American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Bootstrap Estimate of Prediction Error of Simple Linear Regression Models

Sulafah Binhimd*

Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia
Email: shamad@kau.edu.sa

Abstract

Regression analysis is one of the necessary strategies utilized in statistical inferences, that is employed to estimate the relationship between variables. One way to measure the efficiency of the regression model is to estimate the prediction error, the best model is to have the lowest prediction error. During this paper we are going to estimate the prediction error using bootstrap methods, we will use two different bootstrap methods, Efron's bootstrap and Banks' bootstrap methods. They are resampling strategies but in a different manner. We will review them later thoroughly during this paper. We will find that Banks' bootstrap will be a good choice in most cases.

Keywords: Banks' bootstrap; Bootstrap methods; Efron's bootstrap method; Prediction error; Regression models.

1. Introduction

Regression models are one of the most important statistical methods which used to estimate the relationship between dependent and independent variables [1]. Here we will use a simple linear regression models, which implies that is one dependent variable and another independent, and also the relationship between them is linear. This model is used for prediction and prediction error is one of the measures used to verify the model's ability to predict the dependent variable. The prediction error is that the expected square of difference between a future response (dependent variable) and its prediction from the model. During this paper we will use two methods of bootstrap to estimate the prediction error, the Efron's bootstrap and the Banks' bootstrap.

---

* Corresponding author.

The Efron's bootstrap method [2] is a "computer-based" approach for assigning measures of accuracy to statistical estimates and based on independent observations, Efron in that article introduced the bootstrap method and praised it in finding solutions to the problems of estimation. He used the bootstrap method to estimate the variance of the sample median, and explained that it is an excellent alternative to the jackknife method which failed with this estimate. The basic idea of the bootstrap is estimating the properties of the probability distribution for a random variable of interest. The Efron's bootstrap sample is obtained via sampling with replacement from the original sample. This method has been used extensively in numerous statistical inference methods. The Banks' bootstrap method [3] is the smoothed version of Efron's bootstrap, it is smooth the Efron's bootstrap by linear interpolation histospline smoothing among the jump points of empirical distribution. Banks' created *n+1* intervals and then sample the observations from them. Banks' used confidence regions to compare his method to other bootstrap methods. He estimated the confidence region at different values of α and used the chi-square test of goodness of fit to compare between methods. During this paper we will estimate the prediction error using these two methods of bootstrap and discuss the performance of this process. This study uses real valued observations on a finite interval to generate bootstrap samples and estimates the prediction error of the simple linear regression model.

In Section 2 we show an overview of the regression models, and therefore the two methods of bootstrap with how they are used to estimate a prediction error of the simple linear regression model. Section 3 will discuss the approach used in this paper with some results achieved. Section 4 shows the conclusions of this study.

2. Materials and Methods

Regression models are one of the most important methods used in statistical inference. It is used to estimate the relationship between dependent and independent variables and widely used for prediction. Regression models involves the unknown parameters β, independent variable *X* and dependent variable *Y*. In simple linear regression, there are one independent variable and two parameters, for modeling *n* data points:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \; i = 1, \dots, n \qquad (1)$$

where ϵ_i is an error term. In case of dealing with sample, the simple linear regression model is estimated by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad , \quad i = 1, \dots, n \qquad (2)$$

The least squares estimate of β is

$$\hat{\beta} = (x^T x)^{-1} x^T y \qquad (3)$$

The residual $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model and the true value of the dependent variable. For more details about regression models see [1,4]. Prediction error is a way that measure how properly a model predicts the value of dependent variable of a future observation, it is used to select the best model which is the model that has the lowest prediction error. In regression models the prediction error is the expected square of difference between a future response (dependent variable) and its

prediction from the model

$$PE = E(y - \hat{y})^2 \qquad (4)$$

There are many ways that to estimate the prediction error and one of these methods is the bootstrap method. In this paper, two methods of bootstrap will be used, the Efron's bootstrap and Banks' bootstrap, below we will review the two methods generally. In [2,5,6] Efron outlined a bootstrap method that depends on sampling with replacement from the original sample, and used it to estimate the bias and standard error of ant estimator. Efron explained that the bootstrap method is more efficient than the Monte Carlo method. By generating *B* bootstrap samples, the distribution of any statistic will be estimate by calculating the statistic from every bootstrap sample. The Efron's bootstrap method can be used with different statistical inferences, see [7,8], these references give a broad picture of the bootstrap method and it is various applications in all applied statistical and mathematical aspects. Now we show the basic steps of the Efron's bootstrap:

- Construct F_n, the empirical probability distribution by putting probability $1/n$ to each value $x_1, x_2, \dots, x_n$, $F_n(x) = \sum_{i=1}^{n} I(x_i \leq x)/n$. It is the number of elements which are less than or equal to *x* in the sample divided by size of this sample.
- Resample *B* samples of size *n* from the original sample, with replacement.
- Calculate the statistic of interest T_n from each sample to get $T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*$.
- Construct the empirical distribution of $T_{n1}^*, T_{n2}^*, \dots, T_{nB}^*$ by placing probability $1/B$ at each one of them.

Banks [2] described new version of Efron's bootstrap method, which we will refer to by Banks' bootstrap. In this method Banks' smooths Efron's bootstrap by linear interpolation histospline smoothing among the jump points of empirical distribution. Histospline is a smooth density estimate based on the information in a histogram. This procedure is:

- Take *n* observations which are real valued, one dimensional on a finite interval.
- Create *n+1* intervals between the *n* observations $x_0, x_1, x_2, \dots, x_n, x_{n+1}$ where x_0 and x_{n+1} are the end points of the possible data range.
- Put uniformly distributed probabilities $1/(n+1)$ over each interval.
- Sample *n* observations from the distribution.
- Calculate the statistic of interest.
- Repeat the last two steps *B* times, to get *B* Banks' bootstrap samples.

In Banks' bootstrap, the empirical distribution function is smoothed using linear interpolation histospline smoothing among the jump points. It spreads the probability $1/(n+1)$ uniformly over any interval between two values of observations.

3. Results and Discussion

Here we will use the bootstrap methods we mentioned earlier to estimate the prediction error. Efron [7]

described the bootstrap estimate of prediction interval, this method resamples *B* bootstrap samples to estimate the model on each, and then fit the model to the original sample to get *B* estimates of prediction error. The average of *B* estimates is the overall estimate of prediction error. He shows the prediction error when the model (2) estimated from the bootstrap samples is applied to the original sample "error 1" and to the bootstrap sample itself "error 2", but Efron improved the bootstrap estimate of prediction error by focusing on the difference between the two errors, it is called "optimism error" and added it to the average residual squared error, and called it "apparent error" $\sum_{i=1}^{n}(y_i - \hat{y})^2/n$. The total of two errors is the bootstrap estimate of prediction error. In this paper we are studying more than one case, the first one when applying the Efron's method which described earlier, and we will refer to it by "Method 1", see [7]. The second case when applying the same model to the Banks' bootstrap sample instead of original sample and called it "method 2", which means we use two types of bootstrap in one process in "Method 2". And use $\boldsymbol{B = 1000}$ bootstrap samples to fit simple linear regression model with these cases. The Banks' bootstrap method used here to resample the independent variable which has simple linear relationship with dependent variable. Data used here from Uniform (0,1), Beta (2,5) and Beta (5,2), to study the symmetric and skewed data, with different sample sizes. The size of bootstrap sample can be chosen different to the size of original sample but here we use the bootstrap sample the same size as the original sample. The Tables below illustrate the results of this study, each table shows the optimism error, apparent error and their total which represents the bootstrap estimate of prediction error. This was done with Method 1 and Method 2.

Table 1: Prediction error if data is drawn from Uniform (0,1)

	Method 1			Method 2		
n	optimism	apparent	total	optimism	apparent	total
20	0.0053	0.0272	0.0325	0.0039	0.0198	0.0237
50	0.0015	0.0201	0.0216	0.0183	0.0015	0.0198
150	0.0005	0.0227	0.0232	0.0004	0.0170	0.0174
200	0.0004	0.0214	0.0218	0.0003	0.0211	0.0214
500	0.0002	0.0213	0.0215	0.0002	0.0211	0.0213
1000	$8.07e^{-5}$	0.0203	0.0204	$8.3e^{-5}$	0.0202	0.0203
2000	$2.2e^{-5}$	0.0207	0.0207	$2.3e^{-5}$	0.0208	0.0208
5000	$2.4e^{-5}$	0.0209	0.0209	$2.6e^{-5}$	0.0207	0.0207

Given the tables here, we see that judging the quality of the model using one type of the mentioned error types can be very optimistic or vice versa. So, it is best to use the total of the two types of error, optimism and apparent errors, to be the bootstrap estimate of the prediction error as indicated in [7]. We are able to see from the tables that the optimism error, generally, decreases as the sample size increases with all distributions discussed here, whereas the apparent error moves between increase and decrease, making the total follow the same path. In Table 1 we find that the lowest value of apparent error is 0.0015 with *n*=50 when using Method 2, while the lowest value of this error using Method 1 is 0.0201 with *n*=50. Tables 2 and 3 show that the smallest

value of apparent error is at sample size *n*=50 with Method 2 and at *n*=20 with Method 1.

Tables 1 and 2 show that the use of Method 2 gives better results because the bootstrap estimate of prediction error is less in most cases.

This is different from the results shown in Table 3 which show that sometimes Method 1 is better and sometimes the Method 2 is a better. Given Table 1 we find that the smallest value of the total is 0.0204 when using Method 1 and 0.0174 when using Method 2.

In Table 2 the lowest value of the total is 0.0062 when dealing with the two methods, while the Table 3 shows us that the minimum value of the total is 0.0059 when using Method 1 and 0.0051 when using Method 2. This shows that the use of Banks' bootstrap often produces better results most of the time.

Table 2: Prediction error if data is drawn from Beta (2,5)

	Method 1			Method 2		
n	optimism	apparent	total	optimism	apparent	total
20	0.0013	0.0055	0.0068	0.0026	0.0111	0.0137
50	0.0007	0.0064	0.0071	0.0004	0.0062	0.0066
150	0.0002	0.0076	0.0078	0.0002	0.0064	0.0066
200	0.0001	0.0071	0.0072	0.0001	0.0064	0.0065
500	$7.7e^{-5}$	0.0069	0.007	$2.8e^{-5}$	0.007	0.007
1000	$2.9e^{-5}$	0.0065	0.0065	$3.8e^{-5}$	0.0062	0.0062
2000	$1.2e^{-5}$	0.0068	0.0068	$5.2e^{-6}$	0.0065	0.0065
5000	$6.6e^{-6}$	0.0062	0.0062	$9.5e^{-6}$	0.0063	0.0063

Table 3: Prediction error if data is drawn from Beta (5,2)

	Method 1			Method 2		
n	optimism	apparent	total	optimism	apparent	total
20	0.0011	0.0050	0.0061	0.0013	0.0059	0.0072
50	0.0008	0.0102	0.0110	0.0004	0.0047	0.0051
150	0.0002	0.0057	0.0059	0.0002	0.0069	0.0071
200	0.0001	0.0067	0.0068	0.0001	0.0065	0.0066
500	$7.7e^{-5}$	0.0071	0.0072	$4.9e^{-5}$	0.0067	0.0067
1000	$3.7e^{-5}$	0.0065	0.0065	$9.7e^{-6}$	0.0063	0.0063
2000	$1.3e^{-5}$	0.0061	0.0061	$1.1e^{-5}$	0.0067	0.0067
5000	$5.7e^{-6}$	0.0062	0.0062	$7.9e^{-6}$	0.0064	0.0064

4. Conclusion

In this paper we discussed the methods of estimating prediction error of the simple linear regression model using two methods of bootstrap, Efron's bootstrap and Banks' bootstrap. The first method used the Efron's bootstrap to fit the model and then apply it to the original sample, and the second one applies the same model to the Banks' bootstrap sample. In both methods we used the total of optimism error and apparent error to be the bootstrap estimate of prediction error. We found that using Banks' bootstrap sample instead of the original sample often gives better results, this gives us a better way to estimate the prediction error.

5. Recommendations

As we explained earlier, this study discusses estimating the prediction error of the simple linear regression models using bootstrap methods. This study can be extended in different ways, such as using bootstrap samples sizes which differ from the original sample size. Additionally, this study could be applied to different distributions or different regression models, this may require applying of a generalization of the Banks' bootstrap method to work with real valued observations on infinite interval.

References

[1] J. Neter, M.H. kutner, C.J. Nachtsheim, W. Wasserman. Applied linear statistical models. McGraw-Hill, 1996.

[2] B. Efron. "Bootstrap methods: Another look at the jackknife." The Annals of Statistics, vol. 7, pp.1-26, 1979.

[3] D.L. Banks. "Histospline smoothing the Bayesian bootstrap." Biometrika., vol. 4, pp. 673-684, 1988.

[4] A.J. Scott,"illusions in regression analysis." International Journal of forecasting, vol. 28, pp. 689-694, 2012.

[5] B. Efron. "More efficient bootstrap computations." Journal of the American statistical association, vol. 85, pp. 79-89,1990.

[6] B. Efron, G. Gong. "A leisurely look at the bootstrap, the jackknife, and cross validation." The American statistician, vol. 37, pp. 36-48, 1983.

[7] B. Efron, R.J. Tibshirani. An Introduction to the Bootstrap. Chapman and Hall, 1993.

[8] A.C. Davison, D.V. Hinkley. Bootstrap Methods and their applications. Campridge University Press, 1997.