Journal of Mathematics and Statistics 6 (2): 84-88, 2010
ISSN 1549-3644
© 2010 Science Publications

Bootstrap Method for Dependent Data Structure and Measure of Statistical Precision

[1]T.O. Olatayo, [2]G.N. Amahia and [3]T.O. Obilade
[1]Department of Mathematical Sciences, Olabisio Onabanjo University,
Ago-Iwoye, Ogun State, Nigeria
[2]Department of Statistics University of Ibadan, Ibadane, Nigeria
[3]Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria

Abstract: Problem statement: This article emphasized on the construction of valid inferential procedures for an estimator $\hat{\theta}$ as a measure of its statistical precision for dependent data structure. **Approach:** The truncated geometric bootstrap estimates of standard error and other measures of statistical precision such as bias, coefficient of variation, ratio and root mean square error are considered. **Results:** We extend it to other measures of statistical precision such as bootstrap confidence interval for an estimator $\hat{\theta}$ and illustrate with real geological data. **Conclusion/Recommendations:** The bootstrap estimates of standard error and other measures of statistical accuracy such as bias, ratio, coefficient of variation and root mean square error reveals the suitability of the method for dependent data structure.

Key words: Truncated geometric bootstrap, standard error, bias, coefficient of variation, ratio, root mean square error and bootstrap-t confidence interval

INTRODUCTION

Ever since its introduction by Efron (1979), considerable attention has been given to bootstrap methods as an application of theoretical and methodological problems for statistics. The bootstrap method for estimating the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data, are available for implementing the bootstrap and the accuracy of bootstrap estimates depend on whether the data are a random sample from a distribution or a time series process.

A typical problem in applied statistics involves the estimation of an unknown parameter θ. The two main questions asked are (i) what estimator $\hat{\theta}$ should be used? (ii) Having chosen to use a particular $\hat{\theta}$, how accurate is it as an estimator of θ? (Efron and Tibshiran, 1993). The bootstrap is a general methodology for answering the second question. It is a computer based method, which substitutes considerable amounts of computation in place of theoretical analysis.

This study is concerned with application of bootstrap method to stochastic time series process and we proposed a non-parametric bootstrap method called a truncated geometric bootstrap method for stationary time series data. The procedure attempts to mimic the original model by retaining the stationarity property of the original series in the resample pseudo-time series. The pseudo time series is generated by resampling blocks of random size at each truncation, where the length L of each blocks has a truncated geometric distribution with appropriate probability attached to it. This method shares the construction of resampling blocks of observation with replacement to form pseudo-time series of equal or less, with the original series, so that the statistics of interest may be recalculated base on the resampled data set. The method has two major components, the construction of a bootstrap samples and the computation of statistics on the bootstrap samples, through some kind of a loop.

The procedure provides and estimates different measures of statistical accuracy for an estimator $\hat{\theta}$, such as standard error, bias, coefficient of variation and root mean square error. We extended it to other measure of statistical accuracy by application of bootstrap-t confidence interval with a goal to improve by an order of magnitude upon the accuracy of the standard intervals $\hat{\theta} \pm Z^{(\alpha)}_{\hat{\sigma}}$, in a way that allows routine application even to a complicated problems and it produced good approximate confidence interval. Most

Corresponding Author: T.O. Olatayo, Department of Mathematical Sciences, Olabisio Onabanjo University, Ago-Iwoye, Ogun State, Nigeria of the proofs and technical details are omitted, these can be found in the references given, particular (Diciccio and Efron, 1996; Efron, 1984; Efron and Gong, 1983; Efron and Tibshiran, 1986).

We described how the bootstrap works, assessing the accuracy or precision of the sample mean. Efron and Tibshiran (1986) described the accuracy of the sample mean for independent data, while in this study it was extended to dependent data structure. Then, a description of the resampling algorithm is as follows: Let Bi, b = [X_i, X_{i+1}, ---,X_{i+b-1}] be the block consisting of be observations starting from X_i. In the preceding, if j > N, X_j is defined to be X_k, where k = j(mod N) and $X_o = X_N$. Let P be a fixed number in [0,1]. Independent of X_1, ---,X_N, let L_1,L_2, --- be a sequence of independent and identically distributed (iid) random variables having a truncated geometric distribution, so that the probability of the event [L_i = r} is K $(1\text{-}P)^{r-1}$ p for r = 1,2,---,N where K is a constant found, using the condition ΣP(L = r) = 1, to be $1/[1-(1-P)^N]$. Independent of the X_i and L_i, let I_1, I_2, --- be a sequence of iid random variables that have the discrete uniform distribution on {0,---,N}. Now, a pseudo time series $X_1^*, X_2^*, ---, X_N^*$ is generated in the following way. Sample a sequence of blocks of random length by the prescription $B_{I1,}$ L_1, $B_{I2,}L_2$, ---. The first L_1 observations in the pseudo time series $X_1^*, X_2^*, ---, X_N^*$ are determined by the first block $B_{I1,}$ L_1 of observations X_{I1}, ---, X_{I1} + L_1-1; the next L_2 observations in the pseudo time series are the observations in the second sampled block B_{I2},L_2, namely X_{I2}, ---, X_{I2} + L_2-1. This process is stopped once n observations in the pseudo time series have been generated. Once $X_1^*, X_2^*, ---, X_N^*$ has been generated, one can compute the quantity of interest for the pseudo-time series. This method of resampling and generating $X_1^*, X_2^*, ---, X_N^*$ defines conditionally on the original data X_1,---,X_N or probability measure P^* and the number of block b at each truncation. It shares the same properties with the stationary bootstrap method of Politis and Romano (1994), since the average length of these blocks is 1/p, it is expected that the quantity 1/p should play a similar role as the parameter b in the moving blocks bootstrap method of Kunsch (1989). Most common statistical methods were developed in the 1920s and 1930s, when computation was slow and expensive. Now that computation is fast and cheap we can hope for and expect changes in statistical methodology. This study discusses one such potential change and evaluates the statistical accuracy or precision of the estimated parameter.

MATERIALS AND METHODS

The description of the bootstrap estimates, as we applied the algorithm described above to a real geological data from a Batan well at regular interval is presented. They are the principal oxide of sand or sandstone, which is SiO_2 or Silicon oxide. The point is that the bulk of oil reservoir rocks in Nigeria sedimentary basins is sandstone and shale, a product of sill stone Olanrewaju (2007) and Nwachukwu (2007). Therefore, having chosen to use a particular $\hat\theta$, how accurate is it as an estimator of θ? We present and test how accurate it is for dependent data structure.

Bootstrap method for standard errors: The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications and estimating the standard error of $\hat\theta$ by the empirical standard deviation of the replications. The result is the bootstrap estimate of standard error denoted by $\hat{s}_{eB}$, where B is the number of bootstrap samples used. The bootstrap algorithm for estimating standard errors and coefficient of variation is as follows:

- Select B independent bootstrap samples, each consisting of n data drawn with replacement from X
- Estimate the standard error $se_f(\theta)$ by the sample standard deviation of the B replications:

$$\hat{se}_B = \left[\sum_{b=1}^{B} \left(\hat\theta^*(b) - \hat\theta(.) \right)^2 \Big/ (B-1)^{1/2} \right] \qquad (1)$$

Where:

$$\hat\theta(.) = \sum_{b=1}^{B} \hat\theta^*(b) \Big/ B$$

The limit of $\hat{se}B_B$ as B goes to infinity is the ideal bootstrap estimate of $se_f(\theta)$:

$$\lim_{B\to\infty} \hat{se}_B = se_f(\hat{\theta}^*) \qquad (2)$$

The non parametric algorithm has the virtues of avoiding all parametric assumptions, all approximations and in fact all analytical difficulties of any kind. The Coefficient of Variation (CV) of a random variable X, is defined to be the ratio of its standard error to be the absolute value of its mean:

$$CV_f(.) = se_f(\hat{\theta}) / \hat{\theta}_f \qquad (3)$$

This measures the randomness or variability in X relative to the magnitude of its deterministic part θ_f, which refers to variation both at the resampling (bootstrap) level and at a the population sampling level.

Bootstrap estimates of bias: Bias is another measure of statistical accuracy, measuring different aspects of $\hat{\theta}$'s behavior. Bias is the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated:

$$Bias_F = Bias_F(\hat{\theta}\,\theta) = E_F[s(X)] – t(F)$$

We generated the bootstrap samples, evaluate the bootstrap replications $\hat{\theta}^*(b) = s(X^{*b})$ and approximate the bootstrap expectation $E_F[s(X^*)]$ by the average:

$$\hat{\theta}^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b) \Big/ B = \sum_{b=1}^{B} s(X^{*b}) \Big/ B \qquad (4)$$

The bootstrap estimate bias based on the B replications is (4) with $\hat{\theta}^*(.)$ substituted for $E_{\hat{F}}[s(X^*)]$:

$$\text{Bai}\hat{\text{s}}_B = \hat{\theta}^*(.) - E(\hat{F}) \qquad (5)$$

The ratio of estimate bias to standard error, $\text{Bai}\hat{\text{s}}_B / \hat{se}_B$ are also calculated as another measure of statistical accuracy and the smaller the ratio, the higher the efficiency of the estimates. If bias is large compared to the standard error, then it may be an indication that the statistic $\hat{\theta} = s(X)$ is not an approximate estimate of the parameter θ.

Root mean square error: This is another measure of statistical accuracy that takes into account both bias and standard error. The root mean square error of an estimator $\hat{\theta}$ for θ, is $\sqrt{E_F[(\hat{\theta}-\theta)^2]}$. It can be shown that the root mean square equals:

$$\begin{aligned}\sqrt{E_F[(\hat{\theta}-\theta)^2]} &= \sqrt{se_F(\hat{\theta})^2 + bias_F(\hat{\theta},\theta)^2} \\ &= se_F(\hat{\theta}).\sqrt{1+\left(\frac{bias_F}{se_F}\right)^2} \\ &= se_F(\hat{\theta}).\left[1+\frac{1}{2}\left(\frac{bias_F}{se_F}\right)^2\right]\end{aligned} \qquad (6)$$

If $bias_F = 0$ then the root mean square error $\sqrt{MSE}$ = its minimum value se_F.

RESULTS AND DISCUSSION

The summary of our findings on the performances of a truncated geometric bootstrap method for dependent data structure based on the implementation of the prescribed algorithm and for block sizes of (1,2,3,4) and bootstrap replicates of (B = 50, 100, 250, 500 and 1000) is given in the Table 1. In Table 1 the measures for statistical accuracy of an estimator $\hat{\theta}$ from the geological data is presented.

From Table 1 it is observed that the bootstrap estimate of $\hat{\theta}$ is nearly unbiased. The standard error are crude but useful measures of statistical accuracy, if the true sampling distribution F is (0, 1), then the true standard error are in the column SE. The Coefficient of Variation (CV)) in each bootstrap replications at different block sizes are moderate with less bias and ratio. The bootstrap bias estimates and the ratio of estimated bias to standard error are small with $\sqrt{MSE}$ of an estimator $\hat{\theta}$. This moderate minimum values in each column, indicate that in each replication we do not have to worry about the bias of $\hat{\theta}$.

As a rule of thumb, a bias of less than 0.25 standard errors can be ignored, unless one are trying to do careful confidence interval calculation Efron and Tibshiran (1993).

The situation is more complicated when the data are time series, because bootstrap sampling must be carried out in a way that suitably captures the dependence structure of the Data Generation Process (DCP). The block bootstrap is the best known method for implementing the bootstrap with time series data when one does not have a finite dimensional parametric model that reduces the DGP to independent random sampling.

Table 1: Summary statistics for bootstrap estimates of Standard Error (SE), Coefficient of Variation (CV), bias, root means square error ($\sqrt{\text{MSE}}$) and ratio for $\hat{\theta}$

Bootstrap replicates	Ave	SE	CV	Bias	Ratio	$\sqrt{\text{MSE}}$
B = 50:						
Block of 1	55.29069	0.5618	0.0102	0.00060	0.00110	0.5618
b = 2	55.29990	0.6836	0.0124	0.00880	0.01430	0.6836
b = 3	55.36110	0.7603	0.0137	0.00710	0.93400	0.7636
b = 4	55.25310	0.8619	0.0156	-0.03700	-0.00429	0.8630
B = 100:						
b = 1	55.36090	0.6095	0.0110	0.07080	0.11620	0.7852
b = 2	55.31530	0.7816	0.0141	0.02520	0.03220	0.8179
b = 3	55.09860	0.7319	0.0133	-0.19150	-0.26170	0.8552
b = 4	55.07470	0.6809	0.0156	-0.25020	-0.25020	0.8807
B = 250:						
b = 1	55.32420	0.7035	0.0127	0.03410	0.04850	0.7852
b = 2	55.30600	0.8177	0.0148	0.01590	0.01950	0.8179
b = 3	55.16120	0.8454	0.0153	-0.12890	-0.15470	0.8552
b = 4	55.17860	0.8736	0.0158	-0.11500	-0.12760	0.8807
B = 500:						
b = 1	55.20390	0.7789	0.0141	-0.08620	-0.11070	0.7837
b = 2	55.25320	0.7966	0.0144	-0.03690	-0.04630	0.7975
b = 3	55.26640	0.7833	0.0142	-0.02370	-0.03030	0.7837
b = 4	55.28540	0.8705	0.0158	-0.00470	-0.00540	0.8705
B =1000:						
b = 1	55.26330	0.9831	0.0177	-0.02680	-0.02730	0.9835
b = 2	55.26680	0.9752	0.0177	-0.02330	-0.02390	0.9755
b = 3	55.26890	0.6751	0.0122	-0.02120	-0.03140	0.6754
b = 4	55.25310	0.6641	0.0122	-0.03700	-0.05330	0.6651

Table 2: Summary statistics of 90% bootstrap confidence interval for $\hat{\theta}$, when B = 500 and 1000

		B =5 00	B = 1000
		------------------------	-------------------------
	Methods	95% confidence Interval	95% confidence Interval
b = 1	Standard	[53.9226, 56.4852]	[53.6773, 56.7305]
	Boostrap-t	[53.6461, 56.8805]	[53.3364, 57.1902]
b = 2	Standard	[53.9428, 56.7305]	[53.6626, 56.8710]
	Boostrap-t	[53.6919, 56.8145]	[53.3554, 57.1782]
b = 3	Standard	[53.9779, 56.5549]	[53.7311, 56.8017]
	Boostrap-t	[54.1584, 56.3794]	[53.9457, 56.5921]
b = 4	Standard	[53.8457, 56.7174]	[53.5792, 57.5601]
	Boostrap-t	[54.1113, 56.3949]	[53.8927, 56.6135]

The trouble with standard intervals is that they are based on and asymptotic approximation that can be quite inaccurate in practice. We implemented bootstrap -t confidence interval for producing good approximate confidence intervals. The goal is to improve by an order of magnitude upon the accuracy of the standard intervals $\hat{\theta} \pm Z^{(\alpha)}_{\hat{\sigma}}$, in a way that allows routine application even to very complicated problems. The bootstrap-t procedure is a useful and interesting generation of the usual student's t method and it is particularly applicable to location statistics like the sample mean. The method was suggested in Efron (1983), but some poor numerical results reduced its appeal. Hall (1988) study showing the bootstrap-t's good second-order properties has revived interest in its use. Babu and Singh (1983) gave the first proof of second-order accuracy for the bootstrap-t and Diciccio and Efron (1992) showed that they are also second order correct.

A practicable t confidence interval for θ at level 1-α is:

$$[\hat{\theta} - \hat{s}_{\theta} t^{*}_{1-\alpha/2}, \hat{\theta} + \hat{s}_{\theta} t^{*}_{1-\alpha/2}]$$

Where:

$\hat{s}_{\theta}$ = The standard error of $\hat{\theta}$

t^{*}_{σ} = The σ quantile of the bootstrap t statistics

It should not be used if $\hat{s}_{\theta}$ is unreliable, especially if strongly dependent on $\hat{\theta}$. Therefore if $\hat{s}_{\theta}$ is reliable we can use:

$$[\hat{\theta} - \hat{s}_{e^*}(\hat{\theta} t^{*}_{1-\alpha/2}, \hat{\theta} + \hat{s}_{e^*}(\hat{\theta}) t^{*}_{1-\alpha/2}]$$

From the Table 2, it is revealed that bootstrap-t confidence interval at 95% level of significance has a wider range than the standard normal confidence interval. The distributions are positively skewed. A confidence interval is desired for the scale parameter θ.

In this case the bootstrap-t confidence interval based on $\hat{\theta}$ is a definite improvement over the standard interval. Therefore with the above results of different methods of measure of statistical accuracy of $\hat{\theta}$, we can fit a time series model to the available data for effective description and predication purposes.

To determine the bottom bootstrap confidence interval. By applying these methods to estimated $\hat{\theta}$ estimator, we have the Table 2.

CONCLUSION

The truncated geometric bootstrap method for dependent data structure is justified by concentrating on basic ideas and applications rather than theoretical consideration. The bootstrap estimates of standard error and other measures of statistical accuracy such as bias, ratio, coefficient of variation and root mean square error reveals the suitability of the method for dependent data structure.

The bootstrap-t confidence interval also produces good approximate confidence intervals for the estimator $\hat{\theta}$ which is suitable for model fitting and predictive purposes.

REFERENCES

Babu, G.I. and K. Singh, 1983. Inference on means using the bootstrap. Ann. Stat., 11: 999-1003. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176346267

Diciccio, T.J. and B. Efron, 1992. More accurate Confidence Intervals in exponential families. Biometrika, 79: 231-245. http://cat.inist.fr/?aModele=afficheN&cpsidt=4681660

Diciccio, T.J. and B. Efron, 1996. Boostrap Confidence Intervals. Biometrika, 88: 189-212.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Ann. Stat. 7: 1-26. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176344552

Efron, B., 1983. The Jacknife, the Boostrap and other Resampling Plans. Society for Industrial and Applied Mathematics, ISBN: 0898711797, pp: 92.

Efron, B., 1984. Better boostrap confidence intervals Tech-Rep. Department of Statistics, Standford University.

Efron, B. and G. Gong, 1983. A leisurely look at at the boostrap, the Jacknife and Cross-validation. Ann. Stat., 37: 36-48.

Efron, B. and R. Tibshiran, 1986. Boostrap measures for standard errors, confidence intervals and other measures of statistical accuracy. Stat. Sci., 1: 54-77. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1177013815

Efron, B. and R. Tibshiran, 1993. An Introduction to the Boostrap. Chapman and Hall/CRC, London, ISBN: 0412042312, pp: 436.

Hall, P., 1988. Theoretical comparison of boostrap confidence intervals. Ann. Stat., 16: 927-985. http://www.projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176350933

Kunsch, H.R., 1989. The Jacknife and the boostrap for general stationary observations. Ann. Stat., 17: 1217-1241. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176347265

Nwachukwu, U.I., 2007. Organic matter, the source of our wealth. An inaugural lecture delivered at Oduduwa Hall, Obafemi Awolowo University, Ile-Ife, Nigeria.

Olanrewaju, V.O., 2007. Rocks: Their beauty, language and roles as resources of economic development. An Inaugural lecture delivered at Oduduwa Hall, Obafemi Awolowo University, Ile-Ife, Nigeria

Politis, D. and J. Romano, 1994. The stationary boostrap. J. Am. Stat. Assoc., 47: 1303-1313. http://direct.bl.uk/bld/PlaceOrder.do?UIN=022518422&ETOC=EN&from=searchengine