World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Detecting overfitting by examining residuals in autoregressive models

Mohammed M. Elnazali *, Aisha A. Salem and Tarek A. Elghazali

University of Benghazi, Department of Statistics, Faculty of Science, Benghazi, Libya.

World Journal of Advanced Research and Reviews, 2024, 24(02), 1162–1174

Publication history: Received on 24 September 2024; revised on 09 November 2024; accepted on 11 November 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.2.3378

Abstract

The aim of this study is to see how overfitting can be detected using non-rigorous analysis of residuals. The well-known statistical packages *R* was used to simulate data from stationary autoregressive models with Gaussian white noise with mean 0 and variance one. In order to see the effect of realization size on our findings, the sample size 50 was used as an example of small realization and the sample size 500 was used as an example of large realization. The method of maximum likelihood was used in the fitting of autoregressive models to the simulated data which is available in the statistical package R. Interesting and promising results were obtained. Our study seems to suggest that comparing estimates with their standard errors is the only reliable criterion in spotting or detecting overfitting. To make sure that the defect in the behavior of the residuals is due only to the over, we used only the same class of models in the simulation and the fitting.

Keywords: Autoregressive Models; Overfitting Problem; Analysis of Residuals; Simulation.

1. Introduction

1.1. Background

When analyzing data with a statistical model, it's important to check if the model accurately represents the data. Evaluating the autocorrelation of residuals can help determine if the model needs adjustments to capture underlying patterns in the data. This could involve adding variables, changing the model's form, or using more advanced techniques. Assessing the impact of changes on model performance is crucial (1). Overfitting in autoregressive models occurs when the model becomes too complex and fits the training data too closely. This leads to poor performance on new, unseen data. Autoregressive models are commonly used in time series analysis to capture the relationship between an observation and a lagged version of itself. However, if these models become too flexible and try to capture every small fluctuation in the training data, they struggle to generalize well to new data. This phenomenon of overfitting presents a significant challenge in effectively using autoregressive models for predictive tasks (2–4).

1.2. Autoregressive Models

A process X_t is an autoregressive process of order p, also known as an AR (p) process if it satisfies the difference equation,

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \,, \qquad (1)$$

where $\alpha_1, \alpha_2, \ldots, \alpha_p$ are parameters, and ε_t is a purely random process with mean zero and variance σ_ε^2. It is similar to a multiple regression model, but X_t is regressed on past values of X_t instead of separate predictor variables. This is why

* Corresponding author: Mohammed M. Elnazali

it is called 'auto'. Autoregressive time series models are frequently used in finance, economics, and engineering to forecast future values based on past observations (5).

1.3. White Noise Process

The process $\{X_t\}, t = 0, \pm1, \pm2, \ldots$, is termed a purely random process if it comprises uncorrelated random variables, i.e., if for all $s \neq t, cov\{X_s, X_t\} = 0$. This is the most basic of all discrete parameter models and relates to the scenario where the process exhibits "no memory," meaning the process value at time t is uncorrelated with all previous values up to time $(t-1)$ (and, indeed, with all future values of the process) (5,6).

2. Material and Methods

The main objective of this study was to detect overfitting in AR models by utilizing rigorous residuals analysis. In order to conduct the research, we simulated the data using the R language from stationary autoregressive models with Gaussian white noise that had a mean of zero and a variance of one. We studied the effect of different sample sizes on the realization of the models by using two samples of 50 and 500 for small and large sizes, respectively. Our study models were fitted using the maximum likelihood method. We found some promising and interesting results. The study suggests that comparing estimates with their standard errors is the only reliable criterion to detect overfitting. To ensure that the defect in the residuals' behavior is solely due to overfitting, we used the same class of models in the simulation and the fitting.

2.1. A Simulation Study

Here we present our element of research. Our study consists of simulating data from a stationary time series models of certain order and then fitting a model from the same class of higher order to the simulated data and see if we can detect the over fitting by examining the residuals calculated from the fitted model. The celebrated statistical package *R* were used to simulate the data of the study, moreover the simulation commands used by Chan & Cryer (2008)(7) and Shumway & Stoffer (2019)(8) were of great benefit to us in the generation of our data from different stationary AR models. It is worth mentioning that in the simulation process the white noise was taken to be a Gaussian with mean 0 and variance 1. Some simulated data with simulations commands used in the study are given in appendix one. In the diagnostic checking stage, we only relay mainly on the followings:

- Plot of the residuals and their descriptions (to see if there is a trend or non-stability in the variance or outliers).
- Plot of autocorrelation function of the residuals (to test independence).
- Comparing estimates with their standard errors.
- Shapiro test for normality.
- Plot the p-values of Ljung-Box test for independence.

In addition, our simulation study consisted of two cases. The first case (3.1 and 3.2) simulated AR (1) with two sample sizes (50 and 500), while the second case (3.3 and 3.4) simulated AR (2).

3. Results and Discussion

3.1. Simulation from AR (1) Model (Small Sample N=50)

Here we consider simulation from the model

$$X_t = 0.4X_{t-1} + e_t\,, \qquad (2)$$

where, e_t is Gaussian white noise with mean 0 and variance 1.

A realization of size 50 simulated observations from Model (2) is plotted in Fig. 1.

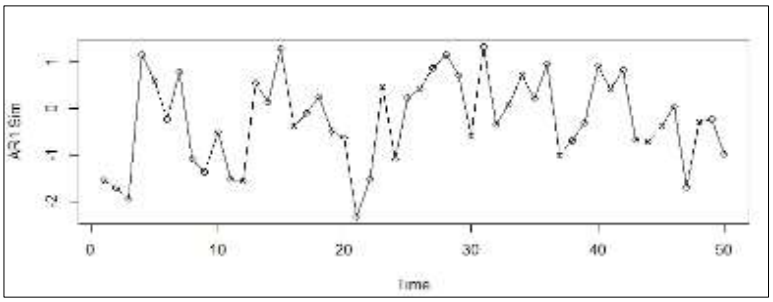

Figure 1 50 observations generated from Model (2)

Then an AR (1) model was refitted to 50 observations from the above model (small sample); the refitted model turned out to be

$$X_t = 0.3375(0.139)X_{t-1} + \hat{e}_t. \quad (3)$$

The residual variance estimate $\hat{\sigma}_e^2 = 0.7722$, $AIC = 133.09$, where the quantity given in the bracket is the standard error of the estimate.

Fitting AR (2) Model: Here we fit autoregressive model of order 2 (AR (2)) to a realization of size 50 simulated from Model (2), the fitted model turned out to be

$$X_t = 0.3467(0.1482)X_{t-1} - 0.0276(0.1485)X_{t-2} + \hat{e}_t, \quad (4)$$

with estimated residual variance = $\hat{\sigma}_e^2 = 0.7716$ is not close to true value used in the simulation and Akaike's information criterion $(AIC) = 135.05$ slightly larger than that of the refitted AR (1) model.

The plot of the standardize residuals calculated from Model (4) and their summary statistics are given in Fig. 2 and Table 1 respectively.

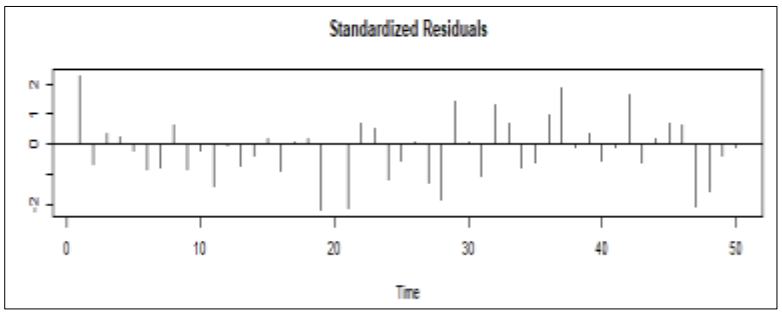

Figure 2 The standardized residuals calculated from Model (4)

It seems that there is no apparent trend and the values are within the expected values of standard normal, however, there is increase in the variance starting from the middle of series to its end.

Table 1 Summary statistics of standardized residuals calculated from Model (4)

Min	1st Qu	Median	Mean	3rd Qu	Max
-1.92302	-0.66325	-0.07736	-0.15358	0.31147	2.00352

Since the mean is smaller than the median the distribution of the residuals is slightly skewed to the left. To test the independence of the white noise, the plot of the sample autocorrelation of residuals (ACF) calculated from Model (4) is given in Fig. 3.

(a)

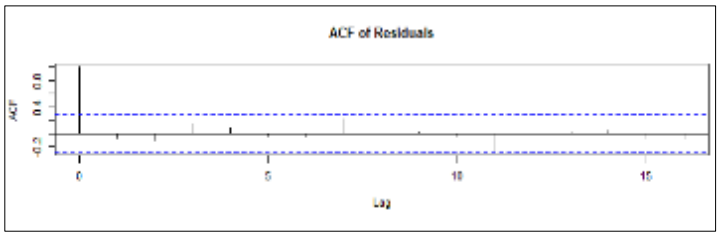

Figure 3 Sample autocorrelation function of standardized residuals calculated from Model (4)

It seems that the sample autocorrelations of the residuals of the fitted AR(2) model indicate that the residuals of the fitted AR(2) model have a small autocorrelation and lie in the interval $(-0.283, 0.283)$ is not significant, except at lag 11, which consistent with a realization of a white noise.

Further test for the independence of errors is given by Ljung-Box test statistic. The p-values for the Ljung-Box test statistics for a whole of values of lags K are given in Fig. 4.

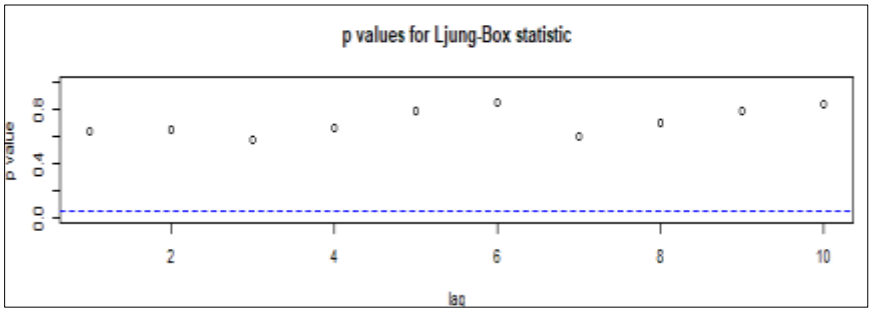

Figure 4 p-values for Ljung-Box statistic

Most of the p-values are greater than 0.05 which means that there is no reason to reject the independence of the residuals.

The normality of the residuals was tested using Shapiro-Wilk test which gave value of W=0.9882, and p-value = 0.8064 which means that the normality of the residuals is not rejected at 0.05 level of significance.

Fitting AR (3) Model: Here we fit an AR (3) model to 50 observations simulated from Model (2); the fitted model turned out to be

$$X_t = 0.3578(0.1439)X_{t-1} + 0.1063(0.1513)X_{t-2} + 0.2519(0.1475)X_{t-3} + \hat{e}_t. \quad (5)$$

The estimated residual variance equal $\hat{\sigma}_e^2 = 0.7266$ and $AIC = 134.25$. The first and second parameter estimates are not significant if we take into consideration their standard errors.

The standardized residuals calculated from Model (5) and their summary statistics are shown in Fig. 5 and Table 2 respectively

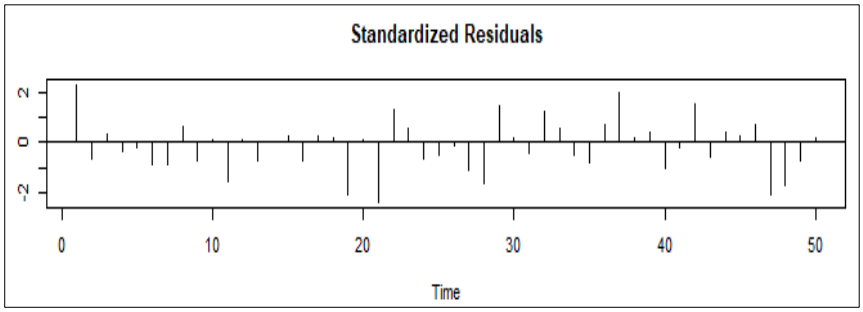

Figure 5 Standardized residuals calculated from Model (5)

There is no apparent trend or outliers, however it seems that the variance is not constant, small variance in the beginning and large variance in the end of the series.

Table 2 Summary statistics of standardized residuals calculated from Model (5)

Min	1st Qu	Median	Mean	3rd Qu	Max
-2.05252	-0.62489	-0.06047	-0.13229	0.29355	1.93182

The values within the range $(-3, 3)$ which consistent with the standard normal, the mean less than the median indicates small skewness to the left.

For the sake of testing the independence of errors, we plot the autocorrelation function ACF of the residuals in Fig. 6 and p-values of Ljung-Box test for a whole range of lag K values in Fig. 7. The horizontal dashed line at 5% helps judge the size of the p-values.

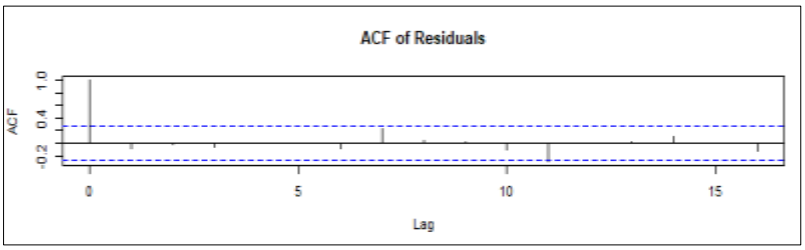

Figure 6 Sample autocorrelation function of standardized residuals calculated from Model (5)

Most autocorrelations are small and lie in interval $(-0.283, 283)$ which indicates independence.

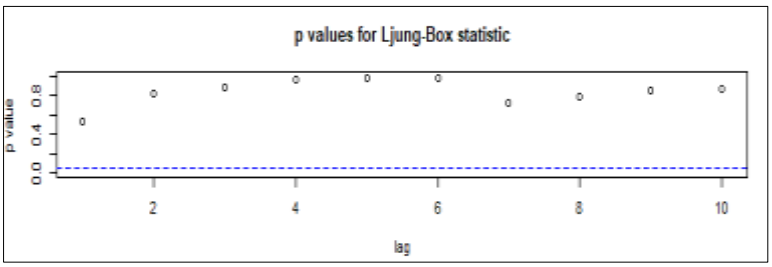

Figure 7 p-values for Ljung-Box statistic

It is evident that all $p - values$ for the Ljung-Box statistics are greater than 0.05 which means that there is no reason to reject the independence assumption of errors at 0.05 level of significance.

Finally, the normality of residuals was tested using Shapiro-Wilk normality test. The test gave a value of 0.98267 corresponds to a $p - value$ of $0.6688 > 0.05$ which means normality of residuals cannot be rejected at 0.05 level of significance.

The sample size used in above study was 50 observations, to see if over-fitting is more detectable for large samples than small samples, we repeated exactly the same study but this time sample size 500 is used.

3.2. Simulation from AR (1) Model (Large Sample N=500)

First AR (1) model was refitted to 500 observations simulated from Model (2) the aim to check the validity of our estimation and to use it for comparison with models of higher order; the refitted model turned out to be

$$X_t = 0.3884(0.0412)X_{t-1} + \hat{e}_t\,, \qquad (6)$$

with residuals variance $\hat{\sigma}_e^2 = 0.9459$ and $AIC = 1397.17$. Note that both the parameter estimate and residual variance estimate are close to the values used in the simulation and this is maybe due to the large sample size used in the fitting.

Fitting AR (2) Model: Here an AR (2) model was fitted to a realization of size 500 observations simulated from Model (2); the fitted model turned out to be

$$X_t = 0.3470(0.044)X_{t-1} + 0.1071(0.0445)X_{t-2} + \hat{e}_t. \quad (7)$$

The value of the estimated residual variance $\hat{\sigma}_e^2 = 0.9386$ and AIC=1393.42.

It is evident that all estimates are significant with small standard errors. In fact, the model is as good as the first order model and its AIC is close in value to the one obtained in the case of AR (1).

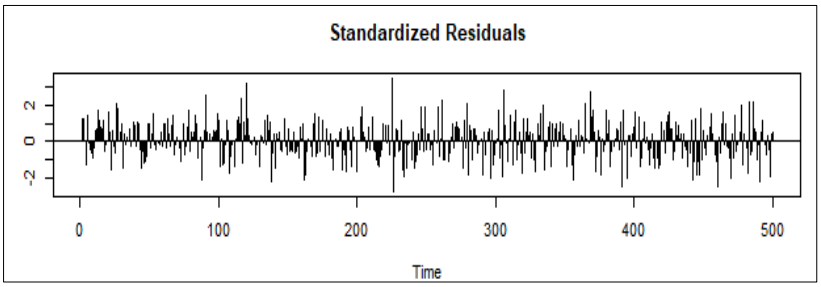

Figure 8 Standardized residuals calculated from Model (7)

No trend exists in the residuals and the values alternate about the mean value 0.

Table 3 Summary statistics of standardized residuals calculated from Model (7)

Min	1st Qu	Median	Mean	3rd Qu	Max
-2.65633	-0.59816	0.04696	0.04736	0.69090	3.30015

The values lie in the range $(-3, 3)$ which is consistent with the standard normal distribution, moreover the closeness between the values of mean and the median indicate both the nonexistence of outliers and symmetry which supports the normality of residuals.

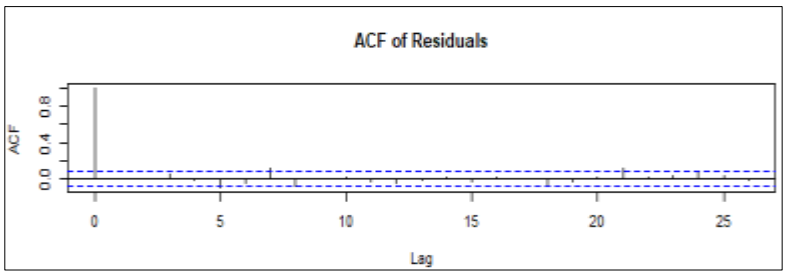

Figure 9 Sample autocorrelation function of standardized residuals calculated from Model (7)

Most correlations are within the interval $(-0.089, 0.089)$ which supports the assumption that errors are independent. To confirm the independence of errors, Box-Ljung was applied for whole range of lag values, and the p-values are plotted for different values of lags in Fig. 10.

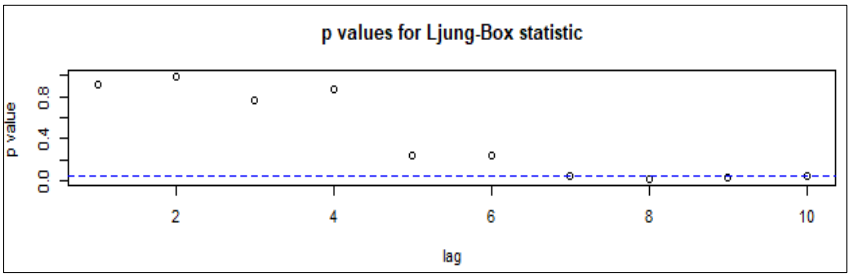

Figure 10 p-values of Ljung-Box statistic

Most p-values exceed 0.05 which indicate that the assumption of errors independence cannot be rejected.

Finally, the normality of residuals was tested using Shapiro-Wilk normality test at 5% level of significance, its value turned out to be $W = 0.99786$ with computed $p - value = 0.7852 > 0.05$ which means that normality of residuals cannot be rejected.

Fitting AR (3) Model: In the case of fitting AR (3) to 500 observations generated from Model (2); the fitted model turned out to be

$$X_t = 0.3446(0.0447)X_{t-1} + 0.0995(0.0471)X_{t-2} + 0.0220(0.0449)X_{t-3} + \hat{e}_t. \quad (8)$$

The estimated residual variance $\hat{\sigma}_e^2 = 0.9381$, $AIC = 1399.18$.

When we compare Model (8) with the fitted AR (1) Model (6), it is evident that the estimate of the first parameter has not changed especially when the size of the standard error is taken into consideration. In addition, the estimate of the second and third parameters are statistically not different from zero. The AIC of Model (8) is larger than that of AR (2) Model (6).

The standardized residuals calculated from Model (8) are plotted in Fig. 10 and their summary statistics in Table 4.

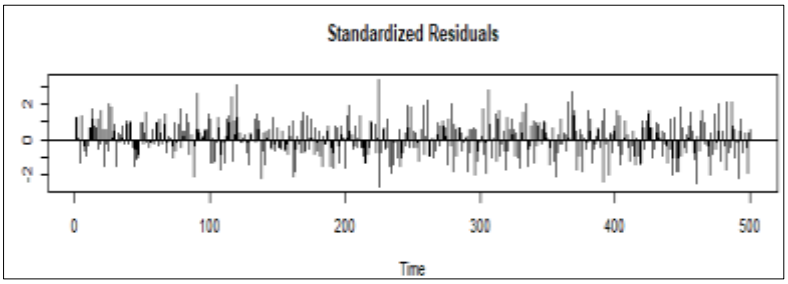

Figure 10 Standardized residuals calculated from Model (8)

It is evident that there is no trend or outliers in the residuals, which means the nonexistence of autocorrelation among them, however, it appears that there is no stability in the variance.

Table 4 Summary statistics of standardized residuals calculated from Model (8)

Min	1st Qu	Median	Mean	3rd Qu	Max
-2.64112	-0.59038	0.04696	0.04618	0.67398	3.30015

The closeness in values between mean and median is an indicator of the symmetry of the distribution and the nonexistence of outliers in the residuals.

The sample autocorrelations of residuals are plotted in Fig. 11 and p values of Ljung-Box test statistic in Fig. 12.

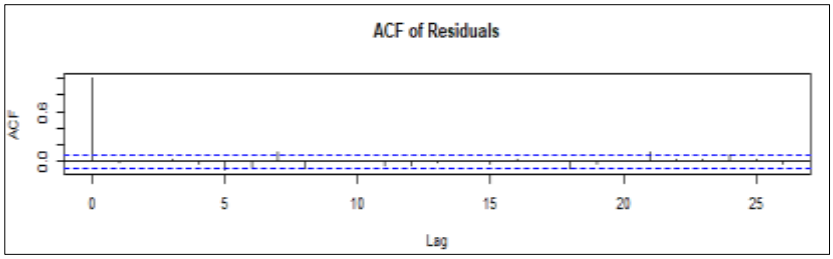

Figure 11 Sample autocorrelation function of standardized residuals calculated from Model (8)

Fig. 11 shows a significant autocorrelation indicate that residuals are not approximate white noise, which suggests that the fitted model lacks aptness.

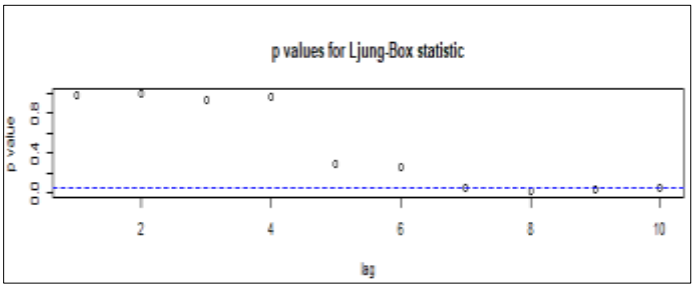

Figure 12 p-values of Ljung-Box statistic

Fig. 12 and the Ljung-Box test of this model gives a chi-square value of 17.715 with 9 degrees of freedom leading to a $p-value$ of $0.01333 < 0.05$ (significant), the assumption of errors independence is rejected at 0.05 level of significance.

3.3. Simulation from AR (2) Model (Small Sample N=50)

Here we simulate a realization of size 50 observations from the stationary AR (2) model

$$X_t = 0.4X_{t-1} - 0.7X_{t-2} + e_t\,, \tag{9}$$

where again the process $\{e_t\}$ is Gaussian process with mean 0 and variance 1.

A realization of size 50 simulated from Model (9) is shown in Fig. 13.

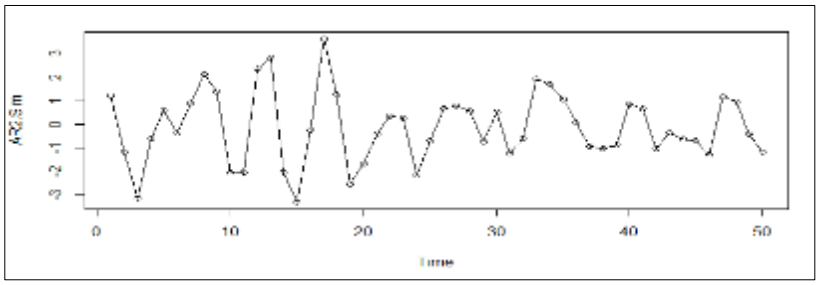

Figure 13 A realization of size 50 simulated from Model (9)

To validate the accuracy of our simulation an AR (2) model was refitted to the above realization, the refitted model turned out to be

$$X_t = 0.4134(0.0996)X_{t-1} - 0.7064(0.094)X_{t-2} + \hat{e}_t \tag{10}$$

The estimated residuals variance $\hat{\sigma}_e^2 = 1.006$ and $AIC = 149.63$. Note that the parameter estimates are very close to the true parameter values and the estimated residuals variance is close to 1, which indicates the reliability of our simulation despite the small sample size.

Fitting AR (3) Model: Here we fit an AR (3) model to 50 observations simulated from Model (9); the fitted model turned out to be

$$X_t = 0.4592(0.142)X_{t-1} - 0.7326(0.1114)X_{t-2} + 0.0655(0.1449)X_{t-3} + \hat{e}_t. \tag{11}$$

The estimated residuals variance $\hat{\sigma}_e^2 = 1.002$ and $AIC = 151.43$.

When we compare these results with the fitted AR (2) model, we see that the estimates of the first and second coefficient have changed very little, especially when the size of the standard errors is taken into consideration. In addition, the estimate of the new third parameter, is not statistically different from zero. The estimate of the residual's variance has not changed much while the AIC has actually increased.

The standardized residuals calculated from Model (11) are plotted in Fig. 14 and their summary statistics in Table 5.

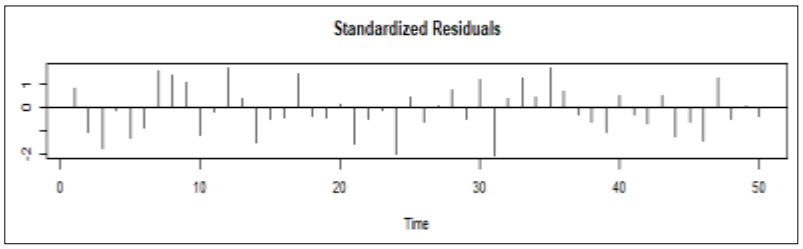

Figure 14 Standardized residuals calculated from Mode (11)

The plot reveals no trend and no outliers and the variance is nearly stable.

Table 5 Summary statistics of standardized residuals calculated from Model (11).

Min	1st Qu	Median	Mean	3rd Qu	Max
-2.57026	-0.62220	0.04658	-0.01365	0.62862	3.09086

To investigate the independence of errors, the sample autocorrelation of residuals is plotted in Fig. 15 and the p-values for Ljung-box statistic is plotted for different values of lags in Fig. 16.

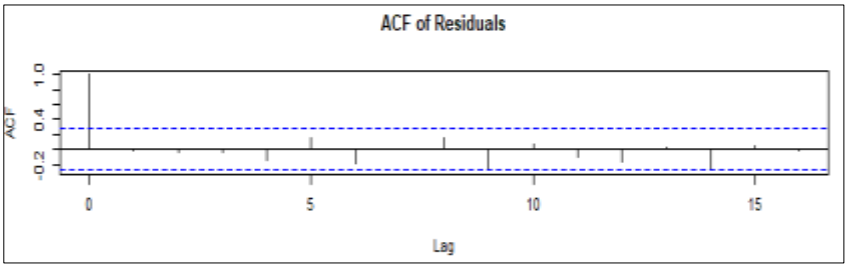

Figure 15 Sample autocorrelation of standardized residuals calculated from Model (11)

Most of the autocorrelations are very small and lie in the interval $(-0.283, 0.283)$ which indicate the independence of the errors.

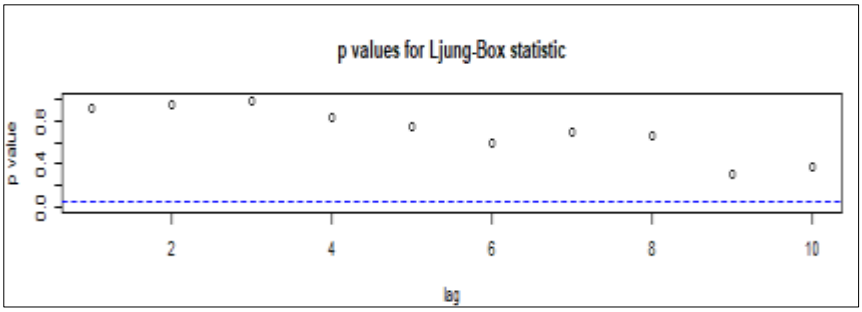

Figure 16 p-values for Ljung-Box statistic

Most of the $p-values$ are greater than 0.05 which means the assumption of error independence cannot be rejected at 0.05 level of significance.

As far as the normality of the errors is concerned the Shapiro-Wilk normality test gave a value of 0.97302 which corresponds to a p-value of $0.3059 > 0.05$ which means that the normality of errors is not rejected at 0.05 level of significance.

Fitting AR (4) Model: An AR (4) model was fitted to 50 observations simulated from Model (9), the fitted model is given by

$$X_t = 0.4615(0.1424)X_{t-1} - 0.7574(0.1528)X_{t-2} + 0.0805(0.1581)X_{t-3} - 0.0342(0.1441)X_{t-4} + \hat{e}_t. \quad (12)$$

The estimated residuals variance $\hat{\sigma}_e^2 = 1$ and $AIC = 153.37$. Again, when we compare Model (12) with the refitted AR (2) Model (10) we see that the first and second coefficients have changed very little, especially when the size of the standard errors is taken into consideration. Moreover, the estimate of the third and fourth coefficients is not statistically different from zero. Note also that the AIC has actually increased than AR (2) and AR (3).

Visual examination of the standardized residuals calculated from Model (12) plotted in Fig. 17, we can see that there is no trend or outliers and their values are consistent with the values of the standard normal distribution.

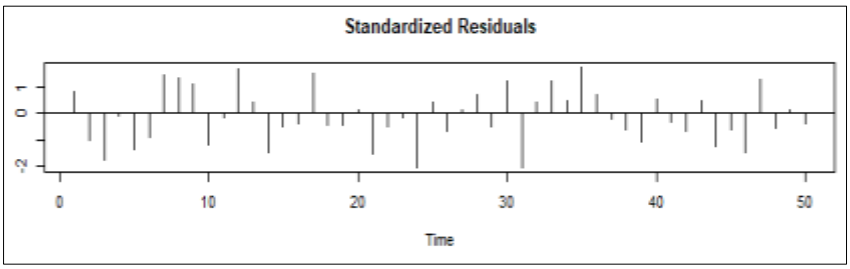

Figure 17 Standardized residuals calculated from Model (12)

Both the autocorrelation function of the residuals (see, Fig. 18) and values (see, Fig. 19) indicate the independence of errors where autocorrelations are small and lie in the interval $(-0.283, 0.283)$ and the p-values for the Ljung-Box test statistic are larger 0.05.

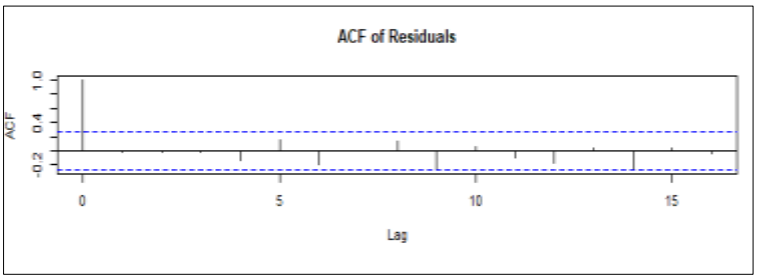

Figure 18 Sample autocorrelation of standardized residuals calculated from Model (12)

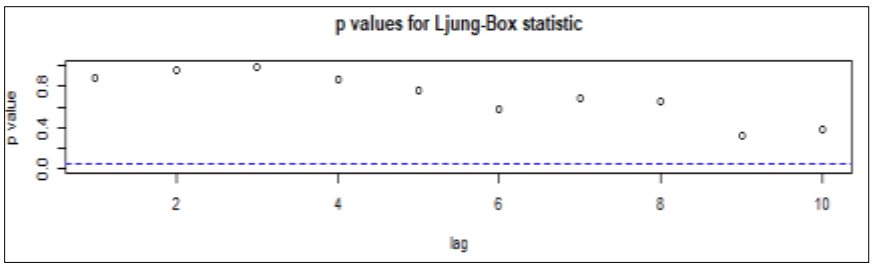

Figure 19 p values for Ljung-Box statistic

3.4. Simulation from AR (2) Model (Large Sample N=500)

Our steps in sub-section (3.3) was repeated using this time sample size 500. First an AR (2) was fitted to 500 observations simulated from Model (9); the fitted model turned out to be

$$X_t = 0.3678(0.0319)X_{t-1} - 0.7022(0.0319)X_{t-2} + \hat{e}_t. \quad (13)$$

The residual variance estimate $\hat{\sigma}_e^2 = 0.9557$, $AIC = 1403.7$. Note that due to the large sample effect standard errors and AIC are smaller than in the case of sample size 50. Again, this result confirms the validity and reliability of our simulation.

Fitting AR (3) model: Here an AR (3) model was fitted to 500 observations generated from the stationary AR (2) Model (9); the fitted model turned out to be

$$X_t = 0.3493(0.0447)X_{t-1} - 0.6926(0.0359)X_{t-2} - 0.0263(0.0448)X_{t-3} + \hat{e}_t \quad (14)$$

The standardized residuals calculated from the Model (14) are plotted in Fig. 20 and their summary is given in Table 6.

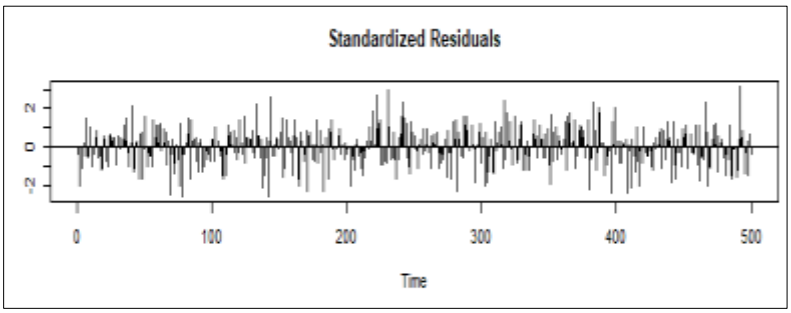

Figure 20 Standardized residuals calculated from Model (14)

Table 6 Summary statistics of standardized residuals calculated from Model (14).

Min	1st Qu	Median	Mean	3rd Qu	Max
-2.010559	-0.627872	0.019024	-0.004471	0.630971	1.953315

For the sake of testing independence, the autocorrelation function of residuals and the $p-values$ for Ljung-Box statistic are plotted in Figs. 21 and 22, respectively.

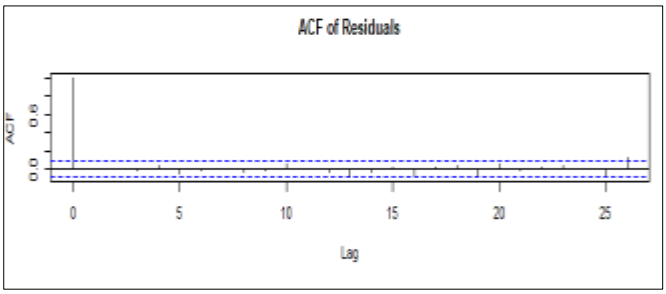

Figure 21 Sample autocorrelations function of standardized residuals calculated from Model (14).

Most autocorrelations lie within the interval $(-0.0894, 0.0894)$ except perhaps at lags 5, 7 and 21.

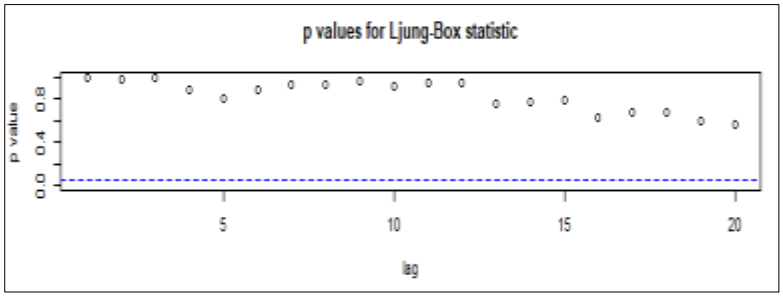

Figure 22 p-values of Ljung-Box test statistic

All p-values are greater than 0.05 which means no justification for rejecting errors independence.

The Shapiro-Wilk normality test applied to the residuals produces a test statistic of $W = 0.996$, which corresponds to a p-value of 0.2747, and we would not reject normality based on this test.

In short, all the remarks said about our results in the case of sample size 50 almost apply in the case of sample size 500.

Fitting AR (4) Model: Here we fit AR (4) model to a realization of size 500 observations simulated from Model (9). The fitted model turned out to be

$$X_t = 0.3495(0.0447)X_{t-1} - 0.6871(0.0475)X_{t-2} - 0.0291(0.0475)X_{t-3} + 0.0079(0.0450)X_{t-4} + \hat{e}_t. \qquad (15)$$

The residual variance estimate $\hat{\sigma}_t^2 = 0.955$, $AIC = 1407.32$.

It is clear that the last two coefficients are not statistically significant if we take in consideration their standard errors. The residual variance estimate is close to the value used in the estimate. The AIC is large that that of the true model AR (2) but smaller than that of Model (12) and this may be to the effect of large sample size.

4. Conclusion

In addition to our marginal comments on chapter three, we conclude with the following general remarks:

- The only crucial and reliable criterion in indicating the possibility of overfitting is that of comparing estimates with their standard errors. This criterion never failed in pointing out the non-aptness of the fitted model, where we can see that all the extra parameters are non-significant.
- Spotting overfitting is affected by the realization size, where we see that overfitting is more apparent in the size 50 than the size 500.
- Overfitting is more easily detected in cases where the difference in order between the true model and the refitted model is appreciable.
- It seems that Box-Ljung test rarely gives significant results (reject the independence assumption of residuals).
- It appears that there is no much difference between the values of the residual variances of the true models used for simulation and the refitted models with higher order, same remarks apply to Akaike's information criterion (AIC) for both models.

It should be mentioned that many diagnostic tools were excluded when analyzing residuals, in particular, the sample partial autocorrelation function of residuals needed with the autocorrelation function of residuals to test errors independence, Q-Q plot for testing normality of the residuals (Shapiro-Wilk test was used instead), hence the title non-rigorous analysis of residuals.

The main conclusion of our research is that "the only important and reliable criterion in detecting the non-aptness of a model due overfitting is comparing estimates with their estimated standard errors in other words estimates should be at least twice the absolute value of its standard error." This criterion should be included in every model diagnostic stage. Other criteria used in the analysis sometimes fail to indicate the non-validity of the fitted model.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Box, George EP and Jenkins, Gwilym M and Reinsel, Gregory C & Ljung GM. Box and Jenkins: time series analysis, forecasting and control. John Wiley & Sons; 2015.

[2] Vandaele W. Applied Time Series and Box-Jenkins Models. New York: Academic Press; 1983.

[3] Nasrabadi & B. Pattern Recognition and Machine Learning. Springer. 2006.

[4] Chatfield C, Xing H. The Analysis of Time Series: An Introduction with R, Seventh Edition. CRC press; 2019.

[5] Priestley MB. Spectral Analysis and Time Series. London: Academic Press; 1981.

[6] Hamilton JD. Time Series Analysis. Princeton, New Jersey: Princeton University Press; 1994.

[7] Chan, K. S., & Cryer JD. Time series analysis. Vol., Springer. 2008.

[8] Shumway, Robert H. &, Stoffer DS. Time Series: A Data Analysis Approach Using R. Time Series: A Data Analysis Approach Using R. 2019. 1–259 p.