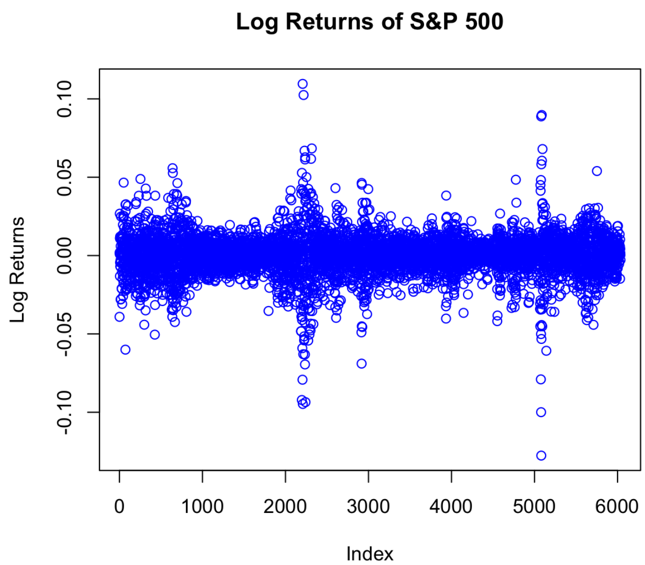
STAT-1050 COURSEWORK

1. Log Returns of the S&P 500 Index

The code begins by retrieving the S&P 500 index data and computing the log returns:

$$R_t = \ln(P_t) - \ln(P_{t-1}),$$

where R_t represents the log return at time t, and P_t and P_{t-1} are the adjusted closing prices of the index at times t and $t-1$, respectively. As explained in the coursework notes, with this transformation, the variance become stabilized and it also prepares the data for time series analysis (Skindilias, 2024).



2. Augmented Dickey-Fuller (ADF) Test

The ADF test evaluates the stationarity of the log returns:

$$\Delta R_t = \alpha + \beta t + \gamma R_{t-1} + \sum_{i=1}^{p} \phi_i \Delta R_{t-i} + \varepsilon_t,$$

where:

- $\Delta R_t = R_t - R_{t-1}$ is the first difference of the log returns,
- α is a constant term,
- βt accounts for a deterministic trend,
- γ tests whether R_t is stationary ($\gamma < 0$ indicates stationarity),
- ϕ_i are lag coefficients,
- ε_t is white noise.

In the analysis, the null hypothesis ($H_0 : \gamma = 0$) was rejectedas $p < 0.05$, confirming stationarity ARIMA (Skindilias, 2024).

Model Selection for ARIMA

For the ARIMA model, the optimal values of p and q were obtained by using **grid search method**. This method systematically evaluates all combinations of p and q within a predefined range $p, q \in \{1, 2, 3, 4, 5\}$), and the reason for this limitation is to avoid overfitting. It compares models, by using selection criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

After evaluating the models, the best model is determined by choosing the minimum AIC and BIC values, which means that the mentioned model balances the goodness of fit and complexity (Data Magic Lab 2023). The results indicated that the optimal combination is $p = 1$ and $q = 0$, as shown in the figure below.

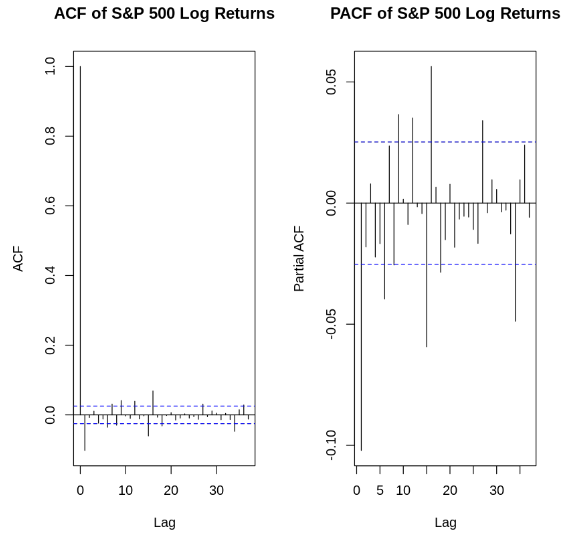


Figure 1: Visualization of Grid Search Results.

Grid Search Explanation

The **grid search** works as follows:

- Define a range for p and q (e.g., $1 \leq p, q \leq 5$),
- Fit an ARIMA model for every combination of (p, q) within this range,
- Evaluate each model using AIC and BIC,
- Select the combination (p, q) that results in the lowest AIC or BIC.

In this way, the model process becomes objective and the best combination of parameters in the specified range is provided.

3. Top 5 Models by AIC and BIC

The ARIMA models evaluated are ranked below:

Model	AIC	BIC
ARIMA(5,0,5)	−35971.57	−35891.10
ARIMA(4,0,3)	−35952.30	−35891.95
ARIMA(3,0,4)	−35951.29	−35890.94
ARIMA(5,0,4)	−35950.74	−35876.98
ARIMA(4,0,5)	−35949.96	−35876.20

4. Auto.ARIMA Model Suggestion

The ‘auto.arima‘ function suggested the ARIMA(1,0,0) model with the following properties:

$$\text{ARIMA}(1,0,0) : R_t = \alpha - 0.1020 R_{t-1} + \varepsilon_t,$$

where:

- Coefficient of R_{t-1}: -0.1020 (s.e. $= 0.0128$),
- $\sigma^2 = 0.0001517$,
- Log-likelihood: 17974.84,
- AIC: -35945.68,
- BIC: -35932.27.

5. Best Model Diagnostics: ARIMA(5,0,5)

The ARIMA(5,0,5) model provided the best fit, with the following equation:

$$R_t = \alpha + \sum_{i=1}^{5} \phi_i R_{t-i} + \sum_{j=1}^{5} \theta_j \varepsilon_{t-j} + \varepsilon_t,$$

where:

$$\begin{array}{lllll} \phi_1 = -0.8245, & \phi_2 = -0.2977, & \phi_3 = -0.5957, & \phi_4 = -0.1058, & \phi_5 = 0.4359, \\ \theta_1 = 0.7257, & \theta_2 = 0.2019, & \theta_3 = 0.5673, & \theta_4 = 0.0208, & \theta_5 = -0.4817, \\ \alpha = 0.0002 & & & & \end{array}$$

The model diagnostics indicate:

- $\sigma^2 = 0.0001508$,
- Log-likelihood: 17997.78,
- AIC: -35971.57,
- BIC: -35891.10.

Model	AIC	BIC	ME	RMSE	MAE	MPE	ACF1
ARIMA(1,0,0)	−35945.56	−35925.45	−0.0000002	0.0123	0.0082	0.6691	−0.0017
ARIMA(5,0,5)	−35971.57	−35891.10	−0.0000004	0.0123	0.0082	0.6693	−0.0019
ARIMA(4,0,3)	−35945.67	−35920.32	−0.0000277	0.0123	0.0082	0.6694	−0.0001
ARIMA(3,0,4)	−35947.43	−35922.12	0.0000126	0.0123	0.0082	0.6694	−0.0000
ARIMA(5,0,4)	−35947.98	−35922.67	−0.0000101	0.0123	0.0082	0.6687	0.0009
ARIMA(4,0,5)	−35946.76	−35921.52	−0.0000133	0.0123	0.0082	0.6695	0.0012

Performance Metrics for Top 5 Models

- **AIC (Akaike Information Criterion)**: A lower AIC indicates a better-fitting model.
- **BIC (Bayesian Information Criterion)**: Similar to AIC but penalizes more for model complexity.
- **Residual Error Metrics**:
  - **Mean Error (ME)**, **Root Mean Squared Error (RMSE)**, and **Mean Absolute Error (MAE)** quantify the prediction error.
  - **(MPE)** and **(MAPE)** provide relative error metrics.
  - **Mean Absolute Scaled Error (MASE)** evaluates error relative to a simple baseline.
- **Residual Autocorrelation (ACF1)**: Measures the first lag autocorrelation of residuals. A value close to zero suggests uncorrelated residuals.

Based on the AIC and BIC values, the best model is ARIMA(5,0,5), achieves the lowest AIC value (-35971.57). While ARIMA(1,0,0) has a simpler structure and a competitive BIC, its AIC is higher, showing a less optimal fit. Additionally, the residual diagnostics for ARIMA(5,0,5) confirm that it captures the underlying patterns effectively, these makes it the preferred model for forecasting.

6. Ljung-Box Test for Residual Diagnostics

The Ljung-Box test was applied to assess residual autocorrelation:

$$Q = 15.979, \quad df = 3, \quad p\text{-value} = 0.001145.$$

This shows that the (H_0) of no autocorrelation in residuals is rejected, and residual autocorrelation may still exist. That suggests that the ARIMA(5,0,5) model does not capture all the patterns in the data. Further improvement may be done, such as refinement of the parameters of the model or feature inclusions.

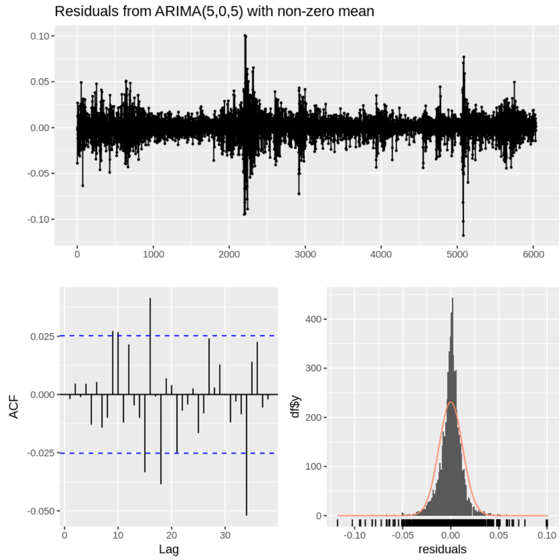
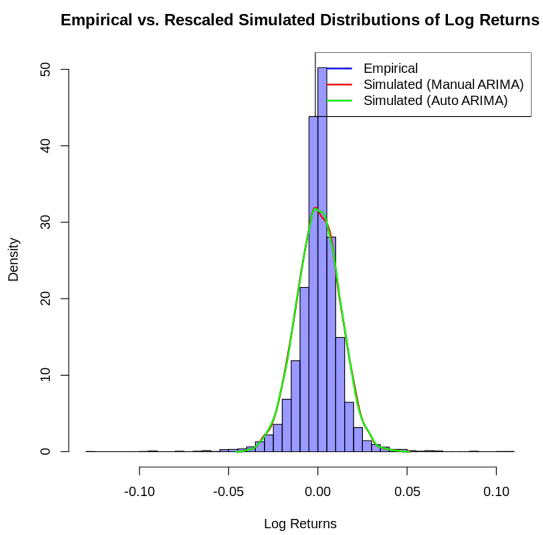


Figure 2: Ljung-Box Test

The Ljung-Box test evaluates whether the residuals of the ARIMA(5,0,5) model are free from autocorrelation. The test statistic Q = 15.979, with 3 degrees of freedom and a p-value of 0.001145, shows an autocorrelation in the residuals, as the p-value is below 0.05. This shows the model does not fully capture all patterns, and improvements may be needed, such as refining the model parameters or incorporating additional features. The red and green lines show simulated distributions based on the fitted parameters of the manual ARIMA and auto ARIMA models, respectively (Skindilias, 2024).



REFERENCES

Skindilias, K. (2024). Week 5 - Moving Average (MA) Models and Stationarity Testing. Available at: https://l24.im/8AZRYbn (Accessed: 24 November 2024).

Skindilias, K. (2024). Week 6 - ARIMA Model Fitting and Diagnostics. Available at: https://l24.im/dci5vN (Accessed: 27 November 2024).

Data Magic Lab (2023). 'Demystifying ARIMA Model Parameters: A Step-by-Step Guide', Data Magic Lab, 30 September. Available at: https://datamagiclab.com/demystifying-arima-model-parameters-a-step-by-step-guide/ (Accessed: 26 November 2024).