THE UNIVERSITY OF TEXAS AT DALLAS

BUAN 6312.S01: Applied Econometrics and Time Series Analysis

Prof. Quanquan Liu

Time series analysis on stock market data using ARIMA

Problem Statement

1 Objective:

Develop predictive time series models for effective analysis of stock market data

2 Dataset Overview:

Historical stock prices for 29 DJIA companies over the last 12 years, excluding 'V' (Visa).

Various formats available for different analytical needs.

3 Challenges and Rewards:

Inherent complexity in analyzing vast financial data.

Potential for significant financial gains through successful model development.

4 Resource Availability:

Dataset serves as a valuable resource for exploring stock market dynamics.

GitHub repository contains scripts for data acquisition and continually updated modeling codes.

Data

Rich Dataset Structure

- Key columns include Date, Open, High, Low, Close, Volume, and Name.
- 'Date' captures the timeline in yy-mm-dd format, detailing stock price movements.
- 'Open,' 'High,' 'Low,' and 'Close' provide comprehensive information on daily stock price dynamics.

Versatile Data Offerings

- Available in extended (13 years) and condensed (past year) versions, accommodating diverse analytical needs.
- Formatted in USD for NYSE data, facilitating consistency and ease of analysis.

Acknowledgments and Adaptation

> - Adapted from 'S&P 500 Stock data' on platforms like Kaggle and GitHub.
> - Scraped from Google finance using 'pandas_datareader' Python library, extending gratitude to Kaggle, Github, and the Market for data accessibility.

Literature review

Time series analysis is a statistical method used to analyze and interpret data points collected over a period of time. The time series data could be collected on a daily, weekly, monthly, or yearly basis. Time series analysis is widely used in economics, finance, engineering, and social sciences to forecast trends, identify patterns, and make predictions based on historical data.

The Comparative Study on Time Series Analysis (Salah et al., Year) evaluates classical methods (CM) against machine learning algorithms (MLA) for demand forecasting, offering nuanced insights into their strengths and weaknesses.

Rob J. Hyndman and George Athanasopoulos' work in "Forecasting: Principles and Practice" (2nd ed), specifically Chapter 8 on ARIMA models https://otexts.com/fpp2/arima.html, and Kuvshinov, D., and Zimmermann, K.'s exploration of stock market capitalization impact https://doi.org/10.2139/ssrn.3236076 add valuable perspectives. Additionally, Hastie, T., Friedman, J., and Tisbshirani, R.'s insights in Chapter 7 of "The Elements of Statistical Learning" (pp. 230–233, Springer) enrich our understanding of statistical learning, data mining, and prediction. Together, these studies highlight diverse time series analysis methods and emphasize the ongoing need for a comprehensive understanding of classical and machine learning-based algorithms in demand forecasting.

Unlocking the Power of Time Series Models

Welcome to the essence of time series analysis! Before diving into the intricate world of predicting stock movements, let's grasp the theory in a nutshell.

1. Auto Regressive (AR) Model

- Essence: Today's value relates to yesterday's.
- Symbol: AR(p) with p lags.

2. Moving Average (MA) Model

- Essence: Variable links to the previous period's residual.
- Symbol: MA(q) with q lags.

3. ARMA Model (Autoregressive Moving Average):

- Essence: A dynamic duo! Combines p auto-regressive and q moving average terms.
- Symbol: ARMA(p,q).

Let's explore individual stocks, uncovering the secrets within their time series—Open, Close, High, Low, and Volume.

Visual Overview of the data

- There are 5 time series in the data provided - (High, Low, Open, Close, Volume).
- We will look at the High values first.

Stationarity

1. A stationary time series is characterized by a consistent mean and variance over time, devoid of any discernible trend. Judging by the appearance, the mentioned time series does not exhibit stationarity. To formally confirm this, the stationarity of the time series was assessed using the Dickey-Fuller test. The outcome of the test indicates the following

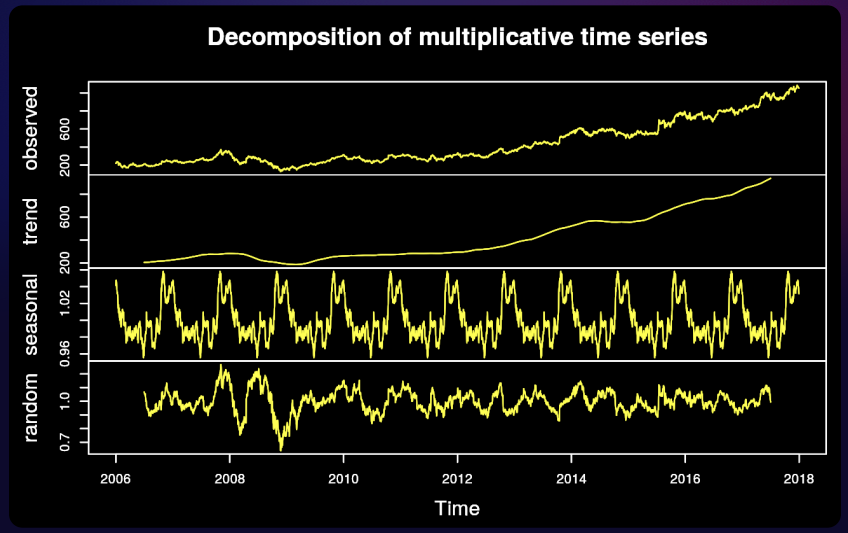
2. Augmented Dickey-Fuller Test:

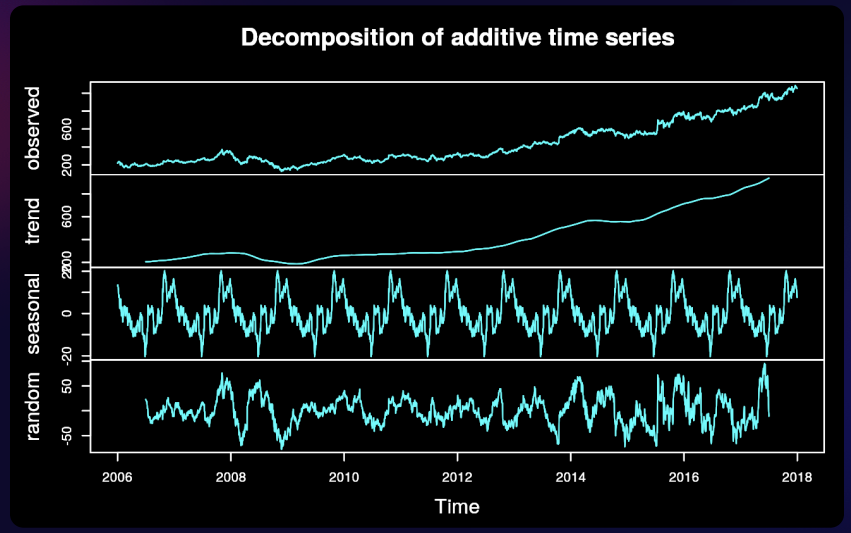
- Test Statistic: -1.3188
- Lag Order: 0
- P-value: 0.8667

3. The test suggests that the time series is non-stationary, failing to reject the null hypothesis of non-stationarity. Therefore, we do not have sufficient evidence to support the alternative hypothesis of stationarity.

Decoding Time Series

Uncover the hidden patterns by breaking down the time series into its trend and irregular components. We explore both additive and multiplicative models for a comprehensive analysis.



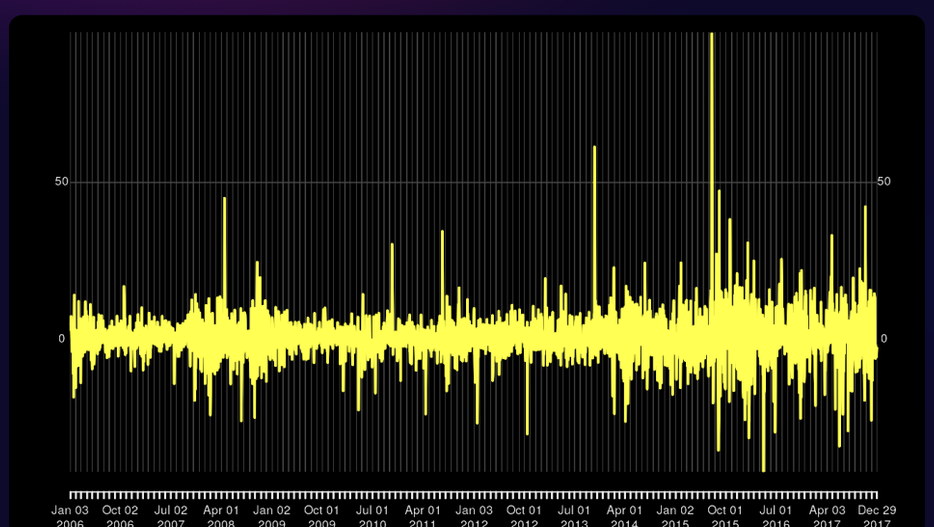


Transforming Time Series through Differencing

1. Differencing stands out as a prevalent technique employed to achieve stationarity in variables. Employing the R function diff, we embark on the differentiating journey. What about exploring fractional differences?
2. Augmented Dickey-Fuller Test

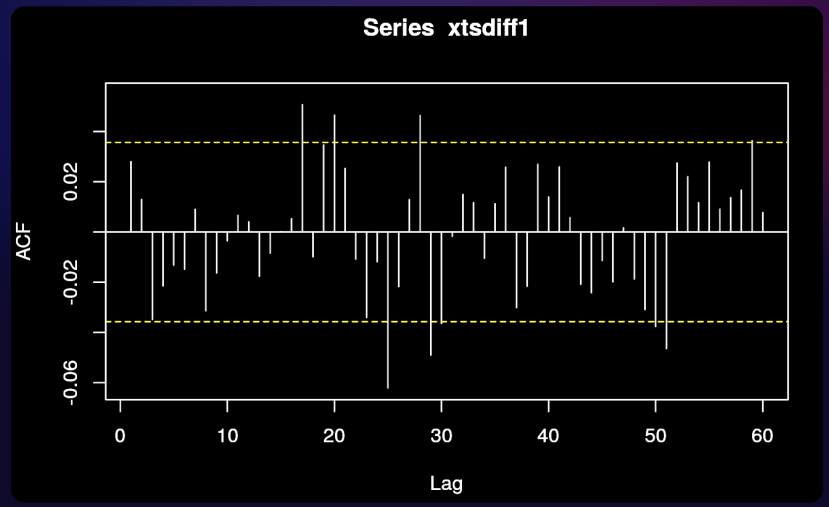
- Data: tsdiff1
- Test Statistic: -53.448
- Lag Order: 0
- P-value: 0.01

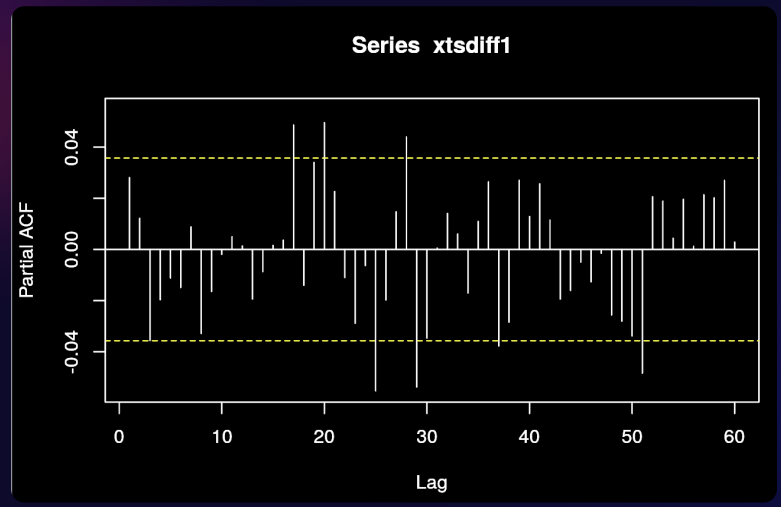
3. The test strongly suggests stationarity, with a remarkably low p-value of 0.01. Hence, we have robust evidence to support the alternative hypothesis of stationarity.



Choosing an ARIMA Model Candidate

Moving forward involves the crucial task of identifying the right ARIMA model, determining optimal values for p and q in an ARIMA(p, d, q) configuration. Typically, this requires a thorough examination of the correlogram and partial correlogram of the stationary time series. In R, this exploration is facilitated by utilizing the acf() and pacf() functions to generate correlograms and partial correlograms, respectively.





Compare sample ACF and PACF to theoretical ARMA models. Use their properties to estimate suitable p, q, and d values or explore an alternative approach discussed next.

Modeling with ARIMA for Optimal Results

1

Embarking on the modeling phase involves leveraging R's powerful tool—auto.arima. This function intelligently selects the best ARIMA model based on criteria like AIC, AICc, or BIC values. The search spans potential models within specified order constraints.

2

Four distinct models are trained using different subsets of training data. For instance, 'tsarima240' is trained on the entire time series, excluding the last 240 daily data points. This strategic approach enhances the model's adaptability and performance.

3

The output and forecast graph for the optimal model have been incorporated.

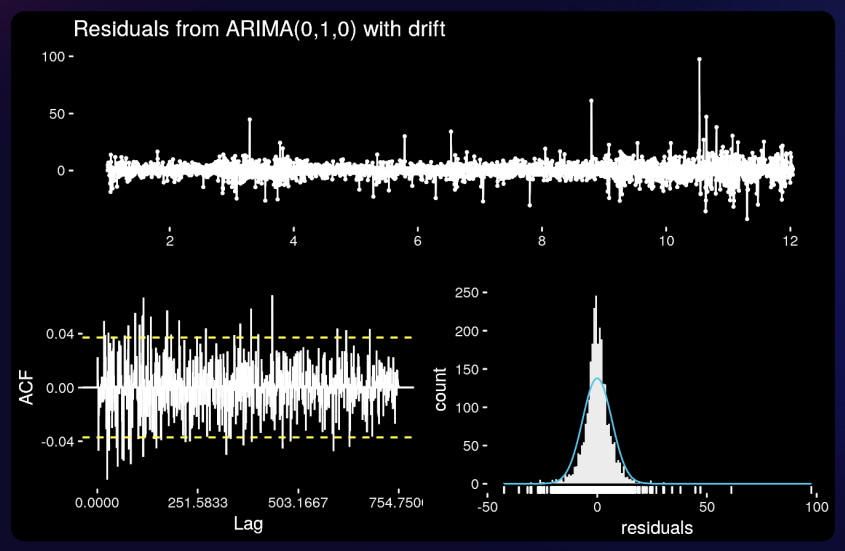
4

ARIMA(0,1,0) with drift

Coefficients:
drift- 0.2200
s.e.- 0.126

sigma^2 estimated as 44.43

log likelihood=-9210.91
AIC=18425.82
AICc=18425.82
BIC=18437.67



- Data: Residuals from ARIMA(0,1,0) with drift
  Q* = 726.5, df = 502.17, p-value = 1.98e-10
- Model df: 1. Total lags used: 503.17

Unraveling Time Series Mysteries with ARIMA

Our journey through time series analysis culminates in a triumph of insights. Employing ARIMA modeling, we transformed raw data into a predictive powerhouse. The selected ARIMA(0,1,0) model, with its drift term, emerged as the beacon of accuracy.

Diligent training on diverse data subsets ensured adaptability, while the Ljung-Box test validated our model's prowess. The forecast graph, a testament to our optimized approach, foretells a future rich in predictive precision.

This project not only demystified time series intricacies but also armed us with a potent tool for data-driven decisions. Let's propel forward, where every data point tells a story!

THE UNIVERSITY OF TEXAS AT DALLAS

Authors:

Anish Gillella – AXG220089

Devansh Pratap Singh – DPS220001

Dimple Neeluri Chandrashekar – DXN210024

Karthik Kotam – KXK210073

Krishna Venkatesan – KXV220007

Kushimithaa Thimmaareddy – KXT220026

Medha Priyanga Saravanan – MXS220057

Sarthak Vajpayee – SXV220020

Yash Agrawal – YXA220007