EUROPEAN ACADEMIC RESEARCH
Vol. VI, Issue 7/ October 2018

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

Bootstrap and multiple imputation under missing data in AR(1) models

ELJONA MILO
University "Fan S. Noli" of Korca, Albania
LORENA MARGO
University "Fan S. Noli" of Korca, Albania

Abstract:

Missing data is a phenomenon detected in many scientific investigations due to the bias often caused and inefficient analysis of the data. Determining the appropriate analytic approach in the presence of incomplete observations is a major question for data analysts. Recently many scientists have developed several statistical methods to address missingness.

Since the introduction by Efron (1979), bootstrap has resulted to be an important method for estimating the distribution of an estimator by applying the resampling of the data. This bootstrap method resulted efficient in the case of independent and identically distributed observations, but in the case of dependent data like time series, classic bootstrap gives incorrect answers. For this reason, to adapt the bootstrap in the case of time series Kunsch (1989) presented a bootstrap method with blocks compounded by a fixed number of observations. Block bootstrap methods developed by researchers resulted suitable in the case of time series and give good results under specific assumptions.

In this paper we will realize simulations with intention to compare the results obtained using block bootstrap combined with missing data mechanisms. We are interested in estimating the parameter in the AR(1) time series model using block bootstrap procedure after filling in the missing values using multiple imputation. We compare the results using several block length also different missing data mechanisms and packages for completing the missing values.

Key words: Bootstrap, time series, missing data, imputation, autoregressive

INTRODUCTION

Missing data is a commonly occurring complication in many scientific investigations. Determining the appropriate analytic approach in the presence of incomplete observations is a major question for data analysts. The analysis of time series data constitutes an important area of statistics. Since, the data are records taken through time, missing observations in time series data are very common. This occurs because an observation may not be made at a particular time from causes like faulty equipment, lost records, or a mistake, which cannot be rectified until later.

When one or more observations are missing it may be necessary to estimate the model and also to obtain estimates of the missing values. By including estimates of missing values, a better understanding of the nature of the data is possible with more accurate forecasting. Different series may require different strategies to estimate these missing values. We emphasize the necessarity of using these strategies effectively in order to obtain the best possible estimates. It is wellknown the importance of both multiple imputations and the bootstrap in deriving confidence bands and critical values for test statistics and also to remove biases in estimators.

In this paper we are focused in the performance of the bootstrap after filling in the missing values in a stationary time series chosen. We generate a time series from an autoregressive model of the first order and conducted simulations to examine the performance of the block bootstrap and multiple imputation.

Missing value mechanisms

There are three important cases to distinguish for the responsible generating processes behind missing values (see Rubin (1987), Rubin and Little (2002)). Let $X=\left(x_{ij}\right),1\le i\le n,1\le j\le p$ denote the data, where n is the number of observations and p the number of observed variables (dimensions), and let $M=\left(M_{ij}\right),1\le i\le n,1\le j\le p$ be an indicator whether an observation is missing $\left(M_{ij}=1\right)$ or not $\left(M_{ij}=0\right)$. The missing data mechanism is characterized by the conditional distribution of M given X, denoted by $f\left(M/X,\phi\right)$, where ϕ indicates unknown parameters. Then the missing values are Missing At Random (MAR) if it holds for the probability of missingness that

$$f\left(M/X,\phi\right)=f\left(M/X_{obs},\phi\right) \quad (1)$$

Where $X=(\mathrm{X}_{obs},\mathrm{X}_{miss})$ denotes the complete data, and X_{obs} and X_{miss} are the observed and missing parts, respectively. Hence the distribution of missingness does not depend on the missing part X_{miss}.

If in addition the distribution of missingness does not depend on the observed part X_{obs}, the important special case of MAR called Missing Completely At Random (MCAR) is obtained, given by

$$f\left(M/X,\phi\right)=f\left(M/\phi\right) \quad (2)$$

If equation (1) is violated and the patterns of missingness are in some way related to the outcome variables, i.e., the probability of missingness depends on X_{miss}, the missing values are said to be Missing Not At Random (MNAR). This relates to the equation

$$f\left(M/X,\phi\right)=f\left(M/(X_{obs},X_{miss}),\phi\right). \quad (3)$$

Multiple imputation and the bootstrap

Multiple imputation is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries

are missing. Application of the technique requires three steps: imputation, analysis and pooling.

In multiple imputation, the imputation process is repeated multiple times resulting in multiple imputed datasets. In this method the imputation uncertainty is accounted for by creating these multiple datasets. The multiple imputation process contains three phases: the imputation phase, the analysis phase and the pooling phase (Rubin, 1987; Van Buuren, 2012).

In the imputation model, the variables that are related to missingness, can be included. That way bias is reduced and estimates are more precise.

In the first phase, the imputation phase, several copies of the data set are created each containing different imputed values. The imputed values are estimated using the means and covariance of the observed data. The specification of the correct imputation model is very important for the performance of multiple imputation. Firstly, it is important to include the correct variables in the imputation process. Accordingly, all variables that are of substantial interest should be included in the imputation model.

Secondly, it is important to have an imputation model that fits the distribution assumptions of the data. In the second phase, the analysis phase, the statistical analysis is carried out. On each imputed dataset, the analysis is carried out that would have been applied had the data been complete. That way as many sets of results are created as the number of imputed datasets created in the imputation phase.

Since the introduction by Efron (1979), bootstrap has resulted to be a very important method in the case of estimating the distribution of an estimator by applying the resampling of the data. The classic bootstrap method of Efron resulted efficient in the case of independent and identically distributed observations, but in the case of dependent data like time series, classic bootstrap gives incorrect answers.

The application of the classic bootstrap is limited by the requirements of independence in the data. In the case of dependent data, Kunsch (1989) presented a bootstrap method with blocks compounded by a fixed number of observations and argued that although individual observations are not independent, blocks of observations can be independent with each other for a proper block length.

The bootstrap and multiple imputations are both computationally intensive methods that have been developed by statisticians in the last twenty years. Both multiple imputations and the bootstrap are powerful techniques used in deriving confidence bands and critical values for test statistics and also to remove biases in estimators. Like many econometric techniques, the validity of these methods is based on asymptotic approximations, and these underlying approximations may be inaccurate for a particular data and model. In spite of these limitations, both of these techniques seems to be an important part of nowadays researches.

We are interested in combining these methods and comparing the results obtained in estimating the parameter in AR(1) time series model chosen.

Our main concern is based in the effect of the performance of block bootstrap in the case of missing data after we fill in the data using multiple imputation and different packages present in R environment.

Basic ideas behind *imputeTS* and *mice* packages

In this paper we used two packages for impute dhe missing value: *imputeTS* and *mice* and compare their performance for impute the missing values in AR(1) process.

The *imputeTS* package is a collection of algorithms and tools for univariate time series imputation. This package specializes on time series imputation. It offers several different imputation algorithm implementations. Beyond the imputation

algorithms the package also provides plotting and printing functions of missing data statistics (see Mortiz (2018)).

MICE (Multivariate Imputation via Chained Equations) is one of the commonly used package by R users, creating multiple imputations as compared to a single imputation takes care of uncertainty in missing values. This package in R, helps us to imput missing values with plausible data values. These plausible values are drawn from a distribution specifically designed for each missing datapoint (see Van Buuren (2011)).

Simulations

As stated above, we are interested in studying the performance of block bootstrap combined with multiple imputation methods used for missing data in the case of estimating the parameter in AR(1) models. We applied suitable functions to obtain missing data in different percentages under the MCAR and MNAR mechanisms in the generated time series. Values have lost in different percentages to better control the performance of these methods.

We generate a time series with size 1000 from the AR(1) model with the coefficient equals to 0.5. Based on the nature of the time series we choose an appropriate block bootstrap method and we conduct simulations using 1000 bootstrap replications for the estimation of the parameter.

First we use the block bootstrap methods to obtain the estimation and we compare the results with the bootstrap estimation in the case of missing data. We lose the values under the MCAR and NMAR mechanisms in different percentages and we fill these missing values using the functions contained in the *mice* package and the *imputeTS* package under the R programming language. Then we apply an appropriate block bootstrap method to obtain the estimation for the parameter.

The simulation results for different block lengths when applying the chosen block bootstrap procedure for estimating

the coefficient in the selected AR (1) model are given in the following tables:

Table 1: Results obtained for the bootstrap estimation before and after imputation for the parameter $\phi = 0.5$ in the case of missingness under the MCAR mechanism (BBE- Block bootstrap estimation, BAI- Block bootstrap after imputation) .

Block length	MICE package	Percent of missing values			
		7%	10%	25%	40%
7	BBE	0.4164267	0.4065379	0.3964842	0.3965801
	BAI	0.401087	0.3792374	0.3547448	0.2893694
10	BBE	0.4171638	0.4171471	0.4170415	0.4171263
	BAI	0.4086008	0.3845318	0.3565449	0.2986512
12	BBE	0.4254202	0.4257027	0.4253197	0.4253977
	BAI	**0.4146571**	**0.3997529**	**0.3796416**	**0.3016322**

From the results obtained presented in the tables above, we note that in the case of low percentage of missing values (7%), the bootstrap estimator before and after the imputation approximates more to the value of the coefficient ϕ. We also notice that by increasing the block length, the block bootstrap estimator after the imputation has a better performance. In the case of higher percent of missing data, the bootstrap estimator after imputation has a lower approximation to the value of the coefficient we estimate.

Following the same idea, the missing data will be impute by the *imputeTS* package and we will compare the results obtained by seeing which method is most appropriate for filling in the missing values in the generated time series.

Table 2 : Results obtained for the bootstrap estimation before and after imputation for the parameter $\phi = 0.5$ in the case of missingness under the MCAR mechanism (BBE- Block bootstrap estimation, BAI- Block bootstrap after imputation)

Block Length	imputeTS package	Percent of missing values			
		7%	10%	25%	40%
7	BBE	0.4395211	0.4413146	0.4273791	0.3966628
	BAI	0.3994554	0.4004721	**0.4113708**	0.3257772
10	BBE	0.4165681	0.4166681	0.4165681	0.4064677
	BAI	0.4194742	0.4207643	**0.4326917**	0.3427908
12	BBE	0.4757582	0.4627742	0.4625403	0.4627742
	BAI	0.4257431	0.4271928	**0.4391727**	0.3478172

From the results obtained in the case when we use *imputeTS* package, we notice that when there is a significant number of missing values (in our case 25%) , this package has quite good performance. Also, from the results obtained, we note that with the increase of block length, we obtain a better approximation. Following the same idea we obtain missing values from the AR(1) series under the NMAR mechanism and then we use both the *mice* and *imputeTS* packages to impute them. Even in this case, we will compare which of the packages used has a better performance.

Table 3: Results obtained for the bootstrap estimation before and after imputation for the parameter $\phi = 0.5$ in the case of missingness under the NMAR mechanism (BBE- Block bootstrap estimation, BAI- Block bootstrap after imputation)

Block Length	MICE package	Percent of missing values			
		7%	10%	25%	40%
7	BBE	0.4519998	0.4356367	0.4332925	0.4036854
	BAI	0.3668479	0.3266193	**0.4614295**	**0.5120746**
10	BBE	0.4600531	0.4254555	0.4674878	0.4361966
	BAI	0.3698929	0.3365399	0.4572336	0.5519208
12	BBE	0.4496224	0.4867121	0.5021096	0.476866
	BAI	0.3882245	0.3500175	0.4404637	0.6352625

Table 4: Results obtained for the bootstrap estimation before and after imputation for the parameter $\phi = 0.5$ in the case of missingness under the NMAR mechanism (BBE- Block bootstrap estimation, BAI- Block bootstrap after imputation)

Block Length	imputeTS package	Percent of missing values			
		7%	10%	25%	40%
7	BBE	0.4112243	0.425108	0.4560708	0.4553541
	BAI	0.3440718	0.3805397	**0.4615858**	0.6916066
10	BBE	0.4558386	0.4555388	0.4321372	0.4852437
	BAI	0.4043438	0.35984	**0.5293976**	0.6446606
12	BBE	0.4260904	0.4369461	0.4884104	0.4632721
	BAI	0.4321072	0.4603563	**0.4975494**	0.6589852

From the results obtained, we note that in the case of using the R *mice* package, the best approximation is in the case of a relatively large number of lost values. The best approximation is obtained when the block length equals 7 in the bootstrap method chosen. While in the case of the *imputeTS* package, as in the first case, the best approximation results in the case of 25% of missing values. We also notice that by increasing the bootstrap block length, we get a better approximation.

CONCLUSIONS

From the simulations conducted for estimating the parameter in the AR(1) time series studied, we noticed that both R packages used for completing the missing values have a good performance. In the case of missing data under the MCAR mechanism, the use of *mice* package has a better performance with the increase of the block length. While the *imputeTS* package has a very good approximation in both cases compared to the bootstrap estimator in the original series. This package has a very good performance even when we have a considerable amount of missing data (in our case 25%).

REFERENCES

1. Allison P. D. (2000), *Multiple imputation for missing data: A cautionary tale: Sociological Methods and Research*, 28, 301-309.
2. Baraldi A. N. and Enders C. K. (2010), *An introduction to modern missing data analyses*, Journal of School Psychology, 48: 5-37.
3. Box G. E. P., Jenkins G. M. and Reinsel G. C. (2008), *Time Series Analysis*, Fourth Edition, John Wiley & Sons, Inc.
4. Carpenter J. and Kenward, M. (2013), *Multiple Imputation and its Application*, 1 ed. Wiley.
5. Chen B, Sumi A, Toyoda S, et al. (2015),*Time series analysis of reported cases of hand, foot, and mouth disease from 2010 to 2013 in Wuhan*, China. BMC Infect Dis ;15:495. [PMC free article] [PubMed]
6. Efron B. (1979), *Bootstrap methods: Another look at the jackknife*, *Ann. Statis.*, **7** (1979), 1-26.
7. Efron B., Tibshirani R.J.,(1993), *An introduction to the bootstrap*, Chapman and Hall, New York.
8. Fung D. S. C. (2006), *Methods for the estimation of missing values in time series*, Edith Cowan University Perth.
9. Honaker J, King G.(2010), *What to Do about Missing Values in Time-Series Cross-Section Data*, American Journal of Political Science. American Journal of Political Science ;54:561-81.
10. Kunsch H. (1989), *The Jackknife and the Bootstrap for general stationary observations*, *Ann. Stat.*, 17, 1217-1241.
11. Little R. J. A. and Rubin D. B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

12. Moritz S. (2018). *CRAN R Package: Time Series Missing Value Imputation* http://steffenmoritz.github.io/imputeTS/
13. Politis D.N., (2003), *The impact of Bootstrap Methods on time series Analysis*, Statistical Sc.
14. Rubin D.E. (1987), *Multiple imputation for nonresponse in surveys*, New York: Wiley.
15. Rubin D. B. & Little, R. J. (2002), *Statistical analysis with missing data*, Hoboken, NJ: J Wiley & Sons.ience, Vol.18, No.2.219-230.
16. Schomaker M., Heumann C.(2018), *Bootstrap Inference When Using Multiple Imputation Statistics in Medicine*, 37(14):2252-226.
17. Van Buuren.S and Groothuis-Oudshoorn.K., (2011), *mice: Multivariate imputation by chained equations in R.Journal of Statistical Software*, 45(3):1-67.
18. Van Buuren S (2012), *Flexible Imputation of Missing Data*, Chapman & Hall/CRC Press, Boca Raton, FL. 342 pages. ISBN 9781439868249.