Big Data Affluence in Statistics Application: A Comparison of Real Life and Simulated Open Data

[*1]Adeboye, Nureni Olawale (PhD) and [2]Olayiwola, M. Oyedunsi (PhD)
[1]Department of Mathematics & Statistics, Federal Polytechnic, Ilaro P.M.B 50, Ilaro, Ogun State, Nigeria
[2]Department of Mathematical Sciences, Osun State University, Osogbo, Osun State, Nigeria
[1]nureni.adeboye@federalpolyilaro.edu.ng

Large data repositories or database management still remain a mirage and tough challenge to accomplish by most developing countries and establishments around the globe. This necessitates the need to improvise on the gathering of suitable data with a good spread to serve as a complement, in the absence of sufficient real-life data. Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. Through classroom activities that involved both sourced real-life and simulated big data in R-environment, models were built and estimates obtained from the adopted techniques revealed the robustness of simulated datasets in a unified observation with improved significant values as reflected in the results. Students appreciated the use of such big data as it enhances their machine learning ability and the availability of sufficient data without delay.

INTRODUCTION

In recent years, there have been many exciting suggestions for teaching statistics in ways that are likely to increase students' understanding and which have been successfully implemented in the classroom, but development of corresponding methods of assessment has been slow. Replacement of timed written examinations by alternative methods of assessment have improved assessment level and understanding (Griffiths and McClone, 1979; Mortensen, 1988; Garfield and Gal, 1999). The impact of big data curriculum in driving home supports for this argument cannot be overemphasized. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from data-sets that are diverse, complex, and of a massive scale (Mashey and Chen, 2016). Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. Question such as: "which one should be trusted more: a model built and validated from limited real life data or model derived from a self-generated data*?* Benefiting from this paradise however required high level of data quality, and this call for efficient data management (Meng, 2018). Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users. Organizations and enterprises are making use of Data more than ever before to inform business decisions and gain deep insights into customer behavior, trends, and opportunities for creating extraordinary customer experiences. (Singh, 1998; John, 2019).

Big data has increased the demand of information management specialists so much that Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole (Huberty, 2015). According to Olavsurd (2016), worldwide income for big data and business analytics will grow from nearly $122 billion in 2015 to $187 billion in 2019, and this presents a projected increase of more than 50 percent over International Data Corporation forecast period. Some areas of its specific applications are in Government, International development, Manufacturing, Healthcare, Education, Media, Information Technology, etc.

It has been suggested by Couldry and Turow (2017) that practitioners in Media and Advertising approach big data as many actionable points of information dissemination. By applying big data principles into the concepts of machine intelligence and deep computing, IT departments can predict potential issues and move to provide solutions before the problems even happen. As big data begin to rise, state of practice

in analytics evolve and there are four major players within this nexus of ecosystem (Xinhua, 2013; Huberty, 2015; Jamborsalamati *et al.,* 2017). These are:

- Data generators. This group belongs to data devices that generate new data about events.
- Data collectors. This group collect data about users and devices with attention to their attributes and attitudes.
- Data aggregators. These organizations glean data from various sources and sell to others.
- Data consumers. This group benefit from the data gleaned and crunched by others.

The industrial shape-shift and the rise of new big data ecosystem has underpinned the foundation of new roles, platforms and analytical methods. The point of concentration in this ecosystem is the emerging roles categorized based on human factors as shown in the figure below:

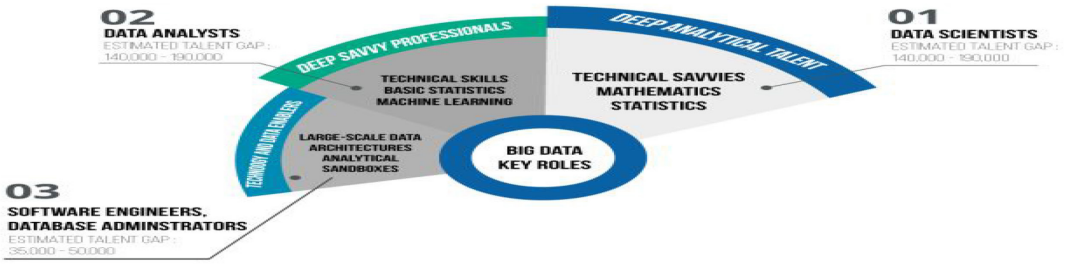

Figure 1: Overview of Big Data Analytical Talents

In furtherance to the above benefits, some researchers have laid emphasis on the superiority of big data in research. Meng (2018), opined that a self-reported administrative data sets covering 80% of population should be more trusted than a 1% survey with 60% response rate. However, estimated results demonstrated data the larger the population size under consideration, the surer researchers fool themselves except data quality is taking into consideration. Adeboye and Agunbiade (2019) carried out Monte Carlo simulations scheme for 81 different variations, of which its design assumed a uniform distribution under a linear heteroscedasticity function. Lagrange Multiplier test employed showed that the tests have good size and power at 5% significant level when compared to limited real life data.

For the purpose of this roundtable discussion, sample activities were carried out using open data of selected 16 West African countries across 10 years sourced from United Nations Educational, Scientific and Cultural Organization (UNESCO) as well as big data (simulated dynamic panel data using R software) with special focus on some macro-economic variables. Estimates obtained from the two datasets revealed the robustness of big data in a unified observation, as regards the superiority of the employed dynamic panel data modelling techniques.

METHOD

The real life data for this article was obtained mainly from United Nations educational scientific and cultural organization (UNESCO) data site which covers a period of 10 years ranging from 2008-2017 for selected West African countries as retrieved in the year 2018 while Monte Carlo simulation scheme was carried out using a data generating procedure specified within a dynamic panel data model.

The data generating procedure (DGP) is giving by

$$y_{it} = \delta y_{i,t-1} + x'_{1it}\beta_1 + x'_{2it}\beta_2 + z'_{1i}\alpha_1 + z'_{2i}\alpha_2 + \varepsilon_{it} + u_{it} \quad (1)$$

$for\ i = 1, \cdots, N\ and\ t = 1, \cdots, T\ where\ \varepsilon_{it} \sim i.i.d.N(0.1), u_{it} \sim i.i.d.N(0.1), \alpha_1 = \alpha_2 = 0. z_{it,}$

u_{it}, ε_{it} are mutually independent random variables. ε_{it} represents the error due to time effect while u_{it} represents that of the cross sectional effect, and $z_{it,}$ the bi − dimensional remainder error term .

The design of Monte-Carlo simulations was demonstrated for final year students of Statistics department, and were carried out to further examine both the effectiveness and finite sample properties of different estimators of parameter α. The cross-sectional unit used was 20 while T=10 is the time dimension used in the study. A balanced panel data was first simulated and the data was made unbalanced (and non-consecutive) by deletion of 2nd time period for some individuals. It was assumed alpha (α) is 0 and there

is no serial correlation effect, while the parameters used are uniformly distributed on the interval [0, 2]. Each of the variation was iterated 1000 times in attempt to generate big data, and the efficiency of the two estimators (i.e. GMM and Instrumental Variable) employed were considered based on some validity checks.

RESULTS

The results of both real life and simulated data are presented in the following tables:

Table 1: Results of Validity Checks on Generalized method of Moments Technique

	Real Life Data		Simulated Data	
Test	Estimates		Estimates	
Test for AR(1) errors	-1.33091	(0.1832)	-0.7451126	(0.4562)
Test for AR(2) errors	-0.47378	(0.6357)	0.1865387	(0.85202)
Sargan over-identification test	87.3077	(0.0001)	14.40015	(1.0000)
Wald (joint) test	6298.11	(0.0000)	0.1324148	(9.1e-8)***

Table 2: Results of Validity Checks on Instrumental Variable Technique

	Real Life Data		Simulated Data	
Test	Estimates		Estimates	
F- statistic	219.195	(0.0000)	823.615	(2e-16)***
Wald (joint) test	18.5464	(0.0001)	226900	(2.2e-16)***
R-Squared	0.69433		0.9997	

Note: p-values are in parenthesis

DISCUSSION

The results in table 1 and table 2 presented the validity checks on the analysis of both real life and simulated data. AR(1) and AR(2) test for first-order and second-order serial correlation in the first-differenced residuals, asymptotically distributed as N (0, 1) under the null hypothesis of no serial correlation (based on the efficient one-step GMM estimator). Since the P-values of all the AR irrespective of orders are greater than 0.05, we therefore accept the null hypothesis and conclude that there is no serial correlation among the variable used. However, comparing the P-values of the two AR for both datasets, the magnitude of difference from the significant values of 0.05 is much higher for simulated data which made the results much more acceptable.

The Sargan test is a test of the validity of instrumental variables. It is a test of the over identifying restrictions. The hypothesis being tested with the Sargan test is that the instrument are uncorrelated to some set of residuals, and therefore they are unacceptable and unhealthy for the real life data as the p value of the test is 0.0001 while that of simulated data is readily acceptable with a P-value of 1.000. **Wald test** is a way to find out if explanatory variables in a model are significant and it also provides justification for the overall goodness of fit for the estimated model. Considering the Wald test in tables 1 and 2, it was shown that the Wald tests were highly significant at 5% level for simulated data while they are only significant for real life data with 0.000 and 0.0001 P- values respectively for the two techniques employed in the analysis. Thus, the models estimated from big data provide a better goodness of fit. This inference is further validated with the results of F-statistic and coefficient of determination presented in table 2.

CONCLUSION

This article has exemplified and emphasized through empirical analysis, the affluence of big data in statistical analysis. Thus statistics application can be greatly enhanced with the use of self-generated big data, hence overcoming the paradox of data availability and its accuracy especially in developing countries of the world, where data collection and manipulation still operate like “Siamese twins”. The adopted techniques work well for both real life and simulated data and Monte Carlo simulations revealed that the two methods have very good finite sample performance. Aside from the immense benefits of big data, Self-generated data is fast becoming inevitable as compliment of real life data, due to some established facts in

statistical survey and these have improved statistical reasoning of students as opined by Garfield and Ben-Zvi (2017). Other benefits are:

- Simulated data can be obtained with minimum resources.
- When the items or units of real life data are destroyed under investigation.
- When the results are required in a short time as in the case of classroom teaching
- When available resources are limited.
- When population under consideration is either constantly changing or in a state of movement.
- When the authorizing bodies for data production are not forthcoming, simulated data can be used to complement the insufficient real life data.

REFERENCES

Adeboye, N. O. & Agunbiade, D. A. (2019a). Monte Carlo Estimation of Heteroscedasticity and Periodicity Effects in a Panel Data Regression Model. *International Journal of Mathematical and Computational Sciences, 13(6), 157-166.*

Couldry, N. & Turow, J. (2017). Media as a Data Extraction; Towards a New Map of a Transformed. Journal of Communication, 68(2), 415-423.

Garfield, J. & Ben-Zvi, D. (2007). The challenge of developing statistical reasoning. *Journal of Statistics Education. 75 (3), 372-396.*

Garfield, J. (2002). The Challenge of Developing Statistical Reasoning. *Journal of Statistics Education, 10(3), DOI:10.1080/10691898.2002.11910676.*

Garfield, J., and Gal, I. (1999). Teaching and Assessing Statistical Reasoning in Developing Mathematical Reasoning in Grades K-12, ed. L. Stiff, Reston, VA: National Council Teachers of Mathematics, 207-219.

Griffiths, H. B. & McClone, R. R. (1979). *Qualities Cultivated in Mathematics Degree Examinations.* A Research report for the Social Science Research Council.

Huberty, M. (2015). Awaiting the second big data revolution: from digital noise to value creation. *Journal of Industry, Competition and Trade, 2015. 15(1): p. 35-47. 3.*

Xinhua, E. (2013). Big Data as a Service: Definition and architecture. *15th IEEE International Conference on Communication Technology (ICCT).*

John, T. W. (2019). *The importance of statistics in management decision making;* Reviewed by Michelle Seidel, B.sc., LL.B., MBA.

Mashey, J. & Chen, H. M. (2016). Big Data as a Service: A Neo-Metropolis Model Approach for Innovation. *49th Hawaii International Conference on System Sciences (HICSS).*

Meng, X. (2018). Statistical Paradises and Paradoxes in Big Data in Big Data: Law of Large Populations, Big Data Paradox and the 2016 US Presidential Election. *Annals of Applied Statistics, 12(2), 685-726.*

Mortensen, P. S. (1988). On a New Approach to an Introductory Course in Statistics for Business and Economics including some Experiences. *In: R Davidson and J Swift (eds) Proceedings of the Second International Conference on Teaching Statistics. University of Victoria, Canada, 415-424.*

Olavsrud, T. (2016). Big Data and Analytics Spending to hit $187 billion. Worldwide Semiannual Big Data and Analytics Spending Guide by International Data Corporation. Retrieved from *https://www.cio.com* on 03/01/2020.

Singh, H. (1998). *Data Warehousing: Concepts, Technologies, Implementations and Management*. Prentice Hall, New Jersey, ISBN 0-13-591793-X, 332p. 3. Wall, L., T.

Jamborsalamati, P., Fernadez, E., Hosain, M. J., Rafi F. H. M. (2017). Design and Implementation of a Cloud-based IoT Platform for data Acquisition and Device Supply Management in Smart Buildings. 2017 Australian Universities Power Engineering Conference (AUPEC), 1-6