Determination of lag threshold
on the measure of collinearity

Agunloye Oluokun Kasali[1], Arnab Raghunath[2] and

Shangodoyin Dahud Kehinde[3]

Abstract

To date, the implementation of lag selection procedures within the context of unit root tests is largely based on application of standard information criteria which are well-known to have lag selection problem. In this paper, we propose a new lag selection procedure based on collinearity measures for an autoregression of differenced lagged values. In empirical analysis, the conventional coefficient of multiple determination denoted by R_C^2 has been widely used in statistical literature for the computation of standard collinearity measures such as variance inflation factor(VIF) and tolerance index(ToL) but given the well-known difference between the underlying assumptions of standard multiple linear

[1] Department of Statistics, Faculty of Social Sciences, University of Botswana,
     Gaborone.
[2] Department of Statistics, Faculty of Social Sciences, University of Botswana,
     Gaborone.
[3] Department of Statistics, Faculty of Social Sciences, University of Botswana,
     Gaborone.

regression and a linear autoregression, the conventional coefficient of multiple determination which is used to compute VIF and ToL   under the framework of standard multiple linear regression may not provide an appropriate collinearity measures under a linear autoregression setting. Hence, the realization of this obvious inadequacy motivated the introduction of our newly proposed statistic called the coefficient of multiple determination for autoregression of differenced lagged values of time series data. Unit root tests are implemented on autoregression made up of differenced lagged values which requires an appropriate choice of optimal truncation lag. However, in the current literature there is dearth of information on possibility of using collinearity statistics as criteria for lag truncation in unit root tests. Using the newly proposed coefficient of multiple determination for autoregression of differenced lagged series denoted by R_d^2, we generate a class of collinearity statistics   called   tolerance index and variance inflation   factor for differenced lagged series denoted   by ToL_d and VIF_d respectively. We used these collinearity statistics as criteria for determining lag threshold for unit root tests. We employed augmented Dickey-Fuller(ADF) and generalized Dickey Fuller (DF-GLS) unit root tests to illustrate   empirically the application of   these collinearity criteria for lag selection and we demonstrated that these criteria select lag-length that is parsimonious thereby avoiding the problem of over-specification of truncation lag-length that is commonly associated with classical lag selection techniques.

1   Introduction

To date the challenges of determining optimal autoregressive lag length for unit root test regressions is crucial in the analysis of time series, since the empirical power property of unit root tests is known to be highly sensitive to the choice of truncation lag parameter. Different choices of optimal truncation lag parameters suggested by different classical lag selection criteria constitute a big challenge to applied practitioners as regards the choice of a particular truncation lag among the array of competing optimal truncation lag parameters suggested by different information criteria. Hence, the choice of correct lag length is an important specification decision. Among popular methods used for determination of truncation lag-length in unit root tests, the standard information criteria such as Akaike Information Criterion (AIC) proposed by $[1]$, Final prediction Error (FPE) due to $[2]$, Schwarz's Bayesian Information Criterion (BIC) introduced by $[3]$ and Hannan-Quinn Information Criteria suggested by $[4]$ are the most widely used lag selection criteria in empirical analysis . However, the major problem with standard information criteria according to $[5]$ is that these widely used lag selection criteria select optimal truncation lag parameter that leads to poor empirical size and power properties. Poor empirical power property of unit root tests is a consequence of over-specification of optimal truncation lag parameter. In this paper, we explore the collinearity property of time series data to provide an alternative procedure for determination of lag threshold for unit root tests based on two important measures of collinearity: variance inflation factor (VIF) and tolerance (ToL). Conventionally, R_C^2 represents the proportion of variance in the $i-th$ independent variable that is associated with the other independent variables in the model. According to $[6]$, R_C^2 gives an excellent measure of the collinearity of the $i-th$ independent variable with other independent variables in the model .The tolerance(ToL) for the $i-th$ independent variable is 1 minus the

proportion of variance it shares with the other independent variables in the analysis. This represents the proportion of variance in the $i-th$ independent variable that is not related to other independent variables in the model. This could serve as a good measure of cut-off for the lag with respect to the variables in the model. The reciprocal of this tolerance may give an important measurement of lag specification for all $k \le \dfrac{N}{4}$, where k is optimal truncation is lag and N is the sample size. In this chapter, our focus is to establish an empirical connection between measures of collinearity and procedure for lag truncation in unit root test regressions using VIF and tolerance index as criteria. Following a study by $[7]$, it is evident that computation of collinearity statistics with conventional coefficient of determination is theoretically and empirically unjustifiable in the context of linear autoregression given the well-known differences between the underlying assumptions of a standard multiple regression model and a linear autoregression. In view of this, we propose a new coefficient of multiple determination for differenced lagged values of time series denoted by R_D^2 as a more credible alternative to conventional coefficient of determination denoted by R_C^2 .Our empirical study shall include estimation of variance inflation factor (VIF) and tolerance (ToL) using the new proposed statistic as well as investigation of the performance and robustness of the proposed statistic using a simulation study. One particular advantage of using this proposed statistic for the computation of VIF and ToL is that the proposed statistic is a function of the conventional coefficient of multiple determination. Since the VIF is a measure of how the variance of a regression coefficient for a particular predictor is inflated, it can be used as a criterion for determining the cut-off point for lag truncation within the framework of linear autoregressive unit root tests.

2   Specification of Tolerance Index and Variance Inflation Factor in Unit Root Testing

Consider the distributed lag representation of ADF and DF-GLS regression models as follows:

$$
\left.
\begin{aligned}
y_t^{(1)} &\equiv \Delta y_t - \rho y_{t-1} = \sum_{j=0}^{k} \gamma_j \Delta y_{t-j} + \varepsilon_{1t} \\
y_t^{(2)} &\equiv \Delta y_t - \alpha - \rho y_{t-1} = \sum_{j=0}^{k} \gamma_j \Delta y_{t-j} + \varepsilon_{2t} \\
y_t^{(3)} &\equiv \Delta y_t - \alpha - \beta t - \rho y_{t-1} = \sum_{j=0}^{k} \gamma_j \Delta y_{t-j} + \varepsilon_{3t} \\
y_t^{(4)} &\equiv \Delta y_t - \alpha - \gamma t - \rho y_{t-1}^d = \sum_{j=0}^{k} \delta_j \Delta y_{t-j}^d + \varepsilon_{4t}
\end{aligned}
\right\}
\tag{1}
$$

We consider the latter part of equation (1) with k regressors and we have

$$
\mathbf{y^{(i)}} = \mathbf{\gamma' \Delta y_L^d}
\tag{2}
$$

where $L = t - j$, $d \geq 0$, $\varepsilon \sim N(0, \sigma I)$ and $j = 1, 2, ..., k \leq \dfrac{N}{4}$

Collinearity is related to the existence of near dependences among the columns of the $\Delta \underline{y}_{t-j}^d$ matrix. The tolerance index on each regressor $\Delta \underline{y}_{t-j}$ serves as measure of lag truncation since it relates the association of the $j-th$ regressor to other lagged regressors in the right-hand side of equation (2).The tolerance is given by:

$$
Tol_C = 1 - R_C^2
\tag{3}
$$

Where R_C^2 is the conventional coefficient of determination obtained in each of the k auxiliary regressions of the form:

$$
\Delta y_{Li} = \delta_0 + \delta_1 \Delta y_{1i} + ... + \delta_{L-1} \Delta y_{(L-1)i} + \delta_{L+1} \Delta y_{(L+1)i} + ... + \delta_k \Delta y_{ki} + u_i
\tag{4}
$$

The Tol_C shows the proportion of variance of $\Delta \underline{y}_L$ that is not accounted for by the remaining $k-1$ regressors and the higher the index , the less the contribution of other $k-1$ regressors and this can be used as an index of measuring the

truncation lag .The variance inflation factor(VIF) can be obtained as the reciprocal of ToL .This is a measure of the increment of sampling variance of the estimated regression coefficient of $\Delta y_L(\delta_L)$ due to linear relationship among the $\Delta y'_L s$

The value of

$$VIF = \frac{1}{ToL_C}$$ (5)

Equation (5) produces a measure of truncation of the lag using the coefficients of regression. To measure the point of truncation of lag, the lag L for which VIF is close to 10 or above and $ToL_i \le 0.1$ will be the point to consider as the best lag for testing the unit root. We shall later examine the range of values for collinearity statistics which we call variance inflation factor and tolerance index for autoregression of differenced lagged series denoted by VIF_D and ToL_D respectively within the framework of unit root tests. The analogous relationship that exist between the regression model described in equation (4) and conventional autoregressive model used in time series analysis (see $[8]$) will allow us to specify the model as:

$$\Delta y_{tL} = \phi_0 + \phi_1 \Delta y_{(t-1)1} + \phi_2 \Delta y_{(t-2)2} + ... + \phi_{L-1} \Delta y_{(t-L+1)L-1} + \phi_{L+1} \Delta y_{(t-L-1)L+1}$$
$$+ ... + \phi_p \Delta y_{(t-p)p} + \varepsilon_{tL}$$ (6)

where ε_{tL} and Δy_{tL}

Equation (6) follows all properties of autoregressive model and could be used to ascertain the optimal lag with respect to Tol_L and $VIF(L)$ defined in equations (3) and (5) above.

3 Main Results : Derivation of coefficient of multiple determination for a differenced autoregressive model

In standard multiple linear regression model, the collinearity statistics such as tolerance index (ToL) and variance inflation factor (VIF) are usually computed using the conventional coefficient of multiple determination denoted by R_C^2. Let us denote these collinearity statistics computed with R_C^2 as ToL_C and VIF_C respectively. This study is based on time series regression involving differenced lagged values. We consider an autoregressive model of the form:

$$\Delta \hat{y}_t = \hat{\phi}_0 + \hat{\phi}_1 \Delta y_{t-1} + \hat{\phi}_2 \Delta y_{t-2} + \ldots\ldots + \hat{\phi}_k \Delta y_{t-k} + \varepsilon_t \tag{7}$$

Where $\Delta \hat{y}_t$ is the differenced series of the observed time series y_t, $\hat{\phi}_0$ is a multiple regression constant, $\hat{\phi}_i, i = 1, 2, \ldots, k$ are regression coefficients to be estimated, $\Delta \hat{y}_{t-i}, i = 1, \ldots\ldots, k$ are the differenced lagged values of $\Delta \hat{y}_t$ and ε_t is error term which is normally and independently distributed with zero mean and variance σ_ε^2

We introduce a new coefficient of multiple determination for differenced time series within the framework of linear autoregression which we define as R_D^2 where

$$R_D^2 = R_{\Delta y_t \Delta y_{t-1} \Delta y_{t-2} \ldots \Delta y_{t-k}}^2 = \frac{\sum \left[(\hat{y}_t - \hat{y}_{t-1}) - \dfrac{\sum \Delta y_t}{n} \right]^2}{\sum \left[(y_t - y_{t-1}) - \dfrac{\sum \Delta y_t}{n} \right]^2} \tag{8}$$

Where $\Delta \hat{y}_t = \hat{y}_t - \hat{y}_{t-1}$ and $\Delta \overline{y}_t = \dfrac{\sum \Delta y_t}{n}$.

Put $\Delta y_t = y_t - y_{t-1}$ into (8), we have

$$R_D^2 = \frac{\sum\left(n\hat{y}_t - n\hat{y}_{t-1} - \sum y_t + \sum y_{t-1}\right)^2}{\sum\left(ny_t - ny_{t-1} - \sum y_t + \sum y_{t-1}\right)^2} \tag{9}$$

From conventional notations, we can have the following representations

$$\tau_1 = n\hat{y}_t - \sum y_t \quad \text{and} \quad \tau_2 = ny_t - \sum y_t,$$

using these two representations in (9) above, we can factor out the conventional coefficient of multiple determination denoted by R_C^2 from R_D^2 as follows:

$$R_D^2 = \frac{\sum\left[\tau_1 + \left(\sum y_{t-1} - n\hat{y}_{t-1}\right)\right]^2}{\sum\left[\tau_2 + \left(\sum y_{t-1} - ny_{t-1}\right)\right]^2} \tag{10}$$

After some algebraic manipulations of equation (10) we have

$$R_D^2\sum\tau_2^2 + 2R_D^2\sum\left[\tau_2\left(\sum y_{t-1} - ny_{t-1}\right)\right] + R_D^2\sum\left(\sum y_{t-1} - ny_{t-1}\right)^2 =$$
$$\sum\tau_1^2 + 2\sum\left[\tau_1\left(\sum y_{t-1} - n\hat{y}_{t-1}\right)\right] + \sum\left(\sum y_{t-1} - n\hat{y}_{t-1}\right)^2 \tag{11}$$

Dividing both sides of (11) by $\sum\tau_2^2$ we have

$$R_D^2 + \frac{2R_D^2\sum\left[\tau_2\left(\sum y_{t-1} - ny_{t-1}\right)\right]}{\sum\tau_2^2} + \frac{R_D^2\sum\left(\sum y_{t-1} - ny_{t-1}\right)^2}{\sum\tau_2^2} =$$
$$R_C^2 + \frac{2\sum\left[\tau_1\left(\sum y_{t-1} - n\hat{y}_{t-1}\right)\right]}{\sum\tau_2^2} + \frac{\sum\left(\sum y_{t-1} - n\hat{y}_{t-1}\right)^2}{\sum\tau_2^2} \tag{12}$$

If we let $\tau_1' = \sum y_{t-1} - n\hat{y}_{t-1}$ and $\tau_2' = \sum y_{t-1} - ny_{t-1}$ in (12) then we have the following:

$$R_D^2 + \frac{2R_D^2\sum\tau_2\tau_2^1}{\sum\tau_2^2} + \frac{R_D^2\sum\left(\tau_2^1\right)^2}{\sum\tau_2^2} = R_C^2 + \frac{2\sum\tau_1\tau_1^1}{\sum\tau_2^2} + \frac{\sum\left(\tau_1^1\right)^2}{\sum\tau_2^2} \tag{13}$$

Factorizing LHS of (13) gives

$$R_D^2\left[1 + \frac{2\sum\tau_2\tau_2^1}{\sum\tau_2^2} + \frac{\sum\left(\tau_2^1\right)^2}{\sum\tau_2^2}\right] = R_C^2 + \frac{2\sum\tau_1\tau_1^1}{\sum\tau_2^2} + \frac{\sum\left(\tau_1^1\right)^2}{\sum\tau_2^2} \tag{14}$$

This reduces to

$$R_D^2 = \frac{R_C^2 + \dfrac{2\sum \tau_1 \tau_1^1}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_1^1\right)^2}{\sum \tau_2^2}}{1 + \dfrac{2\sum \tau_2 \tau_2^1}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_2^1\right)^2}{\sum \tau_2^2}} \tag{15}$$

Since R_C^2 takes on values between 0 and 1, we consider the two extreme values of R_C^2 when $R_C^2 = 0$ and $R_C^2 = 1$. We begin with the first case when $R_C^2 = 0$ in (15), then R_D^2 is given by

$$R_D^2 = \frac{\dfrac{2\sum \tau_1 \tau_1^1}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_1^1\right)^2}{\sum \tau_2^2}}{1 + \dfrac{2\sum \tau_2 \tau_2^1}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_2^1\right)^2}{\sum \tau_2^2}} \tag{16}$$

Similarly, for the second case when $R_C^2 = 1$ in (16), we have

$$R_D^2 = \frac{1 + \dfrac{2\sum \tau_1 \tau_1^1}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_1^1\right)^2}{\sum \tau_2^2}}{1 + \dfrac{2\sum \tau_2 \tau_2^1}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_2^1\right)^2}{\sum \tau_2^2}} \tag{17}$$

If we let $\tau_1 = \tau_1^1$ and $\tau_2 = \tau_2^1$ in equations (16) and (17) respectively, then for the case $R_C^2 = 0$, R_D^2 becomes

$$R_D^2 = \frac{\dfrac{2\sum \left(\tau_1^1\right)^2}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_1^1\right)^2}{\sum \tau_2^2}}{1 + \dfrac{2\sum \left(\tau_2^1\right)^2}{\sum \tau_2^2} + \dfrac{\sum \left(\tau_2^1\right)^2}{\sum \tau_2^2}} = \frac{2R_C^2 + R_C^2}{1 + 2R_C^2 + R_C^2} = \frac{3R_C^2}{1 + 3R_C^2} \tag{18}$$

Similarly, for the case $R_C^2 = 1$, R_D^2 becomes

$$R_D^2 = \frac{1 + \dfrac{2\sum\left(\tau_1^1\right)^2}{\sum \tau_2^2} + \dfrac{\sum\left(\tau_1^1\right)^2}{\sum \tau_2^2}}{1 + \dfrac{2\sum\left(\tau_2^1\right)^2}{\sum \tau_2^2} + \dfrac{\sum\left(\tau_2^1\right)^2}{\sum \tau_2^2}} = \frac{1 + 2R_C^2 + R_C^2}{1 + 2R_C^2 + R_C^2} = \frac{1 + 3R_C^2}{1 + 3R_C^2} = 1 \tag{19}$$

Following equation (19), it is obvious that as R_C^2 approaches 1, R_D^2 also approaches 1 but this is not the case when $R_C^2 = 0$, $R_D^2 \neq 0$ as clearly demonstrated in equation (18). The immediate consequences of the preceding results lead to the following properties of R_D^2 which are easily discernable:

(i)   The values of R_D^2 is non-negative

(ii)   When $R_C^2 = 0$, $0 < R_D^2 < 1$, this implies that the value of R_D^2 can never be zero unlike the conventional coefficient of multiple determination R_C^2 that can be zero

(iii)   As the conventional coefficient of multiple determination R_C^2 approaches 1, the newly proposed coefficient of multiple determination for autoregression of differenced lagged series denoted by R_D^2 also approaches 1

(iv)   If $R_C^2 = 0.9$ this implies that $ToL_C = 0.1$ and $VIF_C = 10$ .Hence, when $R_C^2 = 0.9$, then $R_D^2 = 0.73$ and this gives $ToL_D = 0.27$ and $VIF_D = 3.7$ .This means that for a differenced series, to measure the point of truncation of lag, the lag L for which VIF_D is closed to 4 or above and $ToL_D \leq 0.27$ will be the point to consider as the best lag for testing the unit root.

4   Empirical Analysis of Auxiliary Autoregression

Auxiliary regression is essentially a descriptive regression of each explanatory variable expressed as a function of other explanatory variables. One of the most popular methods to detect collinearity in time series regression is implementation of auxiliary regression. In standard multiple linear regression, the usual convention for determining the seriousness of multicollinearity problem is to compare the coefficient of multiple determination for an auxiliary regression with the coefficient of multiple determination for the main regression. Multicollinearity is a problem if the former is greater than the latter. By the same token, the seriousness of multicollinearity problem within the framework of linear autoregression is also confirmed if the coefficient of multiple determination for a linear autoregression denoted by R_D^2 for an auxiliary autoregression is greater than the coefficient of multiple determination for the main autoregression. We shall examine the auxiliary autoregression for model (6) which is defined as follows:

$$\Delta y_{tL} = \phi_0 + \phi_1 \Delta y_{(t-1)1} + \phi_2 \Delta y_{(t-2)2} + ... + \phi_{L-1} \Delta y_{(t-L+1)L-1} + \phi_{L+1} \Delta y_{(t-L-1)L+1}$$
$$+ ... + \phi_p \Delta y_{(t-p)p} + \varepsilon_{tL} \tag{20}$$

Where maximum lag p is pre-set at 12 for monthly time series data. The associated twelve auxiliary autoregressions are given as follows:

$$\Delta y_{(t-1)1} = \theta_2 \Delta y_{(t-2)2} + \theta_3 \Delta y_{(t-3)3} + \theta_4 \Delta y_{(t-4)4} + \theta_5 \Delta y_{(t-5)5} + \theta_6 \Delta y_{(t-6)6} + \theta_7 \Delta y_{(t-7)7}$$
$$+ \theta_8 \Delta y_{(t-8)8} + \theta_9 \Delta y_{(t-9)9} + \theta_{10} \Delta y_{(t-10)10} + \theta_{11} \Delta y_{(t-11)11} + \theta_{12} \Delta y_{(t-12)12} + u_{t1} \tag{21}$$

$$\Delta y_{(t-2)2} = \theta_1 \Delta y_{(t-1)1} + \theta_3 \Delta y_{(t-3)3} + \theta_4 \Delta y_{(t-4)4} + \theta_5 \Delta y_{(t-5)5} + \theta_6 \Delta y_{(t-6)6} + \theta_7 \Delta y_{(t-7)7}$$
$$+ \theta_8 \Delta y_{(t-8)8} + \theta_9 \Delta y_{(t-9)9} + \theta_{10} \Delta y_{(t-10)10} + \theta_{11} \Delta y_{(t-11)11} + \theta_{12} \Delta y_{(t-12)12} + u_{t2} \tag{22}$$

For $t-3, ..., t-12$ we repeat the same process

Where $\theta_i : i = 1, 2, ..., 12$ are regression coefficients for the explanatory variables

in the auxiliary regressions. The error term u_{tL}, $L = 1, 2, ..., 12$ is a white noise process satisfying the following properties:

i. $E(u_{tL}) = 0$, *for all t*

ii. $\text{var}(u_{tL}) = \sigma^2$ for all $t, \sigma^2 < \infty$

iii. $\text{cov}(u_{tL}, u_{sL}) = 0$ if $t \neq s$

5   Data Analysis

For empirical illustration, we used both simulated and real datasets in order to evaluate the performance of the newly proposed statistic called coefficient of multiple determination for differenced autoregression denoted by R_D^2. This new statistic is used for the computation of tolerance index (ToL_D) and variance inflation factor (VIF_D) for the differenced autoregression. For empirical illustration based on real dataset, we employed US inflation data retrieved from the IMF database. The sample contains 624 monthly observations from January, 1962 to December, 2013.For simulated data, the random number generator module of Excel was used to simulate normally distributed series. The simulated series has mean 30 and variance 10. Prior to unit root testing, we examine the both simulated and real datasets for trends and stationarity using the time plots for simulated data and the US inflation rates respectively.

Figure 1 and Figure 2 show that the simulated data and the US inflation rates are both stationary.
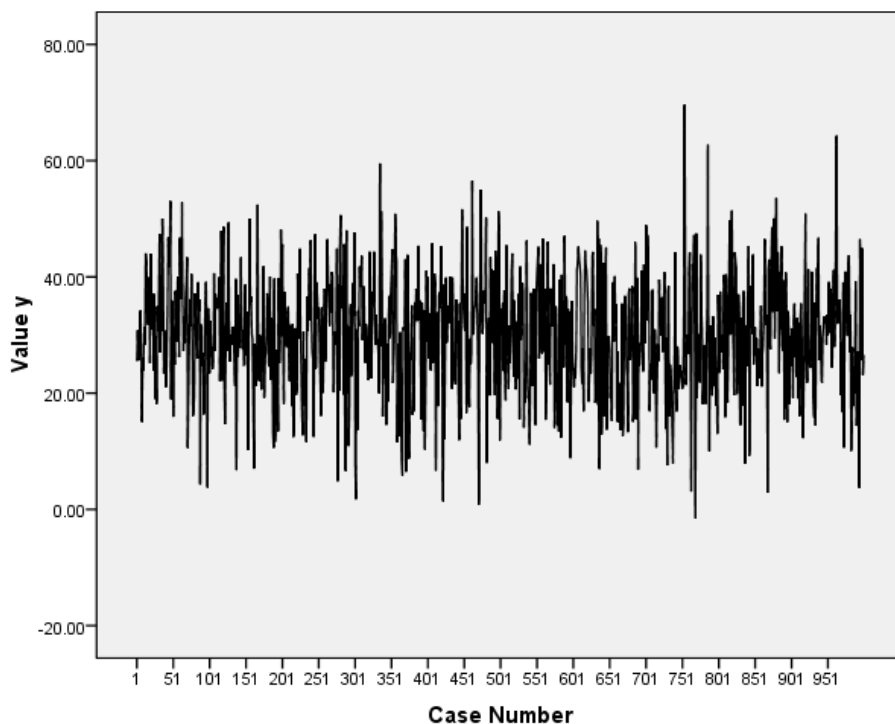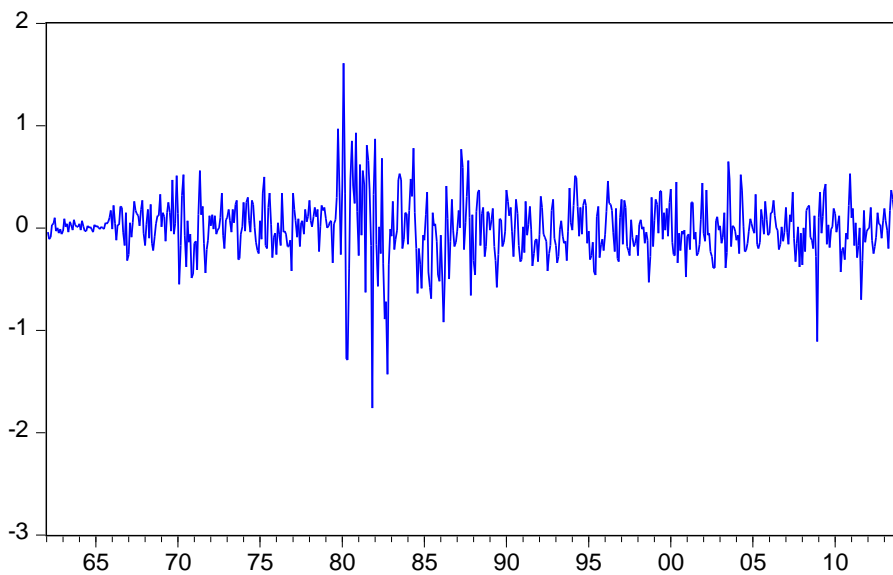
FIGURE 1: SIMULATED DATA AT LEVEL

FIGURE 2 : US INFLATION RATES AT FIRST DIFFERENCE

We now proceed to fit an autoregression of order 12 to our datasets and we apply the collinearity criteria to determine the appropriate lag threshold for unit root testing. Using both simulated and real datasets, the empirical results are presented in Table 1 and 2 below:

Table 1 : Empirical result for simulated data

Model (Dependent Variable)	R	R-Square	Adjusted R-Square	Std. Error of the Estimate	Regression F	Sig	ToL	VIF
Δy_{t-1}	0.650	0.422	0.416	10.58568	64.76	0.000	0.578	1.730
Δy_{t-2}	0.812	0.659	0.655	8.12220	171.27	0.000	0.341	2.933
Δy_{t-3}	0.862	0.743	0.740	7.04066	256.68	0.000	0.225	3.891
Δy_{t-4}	0.880	0.775	0.772	6.59181	279.87	0.000	0.206	4.444
Δy_{t-5}	0.891	0.794	0.791	6.31008	312.67	0.000	0.197	4.854
Δy_{t-6}	0.896	0.803	0.800	6.14481	330.46	0.000	0.197	5.076
Δy_{t-7}	0.896	0.803	0.800	6.14327	329.91	0.000	0.207	5.076
Δy_{t-8}	0.890	0.793	0.790	6.29442	310.40	0.000	0.226	4.831
Δy_{t-9}	0.880	0.774	0.771	6.56866	278.14	0.000	0.258	4.425
Δy_{t-10}	0.861	0.742	0.739	7.02235	233.22	0.000	0.344	3.876
Δy_{t-11}	0.810	0.656	0.652	8.08860	155.02	0.000	0.659	2.907
Δy_{t-12}	0.891	0.794	0.791	6.31008	312.67	0.000	0.197	4.854

In Table 1 above, the collinearity statistics as well as other relevant statistics obtained for the fitted auxiliary autoregressions of the simulated data were presented. The first column of Table 1 above describes dependent variable of the particular auxiliary autoregression being analyzed. The second column through the fourth column presents the values for R, R_D^2 and adjusted R_D^2 respectively. The fifth column through the seventh column presents the values for standard error of the estimates, regression F statistic and significance respectively. The eighth and ninth columns describe the collinearity statistics, tolerance index ToL_D and variance inflation factor VIF_D respectively. Overall, it is clearly evident that all the auxiliary regressions are statistically significant as shown by the values of regression F statistic in the sixth column of the table. The model fit for each auxiliary autoregression is above 0.4. The highest tolerance of 0.578 and the lowest variance inflation factor of 1.730 are achieved at lag 1. This value of tolerance indicates that 57.8% of the variance of the fitted auxiliary autoregression was not accounted for by other regressors and this implies that other regressors have less contribution in the fitted autoregression. Hence, the best point for lag truncation is at lag 1.

Table 2 : Empirical results for us inflation rate series

Model (Dependent Variable)	R	R-Square	Adjusted R-Square	Std. Error of the Estimate	Regression F	Sig	ToL	VIF
Δy_{t-1}	0.416	0.173	0.158	0.27020	11.393	0.000	0.827	1.209
Δy_{t-2}	0.532	0.283	0.269	0.25157	21.477	0.000	0.717	1.395
Δy_{t-3}	0.567	0.322	0.309	0.24457	25.878	0.000	0.678	1.475
Δy_{t-4}	0.573	0.329	0.316	0.24323	26.714	0.000	0.671	1.490
Δy_{t-5}	0.573	0.328	0.316	0.24333	26.640	0.000	0.672	1.488

Δy_{t-6}	0.576	0.332	0.320	0.24258	27.103	0.000	0.668	1.497
Δy_{t-7}	0.575	0.331	0.319	0.24258	26.998	0.000	0.669	1.495
Δy_{t-8}	0.571	0.325	0.313	0.24327	26.322	0.000	0.675	1.481
Δy_{t-9}	0.571	0.327	0.314	0.24301	26.448	0.000	0.673	1.486
Δy_{t-10}	0.577	0.335	0.308	0.24411	25.675	0.000	0.680	1.471
Δy_{t-11}	0.580	0.341	0.268	0.25101	21.340	0.000	0.659	1.517
Δy_{t-12}	0.585	0.352	0.157	0.26940	11.330	0.000	0.648	1.543

The Table 2 above presents the collinearity statistics as well as other relevant statistics obtained for the fitted auxiliary autoregressions for the US inflation rates dataset. The first column of Table 2 above describes dependent variable of the particular auxiliary autoregression being analyzed. The second column through the fourth column presents the values for R, R_D^2 and adjusted R_D^2 respectively. Similarly, the fifth column through the seventh column presents the values for standard error of the estimates, regression F statistic and significance respectively. The eighth and ninth columns describe the collinearity statistics, tolerance index ToL_D and variance inflation factor VIF_D respectively. Overall, it is clearly evident that all the auxiliary regressions are statistically significant as shown by the values of regression F statistic in the sixth column of the table. The model fit for each auxiliary autoregression is less than 0.5. The highest tolerance of 0.827 and the lowest variance inflation factor of 1.209 are achieved at lag 1. This value of tolerance indicates that 82.7% of the variance of the fitted auxiliary autoregression was not accounted for by other regressors and this implies that other regressors have less contribution in the fitted autoregression. Hence, the best point for lag truncation is at lag 1.

6   Conclusion

This study has employed collinearity statistics as criteria for lag selection in unit root tests. We introduced a new statistic called coefficient of multiple determination for autoregression of differenced lagged series. Using both ToL_D and VIF_D the lag threshold for unit root tests was found to be 1.This empirical result shows that our lag selection procedure select parsimonious lag structure that guides against over-specification of truncation lag within the context of unit root tests and these collinearity statistics criteria could be used as an alternative lag selection procedure to cross-check the lag structure that may be suggested by conventional lag selection criteria as a way of safe-guarding against problem of over-specification of truncation lag-length that is commonly associated with classical lag selection criteria such as standard information criteria.

References

[1] Akaike, H., Information theory and an extension of the maximum likelihood principle, in 2nd International Symposium on Information Theory by B. N. Petrov and F. Csaki, eds., Akademiai Kiado: Budapest, (1973).

[2] Akaike, H., Fitting autoregressive models for prediction, *Annals of the Institute of Statistical Mathematics*, **21**, (1969), 243-247.

[3] Schwarz, G., Estimating the Dimension of a Model, *Annals of Statistics*, **6**, (1978), 461-464.

[4] Hannan, E.J., and B.G. Quinn, The Determination of the Order of an Autoregression, *Journal of the Royal Statistical Society*, **41**, (1979, 190-195.

[5] Burke, S.P., Augmented Dickey-Fuller unit root test and the use of information criteria, *Discussion paper*, **10**, Department of Economics, University of Reading, (1993).

[6] O'Brien, A caution regarding rules of thumb for variance inflation factors, *Quality and Quantity*, **41**(5), (2007),673-690.

[7] Moineddin, R; et al, Autoregression as a means of assessing the strength of seasonality in a time series, *Population Health Metrics*, **1**, (2003), 10.

[8] Haggan, V. and Oyetunji, O.B., On the selection of subset autoregressive Time series Models, *Journal of Time Series Analysis*, **5**, (1984),103-113.