BOOTSTRAP METHOD FOR MEASURES OF STATISTICAL ACCURACY

Olatayo,T .O.,Alabi,O.O.,Oguntunde,B.O.,Ogunfolu,O.B. And Odetunde,S.O.
Department Of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria

Abstract

We introduced bootstrap method for dependent data structure and emphasis is on the construction of efficient inferential procedures for an estimator $\hat{\theta}$ as a measure of its statistical accuracy, such as standard error, bias, ratio, coefficient of variation and root mean square error. It was illustrated with real time series data structure.

Key words: Dependent data, Bias, Coefficient of variation and root mean square error.

Introduction

The bootstrap method is not simply computer simulation and bootstrapping is not perfectly synonymous with Monte Carlo. Bootstrap methods rely on using an original sample or some part of it, such as residuals as an artificial population from which to randomly re sampled , Barreto and Howland (2006).

A typical problem in applied statistics involves the estimation of an unknown parameter θ. The two main questions asked are (1) what estimator $\hat{\theta}$ should be used? (2) Having chosen to use a particular $\hat{\theta}$, how accurate is it as an estimator of θ? Davison and Hinkley (1997). The bootstrap method 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. Even for relatively simple problems computer intensive methods like bootstrap are an increasingly good data analytic bargain in an era of exponentially declining computational cost, Efron and Tibshirani (1986). The bootstrap approach, as initiated by Efron (1979), avoids having to derive formulas via difficult analytical arguments by taking advantage of fast computers. And excellent introduction to the bootstrap may be found in the work of Efron and Tibshirani (1993), Hinkley (1988), Dicicco and Romano (1988) and Dimitris (2004). Recently, Kunsch (1989), and Liu and Singh (1992) have independently introduced non parametric versions of the bootstrap that are applicable to weakly dependent stationary observations. Their re sampling procedure has been generalized by Politis and Romano (1994), Bulmann (2002), Andrews (2004) Chang and Park (2003), Davidson and Mackinnon (2004) and Mackinnon (2005), by re sampling 'blocks of blocks' of observation to the stochastic time series process. This article investigates and describes the basis of bootstrap theory as a measure of statistical accuracy for dependent data structure of an estimator $\hat{\theta}$. We will describe how the bootstrap works in terms of a problem, assessing the accuracy of the sample mean for dependent data process.

Suppose that our data consists of a random sample from an unknown probability distribution F on the real line,

$$X_1, X_2, \text{----------}, X_n \sim F \qquad 1.1$$

Having observed $X_1=x_1$, $X_2 = x_2$, ---, $X_n =x_n$, we compute the sample mean $\bar{X} = \sum_1^n x_n / n$ and wonder how accurate it is as an estimate of the true mean $\theta = E_F[X]$. If the second central moment of F is $\mu_2(F) = E_F X^2 - (E_F X)^2$, then the standard error $\sigma(F, n, \bar{x})$, that is the standard deviation of $\bar{x}$ for a sample of size n from distribution F, is

$$\sigma(F) = \left[\mu_2(F)/n\right]^{1/2} \qquad 1.2$$

The shortened notation $\sigma(F) = \sigma(F, n, \bar{x})$ is allowable because the sample size n and statistics of interest $\bar{x}$ are known, only F being unknown. The standard error is the traditional measure of $\bar{x}$'s accuracy. Unfortunately, we cannot actually use (1.2) to assess the accuracy of $\bar{x}$, since we do not know μ_2,(F), but we can use the estimated standard error

$$\bar{\sigma} = \left[\bar{\mu}_2/n\right]^{1/2} \qquad 1.3$$

Where $\bar{\mu}_2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 / (n-1)$, the unbiased estimate of μ_2(F).

AJPS VOL. 1 NO. 2
2008.

Let $\hat{F}$: probability mass $1/n$ on $x_1, x_2, ---, x_n$

1.4

Then we can simply replace F by $\hat{F}$ in (1.2), obtaining

$$\hat{\sigma} = \sigma(\hat{F}) = \left[\mu_2(\hat{F})/n\right]^{1/2},$$

As the estimated standard error for $\bar{x}$. This is the bootstrap estimate.

Since

$$\hat{\mu}_2 = \mu_2(\hat{F}) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n},$$

1.6

$\hat{\sigma}$ is not quite the same as σ, but the difference is too small to be important in most applications. In most cases there is no equivalent to formula (1.2), which expresses the standard error σ (F). As a result, formulas like (1.3) do not exist for most statistics. This is where the computer comes in. It turns out that we can always numerically evaluate the bootstrap estimate $\hat{\sigma}=\sigma(\hat{F})$, without knowing a simple expression for σ(F). The evaluation of $\hat{\sigma}$ is straight forward bootstrap algorithm described in the next section. It effectively gives the statistician a simple formula like (1.3) for any statistic, no matter how complicated. Some authors have described the approach to independent data structure and speculating that it may not be applied to dependent data structure, Efron (1979), Liu and Singh (1992).

In this study, we investigate the application to dependent data structure and justified how accurate is estimator θ for this processes by estimating its standard error, bias, ratio, and root mean square error . 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 hop for and expect changes in statistical methodology.

$$\hat{\sigma}_B = \left(\frac{\sum_{b=1}^{B} \left[\hat{\theta}(b) - \hat{\theta}()\right]^2}{B-1} \right)^{1/2}$$

$$\hat{\theta}() = \frac{\sum_{b=1}^{B} \hat{\theta}(b)}{B}$$

2.3

Material and Method

This section presents a more careful description of the bootstrap estimate of standard error, bias, ratio, coefficient of variation and root mean square error for dependent data structure.

Let σ(F) indicate the standard error of $\hat{\theta}$, as a function of the unknown sampling distribution F,

$$\sigma(F) = \left[\text{var}_F[\hat{\theta}(X_i)]\right]^{1/2} \qquad 2.1$$

Of course σ(F) is also a function of the sample size n and the form of the statistic $\hat{\theta}(X)$, but since both of these are known they need not be indicated in the notation. The bootstrap estimate of standard error is

$$\hat{\sigma} = \sigma(\hat{F}) \qquad 2.2$$

Where $\hat{F}$ by means of a bootstrap algorithm, which depends on the following notation:

$X^* = (x_1^*, x_2^*, ---, x_n^*)$ indicates n independent draws from $\hat{F}$, called a bootstrap sample. Because $\hat{F}$ is the empirical distribution of the data, a bootstrap sample turns out to be the same or less as a blocks of random sample size n drawn with replacement from the actual sample $\{x_1, x_2, ----------, x_n\}$.

The bootstrap algorithm proceeds in three steps.

i) Select B independent bootstrap samples say $X^*_{(1)}, X^*_{(2)}, ---, X^*_{(B)}$, each consist of n data value drawn with replacement from X.

ii) Evaluate the statistic of interest for each bootstrap sample, say $\hat{\theta}(b) = \hat{\theta}(X^*(b))$, b=1,2,---, B

iii) Calculate the sample standard deviation of the $\hat{\theta}(b)$ values

It is easy to see that as $B \to \infty$, $\hat{\sigma}_B$ will approach $\hat{\sigma} = \sigma(\hat{F})$, the bootstrap estimate of standard error.

The bias of $\hat{\theta} = S(X)$ as an estimate of θ, is defined to be the difference between the expectation of θ and the value of the parameter θ.

The bootstrap algorithm is also applicable to evaluate the bias of $\hat{\theta}$:

$$Bias_F = bias_F(\hat{\theta}, \theta) = E_F[S(X) - t(F)] \quad 2.4$$

The bootstrap replication $\hat{\theta}(b) = S(X^{*b})$ will give

$$Bias_F = E_\mu[S(X^{*b}) - t(\hat{F})] \quad 2.5$$

The bootstrap estimate of bias based on the B replication is

$$Bias_B = \hat{\theta}(.) = t(\hat{F})$$

$$\hat{\theta}(.) = \frac{\sum_{b=1}^{B} \hat{\theta}(b)}{B} = \frac{\sum_{b=1}^{B} S(X^{*b})}{B} \quad 2.6$$

Therefore, we calculate both se_B and $bias_B$ from the same set of bootstrap replications. The coefficient of variation and ratio were also computed as a measure of statistical accuracy for estimate of $\hat{\theta}$.

The root mean square error $\sqrt{MSE}$ is also estimated which take into account both bias and standard error. The root mean square error of an estimator $\hat{\theta}$ for θ, is $\sqrt{E_F[(\hat{\theta}-\theta)^2]}$.

For each of the methods of measuring statistical accuracy, a computer program was developed for bootstrap algorithm described and applied to a real dependent data structure for blocks of (1,2,3,4) at bootstrap replication of (B = 50, 100, and 250). This was also accomplished by estimating the coefficient of variation (C.V), ratio and root mean square error $\sqrt{MSE}$ at each replication for different block sizes.

Results and Discussions

The summary of our findings in the bootstrap method as a measure of statistical accuracy for dependent data structure is given in table 1. In table 1 the bootstrap estimates of standard error, (SE) coefficient of variation (C.V), bias, ration and root mean square error $\sqrt{MSE}$ for $\hat{\theta}$ is presented.Coefficient of variation (C.V), bias, ratio, root mean square error $\sqrt{MSE}$ for $\hat{\theta}$.

Bootstrap replicates	Ave	SE	CV	Bias	Ratio	$\sqrt{MSE}$
B=50						
b = 1	55.29069	0.5618	0.0102	0.0006	0.0011	0.5618
b =2	55.2999	0.6836	0.0124	0.0098	0.0143	0.6836
b = 3	55.3611	0.7603	0.0137	0.0071	0.934	0.7636
b = 4	55.2531	0.8619	0.0156	-0.037	-0.00429	0.5630
B=100						
b = 1	55.3609	0.6095	0.0110	0.0708	0.1162	0.6136
b =2	55.3153	0.7816	0.0141	0.0252	0.0322	0.7820
b = 3	55.0986	0.7319	0.0133	-0.1915	-0.2617	0.7570
b = 4	55.0747	0.6009	0.0156	-0.2502	-0.2502	0.6007
B=250						
b = 1	55.3242	0.7035	0.0127	0.0341	0.0485	0.7852
b =2	55.3060	0.8177	0.0148	0.0159	0.0195	0.8179
b = 3	55.1612	0.8454	0.0153	-0.1289	-0.1547	0.8552
b = 4	55.1786	0.7036	0.0158	-0.115	-0.1276	0.6807

B = bootstrap replications and b is block sizes.

From the table 1, it is observed that if the true sampling distribution F is N(0,1), then the true standard error are in the column (SE), the coefficient of variation in C.V in each bootstrap replications at different block sizes are moderate with less bias and ratio. The bias, ratio and root mean square error $\sqrt{MSE}$ of an estimator $\hat{\theta}$ are small. The minimum moderate values in each column indicate that in each replication we do not have to worry about the bias of $\hat{\theta}$ for dependent data. As a rule of thumb, a bias of less than 0.25 standard error can be ignored, unless one are trying to do careful confidence interval calculation, Efron and Tibshirani (1993).

Therefore at each bootstrap replication we can still have a good estimate, but block sizes of 4 at bootstrap replications of 100 will give better information about the autocorrelation structure for dependent data. The higher the block sizes the better, for effective statistical accuracy of an estimator $\hat{\theta}$.

Conclusion

The bootstrap method gives a better and reliable method for measuring statistical accuracy of estimator $\hat{\theta}$ for dependent data structure. The block size of 4 at different replications is to be preferred. It preserves and maintains stationary data structure of the process; and the bootstrap results suggest that the invariance approximately holds for relatively all sample sizes.

Refrences:

Andrews ,D. W.K.(2004). The block-block bootstrap:Improved asymptotic refinements, *Econometrica*,72,673-700.

Barreto ,H. and Howland, F.M.(2006). Introductory Econometrics,Using monte carlo simulation with Microsoft excel. Cambridge University press.

Buhlmann, P.(2002). Bootstrap for time series; Statistical Science, 17,52-72.

Chang, Y.,and Park, J.Y.(2003). A sieve bootstrap for the test of a unit root, *Journal of Time Series Analysis*,24,379-400.

Davison, R., and MacKinnon, J.G. (2004). Econometric Theory and Methods, New York, Oxford University Press.

Davison, A.C. and Hinkly, D.V. (1997). Bootstrap methods and Their Application. Cambridge University Press.

Diciccio, T. and Romano J.(1988). A review of bootstrap confidence intervals (with discussion). *Journal of the Royal Statistical Society*. Ser.B, 50,338-370.

Dimitris, K. (2004). An introduction to Bootstrap Methods, 17th Confrence of Greek Statistical Society.

Efron, B.(1979). Bootstrap methods another look at the Jacknife, Ann.Statist. 7,1-26.

Efron, B. and Tibshirani, R. (1986). Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54-77.

Efron, B. and Tibshrani, R. (1993). An introduction to bootstrap. Chapman and Hall/CRC.London.

Hinkly, D. (1988). Bootstrap methods. Journal of the Royal Statistical Society, B,50,321-337.

Kunsch, H.R.(1989). The Jacknife and the bootstrap for general stationary observations, The Ann. Of Statistics, 17,1217-1241.

Liu,R.Y. and Singh, K. (1992). Moving blocks Jacknife and bootstrap capture weak dependence, Exploring the limits of Bootstrap,eds.R.Lepage and L.

Billard, New York: John Willey.Mackinnon J.G. (2005). Bootstrap Methods in Econometrics. http://www.econ.queensu.ca/ faculty/ mackinnon/

Politis, D. and Romano, J.(1994). The stationary bootstrap. *Journal of American Statistical Association*, 47,1303-1313.