MODELLING OF GROSS DOMESTIC PRODUCT OF SOME SECTORS OF NIGERIA ECONOMY IN THE PRESENCE OF AUTOCORRELATION

Osuolale Peter Popoola, Adekunle A. Araromi, Adesina A. Rafiu, Mattew T. Odusina

Maths and Statistics Department, The Iparapa Polytechnic, Eruwa Oyo State, Nigeria

Corresponding author: O. P. Popoola, osuolalepeter@yahoo.com

ABSTRACT: The research work examined the statistical modelling of Nigeria GDP of some selected sectors in the presence of autocorrelation. It examines the effects and contribution of some economic sectors to the Gross Domestic Production of Nigeria.
The data set in (N million) covered a period of 20 years from 1990 to 2009 for the economic variables of interest. The statistical methods employed were regression analysis, correlation and residual analysis.
The estimated regression equation is given as: Y = -77203 + 6.43816X_1 + 1.26391X_2 + 6.99025X_3 + 1.01914X4. The Coefficient of Determination (R^2) of 0.9806 showed that the four economic variables considered explained about 98% of the variation in the nations GDP.
A further analysis revealed the Variance Inflation Factor with the independent variables been highly correlated which denotes the presence of multicollinearity. Durbin-Watson and Goldfeld-Quandt tests revealed that there was no autocorrelation in the error terms but there was evidence of heteroscedasticity respectively.
It was found that Agriculture sector contribute the highest to the growth of the economy, follow by the manufacturing and oil sector respectively, while the least contribution was experienced from building and construction. The work therefore advice the government to invest more on the agriculture sector and non oil sector in order to increase GDP.
KEYWORDS: GDP, Durbin-Watson, Multiple.

INTRODUCTION

Autocorrelation is the existence of serial correlation among the error term ([GS07]). It is also a problem that is usually associated with time series data, but can also affect cross-sectional data. Autocorrelation can also be defined as correlation between members of observations ordered in time (as in time series data) or space (as in cross-sectional data). For example, a shock to oil prices will simultaneously affect all countries, so one could expect contemporaneous correlation of macroeconomic variables across countries.
When the assumption of lack of independence of error terms is violated, we have autocorrelation problem. If the error terms are correlated in a spherical order, then we also have autocorrelation problem as well.
Autocorrelation of the error terms may occur for several reasons which might be as a result of the following:

i. Omitted explanatory variable
ii. Misspecification of the mathematical form of the model
iii. Interpolation in the statistical observations.
iv. Misspecification of the true random errors (Johnson, 1984).

Classical Linear Regression Model (CLRM) is an approach to modelling the relationship between a dependent variable *y* and one or more independent variables denoted by *X (s)*. In linear regression, models of the unknown parameters are estimated from the data using linear functions. Such models are called a linear model. Most commonly, linear regression refers to a model in which the conditional mean of *y* given the value of *X* is an affine function of *X*. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressor x_i is approximately space linear. This approximate relationship is modelled through a so-called "disturbance term" ε_i an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Where x_i' denotes the transpose, so that $x_i'\ \beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in vector form as

$$y = X\beta + \varepsilon,$$

Where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

In this research, attempt was made to build a GDP model based on building and construction sector, agriculture sector, manufacturing sector and crude oil sector in other to know their contribution to the growth of the nation. The model so designed is given as follow:

GDP = f (BC, A, M, C) + ε Where, GDP = Gross Domestic Product, BC = Building and Construction Sector,A = Agriculture Sector, M = Manufacturing Sector, C = Crude oil Sector, and ε = Error term

Gross Domestic Product (**GDP**): Refers to the market value of all officially recognized final goods and services produced within a country in a given period. GDP per capita is often considered an indicator of a country's standard of living.

The earliest form of regression was the method of least squares (French: *méthode des moindrescarrés*), which was published by Legendre in 1805, and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations; the orbits of bodies about the Sun. Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem.

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean). For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

1. THE GENERALIZED LEAST SQUARES MODEL

Generalized Least Squares (GLS) are technique for estimating the unknown parameter in a Linear Regression Model. The GLS is applied when the variance of the observations are unequal (heteroscedasticity).

In a typical linear regression model we observe data $\{y_i, x_i\}_{i=1}^{n}$ on n statistical units. The response values are placed in a vector $Y = (y_1, ..., y_n)^n$, and the predictor values are placed in the design matrix $X = [[x_{ij}]]$, where x_{ij} is the value of the j^{th} predictor variable for the i^{th} unit. The model assumes that the conditional mean of Y given X is a linear function of X, whereas the conditional variance of Y given X is a *known* matrix Ω. This is usually written as

$$Y = X\beta + \varepsilon \qquad (1.1)$$

$$E[\varepsilon / X] = 0 \qquad (1.2)$$

$$Var[\varepsilon / X] = \Omega. \qquad (1.3)$$

Here β is a vector of unknown "regression coefficients" that must be estimated from the data. Suppose b is a candidate estimate for β. Then the residual vector for b will be $Y - Xb$. Generalized Least Squares method estimates β by minimizing the squared Mahalanobis length of this residual vector:

$$\hat{\beta} = \underset{b}{\arg\min}\,(Y - Xb)'\,\Omega^{-1}(Y - Xb), \qquad (1.4)$$

Since the objective is a quadratic form in b, the estimator has an explicit formula:

$$\hat{\beta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y. \qquad (1.5)$$

The GLS estimator is unbiased, consistent, efficient, and asymptotically normal:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0,(X'\ \Omega^{-1} X)^{-1}). \qquad (1.6)$$

GLS is equivalent to applying Ordinary Least Squares (OLS) to a linearly transformed version of the data. To see this, factor $\Omega = BB'$, for instance using the Cholesky decomposition. Then if we multiply both sides of the equation $Y = X\beta + \varepsilon$ by B^{-1}, we get an equivalent linear model $Y^* = X^*\beta + \varepsilon^*$, where $Y^* = B^{-1}Y$, $X^* = B^{-1}X$, and $\varepsilon^* = B^{-1}\varepsilon$. In this model $\mathrm{Var}[\varepsilon^*] = B^{-1}\Omega B^{-1} = I$. Thus we can efficiently estimate β by applying OLS to the transformed data, which requires minimizing

$$(Y^* - X^*b)'\,(Y^* - X^*b) = (Y - Xb)'\,\Omega^{-1}(Y - Xb) \qquad (1.7)$$

This has the effect of standardizing the scale of the errors and "de-correlating" them. Since OLS is applied to data with homoscedasticity errors, the Gauss–Markov theorem applies, and therefore the

GLS estimate is the Best Linear Unbiased Estimator (BLUE) for β.

2. INTERMEDIATE SECTION

2.1.1 *Correlation Analysis*

This is used to explain the percentage of the variation of Y's which can be explained or is due to the relationship with X. correlation coefficient (r) measures the strength of linear relationship. When r is 0, it implies that there is no linear relationship between X and Y but when r = 1, there is a positive relationship, when r = -1, there is a negative relationship. When a correlation coefficient is calculated from the sample the value is denoted by r which is an estimate of the corresponding parameter called the population correlation ρ (rho). A statistical relation is said to be linear if the point of the scattered plot tends to cluster about a straight line.

$$R_{xy} = \frac{\sum(X-X)(Y-Y)}{\sqrt{\sum(X-X)^2\sum(Y-Y)^2}}$$

2.1.2 *Correlation Ratio*

Correlation ratio is the percentage by which uncertainty (as reflected in variance) about an individual Y-score is reduced by specifying the individual X-score which is denoted by eta. This is defined as follows.

$$\frac{S_Y^2 - S_{y/X=x}^2}{S_Y^2}$$

If the relationship between X and Y is not linear then $r^2 <$ correlation ratio.

2.1.3 *Regression*

Regression analysis is the study of the relationship of dependent variable says Y on the basis of a known measurement of an independent says X.

2.1.4 *Multiple regression*

Given two independent variables X_1 and X_2, we want to fit the equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Where:
β_1 measures the average or expected change in Y when X_1 increased by 1 unit and X_2 remains constant, which is called the partial regression coefficient of Y on X_1.
β_2Measures the average or expected change in Y when X_2 increased by 1 unit and X_1 remains constant, which is called the partial regression coefficient of Y on X_2.

$$\beta_0 = y - \beta_1 X_1 + \beta_2 X_2$$

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Using matrix approach, we have the model

$$Y = \beta X$$

Where:
Y is (n x1) and X= (n x n)
$\beta = (x'x)^{-1}x'y$

2.1.5 *The coefficient of determination*

In the simple linear regression model, the coefficient of determination is identically the square of the simple linear correlation coefficient. Hence, the coefficient of simple determination is conventionally represented as R^2.

$$R^2 = \frac{\sum(\hat{Y}-Y)^2}{\sum(Y-Y)^2}$$

$$R^2 = \frac{\text{Sum of square regresion}}{\text{sum of square total}}$$

2.1.6 *Autocorrelation*

If the classical assumption of the error terms is violated, the problem of serial or autocorrelation arises. It is usually associated with time series data. It can arise as a result of slow economic time series, specification bias resulting from exclusion of important forms, presence of random errors caused by sporadic events etc.

Detection of autocorrelation

- Graphical method: The residual εi are plotted against time n-plot. ε_i against ε_{i-1}. A sample visual examination of OLS residual can give insight about the likely presence of autocorrelation among the error terms.
- Statistical test: the most popular is the Durbin Watson d test

The test statistics d is given an

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

$$D = \frac{2\sum e_t^2 - 2e_t e_{t-1}}{\sum e_t^2}$$

D = 2(1- ρ)

Where $\rho = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$

When d= 0 and ρ=1, there is a perfect positive serial autocorrelation

D= 2 and ρ=0, there is no serial autocorrelation

D= 4 and ρ = -1, there is a negative serial autocorrelation

If 0<ρ<1 or 0<d<2, there is a level of positive autocorrelation

However, if -1<ρ<0 or 2<d<4, there is a level of negative autocorrelation

Remedial measure of Autocorrelation

When ρ is known, we can use the Prais-Winstein transformation to run a generalized or quasi-difference equation. When ρ is not known, we can use the Cochrane-Orcutt iterative procedure to estimate ρ.

3. RESULTS AND DISCUSSION

The output obtain from the SAS analysis, shows that the constant terms of the estimated regression equation gives a negative value of -77203 for the regressed model. The slope of the estimate that represent the parameter of the estimate gives positive values for building and construction (X_1), agriculture (X_2), manufacturing (X_3) and crude oil (X_4) 6.44, 1.26, 6.99 and 1.019 respectively. This implies that increase in the variables will lead to proportionate increase of gross domestic product, a proxy for economic growth. The estimated regression equation is given as:

$\hat{Y}$= -77203 + 6.43816X_1 +1.26391X_2 + 6.99025X_3 + 1.01914X_4

From the model, the results indicate that for every unit (1 million naira) increase in the building and construction, agriculture, manufacturing and crude oil lead to corresponding increase of approximately N6.38016 million, N1.26487 million, N6.99023 million and N1.01799 million respectively in the GDP of the nation. The t-statistics for the analysis shows that there is significant relationship between the contribution of agriculture and gross domestic product. R is the correlation coefficient of the multiple correlations which is equal to 0.99, implies a strong positive between the GDP and all the regressors. R^2 is the coefficient of determination and it is equal to 0.98, which implies that 98 percent of the variation or change in GDP is explained by the explanatory variables (building and construction, agriculture, manufacturing and crude oil). R squared adjusted is used to test the adequacy of the model is also 98 percent which explains that the model is very adequate and statistically significant. This suggests that the multiple regression models are useful in this analysis. Based on standardized coefficient from the results of this research work, it shows clearly that agriculture sector has the greatest contribution to GDP. Durbin Watson statistic, test for first order autocorrelation on the disturbance or error term confirm that there is no autocorrelation among the error term. Hence, there is no dependency or serial correlation among successive value of the error term.

Furthermore, agricultural sector contributes the highest to the growth of the nation's economy, followed by the manufacturing sector and crude oil sector respectively and least contribution came from building and construction. The coefficient of determination R square is 0.98 is found to be statistically significant and hence implies that the explanatory variables explained most of the variation in the GDP. Hence the linear model is appropriate or fits. Also the result from t-test revealed that some of the explanatory variables are significantly different from zero. The correlation coefficient reveals a strong positive relationship among the variables. The variance of error term was tested and it was discovered there are variations. Hence there is presence of heteroscedasticity in the error terms. There are traces of multicollinearity among the explanatory variables. Goodness of fit test (ANOVA) was carried out for testing the adequacy of the model used for forecasting. It was discovered that the economy is improving by the year for the year under consideration because the contribution of each sector to the GDP keep increasing by the year.

REFERENCES

[CO49] **D. Cochrane, G. H. Orcutt** - *Application of Least Square Regression to Relationship Containing Auto-correlation Error term*. Journal of American Statistical Association, 44, 32-61, 1949.

[DW71] **J. Durbin, G. J. Watson** - *Test for Serial Correlation in the Least Squares Regression III*, Biometrika, 58, 1- 42, 1971.

[GS07] **D. Gujarati, N. Sangeetha** - *Basic Econometrics, Fourth Edition*, Mcraw-Hill, New York, 2007.

[Har71] **W. M. Harper** – *Statistics, Second edition*, MacDonald and Evans Limited, 1971.

[HL60] **C. Hildreth, J. Y. Lu** - *Demand Relationship with Autocorrelated Disturbance*. Michigan State University. Agricultural Experiment Statistical Bulletin 276, East Lansing, Michigan, 1960.

[Joh72] **J. Johnston** - *Econometric methods, Second Edition*, New York, McGraw-Hill book, 1972.

[RG69] **P. Rao, Z. Grilches** - *Small Sample Properties of Several Two-Stage Regression Methods in the Context of Autocorrelation Errors*. J. A SA, 64, 251-272, 1969.