CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 7(3) (2022), Pages 345-353
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparisons between Resampling Techniques in Linear Regression: A Simulation Study

Anwar Fitrianto[1,*], Punitha Linganathan[2]

[1,*]Department of Statistics, IPB University, Indonesia
[2]Department of Mathematics, Universiti Putra Malaysia, Malaysia

Email: anwarstat@gmail.com

ABSTRACT

Parameter estimations in linear regression need to fulfill some assumptions. Once the assumptions are not fulfilled, the conclusion is questionable. Bootstraps and Jackknife are resampling techniques that do not require assumptions in estimating the $\hat{\beta}$. The study aims to compare resampling techniques in linear regression. The data used in the study is clean, without any influential observations, outliers, or leverage points. The ordinary least square method was used as the primary method to estimate the parameters and then compared with resampling techniques. The variance, p-value, bias, and standard error are used as a scale to estimate the best method among random bootstrap, residual bootstrap and delete-one Jackknife. After all the analysis, it was found that random bootstrap did not perform well while residual and delete-one Jackknife works quite well. Random bootstrap, residual bootstrap, and Jackknife estimate better than ordinary least square. The study also found that residual bootstrap works well in estimating the parameter in the small sample. At the same time, it is suggested to use Jackknife when the sample size is big because Jackknife is more accessible to apply than residual bootstrap and Jackknife works well when the sample size is large.

Keywords: jackknife; linear; regression; resampling

INTRODUCTION

Regression analysis is a statistical analysis that constructs relationships between dependent or response variables y and independent or regressor variables $(x_1, x_2, \dots, x_k)$. Ordinary least square (OLS) is a traditional way of finding parameter estimates, $\hat{\beta}$ but it relies strongly on assumptions [1]. The reliability and validity of the conclusion in regression analysis are essential ([2], [3]), and they depend on how far the data follows the assumption and on the sample size of the data. It is easier to find the estimated regression coefficient, $\hat{\beta}$ without any assumption or distribution. Bootstrap and Jackknife are resampling techniques that do not need any assumptions in estimating the $\hat{\beta}$,([4]–[6]. Sahinler and Topuz [7] compared the bootstrap and Jackknife methods. Their research discussed strategies for building a regression model using the Jackknife and bootstrap method. The four methods used in their research are bootstrap based on the resampling observations, bootstrap based on the resampling errors, delete-one Jackknife regression and delete-d Jackknife regression. These methods were used to find the parameter estimates, bias, standard errors, and confidence intervals. Their research concluded that

Submitted: December 23, 2021 Reviewed: May 25, 2022 Accepted: July 24, 2022
DOI: http://dx.doi.org/10.18860/ca.v7i3.14550

large bootstrap replicates ensure that the parameter is close to the true parameter. They also suggested that bootstrap replicate is sufficient for estimating the variance and $B = 1000$ for estimating the standard errors. Their research tests the accuracy of bootstrap and Jackknife methods in estimating the distribution of regression parameters with various sample sizes and various bootstrap replicates. Sahinler and Topuz [7] and Li et. al. [8] found that the bootstrap method is appropriate for linear regression and it is usable even when the error is not normally distributed. Algamal and Rasheed [9] further develop resampling in linear regression.

The advantage of bootstrap approximations is that, in general, it needs a smaller sample than the ordinary least square for estimating the parameter. Meanwhile, the disadvantages of bootstrap methods were discussed in Ma et al., [10], Wan et al., [11], [12], and Phaladiganon et al., [13] A few of the disadvantages of the methods are as follows:

a) Bootstrap distribution of is not a good approximation of F, if the sample size is small and with the existence of an outlier,
b) Bootstrap is not suggested to use in dependence structure case like time series, and
c) It is not preferable to use residual bootstrap when the assumptions are violated.

Algamal and Rasheed, [10] concluded that Jackknife method perform quite well when the sample size is large enough ($n \geq 50$). Meanwhile, recent studies by Shao, J., & Tu, D., [14] and Beyaztas, U., & Alin, A., [15] discussed bootstrap and Jackknife in linear regression.

Based on that, the study is aimed to compare parameter estimates of multiple linear regression based on several resampling methods. There are several methods to estimate the $\hat{\beta}$ in bootstrap and Jackknife. The scope of this research is to investigate the bootstrap and Jackknife method with different scenariosThis research considered random bootstrap, residual bootstrap, and Jackknife delete-one observation. The study is limited to multiple linear regression model. First the sample size will be selected with different size and estimate the parameter. The bias and variance will be observed then the relationship between the bias and variance will be investigated. The distribution also will be observed by varying with the increase in the sample size. The value of bootstrap resampling with different bootstrap replicates and sample size gives less bias than ordinary least square. The Jackknife coefficient is calculated by using,

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=n}^{n} \hat{\beta}_{ji} \tag{1}$$

where n is the sample size and $\hat{\beta}_{ji}$ parameter estimate for each sample formed after deleting one of the observations. While the bootstrap coefficient is calculated from

$$\hat{\beta}_b = \frac{1}{B}\sum_{r=1}^{B} \hat{\beta}_{br} \tag{2}$$

$$\hat{\beta}_{br} = \hat{\beta}_{ols} + \left(x'x\right)^{-1} x'e_{br} \tag{3}$$

where $r = 1,2,\dots,B$ is bootstrap replicate, e_{br} is error of the regression,x is the independent variable and $\hat{\beta}_{ols}$ is the parameter estimate from ordinary least square method.

METHODS

Data

The data used in this study is pressure-dropping data, which is available in Montgomery et al., [16]. It has one dependent variable y, and four independent variables, that is x_1, x_2, x_3 and x_4. There are 62 observations in the data. The data was collected from research where the pressure drop was measured for two-phase flow through screen-plate bubble columns. The research was conducted to test the reason of the pressure drop through the bubble cap. A bubble column is used to observe the reaction between the gas and liquid.

The first factor considered in that research is the superficial fluid velocity of the gas. The gas's speed and direction of motion are measured by flow in the column. The second factor is the kinematic viscosity. The friction caused by the thickness of gas when the gas moves through the liquid particles was calculated. Then the distance across the space between two parallel threads was considered. The last factor used in research is the dimensionless number, which is not associated with the physical dimension. It is calculated to relate the gas's superficial fluid velocity and the liquid's superficial fluid velocity. For building the model, the dependent variable y denotes the dimensionless factor for the pressure drop through a bubble cap. The independent variables are x_1 (superficial fluid velocity of the gas (cm/s), x_2 (kinematic viscosity), x_3 (mesh opening, cm), and x_4 (dimensionless number relating the gas's superficial fluid velocity to the liquid's superficial fluid velocity).

Simulation Study Scenarios

The original data will be analyzed using ordinary least square regression data. Then assumptions checkings will be conducted using the residuals of the model. Then, using the sampe original data, resampling techniques using the residuals and random bootstrap resampling will be conducted with four different sample sizes, which are 20, 40,50 and 62. Each sample will be used in three different bootstrap replicates, namely 100, 1000 and 10000.

For the delete-one Jackknife bootstrap, the resampling will be conducted at different sample sizes, namely 20, 40, 50 and 62. The bias, variance, standard error and p-value will be calculated for each method. The best method among this three methods will be chosen according to the value of bias, variance, standard error and p-value.

RESULTS AND DISCUSSION

In this study, full model was used for the reference, which means all independent variables were included in the model regardless the significance of the variables. The fitted full regression model which was obtained based on ordinary least square using SAS software is written as follows:

$$\hat{y} = 5.88839 - 0.48460x_1 + 0.18263\ x_2 + 35.39109x_3 + 5.92695x_4$$

Random Bootstrap Approach

Random bootstrap technique was first used to analyze the data. The resampling was conducted at different sample size 20, 40, 50 and 62. The bootstrap replication were applied in every sample size, namely 100, 1000 and 1000.

Table 1. Summary Statistics for Multiple Linear Regression Using Random Bootstrap at Different Bootstrap Replicates and Sample Sizes for β_0 and β_3

Parameter Estimate	Bootstrap Replicate	Sample Size	Bias	Variance	*p*-value	Standard Error
$\hat{\beta}_0$	100	20	-2.2181	85.6307	0.0001	0.9254
		40	3.3626	22.8120	<.0001	0.4776
		50	1.5437	15.1000	<.0001	0.0469
		62	-0.6707	19.9445	<.0001	0.4466
	1000	20	-1.1044	83.9495	<.0001	0.2897
		40	2.9549	34.4348	<.0001	0.0012
		50	1.4503	18.4197	<.0001	0.1357
		62	-0.8994	19.1731	<.0001	0.1385
	10000	20	-1.2754	203.4686	<.0001	0.0108
		40	2.6252	41.8398	<.0001	0.0647
		50	1.3527	18.8707	<.0001	0.0434
		62	-0.9042	4.9842	<.0001	0.0461
$\hat{\beta}_3$	100	20	2.6410	574.9345	<.0001	2.3978
		40	-5.7656	111.8724	<.0001	1.0577
		50	-5.3883	61.2876	<.0001	0.7829
		62	1.0310	125.2369	<.0001	1.1191
	1000	20	2.4814	629.8633	<.0001	0.7936
		40	-5.4017	211.8649	<.0001	0.4603
		50	-4.5249	73.4070	<.0001	0.2709
		62	1.6356	116.7295	<.0001	0.3417
	10000	20	3.1247	634.0890	<.0001	0.2518
		40	-4.5548	261.6325	<.0001	0.1618
		50	-4.2045	87.0297	<.0001	0.0933
		62	1.8947	37.2858	<.0001	0.1146

Table 1 shows the changes in $\hat{\beta}_3$ and $\hat{\beta}_0$ at different sample sizes and bootstrap replicates. For each parameter estimate, as the sample size changes, the bias changes. More specifically, the bias is getting smaller as the sample size increases.

The variance of $\hat{\beta}_3$ decreases from 574.9345 when the sample is 20 to 61.2876 when the sample size is 50. But, the bias of $\hat{\beta}_3$ increases when the sample is 62 . It can be observed that as the sample size increased from 20 to 62, the variance of parameter estimates decreased. Meanwhile, the bias decreases as the bootstrap replicate increases. For B was set to 100, the intercept shows bias as 1.5437. This value decreases to 1.4503 when the number of bootstrap replicates, B, increases to 1000. When the number of bootstrap replicates was increased to 10000, the bias decreases again to 1.3527. From the results, it can be observed that the bias decreases as the replicate increases. When the bootstrap replicate, B increases from 100 to 1000, the variance decreases from 125.2369 to 116.7295. It decreases further to 37.2858 when B is equal to 10000, which shows 70.23% difference when we compare to 125.2369.

Residual Bootstrap Approach

The second resampling technique that has been used to analyze the data was residual bootstrap. This section displays some results such as parameter estimates, bias, and variances of the parameter estimates using residual bootstrap. The results of $\hat{\beta}_0$ and $\hat{\beta}_1$ are shown in Table 2. In residual bootstrap, the results were more apparent than in random bootstrap. It shows a clear trend of parameter estimates, bias, and variance at different sample sizes and the number of bootstrap replicates. The bias decrease as the sample size increases. When $n = 20$, the bias is 0.2307. Then when the sample increased to 40 the bias became 0.2266 and bias is 0.0684 when the sample size is 50 and at last, when n is 62 the bias became 0.01368. In general, there is a noticeable difference in bias when the sample size increases.

Table 2. Summary Statistics for Multiple Linear Regression Using Residual Bootstrap at Different Bootstrap Replicates and Sample Sizes for β_0 and β_1

Parameter Estimate	Bootstrap Replicate	Sample Size	Bias	Variance	*p*-value	Standard Error
$\hat{\beta}_0$	100	20	1.5277	30.9685	<.0001	0.5565
		40	2.6535	27.0046	<.0001	0.5197
		50	1.5324	19.8861	<.0001	0.4459
		62	-0.3345	15.4073	<.0001	0.3925
	1000	20	0.9635	28.6300	<.0001	0.1692
		40	2.3838	22.7581	<.0001	0.1509
		50	2.0622	20.0467	<.0001	0.1416
		62	0.0035	15.4785	<.0001	0.1244
	10000	20	0.6704	30.9949	<.0001	0.0557
		40	2.2883	24.0894	<.0001	0.0491
		50	2.2491	20.2725	<.0001	0.0450
		62	-0.0193	17.0400	<.0001	0.0413
$\hat{\beta}_1$	100	20	0.2307	0.2037	<.0001	0.0451
		40	0.2266	0.1566	<.0001	0.0396
		50	0.0684	0.1098	<.0001	0.0331
		62	0.0137	0.0819	<.0001	0.0286
	1000	20	0.1630	0.2196	<.0001	0.0148
		40	0.2061	0.1612	<.0001	0.0127
		50	0.0732	0.1322	<.0001	0.0115
		62	-0.0066	0.1025	<.0001	0.0101
	10000	20	0.1547	0.2103	<.0001	0.0046
		40	0.2180	0.1579	<.0001	0.0040
		50	0.0608	0.1338	<.0001	0.0037
		62	-0.0024	0.1071	<.0001	0.0033

The resampling techniques in Table 2 show a clear decrease of the variances when the sample size increases. Let's consider the changes in the variance of $\hat{\beta}_0$ when the bootstrap replicate is 1000. When the sample size is 20 the variance is 28.6300, and the value

becomes 22.7581 when the sample size is 40. Then the variance decrease as the sample size increases to 50 and 62 where the bias become 19.8861 and 15.4785, respectively.

Now let's observe the changes in bias caused by the bootstrap replicate, B, when it is increased from hundred to thousand then ten thousand. For the estimated constant, $\hat{\beta}_0$, when the sample size is 40 the bias changes from 2.6535 to 2.3838, then 2.2883 when B increases from 100 to 1000 then 10000, respectively. The variance also decreases when the bootstrap replicate increases.

Delete-one Jackknife Approach

The third technique that was used in this research is Jackknife delete-one. The method was applied with different sample sizes , which are 20, 40, 50 and 62. Table 3 and Figure 1 display the changes in bias of all parameters for delete-one Jackknife. The bias decreases as the sample size increases. But when sample size equal to the population size the bias shows an increasing state. Using the population as sample size might show this type of result. Plot of variance versus sample size for all parameters are shown in Figure 2. From the plot, it can be seen that the variance also shows a decreased state from sample 20 to sample 62. Small variances give a better estimation in linear regression. The bias and variance also not interrelated in delete-one Jackknife. The *p*-value also shows that all parameter estimates are significant. The standard error also clearly shows that the increase in sample size will give a better estimation.

Table 3. Summary Statistics for Multiple Linear Regression Using Delete-one Jackknife at Different Sample Size .

Parameter Estimate	Sample Size	Bias	Variance	*p*-value	Standard Error
$\hat{\beta}_0$	20	0.5586	2.9683	<.0001	0.3852
	40	2.2937	0.7335	<.0001	0.1354
	50	2.1648	0.3625	<.0001	0.0851
	62	-3.1721	0.2212	<.0001	0.0597
$\hat{\beta}_1$	20	0.1617	0.0161	<.0001	0.0284
	40	0.2182	0.0054	<.0001	0.0117
	50	0.0662	0.0046	<.0001	0.0096
	62	0.6613	0.0029	<.0001	0.0069
$\hat{\beta}_2$	20	0.0249	0.0001	<.0001	0.0017
	40	0.0006	0.0000	<.0001	0.0007
	50	-0.0045	0.0000	<.0001	0.0005
	62	0.0054	0.0000	<.0001	0.0004
$\hat{\beta}_3$	20	-2.0491	18.1473	<.0001	0.9526
	40	-7.3628	3.7708	<.0001	0.3070
	50	-5.6852	1.4059	<.0001	0.1677
	62	-3.7014	0.9218	<.0001	0.1219
$\hat{\beta}_4$	20	0.3589	1.9284	<.0001	0.3105
	40	0.6624	0.4470	<.0001	0.1057
	50	-0.0712	0.3889	<.0001	0.0882
	62	0.7431	0.4339	<.0001	0.0837

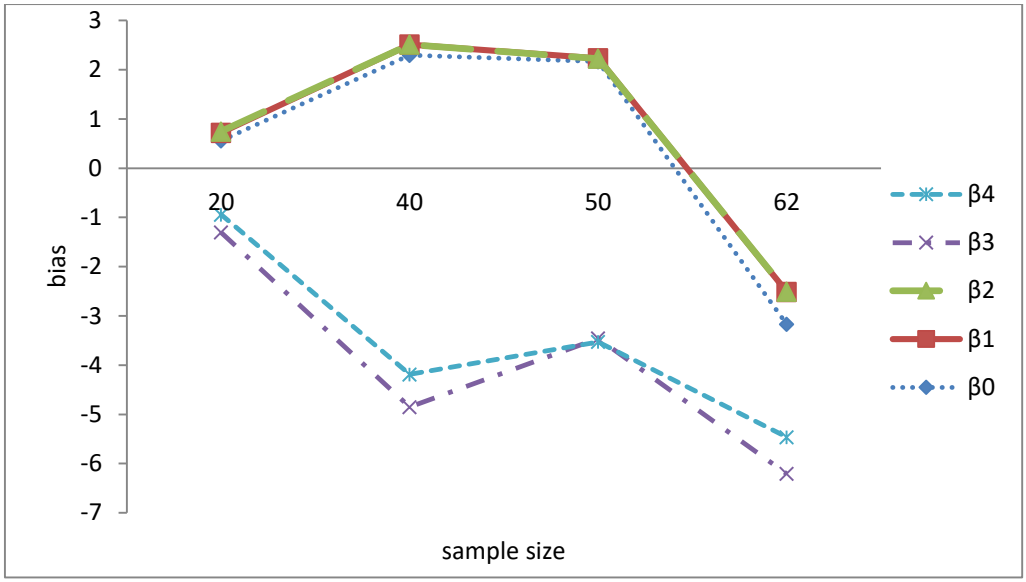

Figure 1. Changes of Bias in All Parameter Estimation when Sample Size Increases in Delete-one Jackknife.

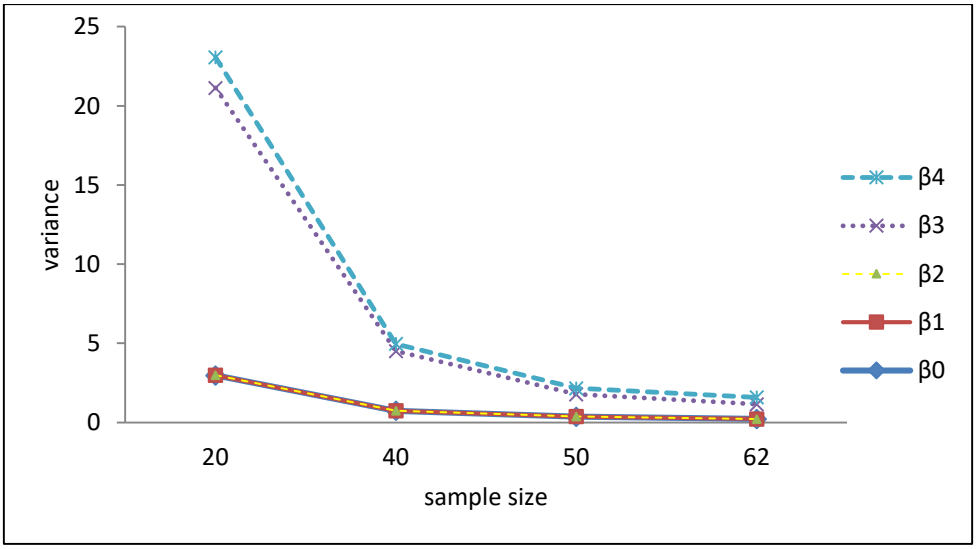

Figure 2. Changes of Variance in All Parameter Estimation when Sample Size Increases in Delete-one Jackknife.

The difference between residual bootstrap estimation and random bootstrap estimation is obvious when the sample size is 20 (small). The residual bootstrap provided better parameter estimation than random bootstrap in bias and variance. This shows that residual has a big influence in linear regression. But, as the sample size increases, both residual and random bootstrap methods show similar results. The increase in bootstraps replicates and sample size gave better parameter estimation in both methods. Jackknife delete-one gave a small variance, but the value of the bias was big when the sample size was small. The bias and variance decrease as the sample size increases.

CONCLUSIONS

Residual bootstrap, random bootstrap, and delete-one Jackknife were compared. Jackknife is not advisable to use when the sample size is small. However, when the sample

size is big enough which is near to population size, it will give better parameter estimation than random bootstrap and residual bootstrap. In a situation where the sample size is small due to cost consideration, it is better to use residual bootstrap than other methods in linear regression. In conclusion, it is advisable to use residual bootstrap when the sample is small. The bigger bootstrap replicates will give better parameter estimation. The Jackknife can be used when the sample size is big enough. This method will be useful when the sample size is too big which may take time to process in both random and residual bootstrap.

In the future, this research can be extended to observe how these methods react when there is an outlier, influential point or leverage point. Moreover, the comparisons may involve other resampling techniques to compare which method works well in multiple linear regression.

REFERENCES

[1] M. Alrasheedi, "Parametric and non-parametric bootstrap: A simulation study for a linear regression with residuals from a mixture of Laplace distributions," *European Scientific Journal*, vol. 9, no. 12, 2013.

[2] R. F. Gunst and R. L. Mason, *Regression analysis and its application: a data-oriented approach*. CRC Press, 2018.

[3] A. Althubaiti, "Information bias in health research: definition, pitfalls, and adjustment methods," *J Multidiscip Healthc*, vol. 9, p. 211, 2016.

[4] M. R. Chernick, "Resampling methods," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 255–262, 2012.

[5] R. E. McRoberts, S. Magnussen, E. O. Tomppo, and G. Chirici, "Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data," *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3165–3174, 2011.

[6] R. G. Clark and S. Allingham, "Robust resampling confidence intervals for empirical variograms," *Mathematical Geosciences*, vol. 43, no. 2, pp. 243–259, 2011.

[7] S. Sahinler and D. Topuz, "Bootstrap and jackknife resampling algorithms for estimation of regression parameters," *Journal of Applied Quantitative Methods*, vol. 2, no. 2, pp. 188–199, 2007.

[8] X. Li, W. Wong, E. L. Lamoureux, and T. Y. Wong, "Are linear regression techniques appropriate for analysis when the dependent (outcome) variable is not normally distributed?," *Invest Ophthalmol Vis Sci*, vol. 53, no. 6, pp. 3082–3083, 2012.

[9] Z. Y Algamal and K. B Rasheed, "Re-sampling in linear regression model using jackknife and bootstrap," *IRAQI JOURNAL OF STATISTICAL SCIENCES*, vol. 10, no. 18, pp. 59–73, 2010.

[10] J. Ma *et al.*, "Probabilistic forecasting of landslide displacement accounting for epistemic uncertainty: a case study in the Three Gorges Reservoir area, China," *Landslides*, vol. 15, no. 6, pp. 1145–1153, 2018.

[11] C. Wan, Z. Xu, Y. Wang, Z. Y. Dong, and K. P. Wong, "A hybrid approach for probabilistic forecasting of electricity price," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 463–470, 2013.

[12] G. A. Nelson, "Cluster sampling: a pervasive, yet little recognized survey design in fisheries research," *Trans Am Fish Soc*, vol. 143, no. 4, pp. 926–938, 2014.

[13] P. Phaladiganon, S. B. Kim, V. C. P. Chen, J.-G. Baek, and S.-K. Park, "Bootstrap-based T 2 multivariate control charts," *Communications in Statistics—Simulation and Computation*®, vol. 40, no. 5, pp. 645–662, 2011.
[14] J. Shao and D. Tu, *The jackknife and bootstrap*. Springer Science & Business Media, 2012.
[15] U. Beyaztas and A. Alin, "Sufficient jackknife-after-bootstrap method for detection of influential observations in linear regression models," *Statistical Papers*, vol. 55, no. 4, pp. 1001–1018, 2014.
[16] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.