimate the standard error of $\underline{\hat{\beta}}$. Define the matrices G

$$G = x'x \quad (6)$$

Then the variance of $\underline{\hat{\beta}}$ is

$$var(\underline{\hat{\beta}}) = var\left((\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}\right)$$

$$= \left((\underline{x}'\underline{x})^{-1}\underline{x}'\right) var(y) \left((\underline{x}'\underline{x})^{-1}\underline{x}'\right)'$$

$$= \left((\underline{x}'\underline{x})^{-1}\underline{x}'\right) var(y) \left(\underline{x}(\underline{x}'\underline{x})^{-1}\right)$$

$$= \sigma_F^2(\underline{x}'\underline{x})^{-1}$$

$$= \sigma_F^2 G^{-1}$$

where the variance of y is $\sigma_F^2\mathbf{I}$ [3]. Then the standard error of $\underline{\hat{\beta}}$ is

$$se(\hat{\beta}_{j-1}) = \sigma_F\sqrt{G^{jj}}; j = 1,2 \quad (7)$$

G^{jj} is the j-th diagonal element of G^{-1}. Practically, σ_F is estimated by

$$\hat{\sigma}_F = \left\{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{n-1}\right\}^{1/2}$$

Results

Using the least square method, we already had the estimation for $\underline{\beta}$. The estimation for error can be written as

$$\hat{\varepsilon}_i = y_i - x_i\hat{\beta} \quad ; \quad i = 1,2,\cdots,n \quad (8)$$

Further the empirical distribution of the estimating error is

$$\hat{F} = P(\hat{\varepsilon}_i) = \frac{1}{n} \quad ; \quad i = 1,2,\cdots,n \quad (9)$$

Let $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*)$ is the n-sized random sample with replacement of $\hat{\varepsilon}_i$, then the bootstrap's response variable is

$$y_i^* = x_i\hat{\beta} + \varepsilon_i^* \quad ; \quad i = 1,2,\cdots,n \quad (10)$$

If the actual data is written as $z_i = (x_i, y_i)$, then the bootstrap data is written as $z_i^* = (x_i, y_i^*)$. The bootstrap regression model is

$$y_i^* = x_i\beta^* + \varepsilon_i^{**} \quad (11)$$

where $E(\varepsilon_i^{**}) = 0$ dan $var(\varepsilon_i^{**}) = \sigma_F^2$. The expected value of y_i^* is $x_i\hat{\beta}$ and the variance of y_i^* is σ_F^2 [3]. Using the least square method, the estimation for the bootstrap regression parameter is

$$\underline{\hat{\beta}^*} = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}^* \quad (12)$$

We can obtain the variance of $\underline{\hat{\beta}^*}$ using the similar way to obtain the variance of $\underline{\hat{\beta}}$.

$$var(\underline{\hat{\beta}^*}) = var\left((\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}^*\right)$$

$$= \left((\underline{x}'\underline{x})^{-1}\underline{x}'\right) var(y^*) \left((\underline{x}'\underline{x})^{-1}\underline{x}'\right)'$$

$$= \left((\underline{x}'\underline{x})^{-1}\underline{x}'\right) var(y^*) \left(\underline{x}(\underline{x}'\underline{x})^{-1}\right)$$

$$= \sigma_F^2(\underline{x}'\underline{x})^{-1}$$

$$= \sigma_F^2 G^{-1}$$

p. ISSN: 1411-1047
e. ISSN: 2503-2364

Because y^* is from the same population with y, then the variance of y^* is equal to the variance of y, σ_F^2. The standard error of $\underline{\hat{\beta}}^*$ is

$$se\left(\hat{\beta}_{j-1}^*\right) = \sigma_F\sqrt{G^{jj}} \qquad (13)$$

As in the linear regression without bootstrap, σ_F is estimated using the bootstrap data.

$$\hat{\sigma}_F = \left\{\frac{\sum_{i=1}^{n}\left(y_i^* - x_i\hat{\beta}^*\right)^2}{n-1}\right\}^{1/2}$$

Both equations (7) and (13) are equal. Therefore we have shown that the formula to estimate the standard error of regression's coefficient without bootstrap can be applied to estimate the standard error on the bootstrap one. The difference between both of them is the estimated value of σ_F.

In summary, the steps to estimate the standard error by bootstrapping the residuals are as follows:

a) Take an n-sized random sample with replacement from the residuals, $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*)$.

b) Compute

$y_i^* = x_i\hat{\beta} + \varepsilon_i^* \;\; ; \;\; i = 1,2,\cdots,n$

therefore we have a data-set

$z^* = (z_1^*, z_2^*, \cdots, z_n^*)$, where

$z_i^* = (x_i, y_i^*)$.

c) Estimate the regression coefficient, $\underline{\hat{\beta}}^* = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}^*$.

d) Repeat a) to c) steps B-times.

e) Estimate the standard error of $\hat{\beta}$,

$$\widehat{se}_B(\hat{\beta}_j^*) = \left\{\sum_{b=1}^{B}\left[\hat{\beta}_{jb}^* - \hat{\beta}_j^*(.)\right]^2/(B-1)\right\}^{1/2}$$

where $\hat{\beta}_j^*(.) = \frac{\sum_{b=1}^{B}\hat{\beta}_{jb}^*}{B}$; $j = 0,1$.

Case Study

In this section, we apply the bootstrap method to estimate the standard error of the regression model based on work's accident's data on PT Artistika Mandiri in 2013. There are 43 workers are being observed about how many accident they have experienced at work, based on their work hour and division. The divisions are: frame, weaving, quality control, and logistic. The estimated regression model is as follows

$$accident = 2{,}886 + 0{,}001019(hours) - 0{,}4953(d_{weaving}) - 0{,}1998(d_{q.control}) - 1{,}5804(d_{log})$$

The estimated standard error of the regression coefficients of the model are shown in the Table 1.

p. ISSN: 1411-1047
e. ISSN: 2503-2364

Table 1. Standard Error of Regression Parameter

Parameter	Estimation	$\widehat{se}$
$\hat{\beta}_0$	2,886	1,0572
$\hat{\beta}_1$	0,001019	0,0004243
$\hat{\beta}_2$	-0,4953	0,911
$\hat{\beta}_3$	-0,1998	0,8758
$\hat{\beta}_4$	-1,5804	1,2373

Based on the equation (14), the estimation of error, which is the residuals, are [,1]

[1,] 0.57033966

[2,] -5.42966034

[3,] 0.57033966

[4,] -0.42966034

[5,] 1.57033966 ……..

[42,] 0.54770105

[43,] -3.84923368

A bootstrap residuals sample is generated 100 times then compute the $\hat{\beta}^*$ and its standard error. The results are in Table 2.

Table 2. Bootstrap Standard Error of Regression Parameter

Parameter	$\widehat{se}_{100}$
$\hat{\beta}_0$	1,0133
$\hat{\beta}_1$	0,0004225
$\hat{\beta}_2$	0,8124
$\hat{\beta}_3$	0,8461
$\hat{\beta}_4$	1,1483

As shown in Tables 1 and 2, the bootstrap estimated standard errors are close to the least square one. If the bootstrap sample size is increased then we believe that $\widehat{se}_\infty(\hat{\beta}_j) \approx \widehat{se}(\hat{\beta}_j)$.

Conclusion

Based on the case study sections, we have seen that estimating the standard error of regression coefficient using the bootstrap method will have the same result if we use the least square one. It means, if we analyze the data using another type of regression which does not have the closed-form formula for the standard error of its coefficient, then the bootstrap method can be applied. We have described its application in a simple linear regression model. Further investigation is needed for other regression models such as logistic, time series, poisson and many more and use the result from the current model as a 'blue print'.

References

[1]Chatterjee, S., Price, B. 1977. Regression Analysis by Example. John Wiley & Sons, inc. New York

[2]Draper, N., Smith, H. 1992. Analisis Regresi Terapan Edisi Kedua. PT Gramedia Pustaka Utama. Jakarta

[3]Effron, B., Tibshirani, R.J. 1993. An Introduction to the Bootstrap. Chapman & Hall. London

[4]Kreiss, J. P., Paparoditis, E. 2015. Bootstrapping Locally Stationary Processes. Journal of The Royal Statistical Society. B, 77, 267-290