Autoregression Modelling on Time Series Data

Mannat Soni
Advanced Financial Technology Coursework 2

I. Introduction

A. Time series

Time series is a special type of data where each data point is at a constant interval from the other [2]. With characterization and modelling, Time series forecasting involves the collection and analysis of historical data in order to develop a model producing statistically significant results [1]. One of the most widely used ways of performing technical, especially on stock prices is using Autoregression [5].

B. Stationarity of Time Series data

A time series analysis largely depends on whether or not it is stationary. Stationarity simply means how much the future behaviour of a time series reflects its past behaviour [2]. For a time series to be stationary it has to satisfy all three conditions [3]: constant mean, constant variance and Co-variance of y_t and y_{t-s} being time-invariant. Many statistical tests like Augmented Dickey-Fuller tests [3], to check for stationarity and KPSS [4] test, to check for trends can be used. In order to make time-series data stationary, some ways are are differencing, taking a log or a square root etc [2].

C. ACF and PACF

In order to perform model identification in Autoregression, ACF and PACF are used as vital statistical tools [1]. Autocorrelation determines the correlation between a time series and its previous lags whose inertia then carries on to subsequent lags. It tells us how a time series with different lags are linearly related to one another. On the other hand, PACF (Equation 2 for lag=2) determines the correlation between two time periods that have a gap of lag k. In order to determine the number of lags required for a time series to be linearly correlated to itself, PACF function is used.

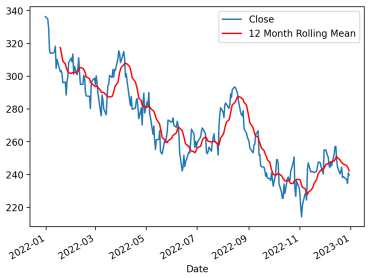

Fig. 1. Mean and Standard Deviation For Closing Time-Series

D. Autoregression

Autoregression very intuitively defines itself, as a regression of its own past values to determine future values. Equation 1 gives us a general form of AR with lags p, where c is the constant, $phi's$ are lag coefficients up to order p, and $epsilon$ is the irreducible error (white noise). This is represented as AR(p). This p-value is determined by the PACF function.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

II. Design and Analysis of the Experiment

A Microsoft stock price dataset is used that has 253 rows and 2 columns. This dataset is a time series data starting from 2021-12-31 to 2022-12-30 giving us information about Closing prices of Microsoft stock data. This experiment is divided into three sections: 1) Data Visualization and Making the data Stationary, 2) Finding Autocorrelation and Partial Autocorrelation functions for model building and 3) Applying Autoregression.

A. Data Visualization and Making the data Stationary

1) Data Visualization: Figure 1 gives us a description of the Closing price of Microsoft data with mean trend over the time series. There is an overall downward trend that is observed with occasional peaks during April, August and December.

On viewing the Figure 1 plot with the naked eye, data appears to be non-stationary as the mean of the data is constantly decreasing and a trend is observed by occasional peaks. Statistical tests, Augmented Dickey-Fuller Test gives us a p-value of 0.098 which is greater than 0.05. This confirms that the time series has a unit node, indicating it is non-stationary.

2) Making data Stationary: A recognized way of making stock prices' time series data stationary is by using Weighted Moving averages [7]. In the weighted moving average more recent values are given a higher weight. Here, exponentially weighted moving average is used where weights are assigned to all the previous values with a decay factor. The time series is made stationary by first taking a log of the values and then subtracting the values with their Exponential Moving Average, with a half-life set to 12. The data is set to be stationary as shown in the figure. The statistical value obtained by Augmented Dickey-Fuller tests is 0.0050. Hence, our time series does not have a root node and is now stationary.

B. Finding ACF and PACF for model building

This part of the experiment focuses on curating Partial Autocorrelation Function from scratch for lags=2. Initially,

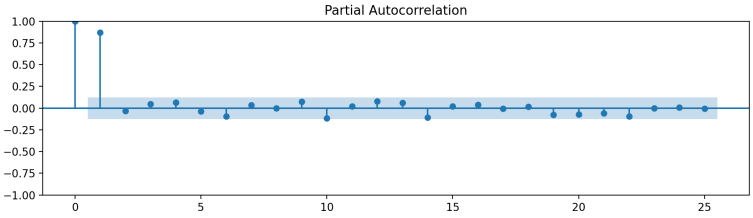

Fig. 2. PACF values with only lag=1 being statistically significant

two columns are added with lag = 1 and lag =2 respectively. These two series are used to find predicted values, using linear regression that is then used to compute residual values. The residual values are computed by finding the difference between the predicted values and actual values which are the two time series needed to compute PACF using equation 2. The residual series gives us the amount of variance in series T which can not be computed by the variance in T-1 with the noise. The numerator value gives us the covariance between the two residual time series and the denominator gives us the standardized covariances [8].

$$\text{PACF}\left(\boldsymbol{T}_i, \boldsymbol{k}=\mathbf{2}\right) = \frac{\text{Cov}\left(T_i \,\middle|\, T_{(i-1)}, T_{(i-2)} \,\middle|\, T_{(i-1)}\right)}{\sigma_{T_i|T_{(i-1)}} \times \sigma_{T_{(i-2)})T_{(i-1)}}} \quad (2)$$

This gives us PACF for Lag=2. On calculating from scratch we get the value -0.035. This value is verified by using the python library which comes out to be -0.034. Plotting ACF and PACF plots determine how many lags are required to perform Autoregression. Figure 2 shows that only lag=1 is statistically significant and the rest of the values for the remaining lags are either 0 or so statistically very close to 0. Thus the time series data has an AR(1).

C. Implementing AR model from Scratch

The dataset is divided into training (70%) and test (30%) sets. The AR model is implemented from scratch. The time series computed using the exponential moving average is used as the dependent variable (values to be predicted=y) and the time series generated by lag=1 (T-1) is used as an independent variable (x). The intercept computed after training the data is obtained as 0.866. This is then used to test our model on our test data (Figure 3). The Root Mean Square error metric is used to test how well our model performed, which came out to be 0.0223.

D. Strengths and Limitations of an AR Model

AR models predict futures on the basis of past values which is one of the simplest ways of analyzing technical stock data. But main limitations are fat-tails where large losses or gains are anticipated at higher probabilities than suggested by the normal distribution and volatility clustering [9]. This can be countered by using an integrated ARMA model where constant variance is assumed and white noise is allowed to be non-gaussian. Another limitation is assumption of stationarity that affects the predicts of an AR model [10]. This is resolved

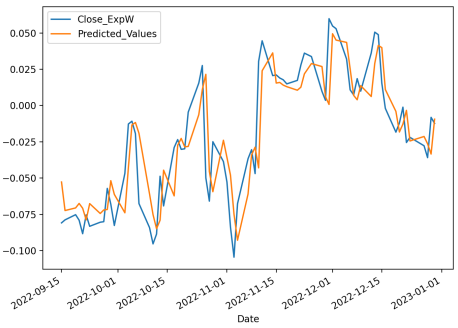

Fig. 3. Predicted values and Original Values for Test Data

using ARMA models that solves this issue by exploiting the concept of returns. Other than just time series data AR can also be used for tasks like Natural Language Processing. They perform polynomial time computation [11]. But this brings limitations in computing next symbol probabilities. In order to resolve that, two new models are recommended which are Alternative include energy-based models and Latent Variable Autoregressive Model [11].

REFERENCES

[1] J. Kaur, K. S. Parmar, and S. Singh, "Autoregressive models in environmental forecasting time series: a theoretical and application review," Environmental Science and Pollution Research, vol. 30, no. 8, pp. 19617–19641, Jan. 2023

[2] A. Nielsen, Practical Time Series Analysis: Prediction with Statistics and Machine Learning. "O'Reilly Media, Inc.," 2019. Accessed: Feb. 24, 2023. [Online].

[3] R. Mushtaq, "Augmented Dickey Fuller Test," papers.ssrn.com, Aug. 17, 2011.

[4] B. Hobijn, P. H. Franses, and M. Ooms, "Generalizations of the KPSS-test for stationarity," Statistica Neerlandica, vol. 58, no. 4, pp. 483–502, Nov. 2004.

[5] J. Fern and o, "What Does Autoregressive Mean?," Investopedia.

[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] S. H. Shih and C. P. Tsokos, "A Weighted Moving Average Process for Forecasting," Journal of Modern Applied Statistical Methods, vol. 7, no. 1, pp. 187–197, May 2008.

[8] "Understanding Partial Auto-correlation And The PACF," Time Series Analysis, Regression and Forecasting, May 27, 2021.

[9] A.-C. Petrică, S. Stancu, and A. Tindeche, "Limitation of ARIMA models in financial and monetary economics.," Theoretical & Applied Economics, vol. 23, no. 4, 2016.

[10] S. Yi, A. Viscardi, and T. Hollis, "A Comparison of LSTMs and Attention Mechanisms for Forecasting Financial Time Series." Accessed: Feb. 24, 2023. [Online].

[11] C.-C. Lin, A. Jaech, X. Li, Matthew, R. Gormley, and J. Eisner, "Limitations of Autoregressive Models and Their Alternatives." Accessed: Feb. 24, 2023. [Online].