PSTAT 174 Final project: Analysis and Forecasting of Quarterly Earnings per Johnson & Johnson Share

Jinglin Yang (jinglinyang@ucsb.edu)

2024-06-11

Abstract

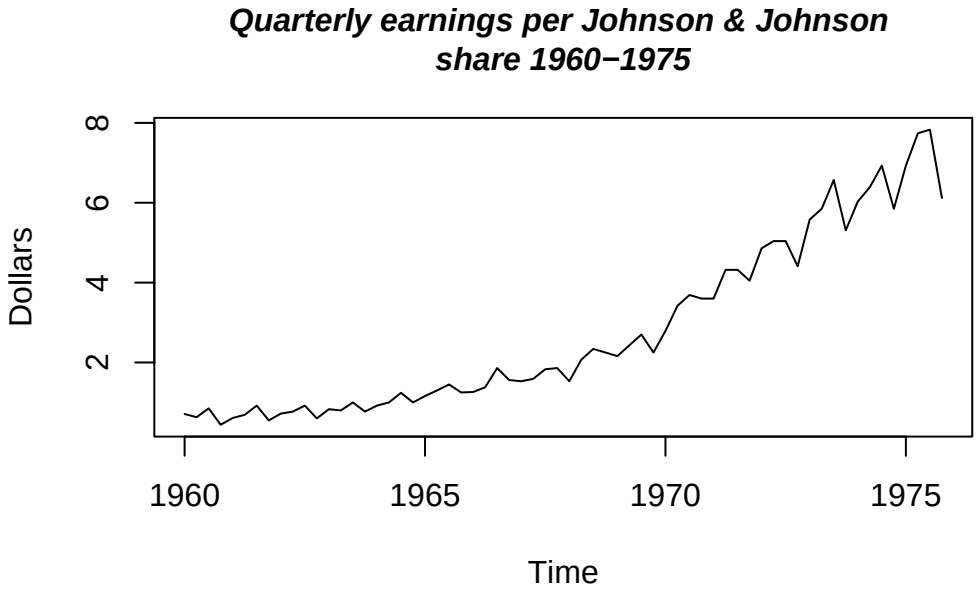
This report explores the time series of quarterly earnings per Johnson & Johnson share in dollars and employs Box-Jenkins method to find a adequate fitted model and forecast the value of future 12 values. In addition, spectral analysis is used to explore the data from a different perspective of cycles, rather than time periods.

1. Introduction

In this report, I chose to analyze the quarterly earnings per Johnson & Johnson share in dollars, *JohnsonJohnson* from the R packages *datasets*, from 1960 to 1980. The purpose of the analysis is to find a good fit, not necessarily perfect, model for the data and predict the future values based on the selected fitted model. Since Johnson & Johnson is a well-known and successful company, I want to study the trend of earnings of its share back in the old days, rather than the recent trend. It would be interesting to examine the change of the share earnings since earnings of the company's share is linked to the potential of that company. I used the Box-Jenkins method to fit the SARIMA model rather than ARIMA model due to its seasonality, and choose the one model according to coefficients test and the comparison of AIC and BIC.

2. Exploratory Data Analysis

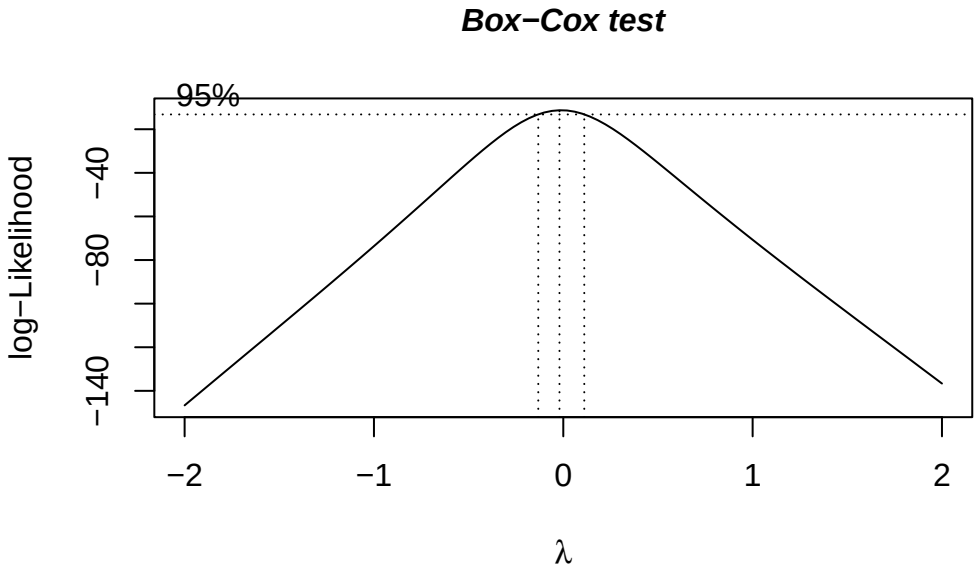
The time series employed for analysis starts from January 1960 to December 1975, and data from January 1976 to December 1978 is employed to compare the predicted values and the real observed values. The data were collected by quarters, and 64 quarters were observed, i.e., the sample include 64 observations. The data were provided (personal communication) by Professor Paul Griffin, https://gsm.ucdavis.edu/profile/paul-griffin, of the Graduate School of Management, University of California, Davis.

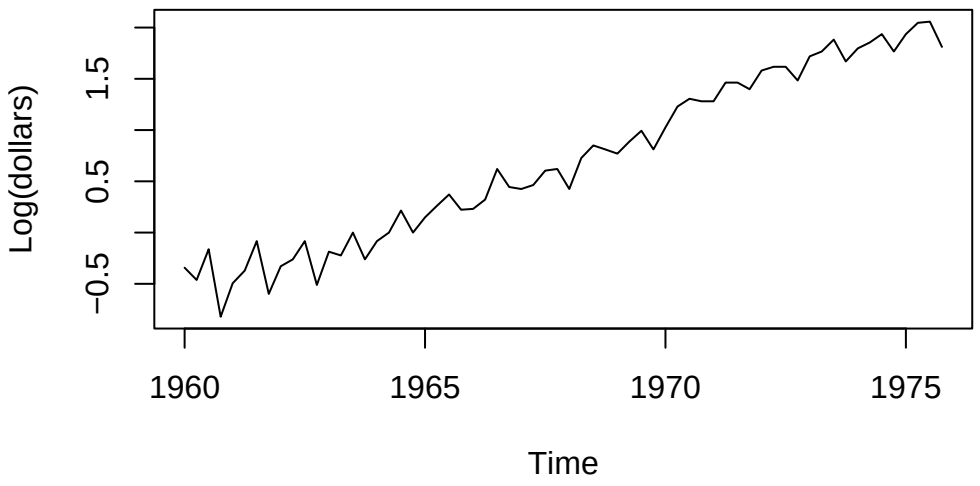


Box-Cox transformation

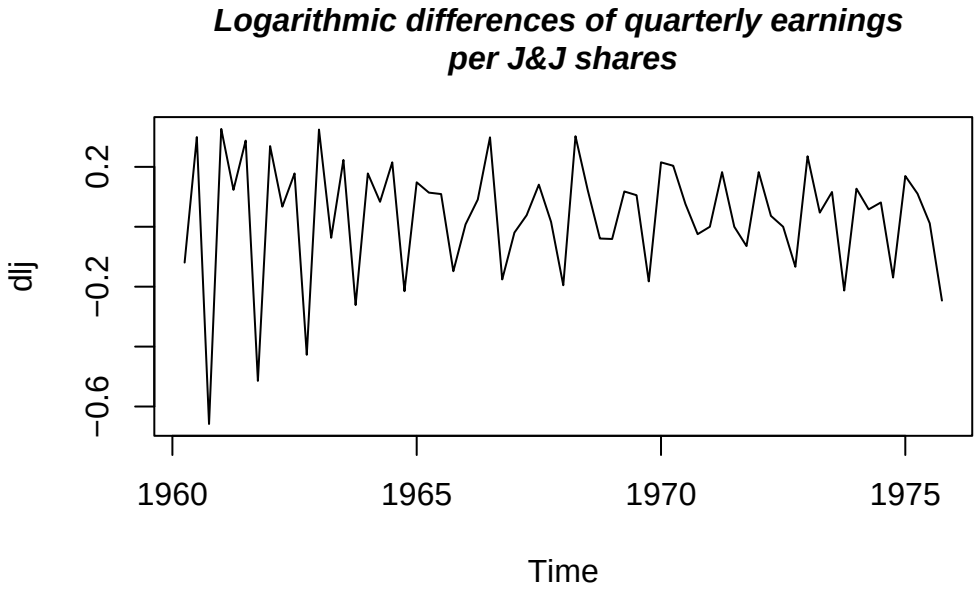
From the plot, the time series is not stationary, and there are obvious trend and seasonality. To attain stationery, Box-Cox test and transformation is performed. According to the optimal λ and the corresponding confidence interval, the optimal λ is approximated to 0, and thus logarithmic transformation is employed.

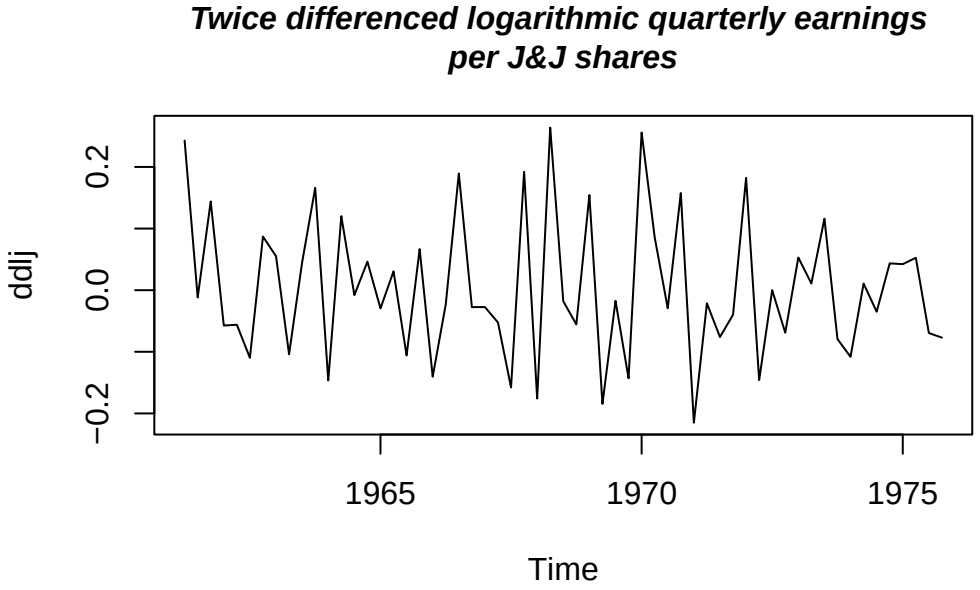
$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ log x_t & \text{if } \lambda = 0 \end{cases}$$





Even though the data is transformed, the trend is still presented, and thus differencing at lag 1 is used to detrend the logarithmic data as a way to obtain stationery. Furthermore, the original data are recorded quarterly and shows some seasonal trend across time, so it is natural to assume the seasonal period $s = 4$. The transformed data is again difference at lag $= 4$ to remove seasonality. $y_t = (1 - B)(1 - B^4)x_t = \nabla\nabla_4 x_t$





According to the Augmented Dickey-Fuller Test(ADF test), at 5% significance level, the p-value is equal to 0.01, and thus we reject the null hypothesis so that the twice differenced logarithmic data is a stationary series and has constant variance over time. Therefore, it can be used for further analysis.

3. Methodology

3.1 $SARIMA(p,d,q)(P,D,Q)_s$ model

Prior to model fitting and selection, we need to check the stationery of the data. That being said, if the data does not appear to be a stationary series, Box-Cox transformation is performed to achieve stationery. With Box-Jenkins method, by examining the plot of ACF and PACF, p, the non-seasonal Auto-regressive parameters(AR), is defined by the cutoff from the PACF, and q, the non-seasonal moving average parameters(MA) is defined by the cutoff from the ACF. If the data has a trend, then differencing, $\nabla = 1 - B$ is used to eliminate the trend and decrease the variance in order to approach stationary. In addition, the Johnson & Johnson data is collected as quarterly data, so assumption for seasonal period $s = 4$ is made, $\nabla_4 = 1 - B^4$. In order to define the seasonal parameters, the ACF and PACF twice differenced data on non-seasonal order of d and seasonal order of D. From the plots, P(SAR) is defined by the cutoff lag from PACF and Q(SMA) is defined by the cutoff lag from ACF.

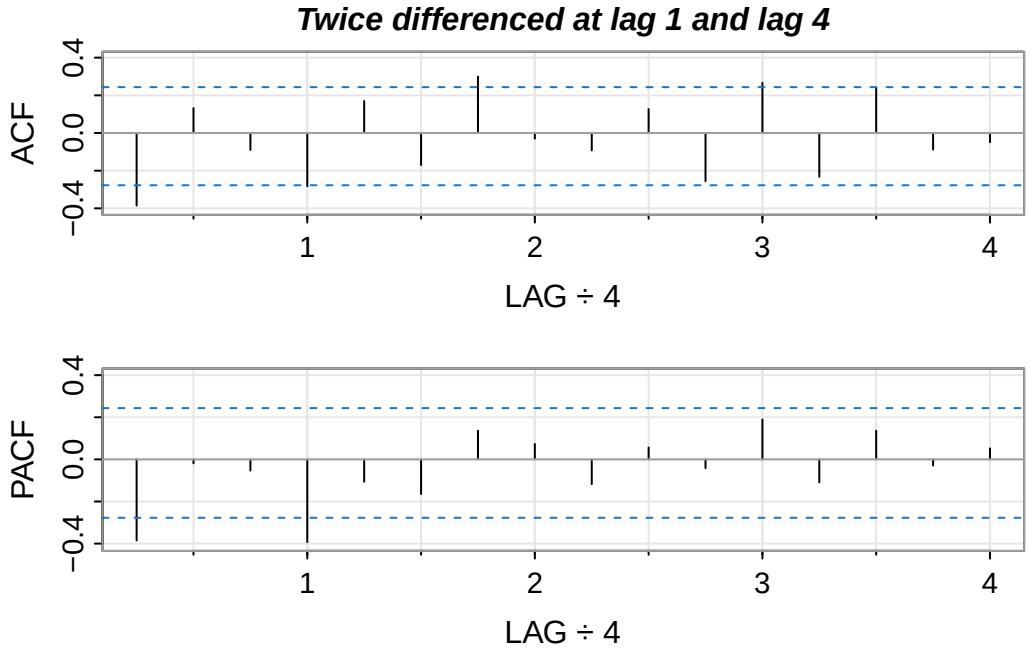
3.2 Spectral Analysis

As a method to analyze time series, the data is examined in terms of periodicity and frequency of periodic process as trigonometric function $x_t = Acos(2\pi\omega t + \phi)$. The spectral analysis does the same work as auto-covariance function, but the difference is that spectral analysis analyzed the time series in terms of frequency, ω, while auto-covariance analyzed in terms of the real timer period, the lags, h. Periodogram explains the power of specific frequency across different frequencies of that data set and is defined by *discrete Fourier transform*(DFT), $I(\omega_j) = |d(\omega_j)|^2$, where $\omega_j = j/n$ is the fundamental frequencies, $j = 1, 2, ..., n-1$ and n is the total number of samples. In this particular case, $n = 64$. Larger the value of periodogram, stronger the power of that frequency and more dominant the frequency is. In addition, it helps detecting the possible seasonal pattern and components of the data and examining the choice of parameters of SARIMA model.

4. Results

4.1 Model selection

Model fitting



From the ACF and PACF of the twice differenced data, d = 1, D = 1, and S = 4 are determined because the data are twice differenced at lag 1, and lag 4 due to seasonality. In addition, there is a spike at lag 1 of PACF, indicating possible P = 1, and there is also a spike at the first tick (corresponding lag 1 at undifferenced data), so we can say that p = 1. In ACF, there are spikes at lag 1 and lag 3, indicating that Q = 1 or 3. Moreover, the first tick of ACF is a spike, indicating that q = 1. In the plot of ACF, the significant values at 7th lag is noticeable. However, in my opinion, it might result in overfitting the model with the number of parameters beyond the length of a season, i.e., fitting the model where q = 7 while s = 4.

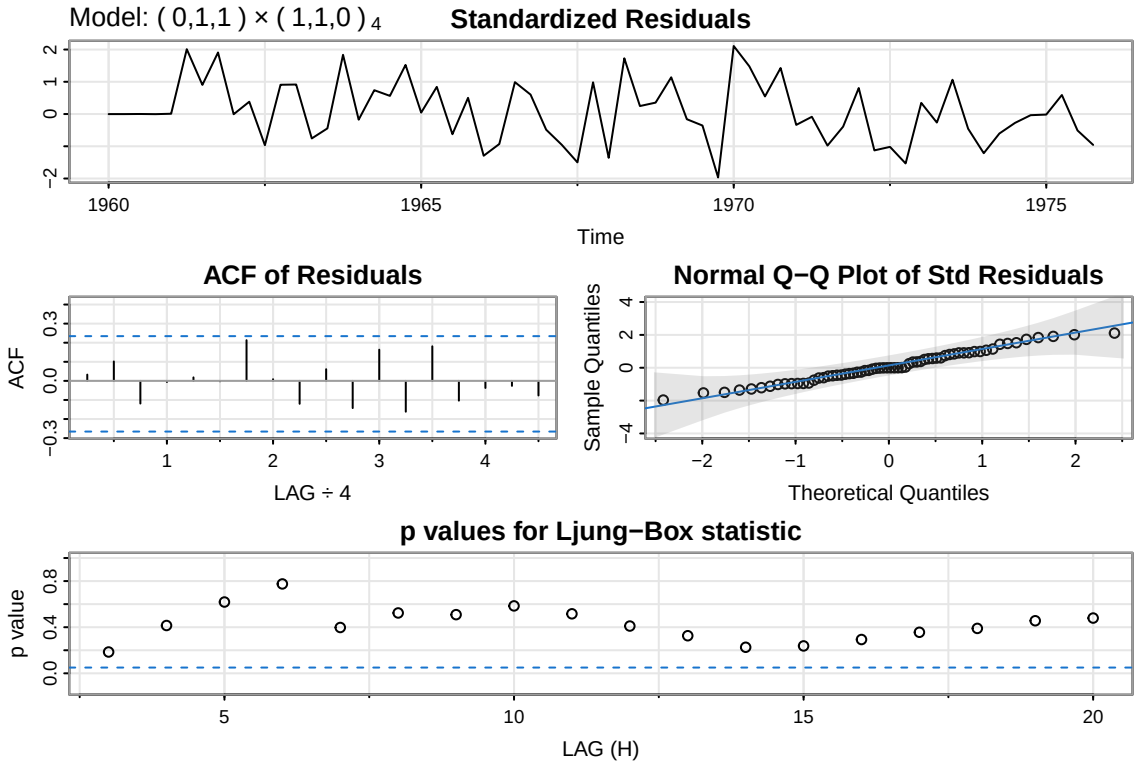
Model diagnostics

The following models are analyzed:

Model	Coefficient Test	AIC	BIC
$SARIMA(1,1,1)(1,1,1)_4$	AR1, SAR1, SMA1 non-significant	-102.02	-89.62991
$SARIMA(1,1,0)(1,1,1)_4$	Seasonal components non-significant	-100.9	-90.58972
$SARIMA(0,1,1)(1,1,1)_4$	Seasonal components non-significant	-103.99	-93.67572

Model	Coefficient Test	AIC	BIC
$SARIMA(0,1,1)(1,1,0)_4$	All components significant	-105.82*	-97.5847*
$SARIMA(0,1,1)(1,1,3)_4$	SMA2, SMA3 non-significant	-103.84	-89.3709
$SARIMA(0,1,1)(0,1,1)_4$	All components significant	-104.98	-96.752
$SARIMA(0,1,1)(0,1,3)_4$	SMA2, SMA3 non-significant	-103.26	-90.87663

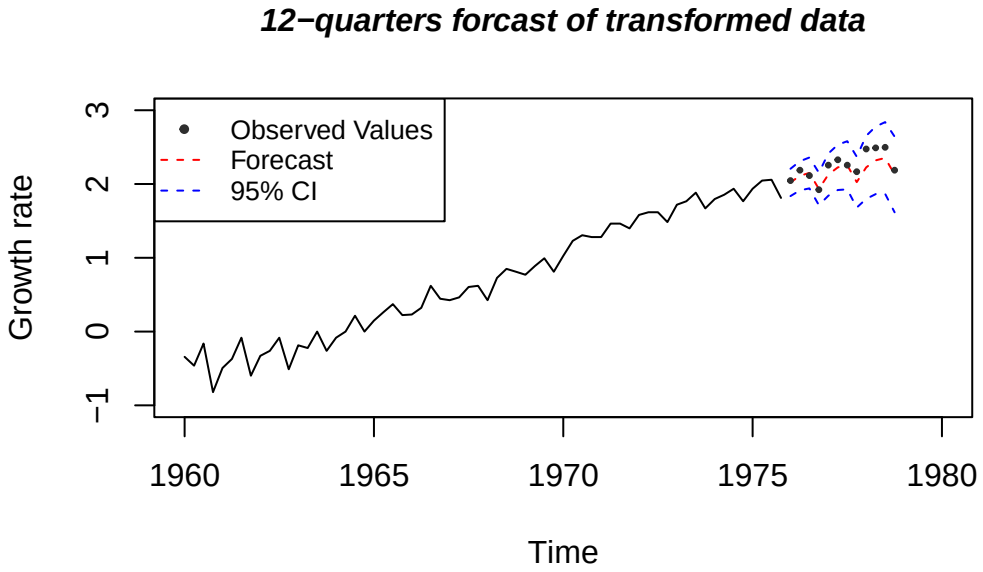
To determine the one that has the best performance, *Akaike Information Criterion*(AIC) a*nd Bayesian Information Criterion*(BIC) are selected as the metrics to make comparisons. Among the seven fitted model, $SARIMA(0,1,1)(1,1,0)_4$ has all components significant and smallest AIC and BIC, and thus it is selected as the best candidate of the seven fitting models.

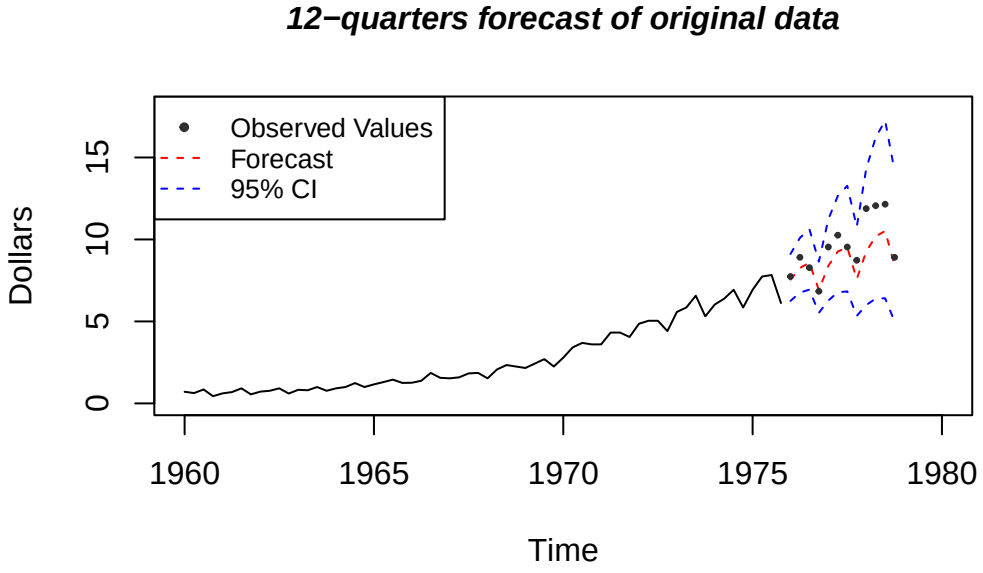


To diagnose the chosen model, $SARIMA(0,1,1)(1,1,0)_4$, there are no obvious pattern in the distribution plot of residuals of the fitted mdoel, and the normal q-q plot of residuals indicates that the assumption of normality is reasonable. Plus, the plot of estimated ACF of residuals does not shows any spikes and apparent departure from the model, and the Q statistics is not significantly different. Thus, we failed to reject the null hypothesis that residuals are white noise, and then concluded that this model can be used as the final model for future predictions.

4.2 Forecast

To forecast the future 12 values, we predict the quarterly earnings per J&J share from 1976 Q1 to 1978 Q4. Since the data recorded the actual observed values for that 12 quarters, we can compare the observed values and the forecast values to examine the accuracy of the fitted model.

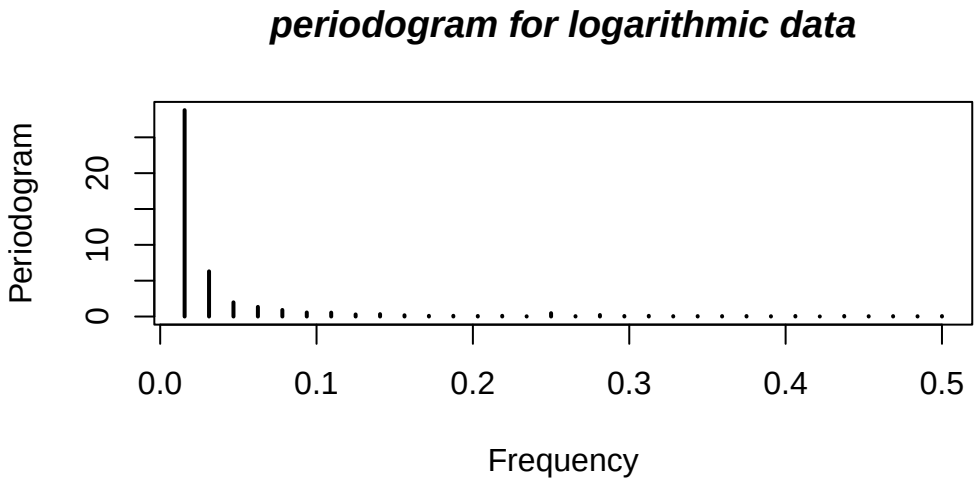


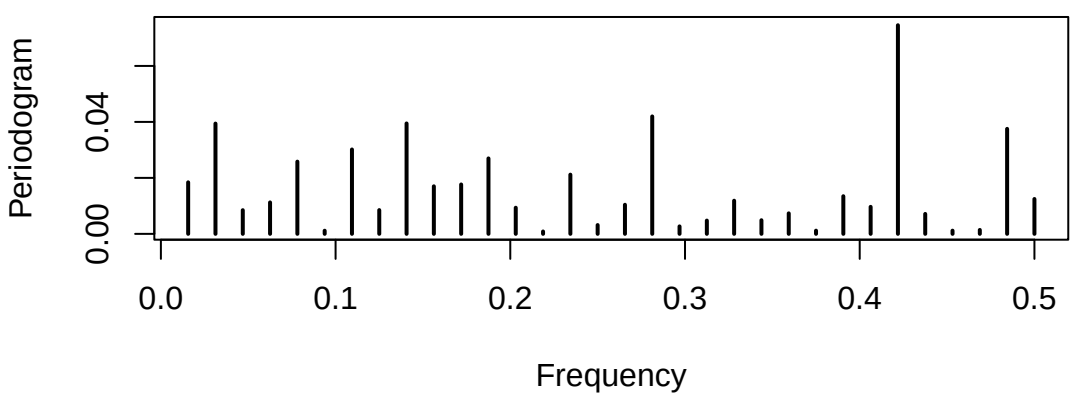


From the plots of forecasting from both transformed and original data, the observed values are mostly fall along the forecast values, except for Q1, Q2, and Q3 in 1978. However, the observed values fall into the 95% confidence interval of forecasting, and thus the model perform adequately.

4.3 Spectral analysis

The periodogram shows a spike at $\omega \approx 0.25 \approx 1/4$, indicating that there is a dominant frequency at 0.25 and a pattern of 4 cycles per year. The frequency matches the seasonality we determined in the previous section, implying that the model correctly identify the seasonal component to fit the data.





The periodogram plot graph the spectral density of the residuals of the fitted model, $SARIMA(0,1,1)(1,1,0)_4$, and it does not show specific dominant frequency and appear to a noise pattern. In this case, the residuals does not appear to have seasonality or trend, which means the model performs well.

Conclusion

This analysis utilizes Box-Jenkins method to identify a time series model for the quarterly earnings per Johnson & Johnson share data set in order to predict the 12 future values. The data set examined include 64 observations, in which observations of 64 quarters were recorded over 16 years from 1960-1975. The original time series contain trend and seasonality, so Box-Cox transformation is performed to obtain a stationary series, and the twice differenced, at lag 1 and lag 4, logarithmic time series is checked by ADF test. The model fitting is based on the plots of ACF and PACF of the transformed time series, and the seasonality is suggested by the frequency the data was collected. From the analysis of AIC and BIC, $SARIMA(1,1,0)(0,1,1)_4$ is selected as the best fit among the seven potential fitted model. To diagnose the adequacy of selected model, the distribution of residuals, the ACF of residuals, the normality of residuals, and the Ljung-Box Test for residuals are checked, and $SARIMA(1,1,0)(0,1,1)_4$ is a good fit. In addition, spectral analysis is performed to first see if the dominant frequency corresponds to the seasonality and second provide credibility for the choice of seasonal component in selecting SARIMA model rather than ARIMA model. The plot of forecasting indicates that the selected model can accurately predict values for future 2 years, since the predicted values fall greater apart from the observed values. However, all the observed values of the predicted period falls into the 95% confidence interval. Overall, the selected model is a good fit for this time series, and there are increasing trend for earnings of Johnson & Johnson share in that period. Therefore, at that time, it is a worth investing company.

References

Shumway, Robert H., and David S. Stoffer. 2017. *Time Series Analysis and Its Applications*. Springer International Publishing. https://doi.org/10.1007/978-3-319-52452-8.

Appendix

```
library(astsa, warn.conflicts = F)
library(tseries, warn.conflicts = F)
library(forecast, warn.conflicts = F)
library(MASS, warn.conflicts = F)
library(TSA)
library(latex2exp)
jj <- JohnsonJohnson[1:64] %>% ts(start = c(1960,1), end = c(1975, 4), frequency = 4)
plot(jj, main = "Quarterly earnings per Johnson & Johnson \nshare 1960-1975", ylab = "Dollars
bc <- boxcox(jj~time(jj), lambda = seq(-2, 2, by = 0.1))
lambda <- bc$x[which.max(bc$y)] #-0.02020202
lj <- log(jj)
title("Box-Cox test", cex.main = 1, font.main = 4)
plot(lj, ylab = "Log(dollars)", main = "Logarithmic quarterly earnings per J&J share 1960-197
# first difference due to trend
dlj = diff(lj)
plot(dlj, main = "Logarithmic differences of quarterly earnings \nper J&J shares", cex.main =

# the variance of second differenced is smaller than the first differenced
#var(dlj)
#var(ddlj)
# second difference due to seasonality
ddlj = diff(dlj, 4)
plot(ddlj, main = "Twice differenced logarithmic quarterly earnings \nper J&J shares", cex.ma
# check if the transformed data is stationary
adf.test(ddlj)# it is stationary
model1 = sarima(lj, 1,1,1, 1,1,1,4) # ma1 significant, AIC = -1.695214  AICc = -1.682659  BIC
```

```
model2 = sarima(lj, 1,1,0, 1,1,1,4) # ar1 significant, AIC = -1.676269  AICc = -1.668873  BIC

model3 = sarima(lj, 0,1,1, 1,1,1,4) # ma1 significant, AIC = -1.728574  AICc = -1.721178  BIC

model4 = sarima(lj, 0,1,1, 1,1,0,4) # All significant, AIC = -1.759615  AICc = -1.755983  BIC

model5 = sarima(lj, 0,1,1, 1,1,3,4) # sma2, sma3 not significant, AIC = -1.726036  AICc = -1

model6 = sarima(lj, 0,1,1, 0,1,1,4) # All significant, AIC = -1.745502  AICc = -1.74187  BIC

model7 = sarima(lj, 0,1,1, 0,1,3,4) # sma2,3 not significant, AIC = -1.716344  AICc = -1.703
m1 = arima(lj, order = c(1,1,1), seasonal = list(order = c(1,1,1), period = 4))

m2 = arima(lj, order = c(1,1,0), seasonal = list(order = c(1,1,1), period = 4))

m3 = arima(lj, order = c(0,1,1), seasonal = list(order = c(1,1,1), period = 4))

m4 = arima(lj, order = c(0,1,1), seasonal = list(order = c(1,1,0), period = 4))

m5 = arima(lj, order = c(0,1,1), seasonal = list(order = c(1,1,3), period = 4))

m6 = arima(lj, order = c(0,1,1), seasonal = list(order = c(0,1,1), period = 4))

m7 = arima(lj, order = c(0,1,1), seasonal = list(order = c(0,1,3), period = 4))
acfddlj = acf2(ddlj, 16, main = "Twice differenced at lag 1 and lag 4", cex.main = 1, font.ma
best <- sarima(lj, 0,1,1, 1,1,0,4) # all significant, AIC = -1.759615  AICc = -1.755983  BIC
a = Arima(lj, order = c(0, 1, 1), seasonal = list(order = c(1, 1, 0), period = 4))
fc = forecast(a, 12)
point_fc = exp(fc$mean)
lower_fc = exp(fc$lower)[1:12,2] # 95% confidence interval lower bound of original
lower_fc = lower_fc %>% ts(start = c(1976,1), end = c(1978, 4), 4)
upper_fc = exp(fc$upper)[1:12,2] #95% confidence interval upper bound of original
upper_fc = upper_fc %>% ts(start = c(1976,1), end = c(1978, 4), 4)
jjpred = JohnsonJohnson[65:76] %>% ts(start = c(1976,1), end = c(1978, 4), frequency = 4) # c
ljpred = log(jjpred) # log transformation on the observed values
ljfc = sarima.for(lj, 12, 0,1,1,1,1,0,4, plot = F)
plot(lj, xlim = c(1960,1980), ylim = c(-1, 3), main = "12-quarters forcast of transformed dat
lines(ljfc$pred, col = "red", lty = 2)
lines(ljfc$pred + 1.96*ljfc$se, col = "blue", lty = 2)
lines(ljfc$pred - 1.96*ljfc$se, col = "blue", lty = 2)
points(time(ljpred), ljpred, col = "grey18", pch = 20, cex = 0.5)
legend("topleft",
```

```
       legend = c("Observed Values", "Forecast", "95% CI"),
       col = c("grey18", "red", "blue"),
       pch = c(20, NA, NA),
       lty = c(NA, 2, 2),
       merge = TRUE,
       cex = 0.8)
plot.ts(jj, ylim = c(0,18), xlim = c(1960, 1980), ylab = "Dollars", main = "12-quarters forec
lines(point_fc, col = "red", lty = 2)
lines(lower_fc, col = "blue", lty = 2)
lines(upper_fc, col = "blue", lty = 2)
points(time(jjpred), jjpred, col = "grey18",pch = 20, cex = 0.5)
legend("topleft",
       legend = c("Observed Values", "Forecast", "95% CI"),
       col = c("grey18", "red", "blue"),
       pch = c(20, NA, NA),
       lty = c(NA, 2, 2),
       merge = TRUE,
       cex = 0.8)
periodogram(lj, main = "periodogram for logarithmic data", font.main = 4)
periodogram(a$residuals, main = TeX("periodogram for residuals of $SARIMA(0,1,1)(1,1,0)_4$",
                                    italic = T, bold = T))
```