The International Journal of

Materials and Engineering Technology

ISSN :
TIJMET, Vol. 01 (01), 2018, (5 pages) ID: 180101001
Type of the Paper: Article
Received: 02.05.2018; Accepted:26.07.2018

PARAMETER ESTIMATION FOR LINEAR REGRESSION USING BOOTSTRAP METHOD

RECEP BINDAK

Technical Sciences School, Gaziantep University, Gaziantep, TURKEY.
E-mail: rbindak@gantep.edu.tr

Abstract

The bootstrap method firstly was introduced by Efron in 1979 as a general method for assessing the statistical accuracy of an estimator. Bootstrap is a computer-based resampling approach and a nonparametric statistical inference method. In this study, the use of the Bootstrap method in the parameter estimation of the linear regression is introduced and given a sample application on a real data set. In addition, if the data set contains outliers the effect that occurs in parameter estimation is examined. Confidence intervals and standard errors have been identified for various bootstrap repetitions numbers. As a result, it has been found that even 200 bootstrap repetations may suffice to obtain proper results.

Keyword: Bootstrap method, Resampling, Linear regression, Parameter estimation

1.Introduction

Linear regression is one of the most popular statistical techniques used in many fields, including engineering and technological research. The model for linear regression is given as

$$Y = X\beta + \epsilon \qquad (1)$$

where **Y** is an nx1 vector of outcomes variable, **X** is an nxk matrix of independent variables, $\boldsymbol{\beta}$ is a kx1 vector of unknown parameter and $\in$ is an nx1 random error term. Error term of $\in$ is independent and identically distributed. It is well known that if the error terms are normally distributed in the regression model (Eq.1), the ordinary least squares (OLS) is the best choice in estimating parameters [2]. According to OLS method, unknown regression coefficients can be estimated by Eq. (2).

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (2)$$

A typical problem in the applied statistic is related to the estimation of an unknown θ parameter. There are two important questions to answer about this: 1) Which $\theta\hat{}$ estimation should be used? 2) Let's say we choose a certain $\theta\hat{}$, how do we decideif this is the right estimator? Bootstrap is a general methodology where to answer the second question [3].

The scientific and statistical inference is clearly built on parametric models and often gives good results. However, the limited scope of parametric models in modern science and the increasing complexity of the studied system pose a risk of incorrect identification of the model. For this reason, alternation based techniques (such as bootstrap recalling) are needed [4].

It is more convenient to use the Bootstrap method because there is no assumption in classical methods when the assumptions required are not fulfilled. It was determined that the bootstrap method is superior to the variance analysis in cases where the sample volume is too small or the sample distribution is not normal. In addition, the results obtained by using the conventional statistical methods and the results of the bootstrap method are similar [5].

A technical term for a sample summary number is sample statistic. For example, the sample means, a summary statistic will variate from sample to another sample and a statistician would like to know the magnitude of this variation around the corresponding population parameter. This is then used in assessing 'tolerance' or 'margin of errors'. The entire picture of all possible values of a sample statistics presented in the form of a probability distribution is called a sampling distribution. A general heuristic method can be applied to any kind of sample statistics. Sometimes, however, the researcher may want to implement his own method to avoid technical complications. Bootstrap is such a method, For example, sample selection from a population is repeated and the sampling distribution can be generated by the statistics calculated from these samples. But there may not always be a real population at all. The logic behind bootstrapping is to extract this information from the "proxy population" at hand[6].

The use of the bootstrap method in regression models has become widespread in recent years. For example, Bootstrap methods for dependent data (such as time series) are discussed in [7]. The bootstrap application is discussed in the functional linear regression and used to create pointwise confidence intervals by Gonzàlez-Manteiga & Martinez-Calvo [8]. In the nonparametric regression, the bootstrap approach was used in confidence intervals [9]. Takma and Atıl [5] found that in their studies, hypothesis testing with classical methods, confidence intervals and the results of applying them in regression analysis were similar to those of the bootstrap method.

A two-step bootstrap model averaging approach has been used to characterize the choice of explanatory variables in a linear regression [10]. The study of Okutan [11] showed that the use of the bootstrap method can give more realistic results than estimating the variance of the OLS estimator for the linear regression model. Amiri and Zwanzig [12] proposed a family of tests based on the bootstrap method about some coefficients of variation and examine its properties using Monte Carlo simulations. In another study, Amiri and von Rosen [13], attempted to depict the applicability of the bootstrap method than the conventional methods of the contingency table. In their study, they shows that nonparametric bootstrap method can be used to improve the result. The study of Baydılı & Sığırlı [14] found that among the most widely used test for homogeneity of variance, was the bootstrap Levene median test with the best performance. The performances of bootstrap-t and percentile bootstrap methods are compared in terms of type 1 error rates by using a different number of bootstrap replications, trimming proportions and population distributions [15]. Amiri and Modarres [16], explore the use of bootstrap for testing independence of two categorical variables. In some study, the Jacknife-after-Bootstrap were used as as a diagnostic tool in linear regression models [17,18] and in logistic regression model [19].

The present study explores the theoretical reasoning behind the bootstrap for linear regression analysis. We consider the nonparametric approaches. In this study, estimation of unknown beta coefficients in Linear Regression by the Bootstrap method is discussed. Bootstrap is a computer-based re-sampling approach. Bootstrap is a nonparametric statistical inference approach. More traditional distribution is used instead of assumptions and asymptotic results. Because the distribution assumptions are not required, the bootstrap allows for more accurate inference if the data distribution is unknown or the sample size is small.

The article is organized as follows. The bootstrap method, backgroud and its implementation are treated in Section 2. Section 3 includes the method and information about the real data which are used. In section 4 gives results for bootstrap resampling method to estimates parameters of linear regression. Finally, the last section summarizes the conclusions of the study.

2.The Bootstrap Method Background

The bootstrap method emerged with Efron 's 1979 horizon opening article.The Bootstrap method requires a few assumptions, It does not contain complex mathematical formulas but has a strong mathematical background and was popular in the 80's with computer usage [20].

Bootstrap uses the sample data to estimate the relevant properties of the population. It empirically constructs the sampling distribution of a statistic by resampling. The resampling procedure is in parallel with sample withdrawal from the population. A bootstrap sample is a sample withdrawal (selected by n times replacement) from the original sample.

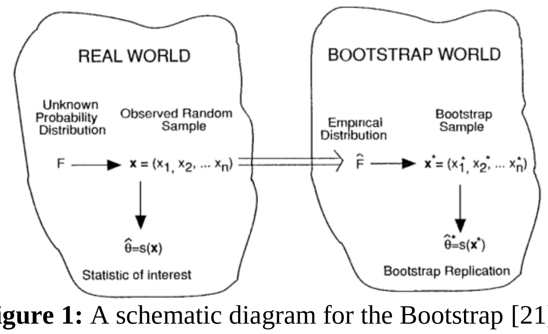

Figure 1: A schematic diagram for the Bootstrap [21]
$x = (x_1, x_2, \ldots, x_n)$ original sample

$x^* = (x_1^*, x_2^*, \dots, x_n^*)$ bootstrap sample
For example, if n=7 then we can get two bootstrap samples, such as
$x^{*1} = (x_5, x_7, x_5, x_6, x_4, x_1, x_3)$ and
$x^{*2} = (x_2, x_1, x_6, x_7, x_7, x_4, x_5)$ [21]
There are two main types of bootstrap: Parametric and Nonparametric. In parametric bootstrap: We know that F belongs to a family of parametric distributions, and we only want to estimate this parameter from the sample. We are sampling F samples to estimate the parameter. In nonparametric bootstrap: We do not know the shape of F, and we estimate it from the empirical distribution we obtain from the data (ie from $\hat{F}$). The general Bootstrap algorithm is as follows:
1.producting x* bootstrap sample (size n) from $\hat{F}_n$
2.calculating $\hat{\theta}^*$ statistic fort this sample
3. Repeating B times step 1 and 2
As a result the $\hat{\theta}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*, \dots, \hat{\theta}_B^*)$ obtained. Thus, the desired quantity is calculated.
Estimate of bootstrap standart error (SE)given by equation (3);

$$SE(B) = \sqrt{\frac{\sum(\hat{\theta}_i^* - \theta^*(.))^2}{B}} \qquad (3)$$

Where $\hat{\theta}_i^*$ (i=1,2,…, B), is the bootstrap value of the i^{th} sample and $\theta^*(.) = \frac{1}{B}\sum \hat{\theta}_i^*$.

3.Materials and Methods

Regression coefficients were estimated using hormone data. By resampling the residuals then the y_i^* values were calculated. Finally, coefficients with (y_i^*, x_i) points were fitted. The bootstrap replicate is taken as B = 50, 100, 200, 500, 1000, 2000, 5000 and means and standard errors of the coefficientsfor each B numbers were calculated. The result is that the bootstrap coefficients were compared with the OLS (original) coefficients.

Data: The hormone data were taken from Efron & Tibshirani [21]. A medical device has been tested on n = 27 subjects to give continuous anti-inflammatory hormone. The dependent variable y_i is the amount of hormone remaining in the device after wearing. The independent variable is the usage time of the device (wear-out) in hours.

Algorithm for Bootstrap regression:
1. By applying OLS to the (x_i, y_i) data, the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ were estimated. Residuals are calculated with $\hat{u}_i = y_i - \hat{y}_i$
2. A sample of n = 27 was selected with resampling from residuals, $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i^*$ values were calculated.
3. $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ estimates were obtained using (y_i^*, x_i).
4. Steps 2 and 3 were repeated B times.
5. For each B situation, mean and standard errors were calculated for $\hat{\beta}_0$ and $\hat{\beta}_1$.
Monte Carlo simulation was used to resample residuals. Microsoft Excel was used for all calculations.

4.Results

Ordinary least squares parameter estimates using original hormone data were calculated. Some statistics for OLS estimates were given in Table 1. Predicted y_i outcome variable and the residuals are given in Table 2.

Table 1: OLS regression estimates for the original sample

	OLS coeff.	Std.Error	t	p-value	95% CI Low	95% CI High
Intercept	34.167528	0.867197	39.3999	<0.00001	32.3815	35.9535
Hrs	-0.057555	0.004464	-12.8683	<0.00001	-0.06664	-0.04825

R^2=.86883 Adj-R^2=.863584 SSexp=936,5355, SSres=141.3911 F=165.593 p<0.0001 White Homoscedasticity test: Test statistic: TR^2 = 4,167089, with p-value = P(Chi-square(2) > 4,167089) = 0,124488

Table 2 and Table 3 show that OLS and bootstrap parameter estimates respectively. The parameter estimates obtained by applying OLS to the original data are 34.16752817 for intercept and -0.05745463 for slope. In the bootstrap method, it is seen that even in 50 repetations, very good results are obtained. As the number of repetitions increases, it is clear that the bootstrap estimates are closer to the original estimates. For B = 200 bootstrap repetations, parameter estimates and standard errors obtained as $\hat{\beta}_0^*$ = 34.156546 (.83349107) ; $\hat{\beta}_1^*$ = -.0573257 (.00423444). These estimates are very close to the ordinary least squares estimates. This shows that 200 repetations for bootstrap in simple linear regression, may be enough for a good estimate.

5. Conclusions

In this study, parameter estimation for linear regression using the bootstrap method and one application with real data set were introduced. The results of the article demonstrate the bootstrap method which one of a resampling approach in estimating the linear regression parameters. To show the applicability, the real data (

hormone data) were used to assess the performance of the method. The bootstrap repetations number were taken as B=50, 100, 200, 500, 1000, 2000 and 5000. Regression coefficients obtained from original observations and bootstrap regression coefficients were found to be close to each other. It seems that estimates are close enough even when the bootstrap is B = 50-100 repeated. The result is that the requisite recount in the bootstrap regression by re-sampling residues can be taken as 200. As a result, because the bootstrap method uses a nonparametric technique, it can be reliably used as an alternative to the OLS in situations where parametric conditions are violated.

Table 2: Caption of table Predicted yi outcomes and the residualsusing original data

Hrs (x_i)	Amount (y_i)	Predicted	Residuals
99	25.8	28.48034	-2.680344602933
152	20.5	25.43569	-4.935690771958
293	14.3	17.33576	-3.035762655591
155	23.2	25.26335	-2.063351875865
196	20.6	22.90805	-2.308053629262
53	31.1	31.12287	-0.022874343024
184	20.9	23.59741	-2.697409213634
171	20.9	24.34421	-3.444211096703
52	30.4	31.18032	-0.780320641722
376	16.3	12.56772	3.732280136313
385	11.6	12.0507	-0.450703175409
402	11.8	11.07412	0.725883902451
29	32.5	32.50159	-0.001585511768
76	32	29.80161	2.198390527021
296	18	17.16342	0.836576240502
151	24.1	25.49314	-1.393137070656
177	26.5	23.99953	2.500466695483
209	25.8	22.16125	3.638748253807
119	28.8	27.33142	1.468581371020
188	22	23.36762	-1.367624018843
115	29.7	27.5612	2.138796176229
88	28.9	29.11225	-0.212253888607
58	32.8	30.83564	1.964357150464
49	32.5	31.35266	1.147340462185
150	25.4	25.55058	-0.150583369353
107	31.7	28.02077	3.679225786648
125	28.5	26.98674	1.513259163206

.

Table 3. The bootstrap parameter estimates obtained by resampling residues.

Number of repetitions (B)	**Mean (std.Error) of $\hat{\beta}_0^*$**	**Mean (std.Error) of $\hat{\beta}_1^*$**
50	34.283072 (.71022693)	-.0583396 (.00385206)
100	34.139065 (.87071901)	-.0557362 (.00447734)
200	34.156546 (.83349107)	-.0573257 (.00423444)
500	34.158258 (.79678649)	-.0574234 (.00407645)
1000	34.187013 (.84154165)	-.0575513 (.00424061)
2000	34.164888 (.83637047)	-.0574991 (.00434638)
5000	34.166802 (.83576703)	-.0574207 (.00431041)

References

1. Efron B. 1979, Bootstrap methods: Another look at the Jackknife. *Annals of Statistics*, **7**: p. 1-26.
2. Kantar, Y.M., I. Usta, and Ş. Acıtaş, 2011. A Monte Carlo simulation study on partially adaptive estimators of linear regression models. *Journal of Applied Statistics*, **38**(8): p. 1681-1699.
3. Efron, B., and R. Tibshirani, 1986. Bootstrap Methods for standat errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, **1**: p. 54-77.
4. Sprenger, J, 2011. Science without (parametric) models: the case of bootstrap resampling. *Synthese*, **180**: p. 65–76.
5. Takma, Ç., and H. Atıl, 2016. Study on bootstrap method and it's application II confidence interval, hypothesis testing and regression analysis with bootstrap method [Bootstrap metodu ve uygulanışı üzerine bir çalışma 2. Güven aralıkları, hipotez testi ve regresyon analizinde bootstrap metodu]. *Ege Üniversitesi Ziraat Fakültesi Dergisi*, **43**(2): p. 63-72.
6. Singh, K., and M. Xie, Bootstrap: A statistical method. In P. Peterson, E. Baker & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed). 2010, Elsevier Science: Rutgers University- NJ, p. 46 –51.
7. Kreiss, J.P., and E. Paparoditis, 2011. Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society* **40**: p. 357–378.
8. Gonzàlez-Manteiga, W., and A. Martinez-Calvo, 2011. Botstrap in functional linear regression. *Journal of Statistical Planning and Inference*, 141: p. 453–461. doi:10.1016/j.jspi.2010.06.027
9. McMurry, T., and D.N. Politis, 2008. Bootstrap confidence intervals in nonparametric regression with built-in bias correction. *Statistics and Probability Letters*, **78**: p. 2463–2469.
10. Buchholz, A., N. Holländer, and W. Sauerbrei, 2008. On properties of predictors derived with a two-step bootstrap model averaging approach-A simulation study in the linear regression model. *Computational Statistics & Data Analysis*, **52**: p. 2778 – 2793.
11. Okutan, D., *The usage of the Bootstrap method in regression analysis and its comparison with other methods*. 2009, Ms Thesis Ege University: İzmir.
12. Amiri, S., and S. Zwanzig, 2011. Assessing the coefficient of variations of chemical data using bootstrap method. *Journal of Chemometrics*, **25**(6): p. 295-300.
13. Amiri, S., and D. von Rosen, 2011. On the efficiency of bootstrap method into the analysis contingency table. *Computer Methods and Programs in Biomedicine*, **104**(2): p. 182-187.
14. Baydılı, K.N., and D. Sığırlı, 2017. Comparison of variance homogenity tests for different distributions. *Turkiye Klinikleri Journal of Biostatics*, **9**, doi: 10.5336/biostatic.2017-5631
15. Özdemir, A. F., and G. Navruz, 2016. The effect of number of bootstrap samples, trimming proportion and distribution to the results in bootstrap-t and percentile bootstrap methods [Bootstrap-t ve yüzdelik bootstrap yöntemlerinde tekrar sayısı, budama yüzdesi ve dağılımın sonuçlara etkisi]. *Nevşehir Bilim ve Teknoloji Dergisi*, **5**(2): p. 74-85.
16. Amiri, S., and R. Modaress, 2017. Comparison of tests of contingency tables. *Journal of Biopharmaceutıcal Statıstics*, http://dx.doi.org/10.1080/10543406.2016.1269786
17. Martin, M. A., and S. Roberts, 2010. Jackknife-after-bootstrapregression influence diagnostics. *Journal of Nonparametric Statistics*, **22**(2): p. 257-269.
18. Beyaztas, U., A. Alin, and M.A. Martin, 2014. Robust BCa–JaB method as a diagnostic tool for linear regression models. *Journal of Applied Statistics*, **41**(7): p. 1593-1610.
19. Beyaztas U, & Alin A. (2014) Jackknife-after-Bootstrap as logistic regression diagnostic tool. Communications in Statistics - Simulation and Computation, 43:9, 2047-2060.
20. Karlis D. 2004, An introduction to Bootstrap Methods. 17th Conference of Greek Statistical Society. http://stat-athens.aueb.gr/~karlis/lefkada/boot.pdf
21. Efron, B., and R. Tibshirani, *An Introduction toThe Bootstrap*. 1993: Chapman&Hall.Inc. USA,p. 87-107.