S&P 500 Index Analysis and Forecast using SARIMA model

Simon Fraser University

STAT 300W: Statistics Communications

STAT 300W Team 7

Japnoop Grewal	#301301391
Mingyuan He	#301385289
Heeju Choi	#301341550

Introduction

Stock indices track the overall health of the stock market as they are a weighted average of individual stocks. Historical index price shows how the market has reacted; therefore, it helps identify risk-reward and forecast the future.

This analysis will focus on one of the most popular financial indices, the S&P 500. The S&P 500 is a capitalization-weighted index representing the US economy, covering 80% of the US stock market with the 500 largest trading companies. This analysis aims to identify the risk and reward profile and forecast for the next three years with the monthly historical S&P 500 index price data since January 1950. The results of the analysis would help the investors make better investment decisions.

The results of the analysis support that the S&P 500 index has a high likelihood of making a profit. The fitted SARIMA model gives the negative value of the risk of -0.05, which implies that the index has a high chance of making a profit. The forecast for the next three years also shows an upward trend where the predicted index prices are increasing.

Statistical Analysis

Study design

Data for the analysis is time series with the monthly index prices collected between January 1950 and March 2022. The overall sampling strategy and baseline study protocol of the S&P 500 index are based on the eligibility criteria of the Index Committee. The time series data with live S&P 500 index price may provide the most accurate analysis as it takes detailed prices. However, the monthly index prices are used in the study for cost and time efficiency. The 867 monthly S&P 500 historical index prices are the average closing prices (*S&P 500 Historical Prices by Month*, n.d.). Also, the prices in the analysis are the actual price value without any adjustment for inflation.

Variables

The analysis of the dataset requires the use of the two key variables. The variables are time and prices of the S&P 500 index. The monthly prices of the S&P 500 index from January 1950 to March 2022 are a response variable. The monthly date is an explanatory variable supporting the index prices.

Analysis Method

Firstly, these data are a collection of observations obtained through repeated measurements over time, meaning they are time series data. Secondly, this study aims to determine the risk-reward profile in the next three years. We should use the model which has the function of prediction. Thirdly, the market moves in cycles, which means seasonality is one of the properties of the market. Therefore, it is a good idea to predict the future value of the S&P index using the SARIMA model. SARIMA is a seasonal ARIMA model, one of the most often employed forecasting techniques for univariate time series data. In addition, the SARIMA model permits the direct modeling of the seasonal component of the series, which is an extension of the ARIMA model (Jason Brownlee, August 17, 2018). Then we start to fit the model using R.

From the ACF plot (plot 1 in appendix), This series is not stationary, as seen by the auto-correlation, which is usually relatively high and is dropping very slowly, comparable to that of a Random Walk Process. With a high p-value of 0.99, the Augmented Dickey-Fuller Test (ADF test), which tests the null hypothesis in a time series, cannot reject that the series is non-stationary at the 95 percent confidence interval. Then we need to transform to ensure the data is stationary.

From plot 2 (in appendix), The λ value is -0.05; the Box-Cox transformation method suggests that natural log transformation is needed. However, because of the random walk and the result of the ACF of our original series, it is good to perform log-differencing to the actual time series.

After we transform the data from the plot of log differenced series (plot 3 in appendix), we can observe that the log-differenced figure exhibits characteristics of a stationary model with constant mean and variance, with slight signs of seasonality. From plot 3, the ACF cuts off at lag 4, lag 20, and lag 21. The PACF cuts off at lag 4 and lags 5. Then we can fit the model as SARIMA(0,0,2) X (3,0,0)12, in which '2' means that q value that the number of PACF cutting off, '3' means that P value that the number of ACF cutting off, and '12' means seasonality. However, when we use the auto.arima() function in the forecast package, which can determine the best parameters of the SARIMA model, the result is SARIMA(0,0,1) X (1,0,0) 12. Then we should perform some diagnostic checking to determine which model is better.

After we used sarima() function in the astsa package, we got the plot 4 (in appendix)which is diagnostic of SARIMA(0,0,2) X (3,0,0)12, and plot 5 (in appendix)which is diagnostic of SARIMA(0,0,1) X (1,0,0) 12. Compared to the ACF of residuals and normal Q-Q plot of standard residuals in plot 4, it is similar to them in plot 5. We cannot determine which model is better by these two plots. Then we compare their AIC values. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. The lower the AIC value is, the better the model fit. The AIC value of SARIMA(0,0,1) X (1,0,0) 12 is -3.910382 which is smaller than that of SARIMA(0,0,2) X (3,0,0)12 which is -3.904259. In addition, we manually changed d=1 to account for differences when forecasting the trend because the data we used in forecasting is transformed to log form. Therefore, The model we fit is SARIMA(0,1,1) X (1,0,0) 12.

We use the risk at risk to evaluate the risk for the risk measurement. Value at risk (VaR) measures how much money could be lost by investing in the S&P 500 over a certain period (Will Kenton, June 03, 2022). We use the sarima.for() function to foresee the expected index price for the next three years.

Results

To determine the risk of the index, we use value at risk to determine the chances of making a profit. The Value at Risk plot is shown below in Figure 1. The plot gives a small negative value being -0.05. This value tells us that the index has a high probability of making a profit. Specifically, it implies the investor will lose a maximum of 500 dollars if they invest 10,000 dollars, where the index has a high probability of making a profit compared to the other financial indices. This high probability shows that it may be sensible to invest in the index.

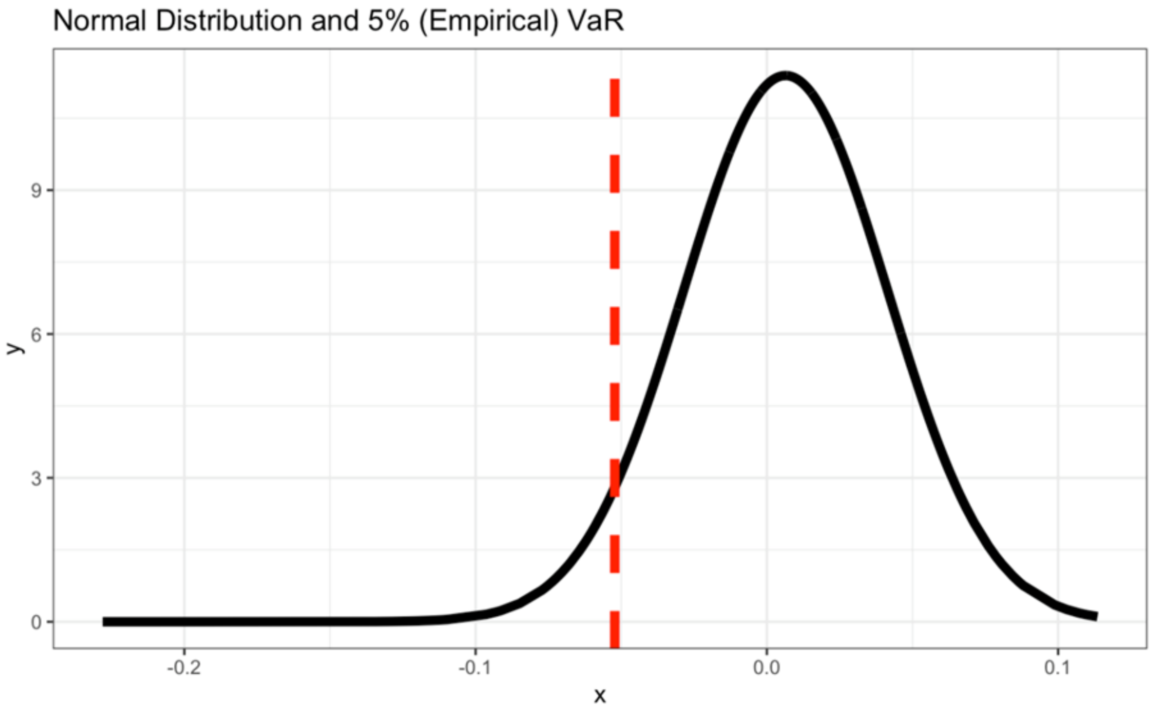


Figure 1 (Value at Risk plot to determine the likelihood of making a profit in the index)

Following the analysis of multiple SARIMA models, we find the best model. We now have what we perceive to be the forecast for the next three years using the model seen in the analysis and diagnostic step. The forecast, with confidence intervals, is shown in figure 2. The forecast shows what we believe the price of the S&P 500 index should be. The grey shaded region and purple outline show the certainty values for the forecast being 95% and 99%, respectively. The forecast shows a general upward trend. This upward trend indicates a strong possibility that the price of the S&P 500 index will rise.

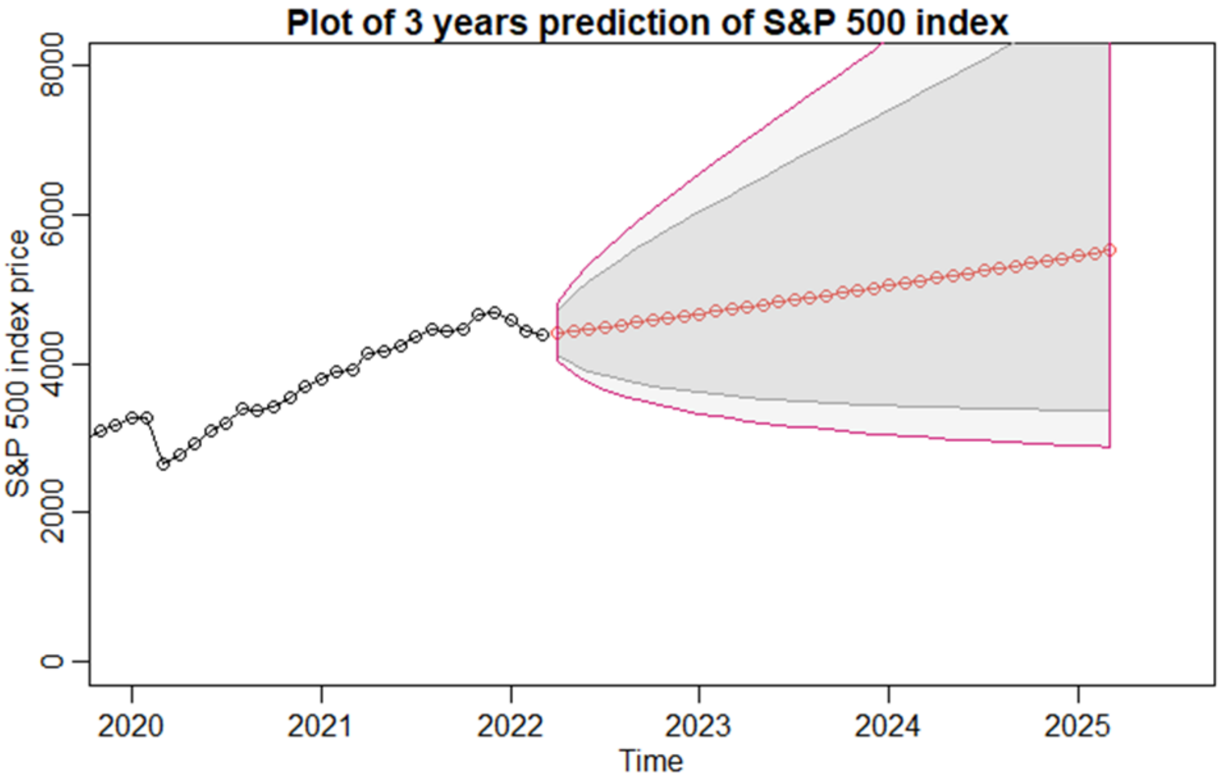


Figure 2 (3-year forecast of the S&P 500 index using SARIMA model)

Conclusion

Our analysis has confirmed that S&P 500 index remains a significant economic health indicator of the United States, as it covers approximately 80% of the US stock market with 500 large companies in capitalization-weight. Thus, it provides a representative and reliable risk-reward profile and forecast on future market turmoil. The analysis investigated the risk-reward profile of the S&P 500 index and forecast for the next three years. The risk and reward profile shows a high likelihood of making a profit. The forecast also shows a general upward trend for the next three years. The overall results imply that the probability of the S&P 500 increasing is high with some risk involved. The analysis emphasizes the importance of understanding the risk and forecast associated with the S&P 500 index, as this knowledge is essential for investors to make wise investment decisions about the S&P 500 index.

Discussion

The analysis demonstrated the statistical modeling with the monthly time series data to observe the risk-reward profile and the forecast. The monthly index may not truly represent the live index price, and its result will turn out differently from the S&P 500 analysis with the live index price. Also, the index is in share price without inflation adjustment. The money value must be different in 1950 and 2022; that is, the increase in the index price may be due to the increase in the money value instead of the rise in the index price. The detailed live index with inflation-adjusted prices can help improve the analysis accuracy and provide more accurate insight into the S&P 500 index.

Reference

1 *S&P 500 Historical Prices by Month*. (n.d.). Multpl.
https://www.multpl.com/s-p-500-historical-prices/table/by-month

2 Jason Brownlee, (August 17, 2018). A Gentle Introduction to SARIMA for Time Series Forecasting in Python
https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

3 Will Kenton, (June 03, 2022). Value at Risk (VaR)
https://www.investopedia.com/terms/v/var.asp

Appendix - Code

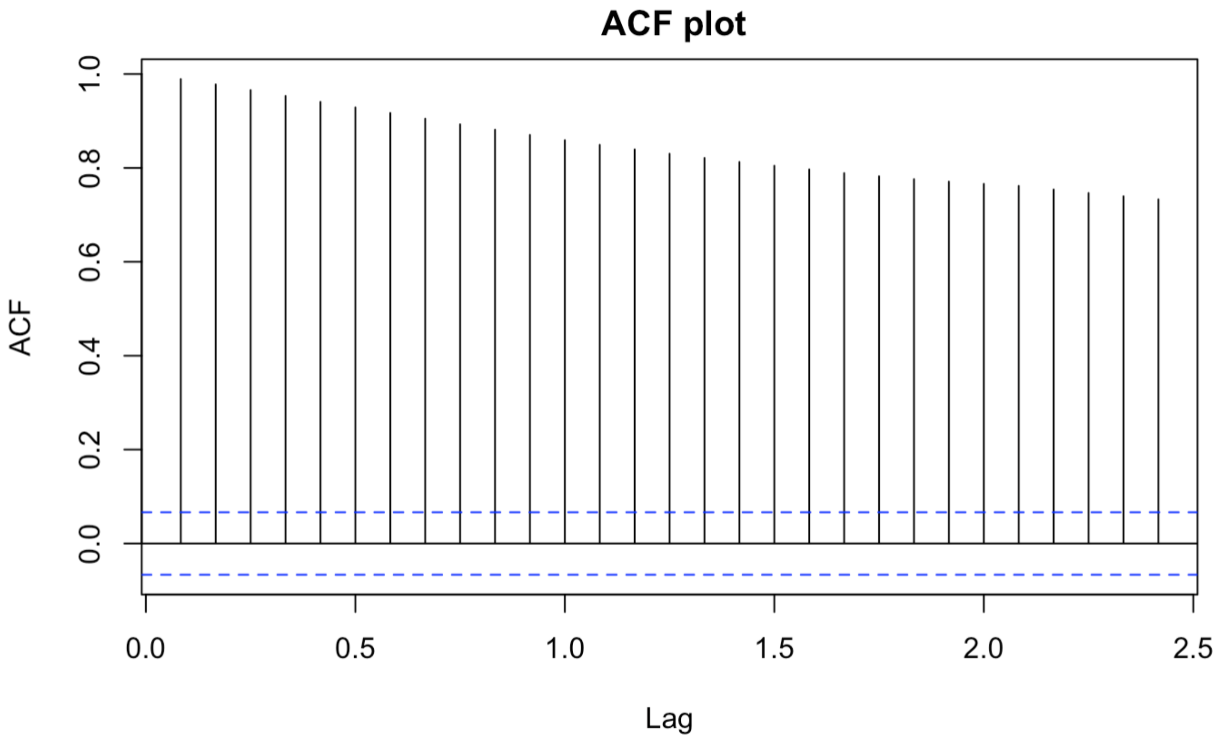
~~~
##Set up
```{r setup, include=FALSE}
library(readxl)
library(TSA)
library(forecast)
library(MASS)
library(tseries)
library(astsa)
library(timeSeries)
library(xts)
```
## Load data
```{r SP500}
y<- read_excel("SP500.xlsx")

sp500 = y[order(nrow(y):1),]
sp500$Date <- as.Date.character(sp500$Date, "%B %d %Y")

ts <- ts(sp500$`S&P 500 Price`,start = 1950, frequency = 12)

plot(ts, ylab = "S&P 500 Price", main = "S&P 500 index from 1950 to 2022")
```
## Initial assessment
```{r Initial Assessment}
##Check stationarity

acf(ts, main = " ACF plot")
adf.test(ts)
# as its p-value is 0.99 which fails to reject the null hypothesis, the data is not stationary. So we
require to transform data to make it stationary
```
## Transformation
```{r Transformation}

#Determine the transformation method
#par(cex=0.8)
box_cox_transformation<-boxcox(ts~1, lambda=seq(-1,1,0.1))
y_values<-as.numeric(box_cox_transformation$y)
~~~

~~~
#extract lambda
lambda<-box_cox_transformation$x[which.max(y_values)] #-0.0505
abline(v=lambda, col=2, lty="dashed")
text(lambda+0.05,max(box_cox_transformation$y), round(lambda,2), cex=0.85)
title(expression(paste("The ML estimate of ", lambda )), cex=0.85)

#tranform the data
diff_ts1 = diff(log(ts))
adf.test(diff_ts1) #confirmed stationary
plot(diff_ts1, main="Plot of log differenced series")
ggtsdisplay(diff_ts1)
```
## Model Diagnostics
```{r SARMIA Model}
#parameter estimate
auto.arima(diff_ts1, trace=TRUE) #ARIMA(0,0,1)(1,0,0)[12]

fit <- sarima(diff_ts1, 0,0,1,1,0,0,12)
fit2 <- sarima(diff_ts1,0,0,2,3,0,0,12)
fit
fit2

#We fit the SARIMA(0,1,1)X(1,0,0)12 model
```
##Risk Measurment
```{r Risk Measurement}
#Value at Risk (VaR) is the most widely used market risk measure in financial risk management
and it is also used by practitioners such as portfolio managers to account for future market risk.
VaR can be defined as loss in market value of an asset over a given time period that is exceeded
with a probability

library(ggplot2)
library(quantmod)
ts_return = dailyReturn(ts, type = "log") #getting same as diff_ts1
den_ts = coredata(ts_return)
distr_ts = dnorm(x = den_ts, mean = mean(den_ts), sd = sd(den_ts))
data_rd = data.frame(den_ts, distr_ts)
# change column names
colnames(data_rd) = c("x", "y")
~~~

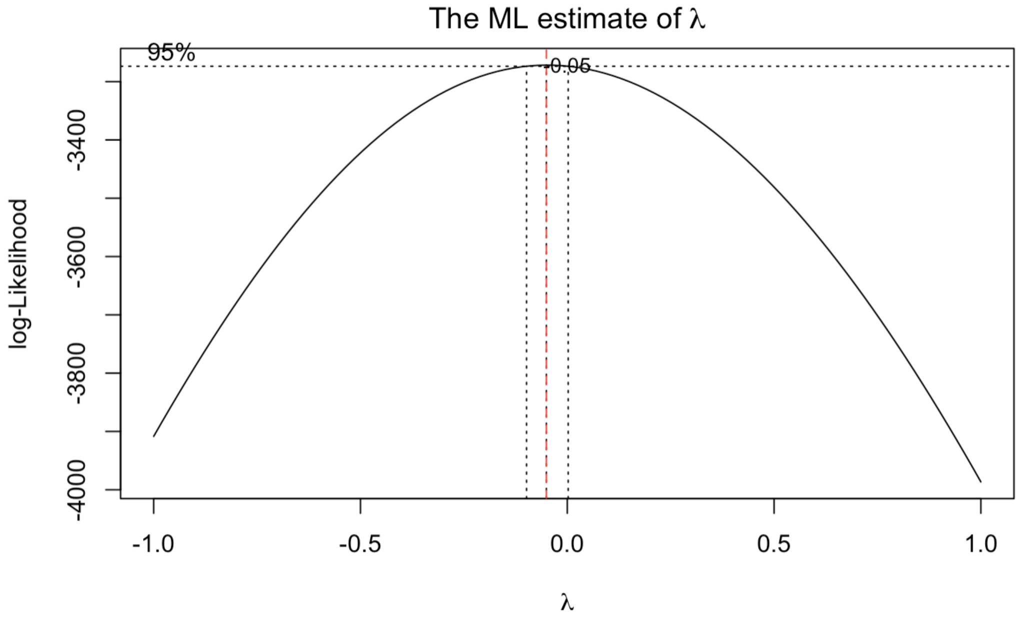
~~~
# normal quantile
var1 = quantile(den_ts, 0.05)
p3 = ggplot(data_rd, aes(x = x, y = y)) + geom_line(size = 2) + geom_vline(xintercept = var1,
    lty = 2, col = "red", size = 2) + theme_bw() + labs(title = "Normal Distribution and 5%
(Empirical) VaR")
p3
```
## Forecasting
```{r Forecasting}
#sarima,for function
forecast <- sarima.for(log(ts),0,1,1,1,0,0,12, n.ahead=36)

#actual return
exp(forecast$pred)

#taking out log-transformed to get actual return

#95% CI P.E.
log_upper = forecast$pred + 1.96 * forecast$se
log_lower = forecast$pred - 1.96 * forecast$se
U = exp( log_upper )
L = exp( log_lower )

#99% CI  P.E.
log_upper3 = forecast$pred + 2.58 * forecast$se
log_lower3 = forecast$pred - 2.58 * forecast$se
U3 = exp( log_upper3 )
L3 = exp( log_lower3 )

plot(ts,xlim =c(2020, 2026),  ylim=c(0, 5500), type = 'o',
main = "Plot of 3 years prediction of S&P 500 index ", ylab="S&P 500 index price")
lines(exp(forecast$pred), col = 2, type = 'o' )

xx = c(time(U), rev(time(U)))
yy = c(L, rev(U))
xx3 = c(time(U3), rev(time(U3)))
yy3 = c(L3, rev(U3))
polygon(xx, yy, border = 8, col = gray(0.6, alpha = 0.2)) #gray
polygon(xx3, yy3, border = 6, col = gray(0.6, alpha = 0.1)) #purple
```
~~~

Appendix - Plot

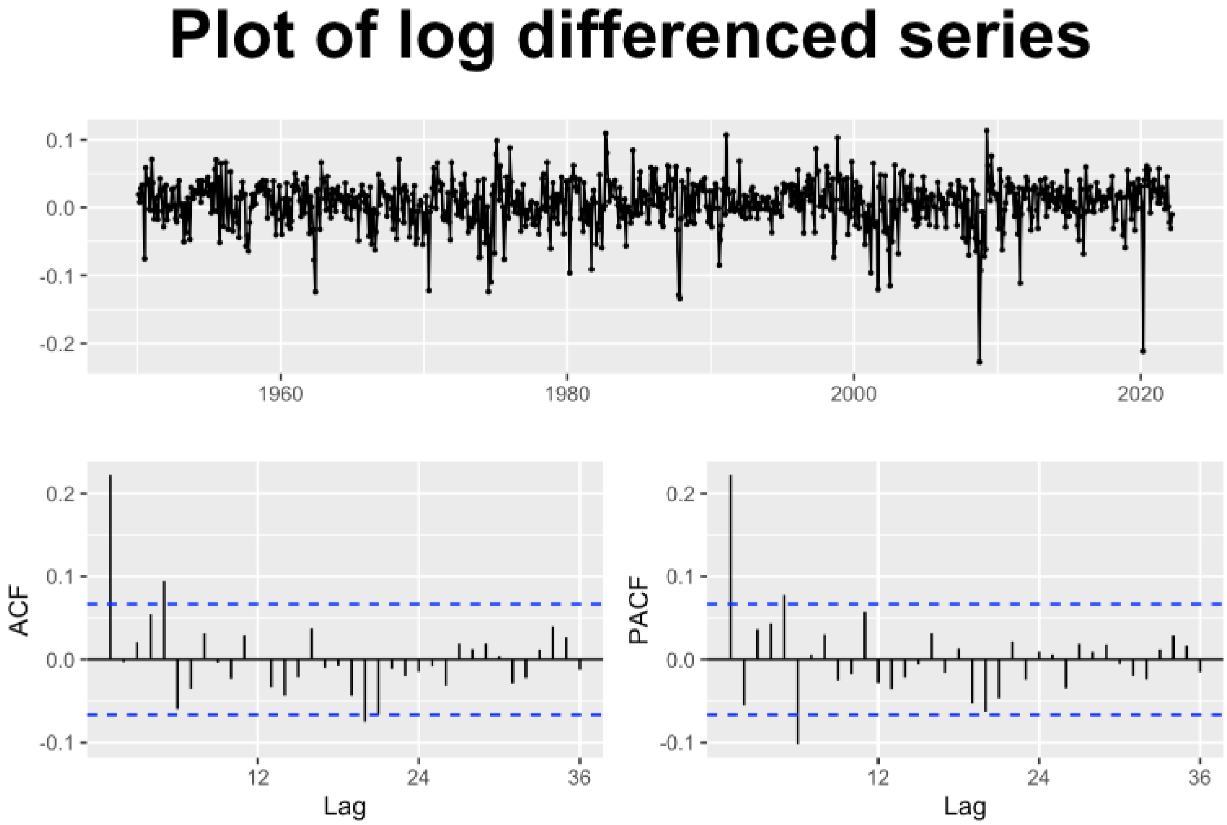
Plot 1 - ACF plot of the original time series data



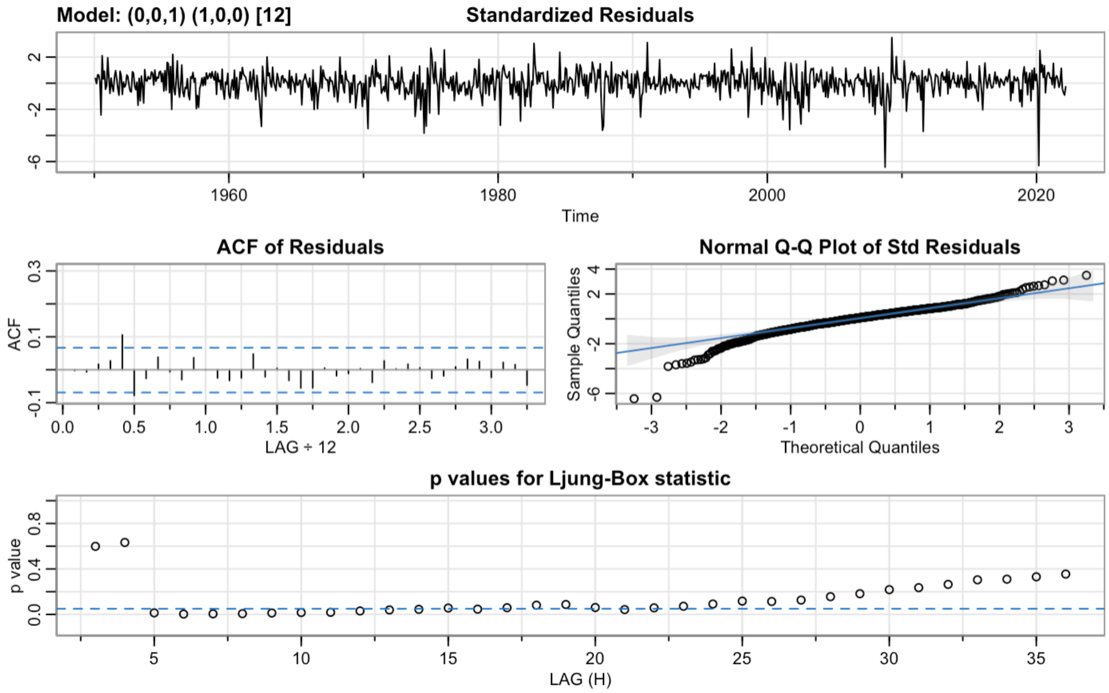
Plot 2 - The ML estimate of λ



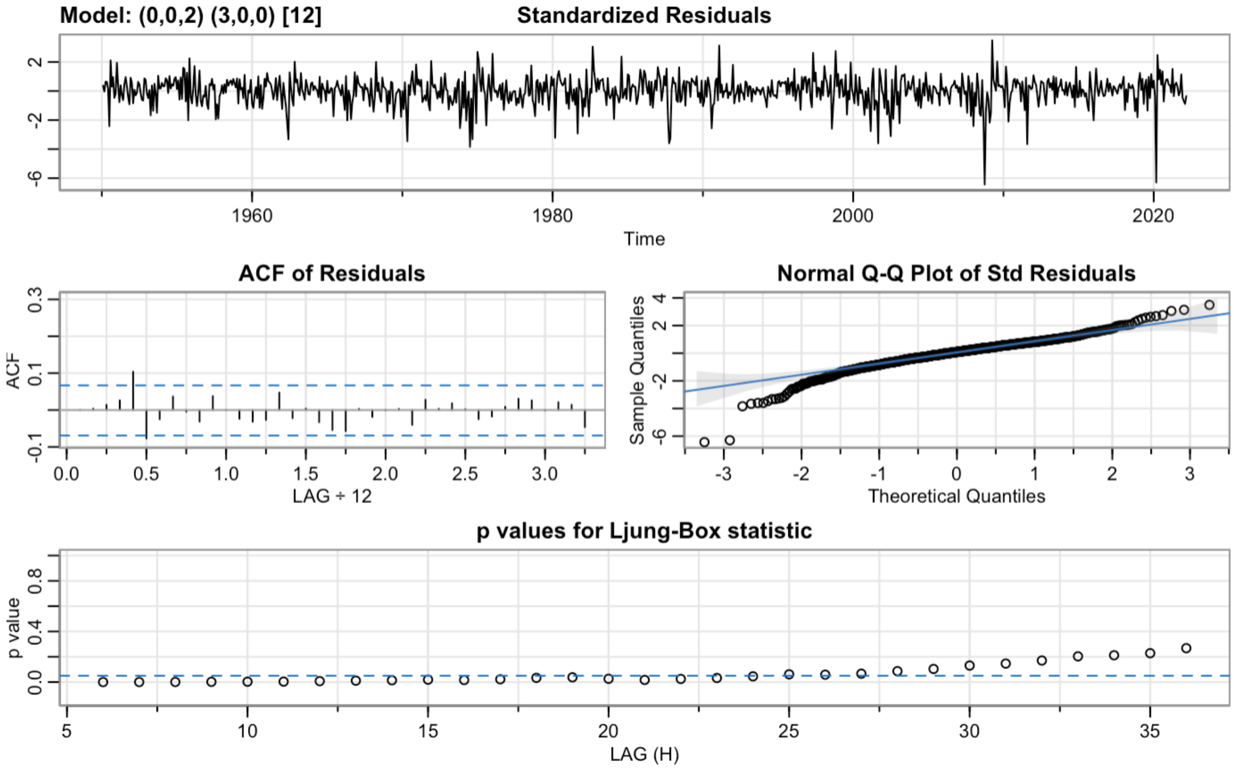
Plot 3 - Time series Plot, ACF plot, and PACF plot for log-differenced data



Plot 4 - Residuals plot of SARIMA (0,0,1) X (1,0,0)12



Plot 5 - Residual plot of SARIMA (0,0,2) X (3,0,0)12


Model: (0,0,2) (3,0,0) [12]
Standardized Residuals
Time
ACF of Residuals
ACF
LAG ÷ 12
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
LAG (H)