Dynamic Time Series Regression: A Panacea for Spurious Correlations

Emmanuel Alphonsus Akpan[*], **Imoh Udo Moffat**[**]

[*]Department of Mathematics and Statistics, University of Uyo, Nigeria
[**]Department of Mathematics and Statistics, University of Uyo, Nigeria

Abstract- The study examined that the linear relationship between Gross Domestic Product (Y_t) and Money Supply (X_t) from 1981 to 2014 is spurious and could be avoided by dynamic regression modeling. The fact that spurious regression always results in misleading correlations between two time series was a big motivation for undertaking this study. Therefore, exploring data from the Central Bank of Nigeria Statistical Bulletin, we found that the linear relationship between the dependent variable (Y_t) and the independent variable (X_t) seemed spurious as the errors of the regression model were found to be autocorrelated. In a bid to correct this problem of spurious regression, we identified lags 0, -1 and -2 of X_t as predictors of Y_t using cross correlation function. Hence, the dynamic regression of current lag and past lags 1, 2 of X_t as predictors of Y_t revealed that the errors are uncorrelated and the coefficient of determination is as low as 0.2086, indicating that Y_t and X_t are totally unrelated.

Index Terms- autocorrelated errors; cross correlation function; dynamic regression; prewhitening; spurious regression.

I. Introduction

It is always common to see in literature that most regression models are reported with very high coefficient of determination, R^2 which indicates that the model is a good fit without considering the problem posed by autocorrelated errors (Granger and Newbold, 1974). Though in classical regression model, the error terms are assumed to be a white noise, that is, a sequence of independent and identically distributed random variables. In practice, however, the error terms always appear to be autocorrelated (Wei, 2006; Box, Jenkins and Reinsel, 2008). But if the autocorrelated error terms are ignored, the regression becomes spurious resulting in misleading correlations, a situation where a significant regression can be achieved for totally unrelated series (Pankratz, 1991; Wei, 2006; Cryer and Chan, 2008). The problem associated with spurious regressions can be avoided by including lagged values of time independent variables in the regression model (Pankratz, 1991; Brockwell and Davis, 2002; Fuller, 1996; Wei, 2006; and, Box, Jenkins and Reinsel, 2008). Different authors use different names to refer to regression models in which the current value of time dependent variable is a function of current and lagged values of time independent variables. For instance, Fuller (1996) referred to such regression models as transfer function models while Pankratz, (1991) called it dynamic regression models. However, prior studies revealed that differencing both the time dependent variable and time independent variables is one approach that spurious regressions could be avoided but failed to take into consideration lagged values of time independent variables, thereby, creating a gap in knowledge that the dependent variable may be related to independent variables with time lags which often results in loss of useful information about the roles of lagged values in explaining movements in the dependent variable. Thus, this paper contributes towards filling the gap by analyzing the relationship between Gross Domestic Product (Y_t) – dependent variable and Money Supply (X_t) – independent variable.

II. METHODOLOGY

Regression Model

Rawlings, Pantula and Dickey (1998) defined a standard regression model as

$$Y_t = \beta_0 + \beta_{1,}X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \varepsilon_t \quad (2.1)$$

where Y_t = dependent variable
β_i = regression parameters, i = 1,…, k
X_{it} = independent variables, i = 1,…, k
ε_t = error term assumed to be i.i.d. N(0, σ_t^2)

Thus, the dependent variable for a time series regression model with independent variables is a linear combination of independent variables measured in the same time frame as the dependent variable. Estimates of the parameters of the model in (2.1) can be obtained by Least Squares Estimation Method (*see for example* Drasper and Smith, 1998; Rawlings, Pantula and Dickey, 1998).

Dynamic Regression Model

Dynamic regression model as specified by Pankratz (1991) is as follows:

$$Y_t = \beta + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_k X_{t-k} + \varepsilon_t \quad (2.2)$$

The intuition is that equation (2.2) is built to take into account useful information about the roles of time (past) lag ($X_{t-1}, \dots, X_{t-k}$) in explaining the movements in Y_t, which is not possible with equation (2.1). The parameters of dynamic regression models are estimated using maximum likelihood method, see Pankratz (1991) for details.

Autoregressive Moving Average (ARMA) Processes

A natural extension of pure autoregressive and pure moving average processes is the mixed autoregressive moving average

$(ARMA)$ processes, which includes the autoregressive and moving average as special cases (Wei, 2006).

A stochastic process $\{X_t\}$ is an $ARMA(p, q)$ process if $\{X_t\}$ is stationary and if for every t,

$$\varphi(B)X_t = \theta(B)\varepsilon_t \quad (2.3)$$

$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$ is the autoregressive coefficient polynomial.

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the moving average coefficient polynomial.

Box, Jenkins and Reinsel (2008) considered the extension of ARMA model in (2.3) to deal with homogenous non-stationary time series in which X_t, is non-stationary but its d^{th} difference is a stationary ARMA model. Denoting the d^{th} difference of X_t by

$$\varphi(B) = \phi(B)\nabla^d X_t = \theta(B)\varepsilon_t \quad (2.4)$$

where $\varphi(B)$ is the nonstationary autoregressive operator such that d of the roots of $\varphi(B) = 0$ are unity and the remainder lie outside the unit circle while $\phi(B)$ is a stationary autoregressive operator.

Thus, (2.4) is called an autoregressive integrated moving average model and can be referred to as an $ARIMA(p, d, q)$ model.

Prewhitening and Cross Correlation Function (CCF)

According to Wei (2002), assuming that independent variable, X_t follows an ARMA(p,q) process,

$$\varphi_x(B)\, X_t = \theta_x(B)\alpha_t \quad (2.5)$$

where α_t is white noise. The series

$$\alpha_t = \theta_x(B)^{-1}\varphi_x(B)\, X_t \quad (2.6)$$

is called the prewhitened series. Applying the same prewhitening transformation to the dependent variable, we obtain a filtered independent series,

$$\beta_t = \theta_x(B)^{-1}\varphi_x(B)\, Y_t \quad (2.7)$$

Let Y = $\{Y_t\}$ be time dependent variable, X = $\{X_t\}$ be time independent variable, and the cross covariance function $\gamma_{t,s}(X, Y) = Cov(X_t, Y_s)$ for each pair of integers t and s. The cross correlation between X and Y at lag k can be defined by $\rho_k(X, Y) = Corr(X_t, Y_{t-k}) = \frac{\gamma_{XY}(k)}{\sqrt{\gamma_X(0)\gamma_Y(0)}}$. In general, the cross correlation function is not an even function since $Corr(X_t, Y_{t-k})$ need not equal $Corr(X_t, Y_{t+k})$. Moreover, the sample cross correlation function (CCF) is useful for identifying lags of independent variable that might be useful predictors of dependent variable (Cryer and Chan, 2008). However, the CCF can be obtained by prewhitening by considering a more general regression model relating X to Y,

$$Y_t = \sum_{-\infty}^{\infty} \beta_k X_{t-k} + \varepsilon_t \quad (2.8)$$

where X is independent of ε. Applying the filter $\pi(B)$ to both sides of (2.8), we have

$$\tilde{Y}_t = \sum_{-\infty}^{\infty} \beta_k \tilde{X}_{t-k} + \tilde{\varepsilon}_t \quad (2.9)$$

where $\tilde{\varepsilon}_t = \varepsilon_t - \pi_1\varepsilon_{t-1} - \pi_2\varepsilon_{t-2} - \cdots$

The prewhitening procedure thus orthogonalizes the various lags of X in the original regression model (Cryer and Chan, 2008).

Model Selection Criteria

For a given data set, when there are multiple adequate models, the selection criterion is normally based on summary statistics from residuals of a fitted model (Wei, 2006).

There are several model selection criteria based on residuals (see Wei, 2006). For the purpose of this study, we consider the well-known Akaike's information criterion (AIC), (Akaike, 1973) defined as

$$\text{AIC} = -2\, ln(\text{likelihood}) + 2(\text{number of parameters})$$

where the likelihood function is evaluated at the maximum likelihood estimates. The optimal order of the model is chosen by the value of the number of parameters, so that AIC is minimum (Wei, 2006).

Model Diagnostic Checking

Box and Pierce (1970) proposed the Portmanteau statistics:

$$Q^*(\text{m}) = \text{T} \sum_{l=1}^{m} \hat{\rho}_l^2 \quad (2.10)$$

where T is the number of observations.

Ljung and Box (1978) modify the Q^*(m) statistic to increase the power of the test in finite samples as follows:

$$\text{Q(m)} = \text{T(T} + 2) \sum_{l=1}^{m} \frac{\hat{\rho}_l^2}{T-l} \quad (2.11)$$

where T is the number of observations.

The decision rule is to reject H_0 if Q(m) > χ_α^2, where χ_α^2 denotes the 100 (1 – α)th percentile of a Chi-squared distribution with m – (p + q) degree of freedom (*see for example* Akpan, Moffat and Ekpo, 2016).

III. DATA ANALYSIS AND DISCUSSION

This study considers the Gross Domestic Product (N' Billion) as the dependent variable and the Money Supply (N' Billion) as the independent variable. The data were obtained from the Central Bank of Nigeria Statistical Bulletin for a period spanning from 1981 to 2014. Each series consists of 34 observations.

First, we regress Y_t on X_t, and obtain the estimated regression model presented in equation (3.1) below:

$Y_t =$ 56.0350 + 4.0717 X_t
s.e (294.2920) (0.5383)
t-value (0.190) (7.564)
(3.1)
p-value (0.85) (1.98e-09)
$R^2 = 0.5709$ [Excerpts from Table 1].

From the fitted model in (3.1), it is observed that the inclusion of the X_t in the model is significant since the p - value = (1.98e-09) < 0.05 level of significance, implying that there is a very strong evidence to conclude that X_t has a significant linear contribution to Y_t. The coefficient of determination (R^2) indicates that the X_t is able to explain about 57.09% of the total variation in Y_t. If the error term is found to be autocorrelated, then the regression model could be termed spurious.

Table 1: Output of Regression Model

```
Call:
lm(formula = Y ~ X)
Residuals:
   Min      1Q  Median     3Q     Max
-3985.9   -72.3    59.2   577.8  3080.8
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.0350   294.2920   0.190     0.85
X             4.0717     0.5383   7.564 1.98e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 1595 on 43 degrees of freedom
Multiple R-squared:  0.5709,        Adjusted R-squared:  0.5609
F-statistic: 57.21 on 1 and 43 DF,  p-value: 1.984e-09
```

In order to check if autocorrelations exist in the residuals obtained from the regression model in equation (3.1), we consider the ACF [Figure 1] of the residuals from the regression model;

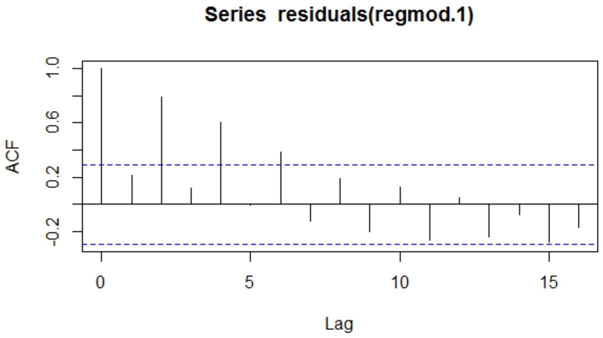

Figure 1: ACF of Residuals from Regression Model

it is observed that there are significant spikes at lags 2, 4 and 6 which are more than 5% of the total lags while all other lags fall within the confidence bounds, as such the residuals from the regression model appear to be autocorrelated and the regression model in (3.1) is termed spurious.

The reason for autocorrelated error term of the regression model in (3.1) is not farfetched. The nonstationarity in Y_t and X_t is more likely the cause of spurious correlations found in the error term of the regression model. To cop-out the menace of spurious correlations in the error term, we therefore, employ the dynamic regression method which takes into account both the lagged and current values of X_t. In order to build the dynamic regression model successfully, we make use of Box and Jenkins three iterative methods; model identification, model estimation and model diagnostic checking.

Model Identification

To identify the nature of the time-lagged relationship between a dependent variable, and current and past values of an independent variable, we examine the cross correlation function (CCF), that is, a smooth tapering pattern in the CCF shows which lags of independent variable we should used. Considering the CCF for Y_t and X_t in [Figure 2], we noticed that it is unclear and misleading. To identify which lags of X_t may predict Y_t, we therefore, apply the prewhitening technique to help us identify the lags of CCF.

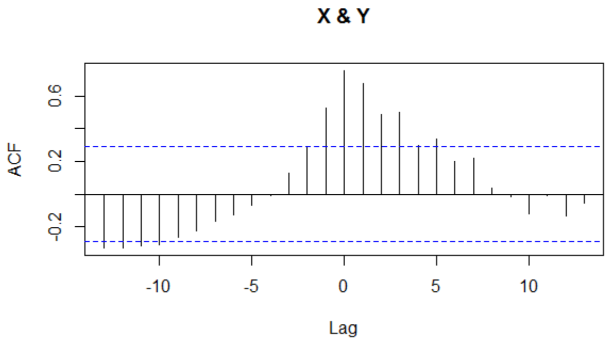

Figure 2: Cross Correlation Function of Money Supply and Gross Domestic Product

By prewhitening, we mean fitting an ARIMA model toX_t and reducing the residuals to white noise. Thereafter, we filter X_t with the fitted ARIMA model to obtain the white noise residual series. Lastly, Y_t is filtered with the same model and then the cross correlation function is computed using the prewhitened Y_t and prewhitened X_t. Now, fitting an ARIMA model to X_t, we allow the data to talk. The ACF and PACF of X_t [Figures 3and 4] respectively, indicate a tentative ARIMA(2,0,2) model alongside ARIMA(2,0,0) ARIMA(0,0,2) models. Both the ACF and PACF of residuals from ARIMA(2,0,2) model, [Figures 5 and 6], ARIMA(2,0,0) model, [Figures 7 and 8], and ARIMA(0,0,2) model, [Figures 9 and 10] respectively, are near white noise.

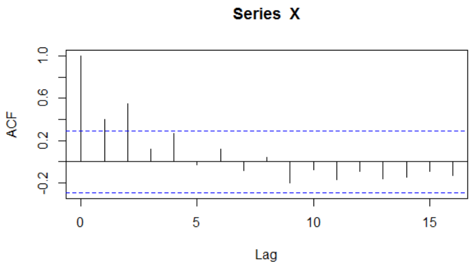

Figure 3: ACF of Money Supply

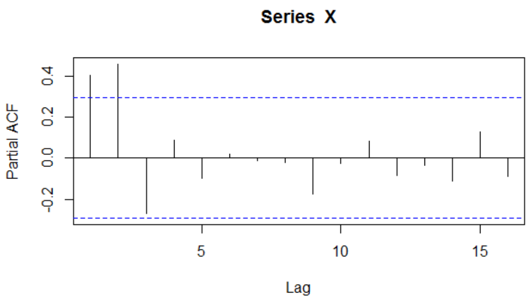

Figure 4: PACF of Money Supply

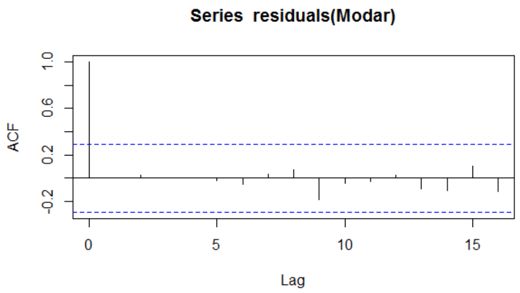

Figure 5: ACF of the Residuals from ARIMA (2,0,2) Model

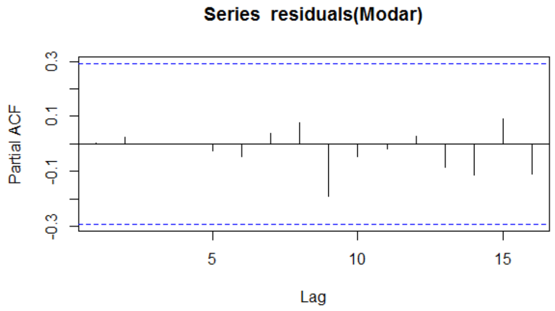

Figure 6: PACF of the Residuals from ARIMA(2,0,2) Model

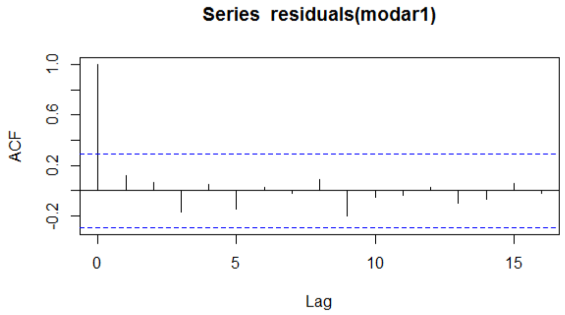

Figure 7: ACF of the Residuals from ARIMA(2,0,0) Model

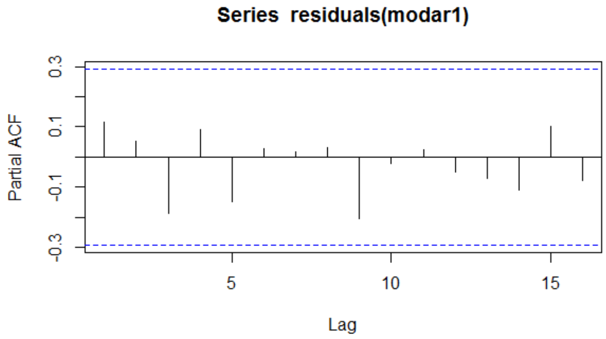

Figure 8: PACF of the Residuals from ARIMA(2,0,0) Model

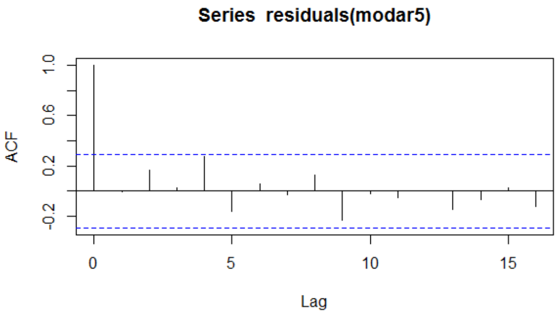

Figure 9: ACF of the Residuals from ARIMA (0,0,2) Model

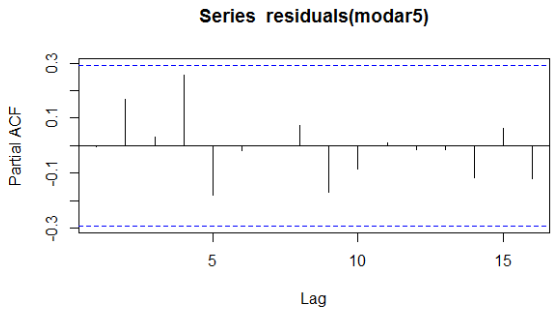

Figure 10: PACF of the Residuals from ARIMA (0,0,2) Model

Since all the three tentative models seem appropriate, their outputs as seen in Table 2, Table 3 and Table 4 for ARIMA(2,0,2), ARIMA(2,0,0) and ARIMA(0,0,2) respectively,

Table 2: Output of ARIMA(2,0,2) Model

```
Call:
arima(x = X, order = c(2, 0, 2))
Coefficients:
         ar1     ar2     ma1     ma2  intercept
     -0.1704  0.5977  0.5434  0.0290   297.0807
s.e.  0.1884  0.1892  0.2383  0.2389   134.1617
sigma^2 estimated as 116088:  log likelihood = -326.69,  aic =
665.38
```

Table 3: Output of ARIMA(2,0,0) Model

```
Call:
arima(x = X, order = c(2, 0, 0))
Coefficients:
        ar1     ar2  intercept
     0.2188  0.4599   281.2463
s.e. 0.1281  0.1289   154.9215
sigma^2 estimated as 126455:  log likelihood = -328.5,  aic = 665
```

Table 4: Output of ARIMA(0,0,2) Model

```
Call:
arima(x = X, order = c(0, 0, 2))
Coefficients:
        ma1     ma2  intercept
     0.3487  0.4537   312.8221
s.e. 0.1588  0.1082    96.1965
```

sigma^2 estimated as 132106: log likelihood = -329.42, aic = 666.84

we observed that both ARIMA(2,0,2) model and ARIMA(2,0,0) model have a smaller information criteria (aic) than ARIMA(0,0,2) model. Parsimoniously, we chose ARIMA(2,0,0) model over ARIMA(2,0,2) model. Thus, the estimated ARIMA(2,0,0) model is presented in (3.2) below;

$$(1- 0.2188B - 0.4599B^2)\,X_t^* = a_t \quad (3.2)$$

where $X_t^* = (X_t - 281.2463)$
[Excerpts from Table 3]
Then, we filter the Y_t series using the model for X_t as shown in (3.3)

$$(1- 0.2188B - 0.4599B^2)\,Y_t \quad (3.3)$$

Now, the cross correlation function (CCF) for the prewhitened X_t (which is the product of the ARIMA(2,0,0) model and its residuals) and the prewhitened Y_t is presented in [Figure 11]. We observed clear spikes at lags 0, -1,-2 and 1. Since we are only interested in the past values of X_t, the spike at lag1 is ignored. Thus, X_t, X_{t-1} and X_{t-2} should be included as predictors of Y_t.

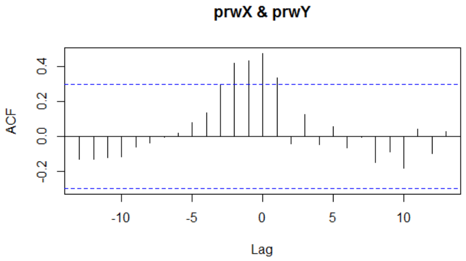

Figure 11: CCF for Prewhitened Money Supply and Prewhitened GDP

Model Estimation

The estimated dynamic regression model is presented in (3.4) below:

$Y_t =$ 185.75550 $-$ $0.13813\mathrm{X}_{t-2} - 0.03104\mathrm{X}_{t-1} -$ 0.14208 X_t
s.e (37.62991) (0.08255) (0.12893) (0.07183)
t-value (4.936) (-1.673) (-0.241) (-1.978) (3.4)
p-value (1.31e-05) (0.1017) (0.8109) (0.0545)
$R^2 = 0.2086$ [Excerpts from Table 5].

From the dynamic regression model in (3.4), the effect of using lagged values in overcoming spurious correlations is clearly seen as both lagged (X_{t-1}, X_{t-2}) and current (X_t) values in the model appear to show no sign of relationship with Y_t since their p - values are all less than 5% level of significance. Moreover, the coefficient of determination ($R^2 = 0.2086$) indicates that there is no linear relationship between X_t and Y_t.

Model Diagnostic Checking

The results from Box-Ljung test indicate that the residuals from the model in (3.4) are uncorrelated since χ^2 = 17.46, df =22 with corresponding p-value = 0.7375 $>$ 0.05 level of significance. [Excerpts from Table 6]

Table 5: Output of Dynamic Regression Model

```
Call:
dynlm(formula = Y ~ XLag2 + XLag1 + X)
Residuals:
   Min     1Q  Median     3Q    Max
-158.84  -80.45  -28.63   20.12  715.01
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 185.75550   37.62991   4.936 1.31e-05 ***
XLag2       -0.13813    0.08255  -1.673   0.1017
XLag1       -0.03104    0.12893  -0.241   0.8109
X           -0.14208    0.07183  -1.978   0.0545 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 156.5 on 42 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.2086,      Adjusted R-squared:  0.1521
F-statistic: 3.691 on 3 and 42 DF,  p-value: 0.01908
```

Table 6: Box – Ljung Test for Residuals from Dynamic Regression Model

```
Box-Ljung test
data:  residuals(regmod01)
X-squared = 17.46, df = 22, p-value = 0.7375
```

IV. CONCLUSION

Although it is well documented in the literature that differencing both the dependent and independent variables offers a good solution to avoiding spurious correlations, our study takes a look at dynamic regression as another approach to avoiding spurious correlations. This was achieved by allowing the dependent variable to be expressed as a function of both lagged and current values of the independent variable. Regressing Y_t on X_t, we found that the errors were correlated being a clear evidence of spurious correlation. However, the cross correlation function of the prewhitened variables indicated that lags 0, -1 and -2 of X_t should be included as predictors of Y_t. Subsequently, we modeled a dynamic regression of Y_t with past lags 1, 2 and current lag of X_t as explanatory variables and the resulting residual series was diagnostically checked using Ljung and Box Q – statistic. The residuals were confirmed to be uncorrelated, revealing that Y_t and X_t are totally unrelated. Therefore, we concluded that the problem of spurious correlations could be avoided by modeling a series using dynamic regression. This study could be extended to include the lagged values of the dependent variable, and both lagged and current values of multiple independent variables as predictors of the dependent variable.

REFERENCES

[1] H. Akaike, A New Look at the Statistical Model Identification,

IEEE Transactions on Automatic Control, 1973, Vol. 19, no 6, pp. 716 – 723.

[2] E. A. Akpan, I. U Moffat and N. B. Ekpo, Arma – Arch Modeling of the Returns of First Bank of Nigeria, European *Scientific Journal*, 2016, Vol.12, no.8, pp. 257 – 266.

[3] G.E.P. Box, G. M Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control.* 3rd Ed., New Jersey: Wiley and Sons, 2008, pp. 5-22.

[4] G. E. P. Box and D. Pierce, Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models, Journal of the American Statistical Association, 1970, 65, pp.1509-1526.

[5] P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting, 2nd Ed,. Springer, 2002.

[6] J. D. Cryer, and K. Chan, *Time Series Analysis with Application in R,* 2nd Ed., Springer, 2008.

[7] N. R. Drasper and H. Smith, *Applied Regression Analysis,* 3rd Ed., New York, John Wiley and Sons, 1998.

[8] W. A. Fuller, *Introduction to Statistical Time Series,* 2nd Ed., New York, John Wiley and Sons, 1996.

[9] C. W. J. Granger and P. Newbold, Spurious Regressions in Econometrics, Journal of Econometrics, 1974, 2, pp.111 – 120.

[10] G. Ljung and G. C. Box, On a Measure of Lack of Fit in Time series models, *Biometrica,* 1978, *Vol.* 2 no.66, pp. 265-270.

[11] A. Pankratz, Forecasting with Dynamic Regressions Models, 3nd Ed., New York, John Wiley and Sons, 1991.

[12] J. O. Rawlings, S. G. Pantula and D. A. Dickey, *Applied Regression Analysis: A*Research Tool, 2nd Ed., Springer, 1998.

[13] W. W. S. Wei, Time Series Analysis Univariate and Multivariate Methods, 2nd Ed., Adison Westley, 2006.

AUTHORS

First Author – Akpan, E. A., Department of Mathematics and Statistics, University of Uyo, eubong44@gmail.com, +2348036200343

Second Author – Moffat, I. U. Ph. D, Department of Mathematics and Statistics, University of Uyo, moffitto@yahoo.com, +2348064497511