Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.5, No.12, 2015

Modeling inflation in Kenya: Comparison of SARIMA and Generalized Least Squares models

Susan W. Gikungu[1*], Anthony Waititu[1], John Kihoro[1, 2]

1. School of Mathematical Sciences, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000-00200, Nairobi, Kenya

2. Co-operative University College of Kenya, Computing and E-learning P.O. Box 24814-00502, Nairobi, Kenya

Abstract

One desire by the policy makers in a country is to have access to reliable forecast of inflation rate. This is only achievable if the right model with high predictive accuracy is used. In this paper, seasonal auto regressive integrated moving average (SARIMA) and Generalized Least Squares regression models are developed to predict Kenya's inflation using quarterly data for the period 1981 to 2013. SARIMA $(0,1,0)(0,0,1)_4$ was chosen as the model with the least Akaike Information Criterion and Bayesian Information Criterion. The parameters were then estimated. The residuals were checked to find out if they follow a white noise process by using residual Q-Q and normality test plots. The Test for normality of residual was also done. Given the high p-values (0.0639237) associated with the statistics as compared to 0.05, we fail to reject the null hypothesis that an error is normally distributed in this residual series. Thus, we conclude that the model provides an adequate fit for the data. In an effort to improve this, inflation was also modeled using Generalized Least Squares regression model.. The data was first checked for heteroscedasticity using the Breusch-Pagan test. Based on the p-value=0.000, which is less than alpha (of 5%), we conclude that there is substantial amount heteroscedasticity in the data. A regression model that forecasts inflation using its lags was constructed. The residuals were checked for normality using q-q plot. Additionally, the Shapiro-Wilk normality test was carried out. Its p-value was 0.08178 and since its greater than 0.05, its concluded that the residuals does not deviate from normality

A comparison was made on the predictive ability of both models. SARIMA $(0,1,0)(0,0,1)_4$ model had the least values of MAPE, MAE and RMSE with the corresponding values given by MAPE=14.155, RMSE=0.2871and MAE=0.23692

Keywords: SARIMA, Generalized Least Squares, Akaike Information Criterion ,Bayesian Information Criterion, Shapiro-Wilk normality test and Breusch-Pagan test

1. Introduction

Inflation as defined by Webster (2000) is the persistent increase in the level of consumer prices or a persistent decline in the purchasing power of money.

Inflation tends to be a relatively persistent process, which means that current and past values should be helpful in forecasting future inflation, Brent and Mehmet (2010). Applying that intuition, the two basic models that exploit information embedded in past values of CPI inflation will be constructed. In 1970, Box and Jenkins introduced autoregressive integrated moving average models, ARIMA. SARIMA model is useful in situations when the time series data exhibit seasonality-periodic fluctuations that recur with about the same intensity periodically, for example, yearly. This characteristic makes the SARIMA model adequate for studies concerning quarterly inflation data.

On the other hand, Generalized Least Squares regression model is a generalization of ordinary least squares in which we explicitly account for correlation in data. GLS is a method for fitting coefficients of explanatory variables that help to predict the outcomes of a dependent random variable.

2. Literature review

Being one of the important areas in economics research, various researchers have done studies on forecasting inflation in different countries. Additionally, GLS and SARIMA models have also been used in different fields. Fannoh et al. (2014) used SARIMA approach to model Liberia's monthly inflation rates which showed that SARIMA model was appropriate for modeling the inflation rates. Otu et al. (2014)used Box-Jenkins methodology to build ARIMA model for Nigeria's monthly inflation. SARIMA $(1, 1, 1)\ (0, 0, 1)_{12}$ model was

developed and used to forecast monthly inflation for the year 2014.

A general frame work for regional analysis and modeling of extreme rainfall characteristics was presented. A GLS regression model that explicitly accounted for inter site correlation and sampling uncertainties was applied for evaluating the regional heterogeneity of the PDS (Partial Duration Series)parameters. The resulting model was used for estimation of rainfall intensity-duration-frequency curves for Denmark, Henrik et al. (2002).

Griffis and Veronica (2007) presented innovative approaches to GLS regression in hydrologic applications. The GLS regression procedure accounted for differences in available record lengths and spatial correlation in concurrent events by using an estimator of the sampling covariance matrix of available flood quantiles.

The problem in obtaining forecasts is to know which model to use while there are different competing ones. Motivated by these researches, I wish to compare forecast efficiency of SARIMA and Generalized Least Squares regression models applied to inflation rate in Kenya.

3. Methodology

3.1 SARIMA model

SARIMA is a Box-Jenkins technique that takes into account time series data and decomposes it into:

-AR (Autoregressive) process- A real valued stochastic process (Y_t)is said to be an AR process of order p, denoted by AR(p) if

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_{1p} y_{t-p} + \varepsilon_t \quad (1)$$

The value of AR(p) process at time t is therefore regressed on its own past p values plus a random shock.

-MA (Moving Average) process- A real valued stochastic process (Y_t) is said to be an MA process of order q, denoted by MA(q) if there exists b_1,.....,b_q and a white noise (ε_t) such that

$$y_t = b_0 \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q} \quad (2)$$

The value of MA(q) process at time t is therefore regressed on its own past errors.

-ARMA - Moving averages MA(q) and autoregressive AR(p) processes are special cases of so called autoregressive moving average processes (ARMA). A real valued stochastic process (Y_t)is said to be an ARMA process of order p,q, denoted by ARMA (p,q) if it satisfies the equation

$$y_t = \varepsilon_t + (a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p}) + (b_0 \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q} \quad (3)$$

This can be re-written as $\phi(z) Y_t = \Theta(z)\, \varepsilon_t$

Where $\phi(z) = 1 + a_1 z + \cdots + a_p z^p$ and

$\Theta(z) = 1 + b_1 z + \cdots + b_q z^q$ are the characteristic polynomials of the AR part and of the MA part of an ARMA (p,q) process (Y_t). z is the back-shift (lag) operator.

-ARIMA process- The Auto Regressive Integrated Moving Average (ARIMA) model, is a broadening of the class of ARMA models to include differencing. A process Y_t is said to be an ARIMA(p,d,q) if $(1-z)^d Y_t$is a causal ARMA (p,q). The corresponding ARIMA equation is

$$\phi\,(z)(1 - z)^d X_t = \Theta_q(z)\, \varepsilon_t \quad (4)$$

-SARIMA process- For a non- stationary time series possibly containing seasonality, that is, seasonal periodic component repeats itself after every s observations,Box-Jenkins (1976) have defined a general multiplicative Seasonal ARIMA model (SARIMA) as

$$\phi_p(z)\Phi_P(z^s)(1 - z)^d(1 - z^s)^D Y_t = \theta_q(z)\Theta_Q(z^s)\, \epsilon_t \quad (5)$$

Where $\phi_p(z)$, $\Phi_P(z^s)$, $\theta_q(z)$and $\Theta_Q(z^s)$ are characteristic polynomials of orders p,P,q and Q respectively and D are the orders of non-seasonal and seasonal differencing respectively. Box Jenkins bases the model selection on three stages ie. Model Identification (specification), estimation of coefficients and diagnostic checking.

Stage1: Model Identification

The objective of this step is to determine the possible SARIMA model that best fit the time series data under consideration. SARIMA model is appropriate for stationary time series therefore stationarity condition must be satisfied. Augmented Dickey Fuller (ADF) test is used to see whether the seasonal differenced series is stationary. The values of p, q, P ,Q are then determined at this step by looking at the patterns of the Autocorrelation function (ACF) and the Partial Autocorrelation Function (PACF).

Stage2: Estimation of coefficients

The parameters are estimated by the maximum likelihood method. In this study, the model with the minimum value of AIC is judged as the best model. It is given by

AIC=-2 In(L) + 2k where $k=(p+q+1)$ and L is the maximized likelihood value.

Stage3: Diagnostic checking

This step involves checking whether the residuals random and ensure that the estimated parameters are statistically significant. This is done by inspecting the i) ACF plots of the residuals ii) the probability plots of the residuals iii) the residual q-q plots . Shapiro-wilk test can also be used to verify normality among the residuals. If the model fails these diagnostic checks then return to the identification stage to find a better model. After choosing a model and checking its fit and forecasting ability, then the model is used to predict.

3.2 Generalized Least Squares

This is a technique for estimating the unknown parameters in linear regression models. This technique is applied when the variances of observations are unequal (heteroscedasticity) or when there is a certain degree of correlation between the observations. A regression model that forecasts inflation using lags of the inflation is constructed. Heteroscedasticity of the data was done using Breusch-Pagan test.

The regression analysis figures out the parameters of that function. In the standard linear regression model,

$$\boldsymbol{y} = \boldsymbol{X\mathcal{B}} + \boldsymbol{\varepsilon} \quad (6)$$

where **y** is the n×1 response vector$\boldsymbol{X}$is an n×p model matrix; $\boldsymbol{\mathcal{B}}$ is a p×1 vector of parameters to estimate; and $\boldsymbol{\varepsilon}$is an n×1 vector of errors. Fox (2002).

Initially, we make the assumption that $\boldsymbol{\varepsilon \sim N_n}(\mathbf{0}, \boldsymbol{\Sigma})$, where the error-covariance matrix $\boldsymbol{\Sigma}$ is symmetric and positive-definite. Different diagonal entries in $\boldsymbol{\Sigma}$ correspond to non-constant error variances, while nonzero off-diagonal entries correspond to correlated errors. Efficient estimation of **β** requires knowledge of $\boldsymbol{\Sigma}$**.** Suppose, for the time-being, $\boldsymbol{\Sigma}$ is known. Then there exists a non-singular matrix **G** such that $\boldsymbol{\Sigma^{-1}} = G^T G$. Consider the following transformed model

$$y^* = \boldsymbol{X^*\mathcal{B}} + \boldsymbol{\varepsilon^*} \quad (7)$$

Where $y^* = \boldsymbol{G}y$, $\boldsymbol{X^* = GX}$ $\boldsymbol{and}$ $\boldsymbol{\varepsilon^* = G\varepsilon}$ **.** The model satisfies the classical assumptions that $(\varepsilon *) = 0$, $Var(\varepsilon *) = I$ and $E(X^{*T}\varepsilon *) = 0$.

The best linear unbiased estimator for the transformed model is

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}^{*T}\hat{\Sigma}^{-1}\boldsymbol{X}^*)^{-1}X^{*T}\Sigma^{-1}y^* \quad (8)$$

In practice, $\boldsymbol{\Sigma}$ is typically unknown so that the GLS estimator cannot be obtained directly, Ayinde (2007). However, $\boldsymbol{\Sigma}$ can be estimated empirically by using the sample data and this makes the GLS estimator a Feasible GLS (FGLS). That is, FGLS is GLS estimation procedure but with estimated covariance matrix, not an assumed one.

3.3 Performance measures

To compare the accuracy of these specifications, the following measures of aggregate error were used:

i)Root Mean Square Error (RMSE)- It is a measure of the differences between values predicted by a model and the values actually observed. A RMSE of 0 indicates a perfect forecasting performance while positive values reflect deviations between the forecast values and the realized values.

ii)Mean Absolute Percentage Error (MAPE)- MAPE measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error. The lower the MAPE value, the better the method of forecasting, Barro (1997)

iii)Mean Absolute Deviation (MAD) - MAD measures the size of the error in units. It is calculated as the average of the unsigned errors. The smaller the MAD, the better the model.

4. Empirical results

The data used in this study is quarterly inflation rate of Kenya from 1981:1 to 2013:4, which is made up of 132 observations. It was obtained from Kenya National Bureau of Statistics. Figure 1 shows the plot of Kenya's quarterly inflation.

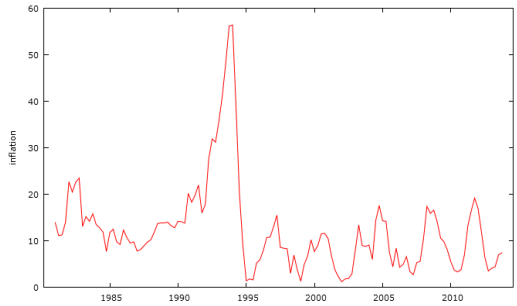

Figure 1: Quarterly inflation rates of Kenya

The plots of the autocorrelation and partial autocorrelation functions are presented in figure 2. The ACF plot dies down in a sine wave pattern which implies that there is a seasonal and a non seasonal component of the series.

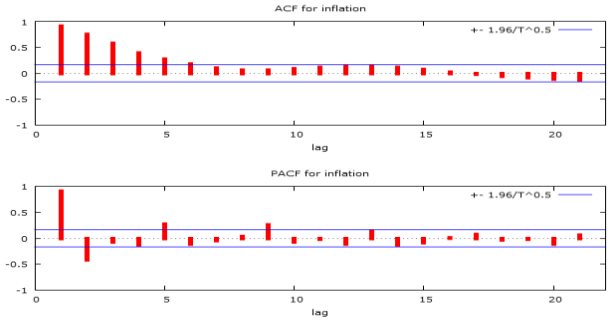

Figure 2:: ACF and PACF

4.1 SARIMA Model

SARIMA modeling requires that the stationarity condition to be satisfied. Augmented Dickey Fuller (ADF) test was used to test for it. The null hypothesis that the inflation rate is not stationary or has unit root. The p-value obtained from the ADF test without a constant is 0.1216. We therefore failed to reject the null hypothesis at 5% level of significance and concluded that the data is non-stationary. This meant that differencing is necessary to make the data stationary. The differenced data was tested again and the value obtained was 1.215e-010 and we concluded that the data was stationary.

The best model was selected by picking the model with the least AIC and BIC values. Model $(0,1,0)(0,0,1)_4$ was selected as the most appropriate. Using the Maximum Likelihood estimator, the model parametersϕ, Θ, Φ and Θ are estimated. The fitted model is given by

$$(1-z)w_t = w_{t-1} + e_t + \Theta_1 e_{t-s} \quad (9)$$

Replacing the coefficients with the corresponding values, the following is obtained;

$$(1-z)w_t = w_{t-1} + e_t + 0.543973 e_{t-s} \quad (10)$$

The residuals are checked to find out if they follow a white noise process. This was achieved by plotting the residual Q-Q and normality test plots as shown in figure 3 and 4 respectively. The Q-Q plot is reasonably straight so normality is okay

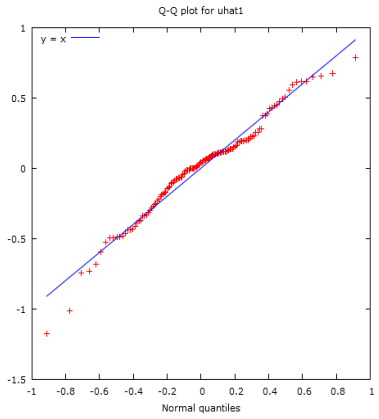

Figure3: SARIMA Residual Q-Q plot

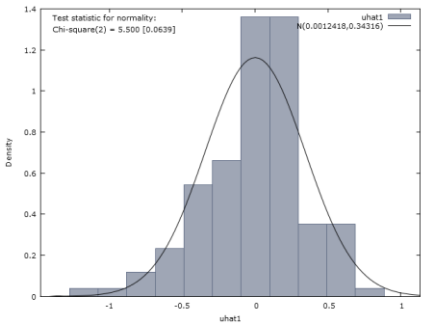

Figure 4:SARIMA Normality test plot

Figure 4 indicates that the normal distribution provides adequate fit for the model. Test for normality of residual was done. Its null hypothesis that an error is normally distributed was not rejected since the p-value=0.0639237 >0.05.

Shapiro test was done on the residuals and the p-value obtained was 0.1174. Since the p-value is > 0.05, it is accepted that the residuals are normally distributed.

4.2 Generalized Least Squares

The model was checked for heteroscedasticity using the Breusch-Pagan test. The test statistic was 19.4957, with p-value = P(Chi-square(4) > 19.4957) = 0.000627899. Based on the p-value, which is less than alpha (of 5%), we conclude that there is substantial amount of heteroscedasticity in the model. The relationship between inflation and its lags is modeled using the GLS model . The R program was used to achieve this. All the independent variables except lag 2 were found to be statistically significant at the 0.05 level.


```
Generalized least squares fit by maximum likelihood
  Model: yt ~ y1 + y3 + y4
  Data: NULL
       AIC      BIC    logLik
  93.87051 111.1673 -40.93526

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.7980999

Coefficients:
                Value  Std.Error   t-value p-value
(Intercept)  1.9828939 0.29557434  6.708613  0.0000
y1           0.2834669 0.08040525  3.525478  0.0006
y3           0.2180377 0.08205238  2.657299  0.0089
y4          -0.3730905 0.08204299 -4.547500  0.0000

 Correlation:
   (Intr) y1     y3
y1 -0.455
y3 -0.394 -0.132
y4 -0.396 -0.131 -0.239

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.89153603 -0.63504212  0.06881144  0.50170933  2.65872396

Residual standard error: 0.5455094
Degrees of freedom: 132 total; 128 residual
```

Figure 5: GLS output

The model was fitted again without the insignificant variable and the final model obtained was

$$y_t = 1.983 + 0.283y_{t-1} + 0.218y_{t-3} - o.373y_{t-4} \qquad (11)$$

The corresponding residual Q-Q plot for the GLS model is as shown in Figure 6. It is clear that most of the points lie on the line and therefore the distribution comes from a normal distribution.

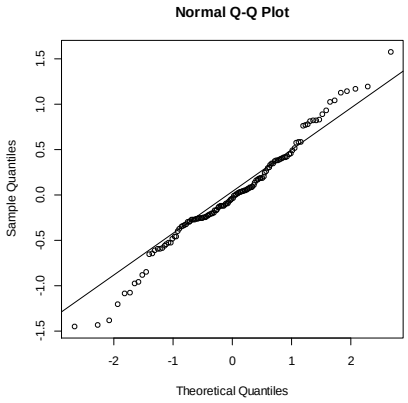

Figure 6: Residual Q-Q-plot for GLS model.

Additionally, the Shapiro-Wilk normality test was 0.08178. Since the p-value>0.05, the data does not deviate from normality.

4.3 Performance measures

	SARIMA	GLS
RMSE	0.2871	0.5501276
MAE	0.23692	0.4161865
MAPE	14.155	30.60809

All the forecast performance measures are lower for the SARIMA model as compared to those of the GLS model.

This suggests that the SARIMA outperforms the GLS model in modeling inflation in Kenya.

5. Conclusion

In this study, we model the inflation rates of Kenya using Seasonal auto regressive integrated moving average (SARIMA) and Generalized Least Squares regression models. The approaches were used to analyze quarterly inflation rates from 1981 to 2013. The best SARIMA model was identified as $(0,1,0)(0,0,1)_4$ with minimum Information Criterion and Bayesian Information Criterion. It was judged as the best model after satisfying the model assumptions.

Inflation was also modeled using Generalized Least Squares regression model.. The data satisfied the heteroscedasticity condition. A regression model that forecasts inflation using its lags was constructed. The residuals were checked for normality using q-q plot. Additionally, the Shapiro-Wilk normality test was carried out. Both tests showed that the residuals do not deviate from normality. All the forecast performance measures were lower for the SARIMA model as compared to those of the GLS model. This suggests that the SARIMA outperforms the GLS model in modeling inflation in Kenya. We recommend that future research is done on the GLS model in handling time series data.

References

[1] M. Henrik , P. Mikkelsen, D. Rosbjreg and P.Harremoes, “Regional Estimation of Rainfall Intensity-Duration-frequency curves using generalized least squares regression of partial duration series statistics,” Water Resources Research, 38: 2002.

[2] A.Otu , O.George, O. Jude , M. Hope and I. Andrew, “Application of Sarima Models in Modelling and Forecasting Nigeria's Inflation Rates,” American Journal of Applied Mathematics and Statistics, 2, pp. 16-28, 2014

[3] V. Griffis and J. Veronica, “The use of GLS regression in regional hydrologic analyses,” Journal of Hydrology, 344, pp. 82- 95, 2007

[4] K. Ayinde, “A Comparative Study of the Performances of the OLS and some GLS Estimators when Stochastic Regressors are Both Collinear and Correlated with Error Terms,” Journal of Mathematics and Statistics, 3(4), pp. 196-200, 2007

[5] M. Brent and P. Mehmet, “Simple Ways to Forecast Inflation: What Works Best?,” Trade Publication, 17, pp. 1-9, 2010

[6] J. Fox,“Time-Series Regression and Generalized Least Squares,” Appendix to An R and S-PLUS Companion to Applied Regression.

[7] R. Webster,New Universal Unabridged Dictionary, Barnes and Noble Books, New York, 2000.

[8] R. Fannoh , G. Orwa and J.Mungatu, “Modeling the Inflation Rates in Liberia SARIMA Approach,” International Journal of Science and Research, 3, pp. 1360-1367, 2014

[9] R. Barro,Macroeconomics, MIT Press, Cambridge, 1997.