An Investigation of Parametric and Nonparametric Bootstrap Data Generating Processes

C. K. Acha

Department of Statistics,
Michael Okpara University of Agriculture,
Umudike, Nigeria.
e-mail: acha.kelechi@mouau.edu.ng

***Abstract*— This study investigates the different performances of parametric and nonparametric bootstrap data generating processes (DGPs), by ascertaining the conditions where the parametric bootstrap DGP method produces minimum error in comparison with the nonparametric bootstrap DGP method in terms of root mean square and other information criteria. In addition, the sampling distribution was identified. The study used secondary data from Central Bank of Nigeria (1987-2015) on the External Sector (ES). Data were analyzed using different bootstrap DGPs with different group proficiency levels, and the kernel density of the empirical distributions that are not too skewed were considered. In this study, 322560 scenarios were replicated 1000 times using bootstrap DGPs and kernel density methods. The results show that across all the assessment bootstrap conditions considered in the study, the parametric bootstrap method performed better than the nonparametric bootstrap models by showing the smallest conditional bias, standard error and root mean square error (RMSE). The kernel density plots revealed that the sampling distribution of the ESis a Chi-square distribution.**

Keywords--*Parametric, External Sector Statistics, Nonparametric, Kernel density, Bootstrap.*

I. Introduction

The continuing development of bootstrap methods has been motivated by the increasing progress in computational speed and efficiency. There are many ways to specify the bootstrap data generating process (DGP) for models as simple as the linear regression model. This is very essential since the estimates based on economic data have significant influence on public policy decisions made concerning a wide range of issues. For example, determination of poverty and inequality within society are vital to formulating economic policies that can affect the lives of millions. Unfortunately, estimates of these types of measures are quite unreliable when based on common statistical methods. Frequently, policies are put in place based on little more than good faith because reliable methods of evaluating the results of existing policies are not available. This boils down to [1] recommendations which called for more research on the parametric bootstrap method and for comparative studies of parametric and nonparametric approaches.

Many researchers have discovered thatthere are many possible extensions on the bootstrap methods and the necessity to carry out more research in this area[2,3].In this study, the various bootstrap DGP approaches were the estimation methods of interest. Therefore, the purpose of this study is to investigate and understand the parametric bootstrap DGP method and to compare the nonparametric and parametric bootstrap DGP methods in estimating the root mean square error (RMSE) and other information criteria under a variety of assessment conditions.

This study will focus on bootstrapping regression models from the external sector with unknown distributions under a variety of assessment conditions. Moreover, the study will be based on examination of the parametric and nonparametric bootstrap DGP functional models and also use the kernel density plots to reveal the sampling distribution of the ESS.

II. Materials and Methods

Recently, the bootstrap has been introduced in more complex and complicated models. Some of the consulted literatures that handled and discussed bootstrap as complex and complicated models are [4-18] among others.

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties when sampling from an approximating distribution, [18]. One standard choice for an approximating distribution is the empirical distribution of the observed data. The schematic bootstrap shows that in the real world, the unknown probability distribution F gives the data $\mathbf{X} = (x_1,x_2,\ldots,x_n)$ by random sampling; from $\mathbf{X}$ we calculate the statistic of interest $\hat{\theta} = s(\mathbf{X})$. In the bootstrap world, $\hat{F}$ generates $\mathbf{X^*}$ by random sampling giving $\hat{\theta}^* = s(\mathbf{X^*})$. There is only one observed value of $\hat{\theta}$, but we can generate as many bootstrap replications $\hat{\theta}^*$ as affordable. The crucial step in the bootstrap process is the process by which we construct from **x** an estimation$\hat{\mathbf{F}}$of the unknown population **F**. The bootstrap must be applied on the right distribution to get accurate statistical inference, [3].There are many bootstrap methods used in econometrics. Among all the bootstrap methods the parametric and nonparametric will be considered in this study. For a detailed discussion of how reliable the bootstrap method is see [19]. Other authors like [4-20] have worked extensively on different aspects of bootstrap.

However, an ideal reference and detailed account on different aspects of bootstrap of the developments with dependent data and independent data, is carried out in [21]. A further discussion was made on the smooth bootstrap which is equivalent to sampling from a kernel density estimate of the data is a small amount of (usually normally distributed) zero-centered random noise. The zero-centered random noise is now added onto each resampled observation while the kernel density is used to extract all the important features of the data set. To ascertain the accurate statistical inference and achieve the set objective in this study; the various assessment conditions like the bias, standard error and RMSE will be considered. Also, two groups of models will be selected to represent the real data sets, after more than 1000 trials within each bootstrap level (B).

III. Analysis

The major aim of bootstrap testing is that, when a test statistic of interest has an unknown distribution under the null hypothesis, that distribution can be characterized by using information in the data set that is being analyzed.
Consider the linear regression model

$$y_t = X_t\beta + u_t, \sim \mathrm{NID}(0,\sigma^2) \qquad (1)$$

where $E(u_t| X_t) = 0$, $E(u_s u_t = 0)\ \forall s \neq t$,, X_t is a row vector of observations on k regressors, n is the number of observations, β is a k-vector and u_t is the error term. The model (1) is a fully specified parametric model, which means that each set of parameter values for β, σ^2 defines just one data generating process (DGP). The first step in constructing a parametric bootstrap DGP is to estimate (1) by ordinary least square (OLS), yielding the restricted estimates$\tilde{\beta}$, and $\tilde{s}^2$. Then the bootstrap DGP is given by

$$y_t^* = X_t\tilde{\beta} + \mu_t^*,\ \mu_t^* \sim \mathrm{NID}(0,\tilde{s}^2), \qquad (2)$$

which is just the element of the model (1) characterized by the parameter estimates under the null, with “asterisk” to indicate that the data are simulated. In order to draw a bootstrap sample from the bootstrap DGP from equation (2), we first draw an n--vector u^* from the N(0, $\tilde{s}^2$I) distribution. For each of the B bootstrap samples, θ_k^*, a bootstrap test statistic $\theta_k^*$is computed from y_k^*in just the same way as $\hat{\theta}$ was computed from the original data, y in (1).

The parametric bootstrap procedure that we have just described, based on the DGP (2), does not allow us to relax the strong assumption that the error terms are normally distributed but if the true error distribution, whether or not it was normal, we could always generate the $\mu^*$from equation (2) (see [15,18]). Alternatively, we know that the empirical distribution function (EDF) of the error terms is a consistent estimator of the unknown cumulative distribution function CDF of the error distribution because the residuals consistently estimate the errors and it follows that the EDF of the residuals is also a consistent estimator of the CDF of the error distribution. Thus, from the fundamentals of statistics, if we draw bootstrap error terms from the empirical distribution of the residuals, we are drawing them from a distribution that tends to the true error distribution as $n \rightarrow \infty$. The value of each drawing must be the value of one of the residuals, with equal probability for each residual. This is precisely what we mean by the empirical distribution of the residuals. On average, each of the residuals appears once in each of the bootstrap samples. If we adopt this resampling procedure, we can write the bootstrap DGP as

$$y_t^* = X_t\tilde{\beta} + \mu_t^*, \qquad \mu_t^* \sim EDF(\tilde{\mu}_t) \qquad (3)$$

where EDF($\tilde{\mu}_t$)denotes the distribution that assigns probability 1/n to each of the elements of the residual vector ($\tilde{\mu}_t$). The DGP from equation (3) is one form of what is usually called a nonparametric bootstrap, although, since it still uses the parameter estimate $\tilde{\beta}$.

IV. Results

Examination of Parametric and Nonparametric Bootstrap DGP Models

Each of the bootstrap forms were represented by using at least one functional model each from real data sets of a particular bootstrap DGP method to illustrate how others were estimated before tabulation;
Here, the original analysis of the data sets will be carried out. Recall (1),

$$y_t = X_t\beta + u_t, \qquad (4)$$

which is now called (4) will be used to estimate original real data sets with fixed sample size is as follows;

Original Model (OM):

GDPt = bo + b_1A+ b_2B+ u_t (5)

- Original Model (OM): B=1999, N(0,0.9), n_1=10000, using (5)

GDPt = 31.450IM + 21.730EX (6)

Standard error (0.023) (0.070)

Bias (0.008) (0.015)

RMSE (0.0009)

The Parametric bootstrap DGP is

$$y_t^* = X_t\hat{\beta} + f(\hat{\mu}_t)v_t^*, \qquad \mu_t^* \sim NID(0,1) \qquad (7)$$

where$f(\hat{\mu}_t) = \frac{\hat{\mu}_t}{(1-h_t)^{1/2}}$

- Parametric Model (PM): B=1999, N(0,0.9), n_1=10000, using (7)

GDPt= 68.710IM + 49.940EX (8)

Standard error (0.003) (0.034)

Bias (0.001) (0.010)

RMSE (0.0002)

Nonparametric bootstrap DGP

$$y_t^* = X_t^*{}_t\hat{\beta} + \mu_t^*, \qquad [y_{t,}^* x_t^*] \sim NID(\bar{x}, s^2) \ (9)$$

- Nonparametric Model (NPM): B=1999, N(0,0.9), n_1=10000, using (9)

GDPt = 51.46IM + 45.140EX (10)

Standard error (0.015) (0.044)

Bias (0.005) (0.012)

RMSE (0.0005)

bootstrap level NPM	Ability Level	Sample Size		OM	PM
		200	**0.0113**	0.0224	0.0480
		1000	0.0328	**0.0131**	0.0323
	N(0,1)	10000	0.0329	0.0168	**0.0162**
		200	0.1116	**0.0365**	0.0812
		1000	0.0344	**0.0204**	0.0397
B=99	N(0,0.9)	10000	0.0162	0.0340	**0.0114**
		200	0.0777	0.0829	**0.0748**
		1000	**0.0345**	0.0356	0.0356
	N(1,0.25)	10000	0.0176	0.0177	**0.0171**
		200	**0.0773**	0.1224	0.0803
	N(0,1)	1000	**0.0328**	0.0598	0.0590
		10000	0.0337	0.0329	**0.0160**
		200	0.1195	**0.0765**	0.1216
B=499	N(0,0.9)	1000	0.0604	0.0597	**0.0344**
		10000	0.0340	**0.0159**	0.0166
		200	0.1227	**0.0748**	0.1240
	N(1,0.25)	1000	0.0601	**0.0345**	0.0599
		10000	0.0342	**0.0171**	0.0177
		200	**0.1063**	0.0813	0.1036
	N(0,1)	1000	**0.0518**	0.0333	0.0485
		10000	0.0297	**0.0172**	0.0273
		200	0.1052	**0.0828**	0.0904
B=1999	N(0,0.9)	1000	0.0506	**0.0332**	0.0480
		10000	**0.0290**	0.0170	0.0188
		200	0.1042	**0.0814**	0.0878
	N(1,0.25)	1000	0.0344	**0.0308**	0.0485
		10000	0.0710	**0.0289**	0.0268

Table 1. Bias of the SLR for Parametric Bootstrap Models in a Real data set

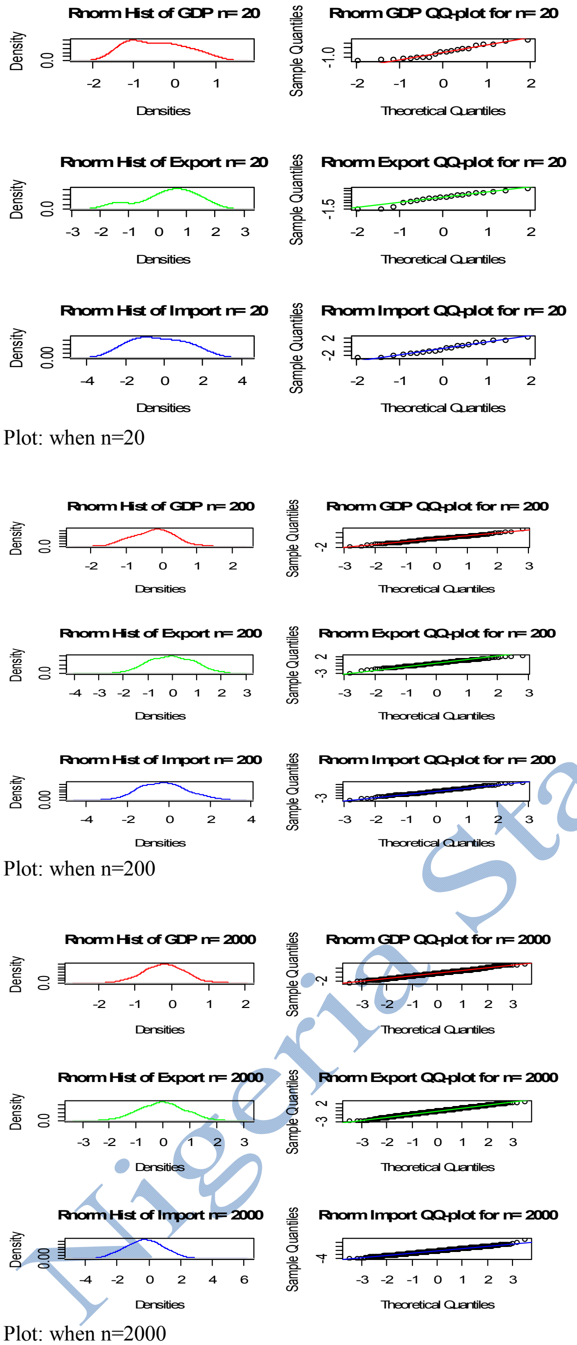
Note. The bold is the smallest value in each row

bootstrap level NPM	Ability Level	Sample Size		OM	PM
		200	**0.0613**	0.1224	0.0680
		1000	0.0328	**0.0319**	0.0323
	N(0,1)	10000	0.0329	**0.0160**	0.0162
		200	**0.0116**	0.0783	0.1195
		1000	**0.0344**	0.0204	0.0397
B=99	N(0,0.9)	10000	0.0165	0.0162	**0.0134**
		200	0.0777	0.0829	**0.0748**
		1000	**0.0345**	0.0356	0.0356
	N(1,0.25)	10000	0.0176	0.0177	**0.0172**
		200	0.0773	0.1224	**0.0703**
	N(0,1)	1000	0.0328	**0.0298**	0.0590
		10000	0.0337	0.0329	0.0160
		200	0.1195	**0.0735**	0.1216
B=499	N(0,0.9)	1000	**0.0304**	0.0597	0.0644
		10000	0.0340	**0.0159**	0.0166
		200	0.1227	**0.0748**	0.1240
	N(1,0.25)	1000	0.0601	**0.0345**	0.0599
		10000	**0.0342**	0.0331	0.0176
		200	0.1063	**0.0813**	0.1036
	N(0,1)	1000	**0.0518**	0.0333	0.0485
		10000	**0.0297**	0.0172	0.0273
		200	**0.1052**	0.0828	0.0904
B=1999	N(0,0.9)	1000	0.0506	**0.0332**	0.0480
		10000	0.0290	**0.0170**	0.0188
		200	**0.1042**	0.0814	0.0878
	N(1,0.25)	1000	**0.0344**	0.0308	0.0485
		10000	**0.0710**	0.0289	0.0268

Table 2: Standard Error of the SLR for Parametric Bootstrap Models in a Real data set.

bootstrap level NPM	Ability Level	Sample Size		OM	PM
		200	0.0713	0.1224	**0.0680**
		1000	0.0328	**0.0319**	0.0323
	N(0,1)	10000	0.0329	**0.0160**	0.0162
		200	0.1116	0.0783	**0.0765**
		1000	**0.0344**	0.0204	0.0397
B=99	N(0,0.9)	10000	0.0162	**0.0159**	0.0340
		200	0.0777	0.0829	**0.0748**
		1000	0.0356	**0.0345**	0.0356
	N(1,0.25)	10000	0.0176	**0.0171**	0.0177
		200	**0.0773**	0.1224	0.0803
	N(0,1)	1000	**0.0328**	0.0598	0.0590
		10000	0.0337	**0.0160**	0.0329
		200	0.1195	**0.0765**	0.1216
B=499	N(0,0.9)	1000	0.0604	0.0597	**0.0344**
		10000	0.0340	**0.0159**	0.0334
		200	0.1227	**0.0748**	0.1240
	N(1,0.25)	1000	0.0601	**0.0345**	0.0599
		10000	0.0342	0.0331	**0.0171**
		200	0.1063	**0.0813**	0.1036
	N(0,1)	1000	0.0485	**0.0333**	0.0376
		10000	0.0297	0.0485	**0.0172**
		200	0.1052	**0.0828**	0.0173
B=1999	N(0,0.9)	1000	0.0506	**0.0332**	0.0177
		10000	0.0290	**0.0170**	0.0188
		200	**0.0814**	0.1042	0.0878
	N(1,0.25)	1000	0.0344	**0.0308**	0.0899
		10000	0.0268	**0.0289**	0.0185

Table 3: RMSE of the SLR for Parametric Bootstrap Models in a Real data set



Plot: when n=20

Plot: when n=200

Plot: when n=2000

Figure 1: Plots with different sample sizes

V. Discussions

In this study, two groups of models were selected to represent the real data sets after more than 1000 trials were carried out within each bootstrap level (B). In fact, 322560 scenarios were replicated more than 1000 times. The selection was based on the fact that as number of trials increase, the models maintain the same pattern, and unless there is change in the pattern another model will not be selected. The two equations above represent each of the groups of models selected; results presented in table (1- 3) will be discussed. This will enable the determination of the effects of the factors (sample size and bootstrap level) on a real data set. Extreme values in the ranges stated above were truncated and special consideration was given to the plotting range and the layout. Even though very low estimates were also observed, results in these ranges are presented in order to demonstrate the
trends and the performance at the lower ends of the distributions for each bootstrap model.

Table 1 shows the conditional bias of the three tests for the bootstrap models when the from real data set, in fact, only the correlation between original values and restricted values were considered. Although the magnitude of bias varied across the bootstrap methods (or models), the pattern of relative effects of these factors was generally consistent within each bootstrap method (or model). It can be seen that sample size and test length of bootstrap level had large effects on bias of the SLR, group proficiency level had relatively small or mixed effects under some conditions bias was smaller for a larger sample size and a shorter test length. Given the same test length, a smaller ratio normally yielded slightly larger bias, especially for the parametric models with the smaller estimate, such as two and three. Although the effect of group proficiency level on bias of the SLR was quite small, it seemed there was an interaction effect between this factor and the bootstrap DGP method. For the nonparametric bootstrap DGP method, as the group differences became larger, the bias of the SLR became somewhat smaller; however, for the parametric methods with a longer test, the bias of the SLR became slightly larger as the groups were more different. There was no evidence showing any effect of the group proficiency level on the parametric method with a short test.

The Tables are presented in the order of tests (1, 2, and 3), sample sizes (200, 1000, and 10000), bootstrap levels (99, 499, and 1999), and group proficiency levels (1, 2, and 3) of the bias, standard error and information criteria for a real data set. It can also be seen that, across different combinations of different test lengths of bootstrap levels group proficiency levels, and the sample size increased, the RMSE obtained from the two bootstrap models, PM and NPM increased at almost all estimated points, which is to be expected because of the property of estimation bias. It can also be noted that although the bias at the two ends of the estimate was large (in absolute value), the curves from different bootstrap models were closer to one another when the sample size was 10000 than when the sample size was 1000. Across all the conditions considered, model-PM yielded much larger bias than model-NPM at almost all the estimates.

In Test 1, across the three different sample sizes and the three different group proficiency levels, the largest bias, standard error and RMSE estimate were always associated with model NPM

Specifically, for group proficiency level 1, 2 & 3, the smallest RMSE was from model PM; while NPM has a small RMSE when the bootstrap levels and sample sizes are large. The same pattern existed for Test 1 in Tables (1 – 3) in terms of the largest/smallest bias and standard error and the rank order of the different models, indicating that including or excluding 2% of the estimated values showed that they had little effect on estimating bias and standard error.

VI. Conclusion

From the tables, it is apparent that, under all bootstrap conditions, the parametric bootstrap functional models produced smaller bias, standard error and RMSE than the non-parametric bootstrap functional models especially when bootstrap level and sample size are large in simple linear equation (SLR).

Finally, this study concludes that parametric bootstrap DGP method produces minimum error in comparison with the nonparametric bootstrap DGP method under several assessment conditions and also the kernel density estimates confirmed that external sector statistics in Nigerian has a chi-square distribution, according to [15]; since the bootstrap distribution created by resampling, also matches the properties of the sampling distribution. This study, therefore, is a stepping stone for further research and prediction in the economic sectors

References

[1] Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

[2] Efron, B., &Tibshirani, R.J. (1993). *An introduction to the bootstrap (Monographs on Statistics and Applied Probability*. New York: Chapman & Hall.

[3] Acha, C. K. (2014a). Parametric Bootstrap Methods for Parameter Estimation in SLR Models. *International Journal of Econometrics and Financial Management,* 2014, 2(5), 175-179. Available online at http://pubs.sciepub.com/ijefm/2/5/2 © Science and Education Publishing DOI: 10.12691/ijefm-2-5-2.

[4] Bergstrom, P. (1999,“bootstrap methods and applications in econometrics - a brief survey”, on-line: http://www.nek.uu.se/pdf/1999wp2.pdf, provided by Uppsala university, department of economics in its series working paper series with number 10.

[5] Bühlmann, P. (2002), ‘Bootstraps for Time Series’, *Statistical Science*, 17, 52–72.

[6] Carpenter M. (2002). Estimation of location extremes within general families of scale mixtures. *J. Statist. Plann. Inf.* 100, 197 – 208.

[7] Carpenter, J. (1999). Test inversion bootstrap confidence intervals. *J. R. Statist. Soc. B* 61, 159 – 172.

[8] Chaubey, Y. P. (2002). Estimation in inverse Gaussian regression: Comparison of asymptotic and bootstrap distributions. *J. Statist. Plann. Inf.* 100 , 135 – 143.

[9] Chen, X., and Fan, Y. (1999). Consistent hypothesis testing in semi parametric and nonparametric models for econometric time series. *J. Econ.* 91, 373 – 401.

[10] Davidson, R. and Flachaire, E. (2001), The Wild Bootstrap,Tamed at Last, GREQAM Document de Travail 99A32, revised.

[11] Davidson, R. and Flachaire, E.,(2008). The wild bootstrap, tamed at last, Journal of Econometrics, Elsevier, 146(1), 162-169.

[12]Flachaire, E. (2002). Bootstrapping heteroskedasticity consistent covariance matrix estimator. *Comput. Statist.* 17, 501 – 506.

[13] Flachaire, E. (2005). More efficient tests robust to heteroskedasticity of unknown form.*Econ. Rev.* 24, 219 – 241.

[14] Acha, C. K. (2014b) Bootstrapping Normal and Binomial Distributions. *International Journal of Econometrics and Financial Management*, 2(6), 253– 256. Doi:10.12691/ijefm-2-6-2.

[15] Acha, C.K. and Acha I.A. (2015) Smooth Bootstrap Methods on External Sector Statistics. *International Journal of Econometrics and Financial Management*, 3(3), 115–120. Doi:10.12691/ijefm-3-3-2.

[16] Acha, I. A. and Acha, C. K. (2011). Interest Rates in Nigeria: An Analytical Perspective. *Research Journal of Finance and Accounting,* 2(3); 71-81 www.iiste.org ISSN 2222-1697 (Paper) ISSN 2222-2847 (Online).

[17] Acha, C. K. and Omekara, C. O. (2016) Towards Efficiency in the Residual and Parametric

Bootstrap Techniques. American Journal of Theoretical and Applied Statistics. 5(5) 285-289. doi: 10.11648/j.ajtas.20160505.16

[18] Acha C. K. (2016). Rescaling Residual Bootstrap and Wild Bootstrap. International Journal of Data Science and Analysis. 2(1): 7-14. doi: 10.11648/j.ijdsa.20160201.12

[19] Davidson, R. and MacKinnon, J.G. (2006), 'Bootstrap Methods in Econometrics', in Patterson, K. and Mills, T.C. (eds), *Palgrave Handbook of Econometrics: Volume 1*

.

[20] Chernick, M. R. (2007). Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition Wiley, Hoboken.

[21] Lahiri , S. N. (2005a). Consistency of the jackknife-after-bootstrap variance estimator for the bootstrap quantiles of a studentized statistic . *Ann. Statist.* 33, 2475 – 2506.

.

.