Dhaka Univ. J. Sci. 67(1): 21-26, 2019 (January)

Over-Differencing and Forecasting with Non-Stationary Time Series Data

Zakir Hossain[1*], Atikur Rahman[2], Moyazzem Hossain[3] and Jamil Hasan Karami[1]

[1]*Department of Statistics, Dhaka University, Dhaka-1000, Bangladesh*
[2]*Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh*
[3]*Department of Statistics, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh*

(Received: 11 June 2018; Accepted: 29 August 2018)

Abstract

In time series analysis, over-differencing is a common phenomenon to make the data to be stationary. However, it is not always a good idea to take over-differencing in order to ensure the stationarity of time series data. In this paper, the effect of over-differencing has been investigated via a simulation study to observe how far or how close the fitted model from the true one. Simulation results show that the fitted model is found to be different and very far from the true model because of over-differencing in most of the scenarios considered in this study. In practice, it may be worthy to consider differencing as well as suitable transformation of the time series data to make it stationary. Both transformation and differencing are used for a non-stationary time series data on average monthly house prices to ensure it to be stationary. We then analyse the data and make prediction for the future values.

Keywords: Autoregressive model, Autocorrelation and partial autocorrelation function**.**

I. Introduction

Time series analysis is usually performed with the stationary series. However, in practice most of the time series is non-stationary. In order to obtain the stationary series, one may consider differencing which is commonly used in modelling time series data[1]. For example, differencing is used in modelling and forecasting with stock price index data[2]. The order of differencing is usually fixed based on the visualization of time series data, autocorrelation function or a statistical test[3]. One may usually consider differencing for the time series data to observe and isolate patterns of the series, such as trend and seasonality that depend on time, and also to stabilize the mean of the process[4].

It is very common to consider differencing once, twice, and even three times or more in case of the non-stationary series to eliminate the variation as well as to ensure the series to be stationary[5]. Differencing is used for the non-stationary integrated series in the study of estimation of the memory parameter of long-memory time series analysis[6]. It is not always guaranteed that the stationary time series can be obtained by taking only differencing of the non-stationary series. The problem of over-differencing is investigated and found to be accountable for the loss of valuable information of the time series and this often affects in the construction of a model [7].

Some previous works are found on the study of the effect of over-differencing. The effect of over-differencing has been investigated with different models e.g., deterministic linear trend and stochastic regression models. Note that in those studies, Monte Carlo techniques are applied. The amount of loss in efficiency incurred is not of much concern in case of estimation and prediction [8,9]. This finding is also supported by another study of the effect of over-differencing [10].

The first order differencing is considered on the log-scaled data of average daily share price index of square pharmaceutical company in Bangladesh to obtain the stationary series[11]. In the study of forecasting crude palm oil prices, the fractionally integrated method is used and also the first-order differencing is adopted to gain the stationary time series[12]. In this paper, we investigate the effect of over-differencing via a simulation study. One may consider both suitable transformation (e.g., natural logarithm) and differencing for the non-stationary time series to be stationary. We show an application of both transformation and differencing techniques to a real world non-stationary time series data.

II. Simulation Study

We investigate the effect of over-differencing via a simulation study. In order to do this, we consider the stationary autoregressive model (*AR*) of order one, i.e.

$$AR\,(1): X_t - \phi X_{t-1} = Z_t\,,$$

where Z_t is a white-noise process with variance σ_Z^2=0.5. Data ($x_1,\ldots,x_n$) are simulated from the above process and then these data are differenced once, giving rise to a series ($y_1,\ldots,y_n$) where

$$y_t = x_t - x_{t-1}\,.$$

We then fit an autoregressive model to the simulated data ($y_1,\ldots,y_n$) and find the order of the autoregressive model which fits the data best according to the minimum Akaike information criterion (AIC)[13] value. Several data sets of different sizes are generated considering different coefficient values of ϕ. We first specify different coefficient values of ϕ as: -0.9, -0.7, -0.5, -0.3, 0.3, 0.5, 0.7 and 0.9. For each of the coefficient values, we consider different sizes of data sets as: *n*=100, 500, 1000, 5000 and 10000. Simulation results of the best fitted *AR* models are summarized in Table 1.

[*] Author for correspondence. e-mail: zakir.hossain@du.ac.bd

Table 1. The best *AR* models of different order (*p*) and corresponding *AIC* values

ϕ	*n*=100		*n*=500		*n*=1000		*n*=5000		*n*=10000	
	p	*AIC*	*p*	*AIC*	*p*	*AIC*	*p*	*AIC*	*p*	*AIC*
-0.9	6	244.87	12	1100.79	12	2244.53	12	11303.37	12	22196.24
-0.7	8	230.49	11	1187.57	12	2184.47	12	11053.67	12	22770.27
-0.5	5	221.13	12	1162.29	12	2237.31	12	11054.32	12	22305.49
-0.3	4	251.50	11	1135.02	12	2229.43	12	11057.21	12	22216.52
0.3	6	238.00	12	1167.65	12	2170.11	12	11036.88	12	22259.96
0.5	**2**	**221.83**	10	1154.80	12	2217.21	12	11032.35	12	22193.84
0.7	2	202.79	12	1145.19	12	2196.82	12	10996.85	12	22186.61
0.9	4	206.77	**2**	**1095.27**	**10**	**2138.97**	**11**	**10876.09**	**12**	**21678.85**

For fixed *n*=100 and different values of ϕ = -0.9, -0.7, -0.5, -0.3, 0.3, 0.5, 0.7, 0.9 the models *AR(6), AR(8), AR(5), AR(4), AR(6), AR(2), AR(2)* and *AR(4)* are found to be the best choices, respectively because of their minimum *AIC* values, those are all different from the true model *AR(1)*. Among these models, *AR(2)* is found to be the best choice for *n*=100 because of its minimum *AIC* value (*AIC=202.79*) with ϕ =0.7, which is not exactly the same but close to the true model *AR(1)*. Simulation results also show that for the different coefficient values and increasing the sample size, for example *n*= 500, 1000, 5000 and 10000; models *AR(2)* [*AIC*=1095.27], *AR(10)* [*AIC*=2138.97], *AR(11)* [*AIC*= 10876.09] and *AR(12)* [*AIC*=21678.85] are found to be the best choices each with coefficient value ϕ =0.9, respectively as their corresponding minimum *AIC* values.

For fixed coefficient value ϕ =0.9, the model *AR(12)* is found to be the best choice in all cases of sample size *n* except the model *AR(6)* for *n*=100, because of the smallest *AIC* values, those are very far from the true model *AR(1)*. Among these models, *AR(6)* is found to be the best choice which is also very far from the true model *AR(1)*. Again, considering different values of *n* in each cases (fixed ϕ), various models are found to be the best, those are all different from the true model. In summary, keeping the sample size *n* fixed, and varying the coefficient ϕ values; as well as keeping the ϕ values fixed and increasing the sample size *n,* in all cases various models are to be the best choices that are all different from the true *AR(1)* which may arise because of differencing. Thus simulation results suggest that it is not always suitable to consider over-differencing in case of the time series analysis to make the data to be stationary. In the next section, we consider a non-stationary time series data set as a real life example, where both differencing and transformation are used in order to make the data to be stationary.

III. Modelling and Forecasting

In this section, modelling and forecasting with the non-stationary time series data are described incorporating with model diagnosis. We now start with source and description of the data that are used in the study.

Source of data

We use a secondary data set of average monthly house prices in the London region from March 1995 to February 2010, taken from the Land Registry. The data are available in the website: https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-january-2017?utm_medium=GOV.UK&utm_source=summary&utm_campaign=section9&utm_term=9.30_21_03_17&utm_content=download_data.

Data description and screening

We first observe the time series plot of the given data. A positive approximate linear trend of average monthly house prices is evident in Figure 1. This indicates that the process is not stationary. We now take differencing once and twice of the given data. The corresponding time series plots of the first and second differenced data are presented in Figure 2.

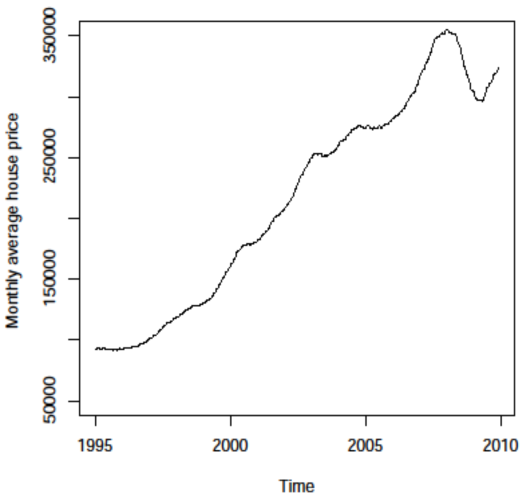

Fig. 1. Time series plot of average monthly house prices from March 1995 to February 2010

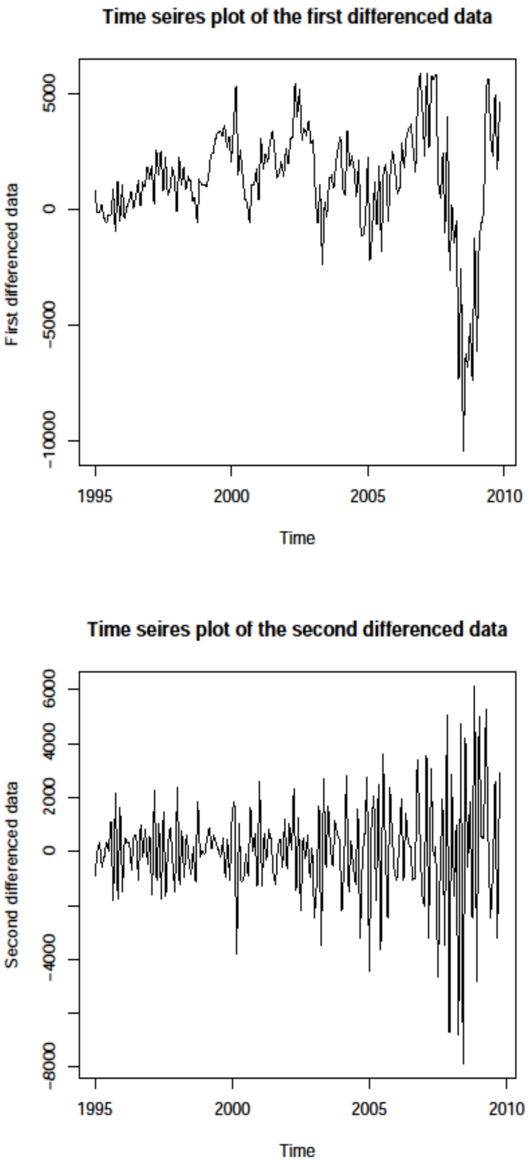

Fig. 2. Time series plots of the first and second differenced data

From Figure 2, it is clear that the process is still non-stationary. Therefore, the natural logarithm transformation of the given data is considered. We then take the first and second differencing of the log-trnasformed data.

Time series plots of first and second differenced log-transformed data are shown in Figure 3. It reveals that the second-differenced log-transformed data seems to be stationary.

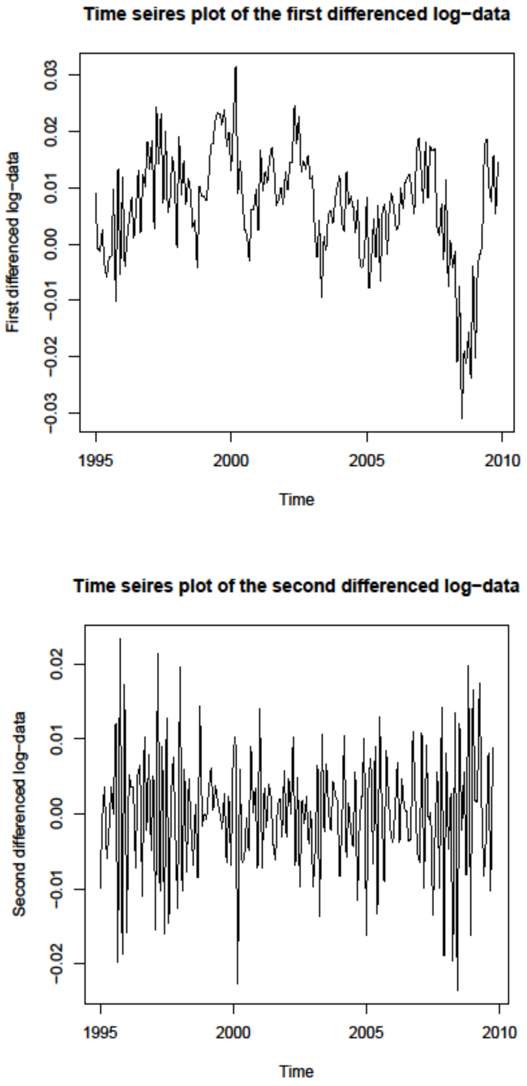

Fig. 3. Time series plots of the first and second differenced log-data

Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the second-differenced log-transformed data are given in Figure 4.

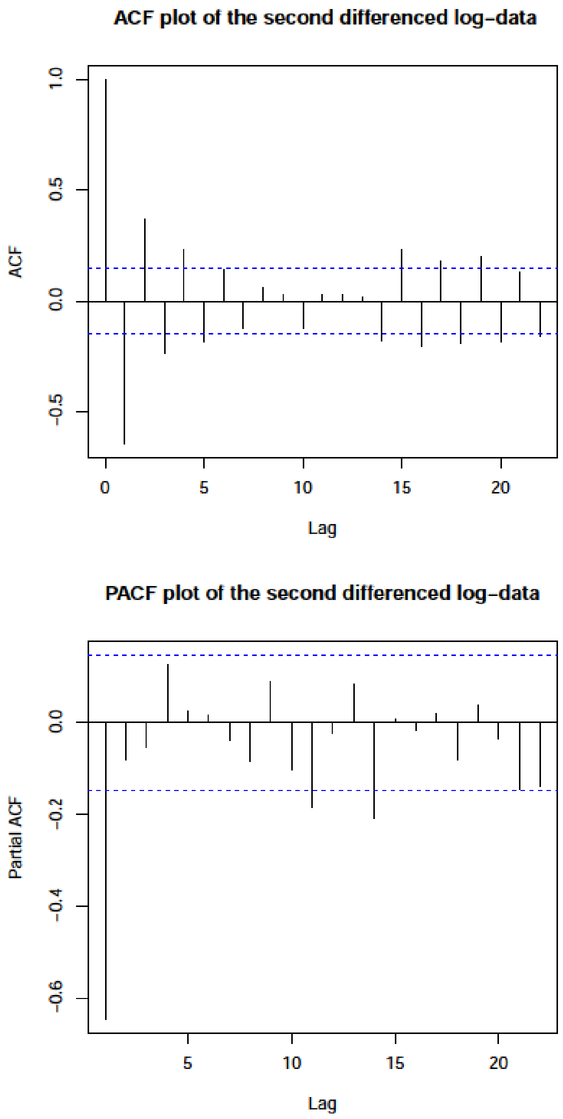

Fig. 4. ACF and PACF plots of the first and second differenced log-data

From Figure 4, it is observed that there is no usual spikes in the ACF plot which indicates no seasonal variation in the data. It is also observed that there are tails off in the ACF plot which determines that the process is autoregressive. From the PACF plot, it is clear that the partial autocorrelation value is higher at lag 1 than others. Moreover, other values are found to be within the approximate 95% confidence interval. Thus it follows that the *AR(1)* model may be the best choice for the process.

Model fitting and diagnosis

The autoregressive process is fitted to the second-differenced log-transformed data where the autoregressive model of order one is found to be the best choice because of its minimum *AIC* value. The estimates of parameters ϕ and σ_z^2 of the model are found to be as $\hat{\phi} = -0.6489$, where the standard error of the estimate is 0.0571, and the estimated variance $\hat{\sigma}_z^2 = 4.466\times10^{-5}$ with minimum $AIC = -1273.25$. The diagnostic checking of the fitted model along with various residual plots are given in Figure 5.

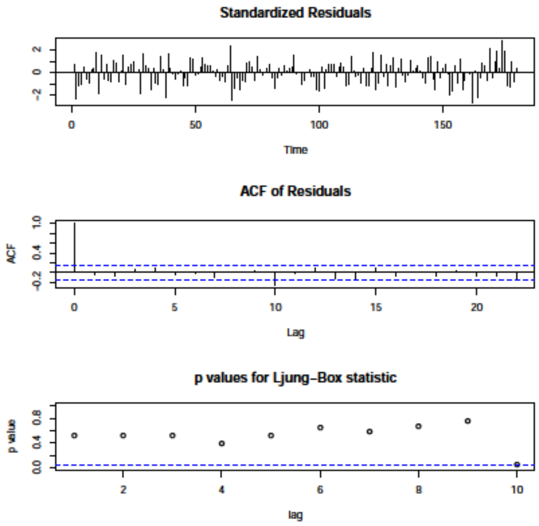

Fig. 5. Model diagnostic plots of the data

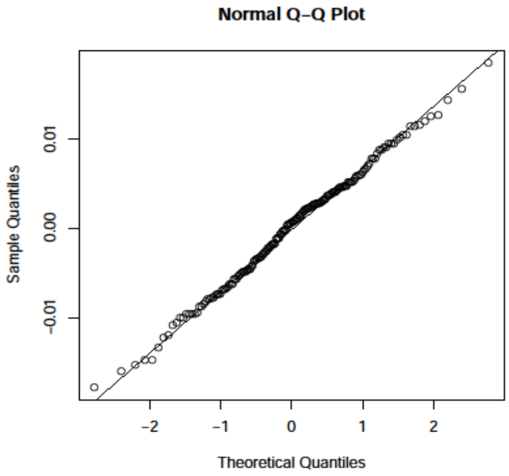

Fig. 6. Normal Q-Q plot of residuals

The time series plot of standardized residuals (1st panel) shows that the residuals are normal with mean zero and constant variance. From the ACF of residuals plot (2nd panel), it is observed that the residuals are not correlated. The *p*-values for testing the null hypothesis of independence of residuals obtained from Ljung-Box test are shown in the third panel. It is found that all *p*-values are well above the dotted line except at lag 10. This indicates that the residuals are independent. Moreover, the residuals are found to be normally distributed as evident from the Q-Q plot in Figure 6.

Forecasting

The prediction is done on average monthly house prices for the next future months of 2010. In Figure 7, the solid line of the top-right corner indicates the future predicted values and the dotted lines represent the corresponding limits of 95% confidence interval. It is clear that the average monthly house prices are expected to be increasing over time.

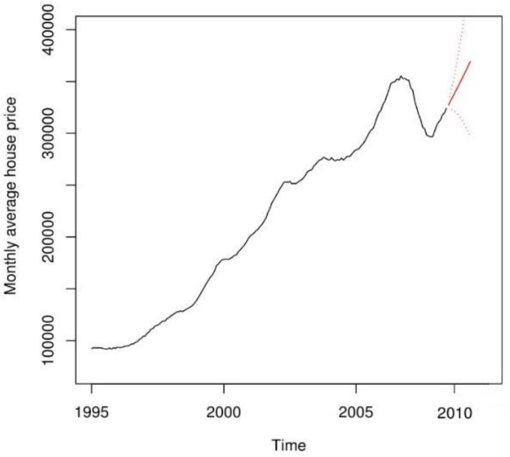

Fig. 7. Time series plot of average monthly house prices with future prediction

IV. Discussion and Conclusion

In this study, an attempt has been made to investigate the effect of over-differencing via a simulation study. Moreover, we consider the model fitting and forecasting with a non-stationary time series data. Simulation study has been conducted using autoregressive model of order one for various choices of the data size and the coefficient value. It is evident from the simulation results that over-differencing is not always suitable for time series data to gain stationarity of the series. On the other hand, it may be a good choice to consider suitable transformation and differencing in case of non-stationary time series data to make it stationary.

We consider an example of non-stationary time series data set of average monthly house prices in London region from March 1995 to February 2010. A positive trend over time is observed in the original data. Therefore, the natural logarithm and differencing have been considered to make the process to be stationary. The autoregressive model of order one has been fitted and investigated by means of some model selection and diagnostics criteria such as AIC, Q-Q plot and ACF plot of standardized residuals. To this end, the autoregressive model of order one is found to be the best choice for analyzing the data considered in this study. Moreover, the increasing trend over time for the predicted values prevails similar to that of the original data.

References

1. Lazim, M.A., 2011. Introductory Business Forecasting: A Practical Approach. 3rd edition, UiTM Press, Malaysia.
2. Erfani, A. and A. J. Samimi, 2009. Long memory forecasting of stock price index using a fractionally differenced ARMA model. *J. Appl. Sci*, Res. **5**, 1721–1731.
3. Ming, C. C. and A. D. David, 1994. Recognizing Over-differenced time series. *Journal of Time Series Analysis.* **15(1)**, 1–18.
4. Hyndman, R. J. and A. George, 2013. Forecasting: principles and practice. Paperback.
5. Cochrane, J. H., 2018. A Brief Parable of Over-Differencing, University of Chicago, http://faculty.chicagobooth.edu/john.cochrane/.
6. Hurvich, C. M.,and B. K. Ray, 1995. Estimation of the memory parameter for non-stationary or noninvertible fractionally integrated processes. *J. Time Series Anal*. **16**, 17–41.
7. Xiu J. and Jin Y., 2007. Empirical study of ARFIMA model based on fractional differencing. *Physica.* **377**, 138–154.
8. Plosser, C. I. and G. W. Schwert, 1977. Estimation of a Non-invertible Moving Average Process: The Case of Over-differencing. *Journal of Econometrics.* **6**, 199-224.
9. Plosser, C. I. and G. W. Schwert, 1978. Money, Income and Sunspots: Measuring Economic Relationships and the Effects of Differencing. *Journal of Monetary Economics*. **4**, 637-660.
10. Harvey A. C., 1981. Finite Sample Prediction and Over-differencing. *Journal of Time Series Analysis.* **2**, 221-232.
11. Paul, J. C., Shahidul H., M. S. Hoque, and M. M. Rahman, 2013. Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd. *Global Journal of Management and Business Research Finance.* **13(3)**, 15-26.
12. Karia, A. A., I. Bujangand I. Ahmad, 2013. Fractionally integrated ARMA for crude palm oil prices prediction: case of potentially over-difference. *Journal of Applied Statistics.* **12**, 2735-2748.
13. Akaike, H, 1973. *Information theory and an extension of the maximum likelihood principle*. Budapest: Akademiai Kiado.