ARIMA Modeling for Stock Price Forecasting

Author: Arunodaya Rajbhandari

Academic Mentor: Andrew Leahy

MATH 399

Knox College

Galesburg, IL

May 28th, 2025

Table of Contents

1. Introduction

Forecasting stock prices has long been a critical challenge in financial markets, as investors, institutions, and governments seek reliable tools to anticipate trends and manage risk. Time series models, especially those grounded in statistical theory, have emerged as a foundational component of this effort. Among them, the Autoregressive Integrated Moving Average (ARIMA) model stands out for its simplicity, mathematical robustness, and adaptability to nonstationary data. ARIMA models, introduced by Box and Jenkins, are widely used for analyzing and forecasting univariate time series, including stock prices, GDP, inflation, and more (Tsay, 2005; Andersen et al., 2001).

In our MATH 399 capstone presentation, the objective was to compare the predictive performance of three time series modeling approaches: ARIMA, GARCH, and LSTM. This paper focuses on the ARIMA modeling section presented by me, offering a detailed overview of its theoretical underpinnings, assumptions, model selection techniques, and practical considerations in real-world financial forecasting.

2. Understanding the Components of ARIMA

The ARIMA model is constructed from three core components (Tsay, 2005):

- **Autoregressive (AR) component:** This part of the model captures the dependency of the current value on its past values. If a time series exhibits persistence or memory, where current values are linearly dependent on previous observations, an AR process is appropriate. An AR model of order p, denoted AR(p), can be expressed as:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t$$

Where a_t are i.i.d $N(0, \sigma^2)$

- **Integrated (I) component:** Many time series are not stationary, meaning their statistical properties change over time. The integration component involves differencing the series *d* times to achieve stationarity. For example, differencing once means:

$$z_t = y_t - y_{t-1}$$

Moving Average (MA) component: This models the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. An MA(q) model has the form:

$$y_t = \delta + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2}$$

These formulations and their properties are discussed extensively in Tsay (2005).

3. Stationarity and Differencing

One of the key requirements for using ARIMA is that the time series must be stationary, meaning its mean and variance are constant over time and it has no seasonal structure (Tsay, 2005). Nonstationary data can yield misleading model estimates and predictions. There are two major types of nonstationarity:

- **Trend nonstationarity:** Solved by differencing the data.
- **Variance nonstationarity:** Solved by transformations such as the logarithm or square root.

For example, stock price data often contains upward or downward trends, requiring first-order differencing to convert it into a stationary series.

4. Autocorrelation and Model Identification

Determining the appropriate order of AR and MA components in an ARIMA model is a crucial step. Two tools that aid in this are:

- **Autocorrelation Function (ACF):** Measures the correlation of the series with its own lagged values. A sharp cutoff after lag *q* suggests an MA(q) model.
- **Partial Autocorrelation Function (PACF):** Measures the correlation between observations separated by *k* lags, controlling for all intermediate lags. A sharp cutoff after lag *p* indicates an AR(p) model.

These concepts and visual tools are essential for model diagnostics and are emphasized in Tsay (2005).

5. Building Mixed Models: ARMA and ARIMA

When both AR and MA components are needed, a mixed model, ARMA(p, q)–is constructed. If the series is nonstationary and requires differencing *d* times, the full ARIMA(p, d, q) model is used. An example of such a model is:

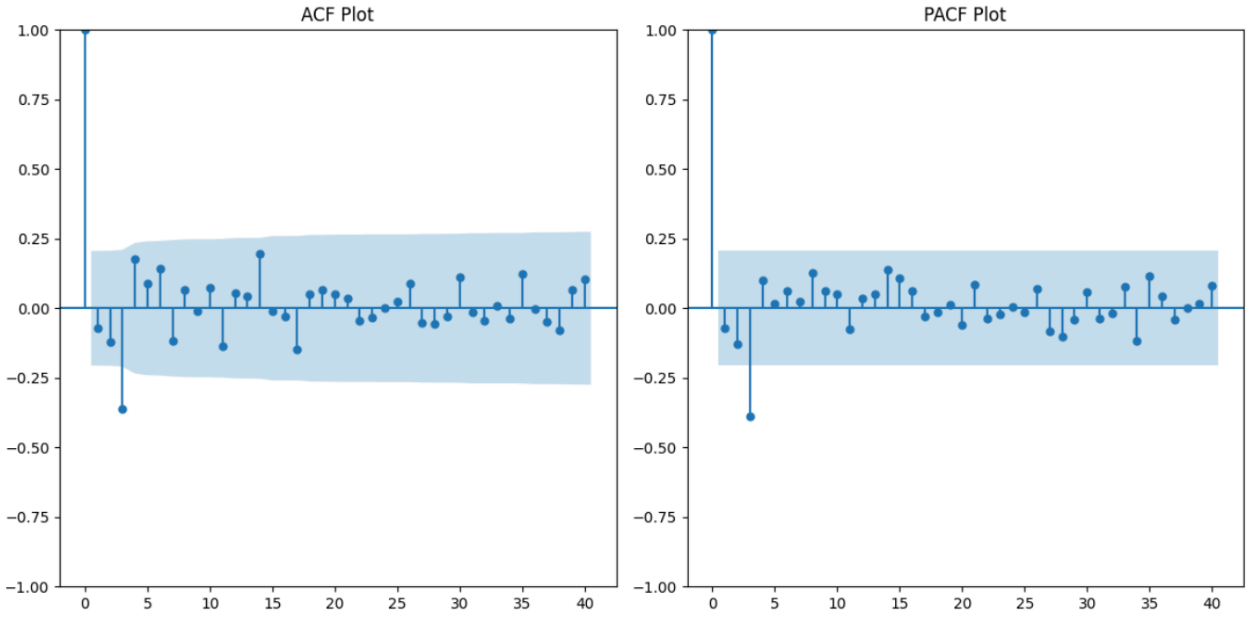
$$z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}$$

In our presentation, we emphasized that using too many parameters can lead to overfitting. Hence, simplicity is preferred, often beginning with ARIMA(1,1,1) and validating its performance (Tsay, 2005).

6. Diagnostic Checks and Model Validation

After estimating an ARIMA model, it's important to verify that the residuals (errors) behave like white noise, uncorrelated and normally distributed. This involves:

- **Plotting residuals**
- **Checking ACF and PACF of residuals**
- **Using statistical tests like the Ljung-Box test**



These diagnostics ensure that no information is left in the residuals and are fundamental to model adequacy checks (Tsay, 2005).

7. ARIMA in Financial Forecasting

ARIMA has been widely applied in financial forecasting, especially in modeling asset returns and price levels. For example, returns on stocks are influenced by numerous random events–government policies, economic news, investor behavior–that cause dependencies across time. ARIMA helps capture such dependencies when no strong external predictors are available (Adebiyi et al., 2014).

However, ARIMA models assume linear relationships and constant variance, which may not hold in financial markets prone to volatility clustering and sudden jumps. For this reason, while ARIMA remains a solid baseline, models like GARCH (for volatility) and LSTM (for nonlinear dynamics) can provide complementary insights (Francq & Zakoian, 2019; Nelson, 1991).

8. Limitations and Extensions

While ARIMA is powerful, it has several limitations (Tsay, 2005):

- Requires large amounts of historical data (at least 50 observations recommended)
- Not ideal for datasets with strong seasonal effects unless extended to SARIMA
- Assumes linearity and constant variance
- Not robust to structural breaks or regime shifts

To overcome these issues, analysts often turn to extensions like SARIMA (for seasonal data), ARIMAX (with exogenous variables), and SARIMAX (seasonal with exogenous variables). In our broader study, these models were compared alongside LSTM and GARCH to assess their forecasting accuracy (Raschka et al., 2022; Fischer & Krauss, 2018).

9. Conclusion

ARIMA remains a cornerstone of time series analysis due to its simplicity and interpretability. Understanding its assumptions, proper model selection techniques, and diagnostics are critical to applying it effectively in real-world forecasting tasks, such as predicting stock prices. As part of our capstone, this foundation enabled a more nuanced comparison with advanced models like GARCH and LSTM. For those entering fields like quantitative finance or data science, mastering ARIMA is a vital first step toward more complex and adaptive forecasting frameworks.

References

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). *Stock price prediction using the ARIMA model*. Proceedings of the 2014 UKSIM-AMSS 16th International Conference on Computer Modelling and Simulation.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). *The distribution of realized stock return volatility*. Journal of Financial Economics, 61(1), 43–76.

Bollerslev, T. (1986). *Generalized autoregressive conditional heteroskedasticity*. Journal of Econometrics, 31(3), 307–327.

Engle, R. F. (1982). *Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation*. Econometrica, 50(4), 987–1007.

Fischer, T., & Krauss, C. (2018). *Deep learning with long short-term memory networks for financial market predictions*. European Journal of Operational Research, 270(2), 654–669.

Francq, C., & Zakoian, J. M. (2019). *GARCH Models: Structure, Statistical Inference, and Financial Applications*. John Wiley & Sons.

Nelson, D. B. (1991). *Conditional heteroskedasticity in asset returns: A new approach*. Econometrica, 59(2), 347–370.

Raschka, S., Patterson, J., & Nolet, C. (2022). *Machine learning with PyTorch and Scikit-Learn: Developing predictive models with Python*. Packt Publishing.

Tsay, R. S. (2005). *Analysis of Financial Time Series* (2nd ed.). John Wiley & Sons.

Zhang, G. P. (2003). *Time series forecasting using a hybrid ARIMA and neural network model*. Neurocomputing, 50, 159–175.