Seemingly Unrelated Regression (SUR) Modeling of Productions and Sales of Carbonated Soft Drinks in Nigeria

W. B. Yahya[1]; S. O. Awoyemi[2]; O. R. Olaniran[3]

[1,3]Department of Statistics,
University of Ilorin,
Ilorin, Nigeria.
e-mail: [1]wb_yahya@daad-alumni.de; [3]rid4stat@yahoo.com

[2]Agricultural and Rural Management Training Institute (ARMTI),
Department of Training Technology,
P.M.B 1343, Ilorin, Nigeria.
e-mail: [2]bodesok2012@gmail.com

***Abstract*— The patterns of sales of manufactured products are sometimes complementary. That is a product may be considered by a consumer in the absence of its preferred close substitute. This consumer behavior often affects the eventual quantity sold or consumed. Modeling simultaneously sales of complementary products usually introduces residual effects because of the contemporaneous correlation existing between the products considered. Therefore, this study aims to develop and evaluate models for predicting the sales of Coca-cola and Pepsi carbonated soft drinks in Nigeria. Models that connected the quantity of carbonated soft drinks produced with the quantity sold were fitted using the SUR, OLS, 2SLS and 3SLS estimators. The results were compared to OLS estimates and evaluated by several statistical measures. We focused only on the assessment criteria for the estimated parameters from the various methods or estimators. Our results showed that the SUR estimator performed better in predicting quantity sold from Pepsi and Coca-cola drinks simultaneously than the other least squares based estimators considered using the estimated absolute biases and mean square errors.**

Keywords- *Ordinary least squares, Seemingly unrelated regression, Two-stage least Square, Three-stage least square.*

I. INTRODUCTION

Much of scientific studies are directed towards discovering the form of relationship between variables and predicting the value of a variable from some functional relationship. Many authors have discussed a number of multivariate regression topics that cut-across model building, forecasting and variable screening methods. In most cases, the position was that a good predictor variable X should be highly correlated with the response variable Y and uncorrelated with other independent variables in regression models. This is intuitively reasonable and usually provides a set of independent variables that may lead to a satisfactory regression model. However, when two or more independent variables are highly correlated with each other, this demonstrate a condition referred to as multicollinearity and once one of the collinear variables is entered into the model, the entry of the second or other variables in the model will demonstrate non-significant results and little, if any increase in the model's R^2 [1,2].

Consider a classical multivariate regression model of the form;

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e} \qquad (1)$$

that involves p independent variables $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ and vector of k response variables $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_k)$. It is generally assumed that vector $\boldsymbol{Y}$ has a multivariate normal distribution with mean $\boldsymbol{X}\hat{\beta}$ and variance-covariance matrix $\sum$. In addition, the model's error term $\boldsymbol{e}$ is assumed to have a multivariate Gaussian density with zero mean and variance-covariance matrix $\sum$. Further assumptions on model (1) are that of homogeneous variance of residuals conditional on predictors; common covariance structure across observations; and independence of observations [3]. If all of these assumptions are met, the least square estimator will be unbiased with minimum variance.

A Seemingly Unrelated Regression (SUR) estimation technique is a method that estimates systems of a set of separate equations that are only correlated through their disturbances. In the literature, such system of equations

that employs SUR technique for its estimation is often referred to as SUR models.

The SUR models have found applications in many real life cases and some of which have been reported in the literature [1]. For example, demand functions can be estimated for different households (or household types) for a given commodity. The correlation among the equation disturbances could come from several sources such as correlated shocks to household income. Alternatively, one could model the demand of a household for different commodities, but adding-up constraints leads to restrictions on the parameters of different equations in this case.

On the other hand, equations explaining some phenomenon in different cities, states, countries, firms or industries provide a natural application as these various entities are likely to be subjected to spillovers from economy-wide or worldwide shocks.

There are two main motivations for the use of SUR model. The first one is to gain efficiency in estimation by combining information on different equations. The second motivation is to impose and/or test restrictions that involve parameters in different equations. Zellner [4] provided a seminar work in this area, and a thorough treatment is available in the books by Srivastava and Giles [5] while additional works on SUR can be found in [6-8].

In this study, the application of the SUR model for predicting the quantity of carbonated drinks sold by two companies (Coca-cola and Pepsi) as determined by their respective quantity produced is demonstrated. The assumption here is that the sales of Coca-cola products at a particular period might influence the sales of Pepsi products, though such influence might not be directly apparent and straightforward. This kind of relationship is referred to as contemporaneous relationship (correlation) since they seem to be unrelated but they are actually related

II. Materials and Methods

A. *The Seemingly Unrelated Regression Estimator*

Consider a complete system of m regression equations of the form;

$$\begin{aligned} y_1 &= X_1\beta_1 + \varepsilon_1 \\ &\vdots \\ y_m &= X_m\beta_m + \varepsilon_m \end{aligned} \qquad (2)$$

where for i = 1, 2, …, m, $\boldsymbol{y_i}$ is an $n \times 1$ vector of observations on the i^{th} response variable, X_i is an $n \times p_i$ matrix of independent variables, β_i is a $p_i \times 1$ vector of regression parameters and ε_i is an $n \times 1$ vector of error for the i^{th} regression with $\varepsilon_i \sim N(0, \sigma_i^2)$. The whole system of m regression models in (2) when stacked together becomes;

$$Y = X\beta + \varepsilon \qquad (3)$$

In (3), $\boldsymbol{Y'} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y}_m)$ is a $mn \times 1$ vector of responses,

$$X = \begin{pmatrix} X_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & X_m \end{pmatrix}$$

is a $mn \times \sum_{i=1}^{m} p_i$ matrix of observations on independent variables, $\beta' = (\beta_1, \beta_2, \ldots, \beta_m)$ is a $\sum_{i=1}^{m} p_i \times 1$ vector of regression parameters while $\varepsilon' = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_m})$ is a $mn \times 1$ measurements on the error terms.

Obviously, all the m regression equations in (2) or (3) appear seemingly unrelated because they, often times, contain different independent variables and parameters. Whereas, these system of equations are actually (contemporaneously) correlated through their error terms $\boldsymbol{\varepsilon}' = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_m})$ with the condition that;

$$E(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}') = \Sigma \otimes I_n \qquad (4)$$

where Σ is an $m \times m$ variance-covariance matrix of the form;

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1m}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2m}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}^2 & \sigma_{m2}^2 & \cdots & \sigma_{mm}^2 \end{pmatrix}$$

The quantity I_n is a $n \times n$ identity matrix and $\otimes$ is a Kronecker product that ensures that each element in Σ is multiplied by I_n. The parameters of the above system of equation (2) or (3) are estimated using the techniques of the Generalized Least Squares (GLS) as;

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \qquad (5)$$

where $\Omega^{-1} = \Sigma^{-1} \otimes I_n$.

In the standard SUR estimations, the vector of disturbances $\boldsymbol{\varepsilon}' = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_m})$ in the system of equation (2) are assumed to be significantly correlated [4]. In particular, Zellner[4] was of the opinion that the contemporaneous correlation should not be less than 0.3 for SUR estimator to be more efficient. If these conditions are present, then the estimator in (2) becomes the SUR estimator given by;

$$\hat{\beta}_{SUR} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y \qquad (6)$$

where $\hat{\Omega}$ is determined from the data.

B. *The Two-Stage Least Square Estimator*

If the regressors of one or more equations are correlated with the disturbances, the Ordinary Least Squares (OLS) and SUR estimates might be biased. This can be circumvented by a two-stage least squares (2SLS) or a three-stage least squares (3SLS) estimation with the use of instrumental variables (IV). The instrumental variables for each equation Z_i can be either different or identical for all equations. They must not be correlated with the disturbance terms of the corresponding equation.

At the first stage, new ("fitted") regressors are obtained by computing;

$$\widehat{X}_i = Z_i(Z_i'Z_i)^{-1}Z_i'X_i \qquad (7)$$

Then, these fitted regressors are substituted for the original regressors in the equation (2) to obtain the 2SLS or 3SLS estimates of β as:

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{\Omega}^{-1}\hat{X})^{-1}\hat{X}'\hat{\Omega}^{-1}Y \qquad (8)$$

C. *The Ordinary Least Squares Estimator*

The OLS estimator of the stacked equation (2) or (3) above is given by;

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \qquad (9)$$

The performances of the three estimators were assessed based on the average Absolute Bias (AB) and average Root Mean Square Errors (RMSE) of parameter estimates.

III. Analysis

Simulation Scheme

The simulation scheme used in this study was adapted from Yahya et. al [1] which considers system of equations containing four distinct linear regression equations here representing quantity produced and quantity sold for Pepsi and Coca-cola carbonated soft drinks. If we let the four equations be distributed as follows:

$$Y_1|X_1 \sim N(X_1\beta_1, \sigma_{11}^2);$$

$$Y_2|X_2 \sim N(X_2\beta_2, \sigma_{22}^2);$$

$$Y_3|X_3 \sim N(X_3\beta_3, \sigma_{33}^2);$$

$$Y_4|X_4 \sim N(X_4\beta_4, \sigma_{44}^2).$$

Thus, with $m = 4$, we have the following system of regression models;

$$y_1 = X_1'\beta_1 + \varepsilon_1$$
$$y_2 = X_2'\beta_2 + \varepsilon_2$$
$$y_3 = X_3'\beta_3 + \varepsilon_3$$
$$y_4 = X_4'\beta_4 + \varepsilon_4$$

where y_1 and y_2 represent the quantity sold of Pepsi products during the raining and dry seasons while y_3 and y_4 represent the quantity sold of Coca-cola products during the raining and dry seasons respectively.

The whole system is assumed to be distributed as;

$$(Y|X) \sim N_4(X\beta, \Sigma \otimes I_n)$$

The covariates and the parameters are therefore defined as follow;

$$X_1' = (X_{10}, X_{11});\ \beta_1' = (\beta_{10}, \beta_{11})$$
$$X_2' = (X_{20}, X_{22});\ \beta_2' = (\beta_{20}, \beta_{22})$$
$$X_3' = (X_{30}, X_{33});\ \beta_3' = (\beta_{30}, \beta_{33})$$
$$X_4' = (X_{40}, X_{44});\ \beta_4' = (\beta_{40}, \beta_{44})$$

In the above representations, X_{11} and X_{22} represent the factory production level of Pepsi products during the raining and dry seasons while , X_{33} and X_{44} represent the factory production level of Coca-cola products during the raining and dry seasons respectively.

The true values of the parameters of the four models as used in the Monte-Carlo studies here and later reported in Tables 1 and 2 were determined from the OLS fitted to the real life data that were collected from the two companies that are producing the two carbonated drinks in Nigeria. Thus, the estimated parameters based on equation-by-equation fit of the least squares model to the real life data were used as the true values of the parameters to simulate dataset for Monte-Carlo study carried out here.

The entire model was, however simulated and estimated using the three estimators considered (SUR, OLS, 2SLS) here at various sample sizes (n); n = 20, 50, 100, 200, 500, 1000 over 1000 replicates in each case. Results of the 3SLS were essentially similar to those of 2SLS, hence, they were dropped from our discussion in this study, except in few cases.

IV. Results

The results of the three estimators at the selected various sample sizes are presented in Tables 1 and 2.

Table 1: Simulation Result For OLS, SUR and 2SLS at Small Sample Sizes $n = 20, 50$. NB: (*) Indicates the true values of the models' parameter as determined from the OLS fitted to the real life data.

Sample Size	Parameters	*TRUE VALUES	OLS	SUR	2SLS
$n = 20$	β_{01}	-16645.4	-16849.3	-18127	-17095.2
	β_{11}	0.98	0.98	0.98	0.98
	β_{02}	74949.08	80429.42	80152.91	80429.42
	β_{12}	0.76	0.75	0.75	0.75
	β_{03}	85507.84	87642.45	86796.87	84874.79
	β_{13}	0.68	0.68	0.68	0.68
	β_{04}	-16310.7	-18451.3	-18608.6	-37753.2
	β_{14}	1.00	1.00	1.00	1.04
$n = 50$	β_{01}	-16645.4	-18287.1	-16496.7	-17444.8
	β_{11}	0.98	0.98	0.98	0.98
	β_{02}	74949.08	74166.02	76520.01	74166.02
	β_{12}	0.76	0.76	0.76	0.76
	β_{03}	85507.84	82451	84986.73	81695.35
	β_{13}	0.68	0.69	0.68	0.69
	β_{04}	-16310.7	-19438	-19088.1	-39260.5
	β_{14}	1.00	1.01	1.00	1.04

Table 2: Simulation Result For OLS, SUR and 2SLS large Sample Sizes $n = 500, 1000$. NB: (*) Indicates the true values of the models' parameter as determined from the OLS fitted to the real life data.

Sample Size	Parameters	*TRUE VALUES	OLS	SUR	2SLS
$n = 500$	β_{01}	-16645.4	-17088.2	-17028.3	-17457.6
	β_{11}	0.98	0.98	0.98	0.98
	β_{02}	74949.08	73989.01	74456.45	73989.01
	β_{12}	0.76	0.76	0.76	0.76
	β_{03}	85507.84	85462.63	85243	84563.95
	β_{13}	0.68	0.68	0.68	0.68
	β_{04}	-16310.7	-14373.8	-15894.6	-16919.4
	β_{14}	1.00	1.00	1.00	1.00
$n = 1000$	β_{01}	-16645.4	-16731.4	-16617.6	-16708.6
	β_{11}	0.98	0.98	0.98	0.98
	β_{02}	74949.08	75021.57	75074.74	75021.57
	β_{12}	0.76	0.76	0.76	0.76
	β_{03}	85507.84	85302	85381.61	85394.58
	β_{13}	0.68	0.68	0.68	0.68
	β_{04}	-16310.7	-16172.3	-16147.6	-16156
	β_{14}	1.00	1.00	1.00	1.00

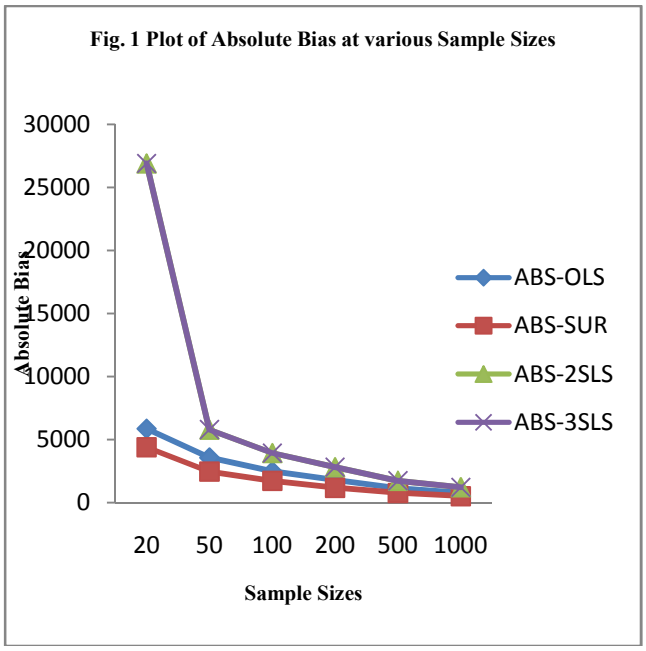

Fig 1: The plots of Absolute Bias at various sample sizes

Table 3: Average Mean Squared Error of The Estimators at Various Sample Sizes

Sample Size	MSE-OLS	MSE-SUR	MSE-2SLS	MSE-3SLS
20	1.57E+10	7.19E+09	1.85E+12	1.85E+12
50	5.59E+09	2.46E+09	4.96E+10	4.96E+10
100	3.12E+09	1.24E+09	1.09E+10	1.09E+10
200	1.39E+09	5.44E+08	4.67E+09	4.67E+09
500	5.38E+08	2.2E+08	1.72E+09	1.72E+09
1000	2.5E+08	96998208	8.35E+08	8.35E+08

V. Discussion

In this work the performances of SUR estimators were compared with other three (OLS, 2SLS) estimators for modelling a system of simultaneous equation were examined. Dataset on quantity produced and sole of Coca-cola and Pepsi carbonated soft drinks were employed to demonstrate different behaviours of the three estimators considered via Monte Carlo experiment.

Various results obtained showed the supremacy of the SUR estimator over others at various sample sizes considered. In all cases, the SUR estimator has the least values of RMSE and AB at all the sample sizes considered.

In Tables 1 and 2, the estimated parameter values of the models as yielded by the three estimators and the true parameter values are provided for small and large samples respectively. The closeness of the estimated parameter values to their true values as provided by the three estimators are provided by their Mean Square Errors (MSEs) as reported in Table 3. The plots of the average Absolute Bias and RMSE of the three estimators at the various sample sizes considered are provided by Figs 1 and 2 where it can be clearly observed that the SUR estimator is most efficient among the estimators considered. However, the graphs showing the performances (Absolute Bias & RMSE) of the 3SLS were equally reported in the two plots. As earlier remarked, its performances were essentially similar to those of the 2SLS as can be observed from Figs 1 and 2.

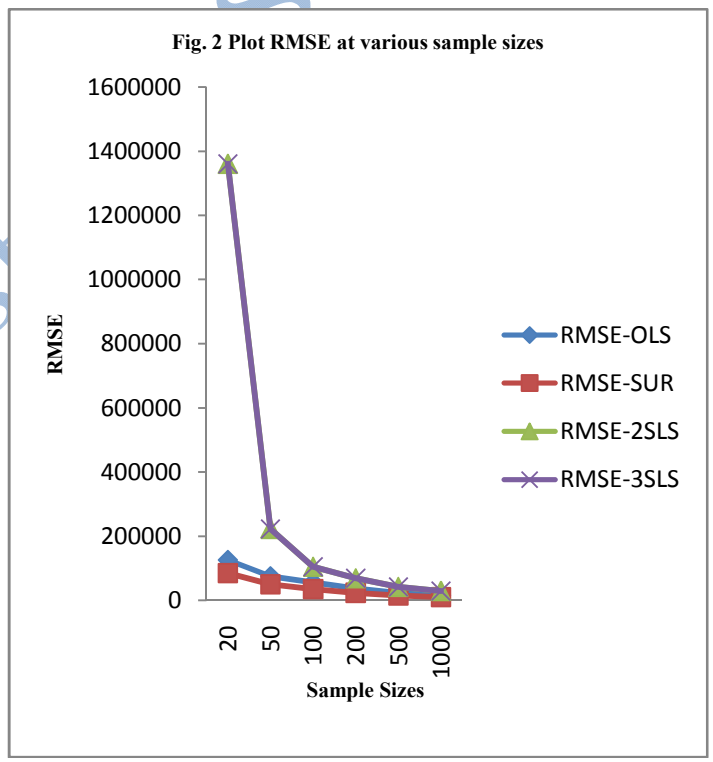

Fig 2: The plots of RMSE at various sample sizes

VI. Conclusion

This study compared the efficiency of three estimators (OLS, SUR and 2SLS) for modeling a system of regression models via Monte-Carlo studies. The results showed that the SUR estimator was most efficient among the three. This good performance of SUR might not be unconnected to the evidence of high contemporaneous relationship between the models' error terms.

Interestingly, the OLS estimator clearly followed the SUR estimator in term of performance with the 2SLS being the worst among the three. However, the good performance of OLS over the 2SLS could be attributed to the fact that each of the models contains only one independent variable. Hence, there is no effect of multicollinearity in the models which naturally enhance the efficiency of OLS estimator.

REFERENCES

[1] Yahya WB, Adebayo SB, Jolayemi ET, Oyejola BA, Sanni OOM (2008). Effects of non- orthogonality on the efficiency of seemingly unrelated regression (SUR) models.*InterStatJournal*,1-29. URL: http://interstat.statjournals.net/.

[2] Yahya WB and Olaifa JB (2014): A note on Ridge Regression Modeling Techniques. *Electronic Journal of Applied Statistical Analysis,* 7(2):343-361.

[3] Blattberg, R.T, and George, E. I., (1991), Shrinkage estimation of price and optional elasticities: Seemingly Unrelated Equations. *Journal of American Statistical Association*. 86: 304-315.

[4] Zellner A. (1962): An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests of Aggregation Bias, Journal of the American Statistical Association, 57: 500-509.

[5] Srivastava, V. K. and Giles D. E. A. (1987): Seemingly Unrelated Regression Equations Models, New York: Marcel Dekker Inc.

[6] Fiebig, D. G. (2001): Seemingly Unrelated Regression, in Baltagi, B. eds, A Companion toTheoretical Econometrics, *Backwell Publishers*, 0101-121.

[7] Timm, N. H. (2000), Applied Multivariate Analysis, Oxford University Press, 120-143.

[8] Zellner A and Theil H (1962). Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica*, 30(1): 54-78.