International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET23102116

Parameter Estimation of Multiple Linier Regression by Comparing Bootstrap and Jackknife Methods

Mika Alvionita S
Program Studi Sains Data, Institut Teknologi Sumatera, Lampung, Indonesia
Email : mika.alvionita@sd.itera.ac.id

ARTICLEINFO

Article History:
Accepted: 05 April 2023
Published: 23 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
604-608

ABSTRACT

This research applies parameter estimation of multiple regression analysis using bootstrap and jackknife resampling methods. bootstrap resampling method is a resampling procedure that draws samples repeatedly randomly with returns. While the jackknife resampling method is a resampling procedure by performing calculations that remove one or more observations from the specified sample. Based on the data in this study, the bootstrap resampling method is better than the jackknife resampling method. This can be seen from the small bias and standard error values in the bootstrap resampling method compared to the jackknife resampling method. Similarly, the bootstrap resampling method's confidence interval is narrower than the jackknife resampling method. In addition, the parameter estimation coefficients between the least squares method and the bootstrap resampling method are almost similar. This shows that the bootstrap resampling method is better than the jackknife resampling method in this study.

Keywords: Boostrap, Jackknife and Method Resampling

I. INTRODUCTION

One of the statistical methods used to determine the relationship between one variable and another is multiple regression analysis. Multiple regression analysis has 2 types of variables, namely independent variables and dependent variables. Dependent variables are variables that depend on other factors, while independent variables are variables that do not depend on other variables or responses. In multiple linear analysis, the independent variable consists of more than one which is symbolized by X and the dependent variable only consists of 1 which is symbolized by Y. The purpose in multiple linear regression analysis is to determine the dependent variable.

The goal in multiple linear regression analysis is to estimate parameters. In estimating the parameters, the method used is the Least Squares Method by minimizing the sum of the residual squares. In estimating parameters, there are several assumptions that must be met, namely that the residuals have a normal distribution, the variance of the residuals is homoskedasticity, and multicollinearity. If these assumptions are met, the parameter estimation will fulfill the Best Linear Unbiased Estimator (BLUE) properties. But sometimes, the samples used in estimating parameters have residuals that are not normally distributed.

One method that can be used to overcome the assumption of non-normally distributed resiaduals is the resampling method. Resampling methods that are often used are bootstrap resampling and jackknife. The bootsrap resmpling method is the withdrawal of samples repeatedly by sampling returns [4]. While the jackknife resampling method is the withdrawal of samples repeatedly without sample returns. The purpose of the bootstrap and jackknife resampling methods is to estimate the standard error and confidence interval of the parameters.

Based on that, this research will determine the parameter estimation of multiple linear regression models with the assumption of normally distributed residuals is not met. This model will apply the bootstrap and jackknife resampling methods. Previous research is Estimation of Parameter Regression Model using Bootstrap and Jackknife written by Hedi[3]. In this research, Hedi conducted simulations that resulted in the bootstrap resampling method being better than the jackknife resampling method. This research will use data from the food industry, namely Sanjai Nitta. The variables used in this research are storage costs, labor costs and production quantities. Therefore, parameter estimation will be carried out using multiple regression analysis models with bootstrap and Jackknife resampling methods.

II. METHODS AND MATERIAL

Multiple Regression Analysis

Multiple regression analysis is performed to estimate the parameters of a population. This parameter estimation is done by the Least Squares Method[5]. The model in multiple linear regression analysis can be written as follows.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad (1)$$

Description :

Y	: dependent variable
$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$	: parameters
x_1, x_2, x_n	: independent variable
ε	: error or residual ; $\varepsilon \sim N(0; \sigma^2)$

Least Squares Method

Equation 1 can be written into matrix notation form as follows :

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2)$$

Description :

$\boldsymbol{y}$	: dependent vector of size n×1
X	: n×k matrix of independent variables
$\boldsymbol{\beta}$	: k×1 vector of independent estimators
$\boldsymbol{\varepsilon}$	: error vector of size n×1

From Equation (2), the expected value of the error denoted by uncorrelated E(ε)=0 and the variance of the error denoted by $var(\boldsymbol{\varepsilon}) = \sigma^2 I$ I are obtained. Since E(ε)=0 then E(y)=Xβ, so the square of the error becomes :

$$\varepsilon'\varepsilon = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})$$

By minimizing the sum of squared errors, $\hat{\beta}$ is obtained as follows:

$$\hat{\beta} = (X'X)^{-1}X'y$$

With $\hat{\beta}$ is an estimator that meets the Best Linear Unbiased Estimator (BLUE) properties.

The purpose of regression analysis is to test the hypothesis of the regression coefficient. This test usually uses the t test, in this case what will be tested is β_0 as follows :

$$H_0: \beta_0 = \beta^*$$
$$H_0: \beta_0 \neq \beta^*$$

Explicitly, the test of this hypothesis is based on the following test statistics

$$t_{hit} = \frac{\hat{\beta}_0 - \beta^*}{s_{\hat{\beta}_0}}$$

The decision-making rule in this test, if the real level is α, is to reject Ho if :

$$|t_{hit}| > t_{(1-\frac{\alpha}{2};n-2)}.$$

Classical Assumption Testing

Normality Test

Normality testing aims to test whether the residual variables in the regression model are normally distributed or not. The residual value is normally distributed when it forms a bell-like curve whose two sides widen infinitely[2]. Normality tests are carried out with the skewness test, Shapiro Wilk test and Kolomogorov-Smirnov test. In this research, the test used is the Shapiro Wilk test, if the p-value> 0.05 then the residuals are normally distributed.

Heteroscedasticity Test

Heteroscedasticity testing aims to test whether in the regression model there is diversity in the variance of the residuals while the variance of the residuals is still called homoscedasticity. A good regression model is a model that fulfills homoscedasticity. The heteroscedasticity tests used are the glejser test, scatterplot test, park test, Breusch Pagan test and white test. In this research, the Breusch Pagan test was used, if the p-value >0.05 then the residuals are homoscedasticity.

Multicollinearity Test

Multicolinerity testing aims to see if the residuals have a correlation between the independent variables in the regression model. The multicolinerity assumption is seen from the variance inflation error (VIF) value. If the VIF value is smaller than 10 or the tolerance value is greater than 0.1, the regression model does not have multicollinearity and vice versa.

Bootstrap Method

Bootstrap comes from the phrase "pull oneself up one's bootstrap" which means that little data, data that deviates from certain assumptions and data that does not have any assumptions about the distribution of the population [1]. Bootstrapping requires a large number of repeated calculations to estimate the shape of the statistical sampling distribution. Bootstrap is a resampling procedure that draws samples repeatedly at random with returns.

Jackknife Method

Jackknife comes from the term folding knife that can be carried everywhere. The term jackknife is an approach to testing hypotheses and calculating confidence intervals. Jackknife is used in statistical inference to estimate the bias and standard error of statistics. Jackknife is a resampling procedure by performing calculations that remove one or more observations from the specified sample.

Method

The data used in this research come from secondary data sourced from West Sumatra, namely Sanjai Nitta Bukittinggi from January 2021 to December 2022. The dependent variables used in this research are storage costs in rupiah (x_1), labor costs in rupiah (x_2), and the amount of production in pcs (Y). The analysis used in this research is Multiple Linear Analysis using bootstrap and jackknife resampling methods. The steps in this research are as follows :

1. Collecting data from UMKM Sanjai Nitta Bukittinggi from January 2021 to December 2021.
2. Estimating parameters with multiple regression analysis using R software.
3. Performing classical assumption tests, namely normality test, heteroscedasticity test and multicolinerity test.
4. Estimating regression parameters using the bootstrap resampling method, as follows
   a. Taking samples from the data obtained by returning as much as n data and then repeating it B times.
   b. Estimating parameters β_0, β_1, and β_2 with the least squares method.
5. Estimating regression parameters with the jackknife resampling method, as follows

a. Take one random sample without return. The random sample is then regressed n times, where n is the amount of data by eliminating the i-th pair of data for each regression, so that n-1 is formed in each regression.

b. Estimation of parameters β_0, β_1, and β_2

6. Comparing the results of the bootstrap and jackknife resampling methods with the results of estimating regression parameters using the least squares method.
7. Interpreting the method performed.

Bootstrap Algorithm

1. Input regression data.
2. Calculate parameter estimation coefficients with the Least Squares Method.
3. Perform sampling one by one with a return of n times.
4. Calculate the new parameter estimation coefficient value on the new sample.
5. Repeat Step 3-4 for B times.
6. Determine the bias value, upper and lower bounds of confidence interval and standard error for each parameter according to the significance level α.

Jackknife Algorithm

1. Input regression data.
2. Calculate parameter estimation coefficients with the Least Squares Method.
3. Remove the i-th data pair to obtain the i-th new sample.
4. Calculate the new parameter estimation coefficient value on the new sample.
5. Repeat Steps 3 and 4 up to n times.
6. Determine the bias value, upper and lower bounds of confidence interval and standard error for each parameter according to the significance level α.

III.RESULTS AND DISCUSSION

This research conducted multiple regression analysis to estimate parameters. This analysis also wants to see if there is a relationship between storage costs and labor costs on the amount of production. The results of parameter estimation using the least squares method are as follows

$$y = 3.407e^{+03} - 2.323\, e^{-03} x_1 + 6.414e^{-05} x_2 + \varepsilon$$

Descriprion :

y : Production quantity (pcs)

x_1 : storage cost in rupiah

x_2 : labor cost in rupiah

After parameter estimation, a classical assumption test is carried out on the residuals. The first classical assumption test carried out is the normality test using the Kolomogorov-Smirnov test. Based on the Kolomogorov-Smirnov test, the p-value is 0.0002076 which is smaller than the 0.05 significance level. So it is concluded that the data is not normally distributed.

Next, the bootstrap and jackknife resampling methods were applied to the data using R software. The parameter estimation results obtained are shown in Table 1.

Table 1

Parameter estimation results using MKT, bootstrap resampling and jackknife

Parameter	Metode Kuadrat Terkecil		Resampling bootstrap		Resampling jackknife	
	Parameter estimation	Standard Error	Parameter estimation	Standard Error	Parameter estimation	Standard Error
β_0	$3.407e^{+03}$	$1.837e^{+02}$	$3.4066e^{+03}$	$2.8523e^{+02}$	$3.392e^{+03}$	198.5
β_1	$-2.323\, e^{-03}$	$2.997e^{-03}$	$-2.323\, e^{-03}$	$4.3442e^{-02}$	$-9.660e^{-04}$	114.6
β_2	$6.414e^{-05}$	$1.206e^{-05}$	$6.414e^{-05}$	$3.2316e^{-05}$	$6.339e^{-5}$	114.6

Based on the parameter estimation results in Table 1, the parameter estimation coefficient values are almost similar between the least squares method, bootstrap resampling method and jackknife. This proves that the value of the parameter estimation coefficient is BLUE. Of the three methods used in estimating parameters, it is found that the least squares method is better than the bootstrap and jackknife resampling methods as seen from the smallest standard error. While the bootstrap resampling method is better than the jackknife resampling method based on its standard error.

Table 2 Bias, confidence interval and parameter estimation with bootstrap and jackknife resampling

Method	Parameter	Estimation	Bias	Lower Limit	Upper Limit
Bootstrap	β_0	$3.4066e^{+03}$	$-6.5193e^{+01}$	$2.9128e^{+03}$	$4.0308e^{+03}$
	β_1	$-2.323\ e^{-03}$	$8.8659e^{-03}$	$-9.6333e^{-02}$	$7{,}3954e^{-02}$
	β_2	$6.414e^{-05}$	$-2.8004e^{-06}$	$3.6047e^{-06}$	$1.3028e^{-04}$
Jackknife	β_0	$3.392e^{+03}$	$1.420e^{+01}$	3003.3	3781.5
	β_1	$-9.660e^{-04}$	$-1.357e^{-03}$	-224.6	224.6
	β_2	$6.339e^{-5}$	$7.516e^{-07}$	-224.6	224.6

Based on Table 2, the bias value using the bootstrap resampling method of β_0 and β_2 is smaller than the jackknife resampling. As for β_1, the bias value of the jackknife resampling method is smaller than that of the bootstrap resampling method. In the confidence interval, it can be seen that the upper limit and lower limit values in the bootstrap resampling method are narrower than the jackknife resampling method. Thus, in research using data from Sanjai Nitta, it is concluded that the bootstrap resampling method is better at estimating multiple linear regression parameters than the jackknife resampling method

IV.CONCLUSION

In this research, the application of bootstrap and jackknife resampling methods in estimating parameter coefficients is seen based on standard error, bias, and confidence interval. The results obtained by the bootstrap resampling method are better than the jackknife resampling method. This can be seen from the small bias and standard error values in the bootstrap resampling method compared to the jackknife resampling method.

Similarly, the bootstrap resampling method's confidence interval is narrower than the jackknife resampling method. In addition, the parameter estimation coefficients between the least squares method and the bootstrap resampling method are almost similar. This shows that the bootstrap resampling method is better than the jackknife resampling method in this research.

V. REFERENCES

[1] Efron B. dan Tibshirani RJ.1993. An Itroduction to Bootstrap. London (UK);Chapman &Hall.

[2] Ghozali I. 2016. Aplikasi Analisis Multivariat. Semarang(ID):Badan Penerbit Undip.

[3] Hedi. 2012. Estimation of Parameter Regression Model Using Bootstrap and Jackknife. Vol.4 No.2. Sigma-Mu

[4] Rodliyah I.2016. Perbandingan Metode Bootstrap dan Jackknife dalam mengestimasi Parameter Regresi Liner Berganda. Jurnal Matematika dan Pendidikan Matematika.1(1):76-86

[5] Walpole,R,E.,Myers, R.H., Myers, S.L., & Ye,K. 2011 Probability Statistics for Engineers &Scientists 9th Ed.USA :Pearson.

Cite this article as :

Mika Alvionita S, "Parameter Estimation of Multiple Linier Regression by Comparing Bootstrap and Jackknife Methods ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 2, pp. 604-608, March-April 2023. Available at doi : https://doi.org/10.32628/IJSRSET23102116

Journal URL : https://ijsrset.com/IJSRSET23102116