FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 3, June (Special Issue), 2023, pp 10 - 15
DOI: https://doi.org/10.33003/fjs-2023-0703-1847

ROBUST ORDER IDENTIFICATION OF ARIMA AND GARCH MODELS: STATIONARY AND NON-STATIONARY PROCESS

***A. A. Abdullahi, E. L. Kazeem, U. F Abbas, M. Hassan**

Department of Mathematical Sciences, Faculty of Science, Abubakar Tafawa Balewa University, Bauchi, Nigeria

*Corresponding authors' email: ahmdy4real@gmail.com

ABSTRACT

Identification is the most important stage of all the stages of the modeling process. This research identifies a suitable order for the two different time series models ARIMA and GARCH. For GARCH two different distributions that is GARCH-STD and GARCH-GED with different sample sizes in fitting and forecasting stationary and non-stationary data structures was considered. The study recommends the use smallest information criterion like AIC and BIC to select the order of the model.

Keywords: Stationary, Non-stationary, Time series, Unit root test, ARIMA model, GARCH model

INTRODUCTION

Identification is the most important stage of all the stages of the modeling process as the other two directly depends on it. The identification attempt to provide a discernment of the characteristics of the stochastic of the series under investigation. Hypothesis testing regarding the generating mechanism, prediction of future values and related inquiries so that the general linear model assumption is achieved (MODEL, 2016). proposed AIC and BIC, Predictive Least Squares (PLS) and Sequentially Normalized Least Squares (SNLS) for model selection Stadntska. T (2008). in a study comparison of automated procedures for ARMA model identification reported that model identifications are more precise for big dependency processes; SCAN and ESACF are superior to MINIC for mixed (1,1) models; the positive effect of simple size is more pronounced for MINIC than for SCAN and ESACF, SCAN and ESACF tend to select higher order mixed structures in larger samples. Their conclusions are confined to stationary non-seasonal time series. The reported findings of their Monte Carlo experiments could help in choosing an appropriate identification procedure if some knowledge about properties of the stochastic process under study is available. The evaluated methods were superior to subjective judgments, for some models and parameterizations their accuracy remained disappointing. Moreover, precise model identification is not guaranteed, even in very large samples. The autoregressive integrated moving average (ARIMA) has been commonly used in the field of social, management and behavioral sciences, Fortes, M., & Delignieres. (2005). demonstrate the procedure for model selection in production system with random output via the use of Adjusted Coefficient of Determination (*R*2), Akaike and Schwarz criteria tools.

The main objective of this research is to determine the best order of some time series models such ARIMA and GARCH using simulation technique under different sample sizes. The rest of the article is organized as follows. Evaluation of the performance of the robust order selection using AIC and BIC is provided. Application of a real dataset in order to find the best order of the model in fitting and forecasting has also been provided. And finally, the conclusion is provided.

MATERIAL AND METHODS

Model Identification Procedures

It is very pertinent at this juncture to explain some basic concepts that are very prominent in time series modeling, they are as follows

Analetic Frame Work of the Study

An autoregressive model is simply a linear regression of the current value of the series against one more prior value of the series. The value of (p) is called the order of the AR model. AR models can be analyzed with one of various methods, including standard linear least squares techniques (see Cryer and Chan, 2008 and the references therein for more details).

Assume that a current value of the series is linearly upon its previous value, with some error. Then we could have linear relationship.

$$X_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + e_t \quad (1)$$

Where $\alpha_1, \alpha_2, \ldots \alpha_p$ are autoregressive parameter and $\boldsymbol{e_t}$ is a white noise process with zero mean and variance (σ^2).

Autoregressive are as their names suggest regressions on themselves. Specifically, apth-order autoregressive process (x_t) satisfied the equation 3.1.0. The current value of the series Y_t is a linear combination of the p most recent past values of itself plus an 'innovation' term e_t that incorporates everything new in the series at time t that is not explained by the past value. Thus, for every t, we assume that e_t is independent of $X_{t-1}, X_{t-2}, X_{t-3}, \ldots$

Consider now the pth- order autoregressive model

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \ldots \emptyset_p Y_{t-p} + e_t \quad (2)$$

With AR characteristic polynomial

$$\emptyset(x) = 1 - \emptyset_1 x - \emptyset_2 x^2 \ldots - \emptyset_p x^p \quad (3)$$

And the corresponding AR characteristic equation

$$1 - \emptyset_1 x - \emptyset_2 x^2 = 0 \quad (4)$$

$$1 - \emptyset_1 x - \emptyset_2 x^2 \ldots - \emptyset_p x^p = 0 \quad (5)$$

Assume that e_t be independent of $Y_{t-1}, Y_{t-2}, Y_{t-3}\ldots$ a stationary solution to the Equation exists if and only if the p roots of the AR characteristic equation each exceed 1 in absolute value (modulus). Other relationships between polynomial root and coefficient may be used to show that the following two inequalities are necessary, but not sufficient, that both

$$\emptyset_1 + \emptyset_2 + \ldots + \emptyset_p < 1 \quad (6)$$

$$|\emptyset_p| < 1$$

Assuming stationary and zero means, we may multiply Equation by Y_{t-k}, take expectations, divide by $\gamma\emptyset$, and obtain important recursive relationship

$$\rho_k = \emptyset_1 \rho_{k-1} + \emptyset_2 \rho_{k-2} + \ldots \emptyset_p \rho_{k-p} \ldots for\ k > 0 \quad (7)$$

Putting k= 1, 2…, and p into Equation (3.1.4) and *using* $\rho 0 = 1$ and $\rho - k = \rho k$, we get the general Yule-Walker equations

$$\rho_1 = \emptyset_1 + \emptyset_2 \rho_1 + \emptyset_3 \rho_2 + \ldots \emptyset_p \rho_{p-1} \quad (8)$$

$$\rho_2 = \emptyset_1 \rho_{-1} + \emptyset_2 \rho_0 + \emptyset_3 \rho_1 + \ldots \emptyset_p \rho_{p-2} \quad (9)$$

.

.
.

$$\rho_p = \emptyset_1\rho_{p-1} + \emptyset_2\rho_{p-2} + \emptyset_3\rho_{p-3} + \ldots\emptyset_p \quad (10)$$

Given numerical values for $\varphi_1, \varphi_2 \ldots, \varphi_p$, these linear equation can be solved to obtain numerical values for $\rho_1, \rho_2, \ldots, \rho_p$. Then Equation can be used to obtain numerical values for ρ_1 at any number of higher lags.
Nothing that

$$E(e_tY_t) = E(e_t(\emptyset_1Y_{t-1}+\emptyset_2Y_{t-2}+\ldots\emptyset_pY_{t-p} + e_t)) = E(e_t^2) = \sigma_e^2 \quad (11)$$

RESULT AND DISCUSSIONS

In this section, we have identified suitable order identification for two different time series models with one considering two different distributions with four different sample sizes in fitting and forecasting stationary and non-stationary data structures. A simulation study is performed to generate a stationary and non-stationary dataset.

Table 1: ADF unit root test with respect to the locations

Sample size	Test values	Lag order	p-value	Hypothesis	Decision	Remark
20	-6.215	2	0.01	Unit root	Reject (H_0)	Stationary
60	-5.525	3	0.01	Unit root	Reject (H_0)	Stationary
100	-4.458	4	0.01	Unit root	Reject (H_0)	Stationary
140	-5.037	5	0.01	Unit root	Reject (H_0)	Stationary

Table 1. Gives the ADF test for stationarity were presented in the table above. A stationary data was simulated from a normal distribution at different sample size of 20, 60, 100 and 140 to investigate empirically if they are stationary. The assumption of stationarity was confirmed in the simulated data. we can see that the unit root test for the simulated data with respect to the location at different sample sizes, the test reveals that the unit root doesn't exist for the data at every location therefore we reject the null hypothesis n conclude the data is stationary. The figures show the auto correlation and partial auto correlation for the data respectively.

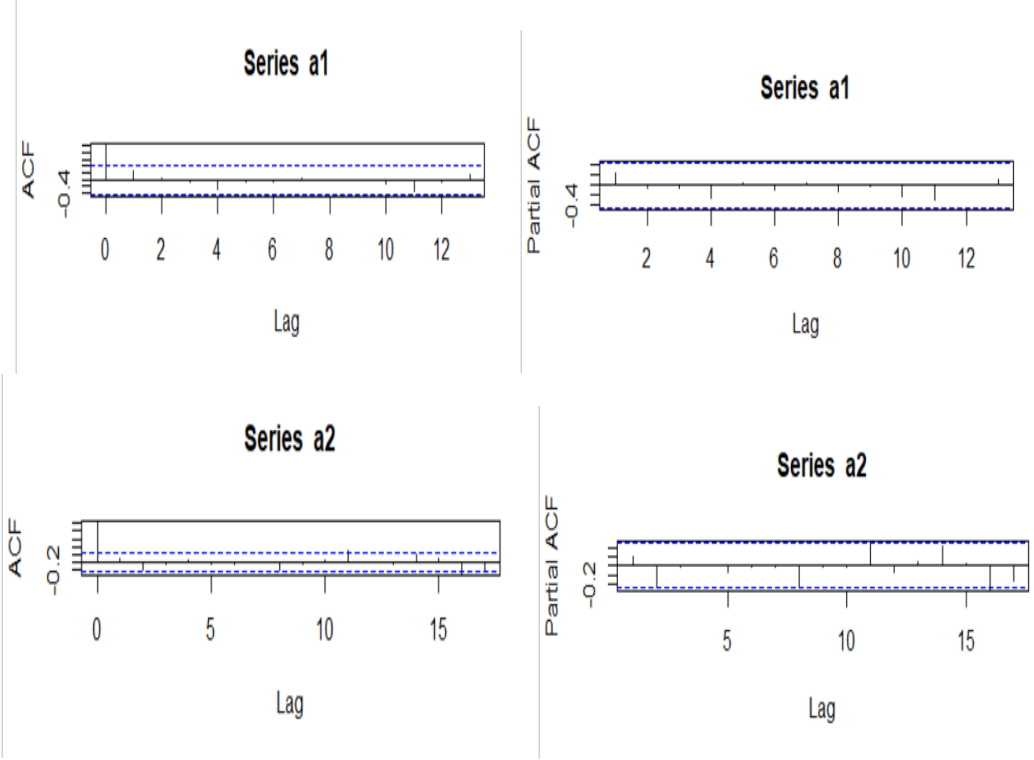

Figure 1: ACF and PACF plot for the simulated data

The figures above show the ACF and PACF behavior graphically at different sample sizes.

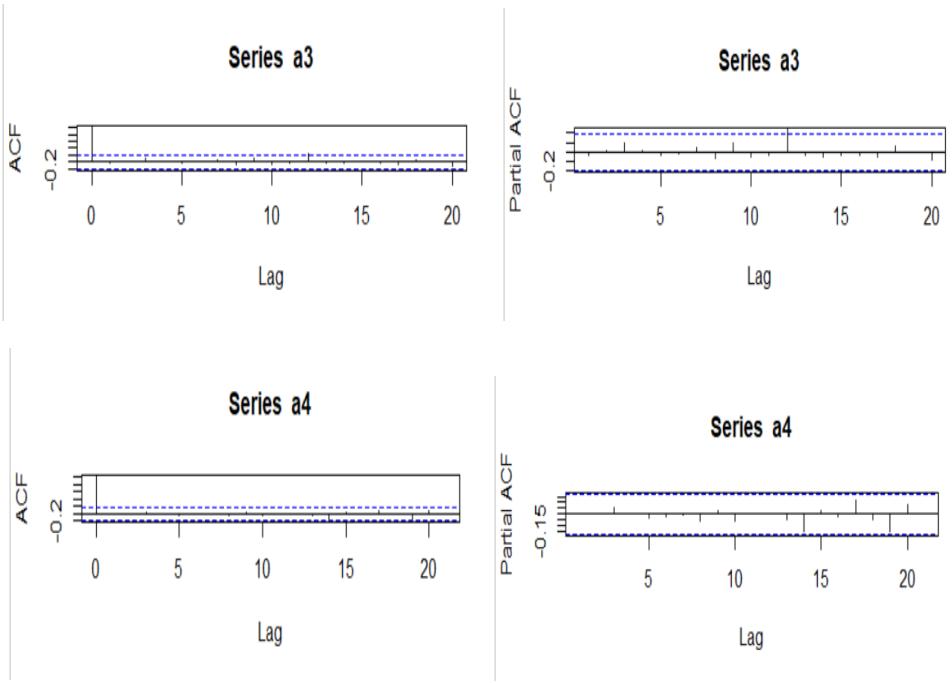

Figure 2: ACF and PACF plot for the simulated data

The figures above show the ACF and PACF behavior graphically at different sample sizes.

Table 2: AIC and BIC values for ARIMA (p,d,q) model

	AIC				BIC			
Sample size	ARIMA (1,1,1)	ARIMA (1,1,2)	ARIMA (2,1,1)	ARIMA (2,1,2)	ARIMA (1,1,1)	ARIMA (1,1,2)	ARIMA (2,1,1)	ARIMA (2,1,2)
20	**62.99**	**63.10**	**64.91**	**65.05**	**65.82**	**66.88**	**68.69**	**69.8**
60	178.87	178.26	178.32	180.31	185.9	186.57	186.63	190.69
100	293.34	295.37	295.29	296.83	301.13	305.75	305.67	309.81
140	396.73	397.17	398.71	400.61	405.52	408.9	410.45	415.28

Table 2: shows the AIC and BIC for the ARRIMA model at different sample sizes where the bolded one are the information criterion with least values. And we found the ARIMA (1,1,1) has the smallest value of all of the models therefore we selected the ARIMA (1,1,1) as a best model.

Table 3: AIC and BIC values for GARCH (p,q) model with GED

	AIC				BIC			
ample size	GARCH (1,1)	GARCH (1,2)	GARCH (2,1)	GARCH (2,2)	GARCH (1,1)	GARCH (1,2)	GARCH (2,1)	GARCH (2,2)
20	3.10	3.20	3.2	3.03	3.30	3.45	3.45	3.33
60	2.97	3.03	3.01	3.03	3.11	3.18	3.02	3.25
100	2.91	2.94	2.94	2.95	3.22	3.16	3.06	3.32
140	**2.92**	**2.90**	**2.93**	**2.96**	**2.92**	**2.93**	**2.95**	**2.98**

Table 3: above shows the goodness-of-fit for the GARCH (p,q) model where p,q=1,2. Four GARCH (p,q) models with the average values of AIC and BIC of 140 replications simulated from each model at various sample sizes. The bolded AIC and BIC are the criterion with minimum information values.

Table 4: shows the estimated parameters and diagnostic of GARCH (1,1)-GED model.

Parameters	**Generalized error distribution**	**p-values**
Ω	1.439e^-01	0.0374*
α_1	1.000e^-04	0.0472**
α_2	0.420181	0.0411*
β_1	1.000e^+00	<2e^-16***
β_2	3.6338	0.0002**

ARCH(1)- LM test	5.893	0.0012
Q^2 (15)	27.6709	0.0037

Note: (*), (**) and (***) denote significance at 1%, 5% and 10% respectively.

Table 4. Shows the diagnostic check for the GARCH (1,1)-GED model from the table we found that all the values are significant at its significance level i.e., 1%, 5%, and 10% represent (*), (**) and (***) respectively.

Table 5: AIC and BIC values for GARCH (p,q) model with StD

	AIC				BIC			
sample size	GARCH (1,1)	GARCH (1,2)	GARCH (2,1)	GARCH (2,2)	GARCH (1,1)	GARCH (1,2)	GARCH (2,1)	GARCH (2,2)
20	3.29	3.39	3.85	3.48	3.49	3.63	3.63	3.78
60	2.99	3.03	3.03	3.07	3.13	3.21	3.21	3.27
100	2.91	2.93	2.93	2.95	3.02	3.06	3.06	3.11
140	**2.81**	**2.83**	**2.83**	**2.84**	**2.90**	**2.93**	**2.93**	**2.96**

Table 5: above shows the goodness-of-fit for the GARCH (p,q) model where p,q=1,2. Four GARCH (p,q) models with the average values of AIC and BIC of 140 replications simulated from each model at various sample sizes. The bolded AIC and BIC are the criterion with minimum information values.

Table 6: shows the estimated parameters and diagnostic of GARCH (1,1)-StD model.

Parameters	**Generalized error distribution**	**p-values**
Ω	1.739e^-01	0.0271*
α_1	1.230e^-06	0.0322*
α_2	0.650181	0.0021*
β_1	1.10e^+04	<2e^-16***
β_2	4.6365	0.0044**
GARCH(1)- LM test	6.497	0.0065
Q^2 (15)	22.071	0.0143

Note: (*), (**) and (***) denote significance at 1%, 5% and 10% respectively.

Table 6 shows the diagnostic check for the selected model i.e. GARCH (1,1) StD and we found that all its parameters are significant at its level.

Volatility Fitting and forecasting of the Selected Model with real data set

Table 7: Stationarity Test for the Data

Difference	**Test values**	**Lag order**	**p-value**	**Hypothesis**	**Decision**	**Remark**
0	-2.8908	3	0.2272	Unit root	Reject (H_1)	Not Stationary
1st	-2.7739	3	0.2732	Unit root	Reject (H_1)	Not Stationary
2nd	-3.6421	3	0.0447	Unit root	Reject (H_0)	Stationary

Table 7. Shows a stationarity check for the real data set, we found that the data is not stationary without the differencing and at the first difference ADF R-value is 0.2272 and 0.2732 respectively but after the 2nd difference the data turn to stationary therefore we will proceed with the analysis at the second difference.

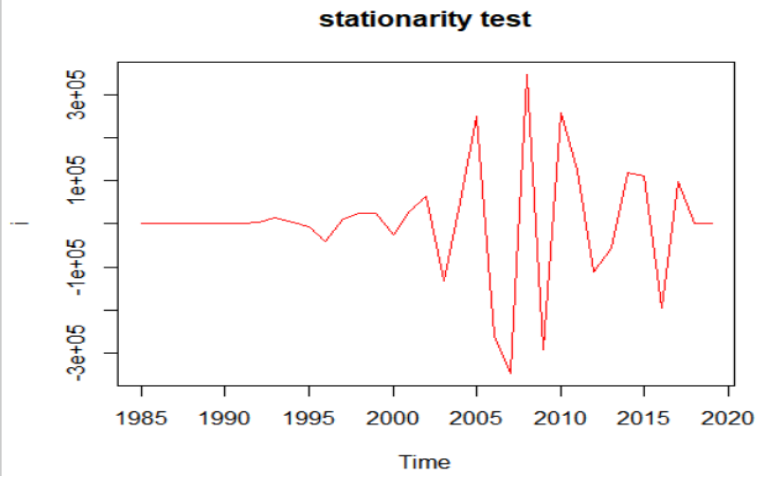

Figure 3: ACF and PACF plot for the Nigerian Meteorological Agency data

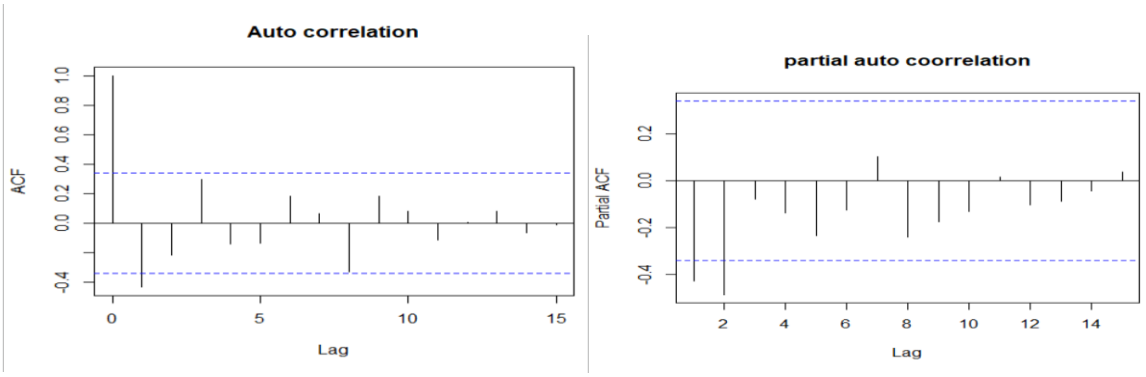

Figure 4: ACF and PACF plot for the Nigerian Meteorological Agency data

Table 8: Estimated parameters for the selected models for ARIMA GARCH with generalized normal Distribution and Student t-Distribution

Model	Parameters				Information criterion	
	μ	α_1	β_1	p-values	AIC	BIC
ARIMA (1,1,1)	119.09+05	-0.406	-1.000	0.0436	854.5	858.9
GARCH(1,1)-GED	1.599e+02	7.448e-01	1.830e-07	0.0005	9.082	9.188
GARCH(1,1)-StD	1.600e+02	9.0480e-01	1.875e-02	0.0030	9.068	9.099

Table 8 is the estimated parameter and the information criteria for the selected models.
Where ARIMA (1,2,1) has a p-value of 0.043 GARCH (1,1)-N 0.0206, GARCH (1,1)-GED is 0.0005, GARCH (1,1)-StD is 0.0030 and EGARCH (1,1) 0.0208 respectively. Therefore, we fail to reject the null hypothesis in favor of alternative for all of the models.

Table 9 Diagnostic check of the selected model of the real dataset

Model	ARIMA (1,1)		GARCH(1,1)-GED		GARCH(1,1)-StD	
Test statistic	**Values**	**P-value**	**Values**	**P-value**	**Values**	**P-value**
Jarque-Bera Test	12.20	0.007	0.995	0.020	1.474	0.079
Shapiro-Wilk Test	2.800	0.002	0.042	0.000	0.801	0.000
Ljung-Box Test (R^2)	12.55	0.024	14.15	0.003	13.782	0.003
LM Arch Test	15.36	0.223	15.38	0.012	20.298	0.061

Table 9 is a diagnostic check for the validation of the above table and we found that all the model is validated with the p-values i.e. they are less than the critical values. And the figure below is the forecasted model for the real data set where the shaded zone is the forecasted zone and the line is the fitted one.

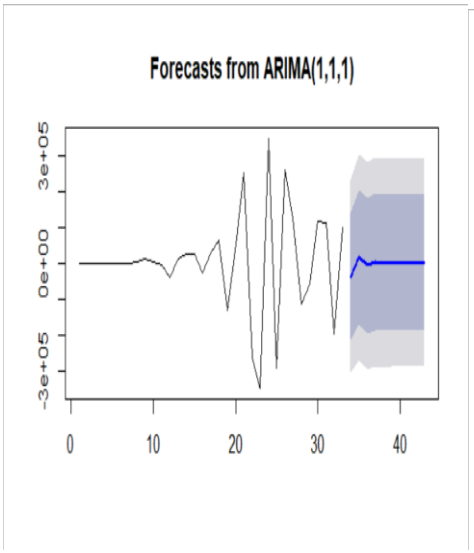

Figure 5: ARIMA (1,1,1) forecast

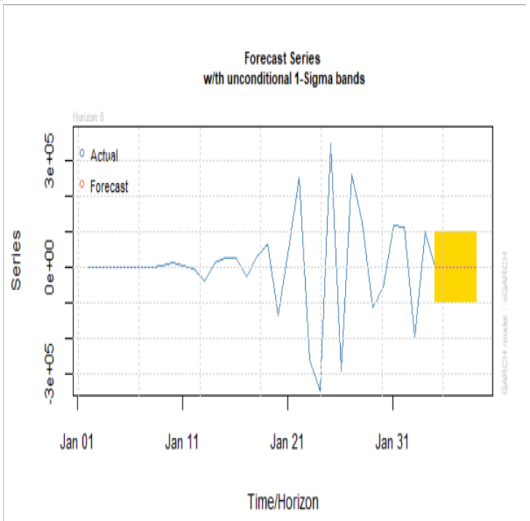

Figure 6: GARCH(1,1)-GED forecast

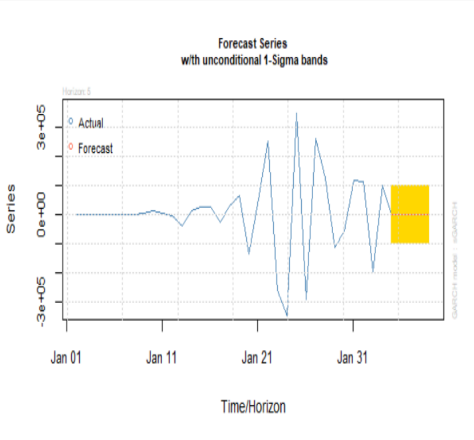

Figure 7: GARCH(1,1)-StD forecast

CONCLUSION

This study establishes the procedure for selecting the order of the model in order to get a good result in fitting and forecasting a time series data, the procedure was tested using a simulation with different sample size and validated with the diagnostic check after, the procedure was used to fit and forecast the real data set, which give us a good result. Indeed, this procedure is the best for selecting a time series order of a model.

RECOMMENDATION

Based on this study, we recommend the following procedure

i. Use smallest Information Criteria to select the order like AIC and BIC
ii. The coefficients must be significant
iii. Use LJuag-Box Q^2 statistic and Lm test as the diagnostic test.

REFERENCES

Fortes, M., Ninot, G., &Delignieres, D. (2005). The Auto-regressive Integrated Moving Average Procedures Implications for adapted physical activity research. Adapted Physical Activity Quarterly, 22, 221-236.

S. Matta L.H.R. speech 2016: continuity from prelinguistic communication to later language abilit a follow up study from infancy to early stage (pp. 581-606). New York: Wiley

Stadntska T. Braurr&werner J. (2008) comparison of automated procedures for ARMA model identification behavior research method 2008, 40(1) 250-262

Velicer, W.F., & Colby, S. M (1997). Time series analysis for prevention and treatment research. In K. J. Bryant, M windle, & S.G west (Eds), the science of prevention research (pp. 221-249) American psychology association

Velicer, W. F.,& Colby, S. M. (1997). Time series analysis for prevention and treatment research. In K. J. Bryant, M. Windle, & S.G. West (Eds.), The science of prevention: Methodological advances firm alcohol and substance abuse research (pp. 211-249). Washington, DC: American Psychological Association.

Velicer, W. F., & Fava, J. L. (2003), Time Series analysis. In J. Schinka & WE Velicer (Eds,), Research methods in psychology (pp. 581-606). New York: Wiley.