STATS 320 Final Project: Model U.S. Stock Performance

Yicheng Shen, Mithcell Wang & Muyang Shi

6/1/2021

Abstract

Stock price forecasting is a popular and intriguing subject in both financial and statistical studies. Time series analysis is commonly perceived as the most applicable and fundamental methodology used to approach this task. In this project, we aim to combine the conventional time series models, ARIMA model in specific, with market and industrial movements to predict the prices of three major U.S stocks. The results of our analysis, which offer reasonable forecast of the stock behaviors, reveal significant and varying relationships between changes in daily stock price and its recent fluctuations, along with the overall market and industrial trends.

Introdction

In this final project, we are interested in using knowledge and tools learned in Time Series Analysis to explain and potentially forecast the stock performance for selected companies in the financial sector (Bank of America, denoted `BAC`), the pharmaceuticals (Moderna, denoted `MRNA`), and information technology (Zoom, denoted `ZM`). Our raw data source comes from Yahoo finance, a website that provides financial news, data and commentary including stock quotes, press releases, financial reports, and other original contents. The unit of analysis is the adjusted stock price of a single trading day and the primary variable of interests is the stock's adjusted daily closing price, defined as the closing price after appropriate adjustments for all applicable splits and dividend distributions adhering to Center for Research in Security Prices (CRSP) standards.

The analysis of stock performance has always been fascinating to researchers and investors. The current literature suggests that there are two main schools of thought in the study of the financial market, technical analysis and fundamental analysis (Xu, 2014). Technical analysis looks at the price movements of a stock and the market, and uses the existing data to predict its future price movements. Fundamental analysis, on the other hand, attempts to determine a stock's value by focusing on underlying factors that affect a company's actual business and its future prospects. We consider the technical approach to be more applicable that it could be performed on industries or the economy as a whole given our available data.

The Bank of America Corporation is an American multinational investment bank and financial services holding company headquartered in Charlotte, North Carolina. Its ticker, or stock symbol, at the New York Stock Exchange is designated as `BAC`. Since it first appeared in 1923, the company has been the representation of large market share, business activities, and economic impact on the American financial network (Qu, 2020). During the COVID-19 pandemic, the corporation has been able to withstand the economic downfall with their digital reformation. It saw a 67% boost in its digital platform after the company purchased Axia Technologies. Today, for a financial institution with nearly $3 trillion on its balance sheet, `BAC` is an important case to study in order to acquire a better understanding of the U.S. banking and financial industry.

Moderna is an American pharmaceutical and biotechnology company based in Cambridge, Massachusetts. It focuses on vaccine technologies based on messenger RNA (mRNA). Its flagship vaccine has also become the second vaccine to be approved by the FDA during the 2020 Covid-19 outbreak. Over the course of its developing phrase of the vaccine, the company has received a significant amount of attention, especially from avid investors who believe in mRNA's future prospects (Mahase, 2020). Additionally, individual investors, who are deemed to be more "technical" (focuses on day/week trends) than "fundamental" (focuses on future prospects, earnings, profitability, etc.) have rushed in to Moderna at the start of the pandemic, providing an additional reason supporting the probable temporal correlation within the stock price.

Zoom Video Communication, Inc. is an American communications technology company headquartered in San Jose, California, and it provides videotelephony and online chat services through a cloud-based software platform and is widely used for teleconferencing and telecommuting. (Lorenz, et al. 2020) The stock for Zoom is listed in NASDAQ, with ticker `ZM`. The stock price of the company has seen quite an major increase since it was listed in April 2019, especially after the onset of the pandemic. As people who have been relying heavily on using the Zoom software for daily school work, we'd love to apply knowledge of Time Series to analyzing the stock price of Zoom.

Data & Explortary Data Analysis

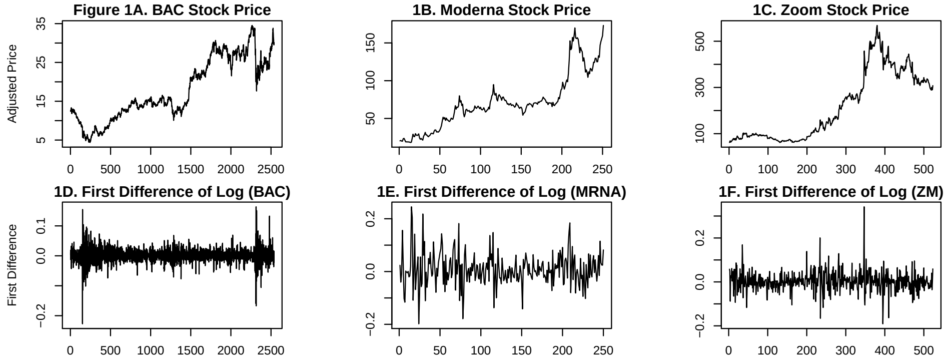
The history of the our stocks can be traced back to years or even decades ago. We are, however, mainly interested in the trend and performance in recent years. Therefore, the time span that we selected for `BAC` is after the recovery from the 2008 economic crisis, roughly incorporating the decade from 2011.01.01 to 2021.05.07. For `MRNA` is from 2020.02.01 to 2021.05.07, the period when the pandemic exerts the greatest influence, and finally for `Zoom`, we took from 2019-04-18, its earliest trading day, until 2021-05-14.

In our exploratory data analysis, we found that the three stocks we selected have different and unique trends over the time periods that we are interested in as displayed in Figure 1 and Table 1.

Table 1: Five Number Summary of Stock Price Across Time

Variable	Min	First_Qu	Median	Mean	Third_Qu	Max
Bank of America	4.38	12.13	14.96	17.73	25.19	34.54
Zoom	62.00	84.63	156.72	216.17	338.74	568.34
Moderna	18.23	48.02	66.30	66.33	74.10	169.86

The `BAC`'s prices over the decade generally exhibit an upward trend through the time series plot. The lowest point was $4.379 on 2011.12.19 and the highest price was $34.538 on 2020.01.02. The mean price ($17.728) is higher than the median ($14.961), suggesting that the distribution of the prices is right-skewed. Figure 1A also indicates nonstationarity of the adjusted prices, with changing mean and covariance as time moves on. Taking the first difference of the logged data gives us a more stationary process with stabilized variance, although there still seems to be big spikes in variances recently due to the economic havocs. This suggests that our subsequent analysis should focus on the modeling the first difference of the data, which appears to be more stationary. Similarly, in order to discern and analyze the roughly stationary processes of the data, our EDA in Figure 1 points to using first difference after a log transformation of `MRNA` and `ZM`.



In Figure 2A, the two potential explanatory variables for `BAC` are `SP500`, which is SPDR S&P 500 Trust ETF representing the market trend, and `XLF`, which is Financial Select Sector SPDR Fund representing the financial industry. We can observe strong positive correlation of the `BAC` with these two explanatory variables. Optimistically, all variables show consistently upward trend over time, but the pandemic in 2020 caused a apparent drop in these three variables.

The two potential explanatory variables for Moderna are again the Standard & Poor 500 index and the corresponding industry ETF, in this case, `XPH`, which is Pharmaceutical Select Sector SPDR Fund representing the Pharmaceutical industry. The correlation between Moderna and `SP500` or `XPH` seen in Figure 2B is quite strong, indicating that they might be potentially good explanatory variables. Additionally, the curvature in both graphs suggests potential non-linear relationships between the variables. Hence, log-transformation might be needed in the following procedures.

Figure 2A. BAC v.s Market & Industry Index

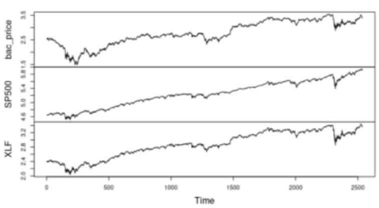


Figure 2B. Moderna v.s Market & Industry Index

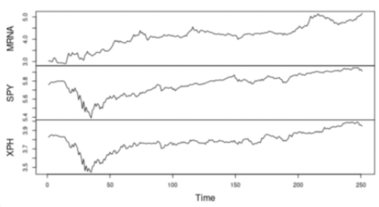
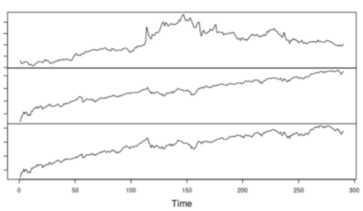


Figure 2C. Zoom v.s Market & Industry Index



The adjusted stock prices of Zoom presents an generally upward trending curve since the stock was listed, but has a downward trending curve starting around November 2020. The first difference of the raw price also exhibits non-stationarity, as there appears to be more variance since February of 2020, and become drastically more volatile since November 2020. To account for that, we performed logarithmic transformation on the stock price, and the resulting first difference of the logged stock price appears to be roughly stationary.

In our exploration of the covariates – the logged market index (`SP500`) and the technology sector index (`XLK`) – we find that both series appear to be generally positively correlated with the stock price of `ZM`, since 2020-03-23. Note that the stock price of Zoom has been rising, but the market index and technology sector index has seen quite a drop during the first few months of 2020 – we think that this is contributed by the fear due to the World Health Organization listing COVID-19 as a global pandemic. Therefore, we decided to conduct our analysis on the series only after 2020-03-23 in order to exclude the effect of the panic.

Model Fitting and Results

Bank of America

After log transformation on each variable, we first fit the `BAC` with two explanatory variables, `SP500` and `XLF`, through an ordinary least squares (OLS) model. The fitted linear model is

$$\hat{E}[\log(\text{BAC})] = -0.02108 - 0.43464 \times \log(\text{SP500}) + 1.76828 \times \log(\text{XLF})$$

The OLS model suggests that both `SP500` and `XLF` retain high statistical significance to `BAC`.

Subsequently, we examined the residuals of the OLS model and aimed to model the temporal errors using proper time series models. After accounting for these two variables, the residuals across time seem generally stationary and do not exhibit any strong linear trend. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) display that in the first 10 lags, there is no strong correlation that could be considered significant. Only lag 2 and 4 have weak correlations that might be significant.

Figure 3A. OLS Residual

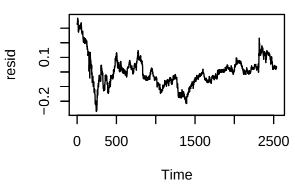


Figure 3B. 1st Difference of Residual

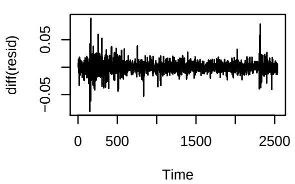


Figure 3C. ACF

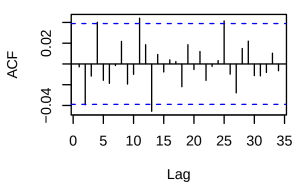
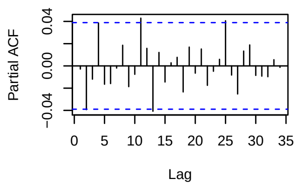


Figure 3D. Partial ACF

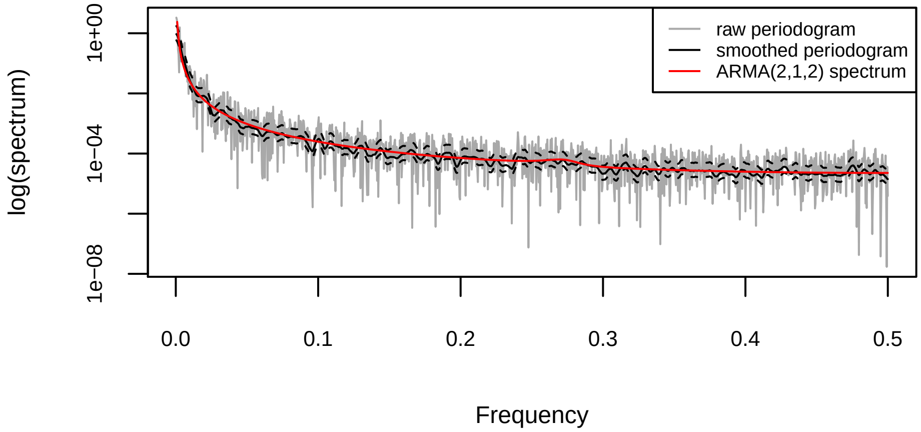


The raw periodogram of the residuals shown in Figure 4 is a decaying, low pass linear filter, with lower frequencies being retained and emphasized and higher frequencies are deemphasized and attenuated. This indicates positive correlation of the residuals across time, with more variability at low frequencies (long term) than higher frequencies (short term), which is consistent with several long meanders and bigger long term differences in the time series plot of residuals.

Since our primary modeling goal is to be able to establish a reasonable connection between `BAC` and variables of interests for prediction, we wanted our time series model of temporal residuals to be suitable for forecasting. Therefore, we prioritized the corrected Akaike Information Criterion (AICc), a likelihood-based selection method, as the important standard to consider. We found that the ARIMA(2,1,2), is the most preferable model in terms of optimizing AICc. Alternatively, we can view this model as using a ARMA(2,2) process to fit the first difference of the temporal residuals.

ARIMA(4,1,1) is another potential model speculated from the ACF plot that could be reasonable. Therefore we proceed with the diagnostics to find a better model. The ARIMA(2,1,2) model's residuals are not significantly correlated at small lags as presented in Figure A of Appendix. They pass our diagnostics, but it exhibits certain inadequacy in Figure A3 during the hypothesis testing using the LjungBox test, which deems ARIMA(4,1,1) to be more adequate. The estimated spectrum of the model (spectrum of $ARMA(2,2)\times|C(f)|^2$, which is the power transfer function) fits reasonably well with the smoothed periodogram's 95% confidence interval, further validating the choice of this particular model.

Figure 4. BAC: Raw Periodogram v.s Model Spectrum



Although the LjungBox test indicates that ARIMA model is not perfect in describing the true behaviors of the data, for simplicity and better forecast objectives, we decided to use ARIMA(2,1,2) to model the temporal residuals of the regression of `BAC`. Therefore, the ARIMA(2,1,2) model is

$$(1-\sum_{i=1}^{2}\phi_i B^i)\nabla[\log(BAC) - 0.6660 \times \log(SP500) + 1.8194 \times \log(XLF)] = (1+\sum_{i=1}^{2}\theta_i B^i)\epsilon_t$$

The results of our ARIMA model for `BAC`, along with a comparison with the OLS model, are shown below.

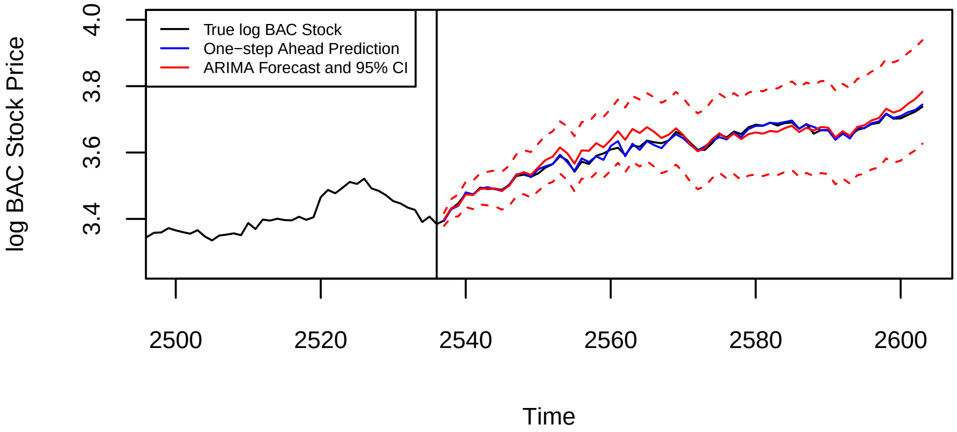
Table 2. BAC Model Parameter Estimations

	ϕ_1	ϕ_2	θ_1	θ_2	log(SP500)	log(XLF)	Intercept
ARIMA(2,1,2) Model	-0.2841	-0.8226	0.2962	0.7879	-0.6660	1.8194	
s.e.	0.1332	0.1012	0.1433	0.1121	0.0396	0.0296	
White Noise Model					-0.4346	1.7683	-0.02107582
s.e.					0.02411485	0.0245	0.06018

From the table above we can see that the standard errors for the estimated predictors in the regression with autocorrelated errors have increased compared with the white noise model, meaning that more variation in OLS should be considered due to temporary correlation.

When employed on the testing data in Figure 5, the one-step ahead prediction of the ARIMA(2,1,2) model is very coherent with true performance of the stock: The true prices in the testing data stay within our 95% confidence interval, which expands as forecast time increases.

Figure 5. Forecasting BAC Stock using Regression with ARIMA(2,1,2)



Moderna

Let's then turn to Moderna. We first fit the `MRNA` with two explanatory variables, `SP500` and `XPH`, through an ordinary least squares (OLS) model with log transformation. The fitted linear model is

$$\hat{E}[log(\text{MRNA})] = -13.9671 + 3.5349 \times \log(\text{SP500}) - 0.6089 \times \log(\text{XPH})$$

The OLS model suggests that `SP500` retain high statistical significance to `MRNA` while `XPH` does not. Hence, XPH is dropped and the model is fitted again. The fitted linear model is

$$\hat{E}[log(\text{MRNA})] = -13.3844 + 3.0346 \times \log(\text{SP500})$$

Subsequently, we examined the residuals of the OLS model and aimed to model the temporal errors using proper time series models. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) display that in the first 15 lags, there are strong correlation with lag 6, 7, and 8.

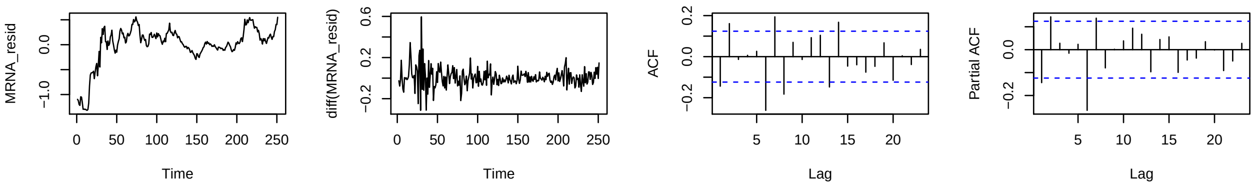


Figure 6A. OLS Residual

Figure 6B. 1st Difference of Residual

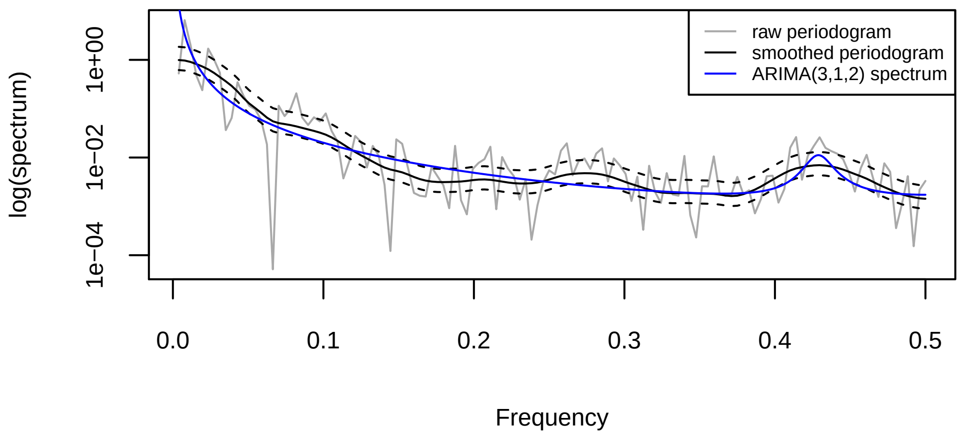
Figure 6C. ACF

Figure 6D. Partial ACF

Our main focus is forcasting, hence we prioritized the corrected Akaike Information Criterion (AICc), We found that the ARIMA(3,1,2) is the most preferable model in terms of optimizing AICc.

The raw periodogram resembles that of `BAC`, with lower frequencies being retained and emphasized and higher frequencies are deemphasized and attenuated. This indicates positive correlation of the residuals across time, with more variability at low frequencies (long term) than higher frequencies (short term). The ARMA(3,1,2) generally passes our diagnostics and hypothesis testing using the LjungBoxPlot test, with a couple some potential violations at longer lags. Looking at the spectrum, the ARIMA(3,1,2) model's spectrum fits the smoothed periodogram quite well, only outside of the confidence interval at a some interval of low frequencies.

Figure 7. Moderna: Raw Periodogram v.s Model Spectrum



We decided to use ARIMA(3,1,2) to model the temporal residuals of the regression of `MRNA`, which is written in backshift notion here

$$(1 - \sum_{i=1}^{3} \phi_3 B^i)\nabla[\log(MRNA) + 0.0514 \times \log(SP500)] = (1 + \sum_{i=1}^{2} \theta_i B^i)\epsilon_t$$

A comparison of our final time series model and the OLS model are shown below.

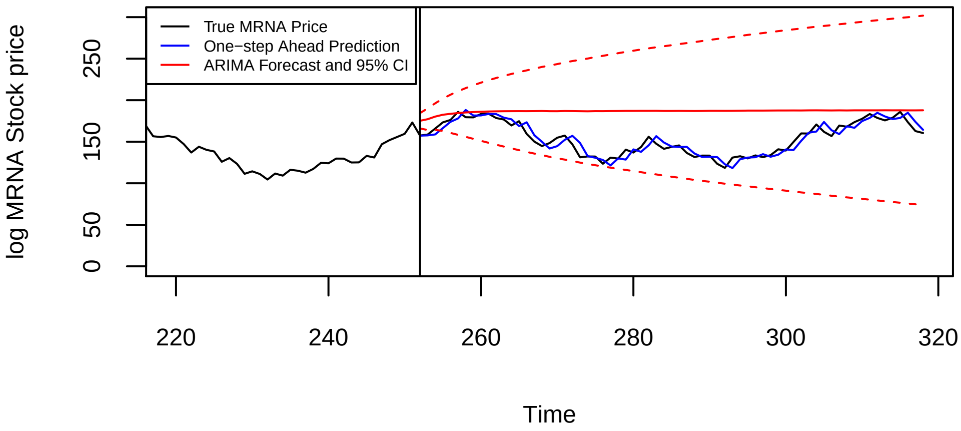
Table 3. Moderna Model Parameter Estimations

	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	log(SP500)	Intercept
ARIMA(3,1,2) Model	-0.7324	-0.3310	0.1805	-0.7473	0.3307	0.0514	
s.e.	0.3234	0.3615	0.0758	0.3255	0.3735	0.0473	
White Noise Model						3.0346	-13.3844
s.e.						0.2316	1.3325

The estimated coefficients of `SP500` , as well as the standard error/variance for explanatory variables have decreased when we considered the regression with autocorrelated errors, meaning that more variation in OLS should be considered due to temporary correlation.

When forecasting the testing data, the one-step ahead prediction of ARIMA(3,1,2) model is coherent with true performance of the stock. The actual stock price is almost always within the 95% interval bound.

Figure 8. Forecasting MRNA Stock using Regression with ARIMA(3,1,2)



Zoom

We initiated our analysis with first fitting an OLS model of logged zoom stock price over the logged market index and logged industry index (coefficients and s.e. estimates shown in table 4). Then, we look at some diagnostics on the OLS model residuals. It suggested that we should take first difference (from the long meanders shown in the residual plot in the middle time range) and AR factors (from the PACF plot) to account for the temporal correlation between the residuals.

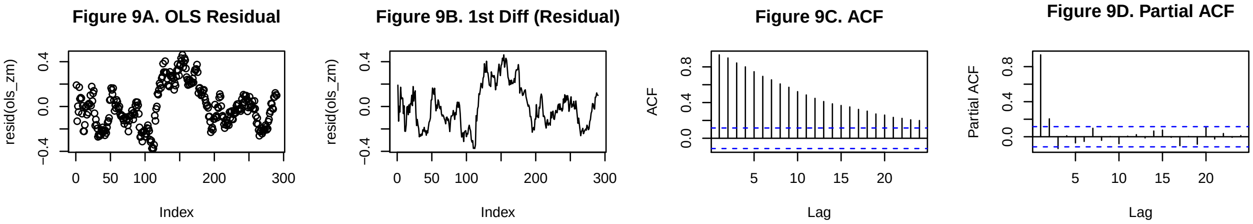
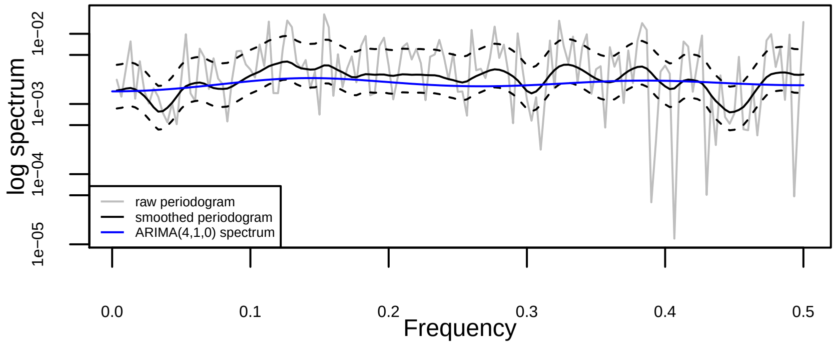


Figure 9A. OLS Residual
Figure 9B. 1st Diff (Residual)
Figure 9C. ACF
Figure 9D. Partial ACF

As previously discussed, we mainly look for a model that optimizes the AICc and we ended up with an ARIMA(4,1,0) process for the OLS residuals. We then refitted the full model with the two covariates and the ARIMA(4,1,0) process and performed residual diagnostics on the ARIMA model residuals. The residuals appeared to be homoscedastic, linear, and uncorrelated; both indicated by the visuals and the LjungBox test. Therefore, the ARIMA(4,1,0) model passes the residual diagnostics.

In addition, we also look at the log periodogram of the time series for Zoom's stock price. The smoothed periodogram (using Danielle kernel with span of 3 applied twice) is relatively flat with no particular spikes of power, or obvious suppressing power across frequencies. When looking at the raw periodogram, one could argue that there is a little suppression on the high frequencies – indicating that the times series could presents less short term variability and somewhat long meanders (long term volatility), which might correspond to weak positive correlation of the residuals across time. The spectrum of an ARIMA(4,1,0) model generally fits well and is coherent with the 95% confidence interval of the smoothed periodogram.

Figure 10. ZM: Raw Periodogram v.s Model Spectrum



Therefore, as the ARIMA(4,1,0) model satisfied the residual diagnostics,

$$(1-\sum_{i=1}^{4}\phi_i B^i)\nabla[\log(ZM) - 2.9723 \times \log(SP500) + 2.6364 \times \log(XLK)] = \epsilon_t$$

We used it for prediction of the Zoom stock price. The model coefficients estimation is shown in table 4.

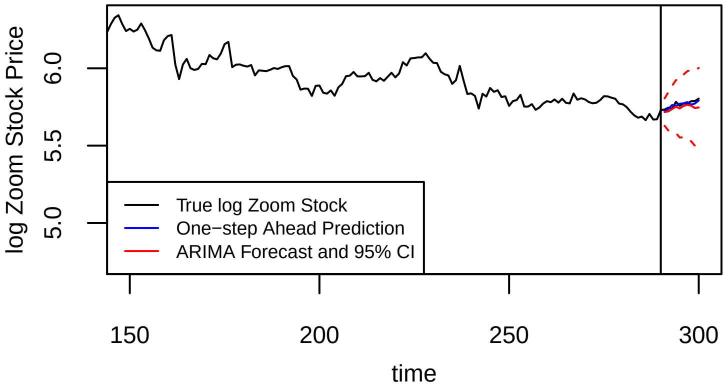
Table 4. 'ZM' Model Parameter Estimations

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	log(SP500)	log(XLF)	Intercept
ARIMA(4,1,0) Model	-0.0156	-0.0211	-0.0388	-0.0618	-2.9723	2.6364	
s.e.	0.0622	0.0624	0.0607	0.0614	0.4499	0.3557	

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	log(SP500)	log(XLF)	Intercept
White Noise Model					-6.8908	7.7292	9.2266
s.e.					0.4931	0.4213	0.9616

We used this ARIMA model to predict the log of Zoom's adjusted stock price – and we did a one-step-ahead prediction as well as a direct forecast on 10 trading days over the period from 2020-05-17 to 2021-05-28, and compared it against the true log adjusted stock price of Zoom over the 10 trading days (using the true values for the covariates over the 10 trading days). The forecast did relatively well as the 95% confidence interval captures the true log adjusted stock price of Zoom.

Figure 11. Forecasting ZM Stock using Regression with ARIMA(4,1,0)



Discussion

In this study, we tried to use market index, industrial indices (as covariates) and time series process (for the temporal correlated errors) to predict the behaviors of some target stocks – Bank of America (`BAC`), Moderna (`MRNA`), and Zoom (`ZM`). We want to specifically discuss three aspects of this study: covariates, model prediction, and its implications and practicality in the field of finance.

The two covariates that we have considered for our models are the market index and the corresponding industry index (XLF, XLK or XLH). There are three caveats that we want to address and elaborate here: the strong correlation between the two covariate and how each covariate is constructed. Note that despite a relatively high positive correlation between each industry specific index and the market index, we chose them to include in our model if the covariate is statistically significant. The market index is constructed through taking the 500 biggest companies listed in the US stock market and weighting them based on their market capitalization. The industry indexes are constructed the same way but instead take the biggest companies listed under the specific industry. Hence, the market index and industry index are intrinsically reliant on individual stock's performance. However, we deem that the reliance is insignificant due to the large number of individual stocks incorporated in each index. This is a compromise that we made in order to achieve valuable covariates for predicting the target stock price.

Secondly, we also need to note the difference in coefficients estimated for the covariates between the OLS and ARIMA models. Note that the estimates are very close for the stock price of Bank of America, but are quite different for Moderna and Zoom. The OLS estimates are still consistent, but not very efficient. Therefore, for `BAC`, which we have decades of data, OLS model can make relative effective estimations of the coefficients and are close to the coefficients given by its ARIMA model; for Moderna and Zoom, for which only limiting data were available, the estimations given by OLS could be more accurate if more data were given. Albeit the difference in the coefficients, we still see a widening of the confidence intervals (bigger standard errors relative to the coefficent estimates), as expected to be the impact of using ARIMA process to account for the temporal correlated errors.

With the understanding of the covariates, we can analyze the prediction of target stocks' future performance. In our study, we were able to partition our dataset to generate a testing portion – in which we know the true values of the covariates for the testing dates. Testing of the models are performed accordingly. Generally speaking, all three models performed quite well. According to the results from the testing phase, the `BAC` model predicted the rise or drop of the price movement (not amplitude) 88.06% of the time, compared to that of the `MRNA` model 54.4% of and `ZM` model of 50%. The result suggests that in terms of direction predicting, all three models have a promising forecast performance. The reason why `BAC`'s model performed the best might be attributed to both a significantly longer period of training data than the two and also a relatively more stable/stationary price movement.

One limitation of this prediction model is that we could not access the two values of the two covariates ahead of the prediction, e.g. when predicting stock price for Jan 2nd, we do not have the true value of the covariates at Jan 2nd. This could be an area of future research where we develop models using "lagged covariates", regressing the target stock prices on the n-day-before covariate values, so that we can use the current covariate values to make predictions for the future n-day.

Finally, the models do provide significant insights for finance analysis. As shown, the time series models do provide a stronger estimate than the normal OLS models, indicating that there is a "time" or so called "momentum" factor in predicting stock prices. Hence, this analysis supports that there is some room for pure technical analysis (entirely based on trend or momentum instead of company fundamentals). Additionally, through the use of hedging (e.g. buy individual stocks and short(sell) market/industry indexes), individuals could construct a strategy based on our model without knowing the true market covariates and still capitalize on the strategy.

Reference

Lorenz, Taylor; Griffith, Erin; Isaac, Mike (March 17, 2020). "We Live in Zoom Now". The New York Times. ISSN 0362-4331. Archived from the original on March 23, 2020. Retrieved March 23, 2020.

Mahase, E. (2020). Covid-19: Moderna vaccine is nearly 95% effective, trial involving high risk and elderly people shows. *BMJ: British Medical Journal* (Online), 371.

Qu, P. (2020). Bank of America Stock Price Research. *Journal of Financial Risk Management, 9*(2), 126-140.

Xu, S. Y. (2014). Stock price forecasting using information from Yahoo finance and Google trend. *UC Brekley.*

Yahoo Finance (2021). Yahoo Finance - Stock Market Live, Quotes, Business & Finance News. *Yahoo! Finance.* https://finance.yahoo.com/.

Supplemental Appendix

Residual Diagnostics of ARIMA(2,1,2) Model for `BAC`

Figure A1. Model Residual

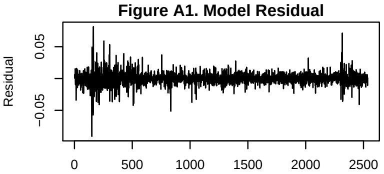


Figure A2. Lag 1 Residuals

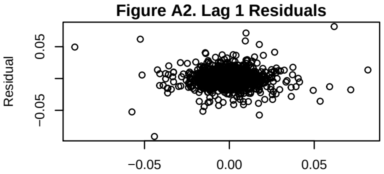


Figure A3. LjungBoxPlot

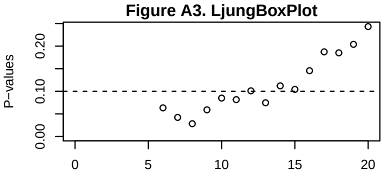
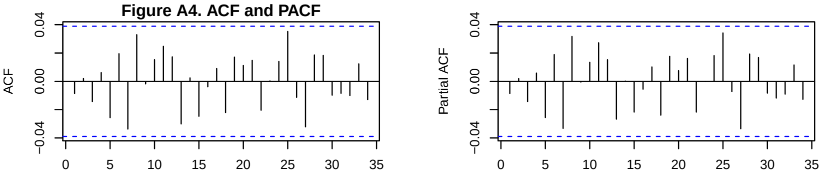
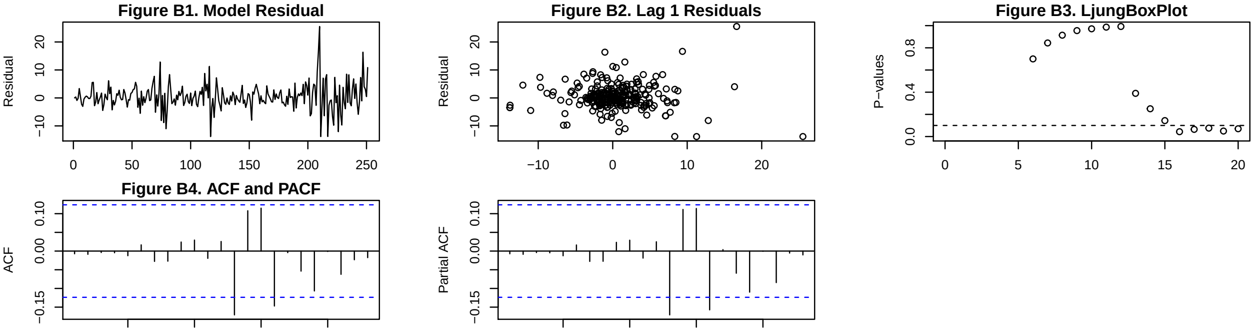


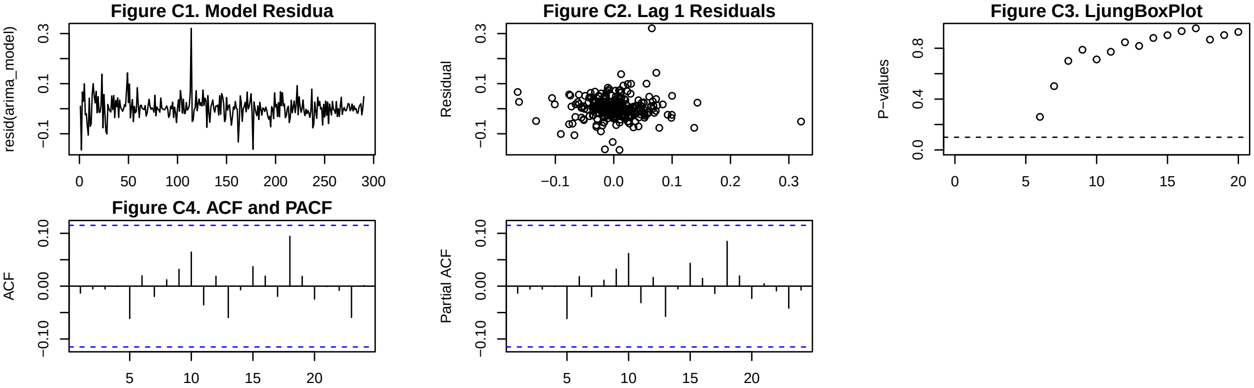
Figure A4. ACF and PACF



Residual Diagnostics of ARIMA(3,1,2) Model for `MRNA`



Residual Diagnostics of ARIMA(4,1,0) Model for `ZM`



~~~
---
title: "Project Code Appendix"
author: "Yicheng Shen, Mithcell Wang & Muyang Shi"
output: pdf_document
---

```{r setup, include=FALSE, message = F,warning=F}
knitr::opts_chunk$set(echo = TRUE, cache = TRUE,warning=F, out.width='50%')
library(generics)
library(timetk)
library(tidyquant)
library(ggplot2)
library(stats)
library(tidyverse)
library(TSA)
library(forecast)
library(kableExtra)
library(knitr)

LjungBoxPlot <- function(model, max.lag = 20, plot = TRUE, tol = 0.1,...){
  N <- nobs(model)
  numParams <- length(coef(model)) - 1 # not counting the mean
  if(max.lag < numParams){
    error("max.lag is smaller than the number of parameters")
  }

  lags <- (numParams+1):max.lag
  pvals <- numeric(length(lags))
  for(k in 1:length(lags)){
    pvals[k] <- Box.test(resid(model), lag = lags[k],
                         fitdf = numParams)$p.value
  }
  if(plot == TRUE){
    plot(lags, pvals, xlim = c(0, max.lag), ylim = c(0, max(pvals)),
         xlab = "Lag", ylab = "P-values",...)
    abline(h = tol, lty = 'dashed')
  }
}

AICc <- function(model){
  k <- length(coef(model))
  n <- nobs(model)
  return(AIC(model) + 2 * (k + 1) * (k + 2) / (n - k - 2))
}
```


## R Codes for Bank of America Analysis

```{r read in data and EDA, fig.height=4, fig.width=8.5}
bankofamerica <- tq_get("BAC",from = '2011-01-01',to = '2021-01-31',get = 'stock.prices')
# 10 years; 2536 trading days
bac_price<-ts(bankofamerica%>%select(adjusted))
summary(bac_price)
par(mfrow=c(2,2))
plot(bac_price, main = "BAC Stock Price (2011-2021)")
plot(diff(bac_price), main = "First Difference of BAC Stock Price")
plot(log(bac_price), main = "Logged BAC Stock Price (2011-2021)")
plot(diff(log(bac_price)), main = "First Difference of Logged BAC Stock Price")
```

```{r read in explantatory variables, fig.height=5, fig.width=8.5}
SP500 <- ts(tq_get("SPY",from = '2011-01-01',to = '2021-01-31',get = 'stock.prices')
        %>%select(adjusted)) # # Market Trend: SPDR S&P 500 Trust ETF
~~~

~~~
XLF <- ts(tq_get("XLF",from = '2011-01-01',to = '2021-01-31',get = 'stock.prices')
       %>%select(adjusted)) #Financial Sector Trend: Financial Select Sector SPDR Fund
par(mfrow=c(2,2))
plot(SP500)
plot(diff(log(SP500)))
plot(XLF)
plot(diff(log(XLF)))
```

```{r log tranformation, out.width='30%'}
bac_sp500 <- data.frame(bac_price, SP500,XLF)
bac_sp500<-rename(bac_sp500, bac.price = adjusted,
                            SP500 = adjusted.1,
                            XLF = adjusted.2)
ggplot(bac_sp500, aes(log(bac.price),log(SP500))) +
  geom_point() + theme_classic()# positive correlation between BAC stock price and SP500
ggplot(bac_sp500, aes(log(bac.price),log(XLF)))  +
  geom_point() + theme_classic() # positive correlation between BAC stock price and XLF
```

```{r EDA plot}
log_bac_sp500 <- data.frame("bac_price" = as.vector(log(bac_price)),
          "SP500" = as.vector(log(SP500)), "XLF" = as.vector(log(XLF)))
log_bac_sp500 <- ts(log_bac_sp500)
plot(log_bac_sp500,yax.flip=T,main='Logged BAC Price v.s Market & Industry Index')
```

```{r OLS}
log_bac_sp500 <- data.frame("bac_price" = as.vector(log(bac_price)),
       "SP500" = as.vector(log(SP500)),"XLF" = as.vector(log(XLF)))
ols_bac_sp500 <- lm(bac_price~SP500+XLF,data=log_bac_sp500)
summary(ols_bac_sp500) # Fit OLS; Significant Variables
```

```{r OLS residuals, fig.height=4, fig.width=10, fig.height=3}
resid <- ts(resid(ols_bac_sp500))
par(mfrow=c(1,4))
plot(resid) # Plot residual
plot(diff(resid))
acf(diff(resid))
pacf(diff(resid))
eacf(diff(resid))
```

```{r Raw Perioddogram}
I <- spec.pgram((resid), demean = TRUE, detrend = FALSE, plot = T, col= "grey") # Higher
power in low frequencies
```

```{r Find Best AICc model}
auto.arima(resid, stepwise = FALSE, ic = "aicc", trace = FALSE) # ARIMA(2,1,2) seems to be
the winner
```

```{r Spectral Analysis}
resid.arima_1 <- arima(diff(resid), c(2,0,2))
resid.arima_2 <- arima(diff(resid), c(4,0,1))
I <- spec.pgram(diff(resid), demean = TRUE, detrend = FALSE, plot = FALSE)
Sbar <- spec.pgram(diff(resid), demean = TRUE, detrend = FALSE, plot = FALSE,
                   kernel("modified.daniell", c(4,4)))
IMASpec <- ARMAspec(model = list(ar = resid.arima_1$model$phi, ma =
resid.arima_1$model$theta,
                                 sigma2 = resid.arima_1$sigma2), plot = FALSE)
ARIMASpec <- ARMAspec(model = list(ar = resid.arima_2$model$phi, ma =
resid.arima_2$model$theta,
~~~

~~~
                                  sigma2 = resid.arima_2$sigma2), plot = FALSE)
plot(I$spec*exp(-digamma(1)) ~ I$freq, type='l', col = 'darkgray',
     xlab = "frequency", ylab = "spectrum", log  = "y",
     main ="First Difference Periodogram v.s Model Spectrum")
lines(Sbar$spec ~ Sbar$freq, col = 'black')
lines(Sbar$spec*Sbar$df / qchisq(0.975, df = Sbar$df) ~ Sbar$freq, col = 'black', lty =
'dashed')
lines(Sbar$spec*Sbar$df / qchisq(0.025, df = Sbar$df) ~ Sbar$freq, col = 'black', lty =
'dashed')
lines(IMASpec$spec ~ IMASpec$freq, col = 'red')
lines(ARIMASpec$spec ~ ARIMASpec$freq, col = 'blue')
legend("bottomright", legend = c("raw periodogram", "smoothed periodogram",
    "ARMA(2,2) spectrum", "ARMA(4,1) spectrum"), col = c("darkgray", "black", "red",
"blue"), lty = 1)

power_transfer <- function(spec,freq){spec/(2 - 2*cos(2*pi*freq))}
I <- spec.pgram((resid), demean = TRUE, detrend = FALSE, plot = FALSE)
Sbar <- spec.pgram((resid), demean = TRUE, detrend = FALSE, plot =
FALSE,kernel("modified.daniell", c(5,5)))
plot(I$spec*exp(-digamma(1)) ~ I$freq, type='l', col = 'darkgray',
     xlab = "frequency", ylab = "spectrum", log  = "y", main ="Raw Periodogram v.s Model
Spectrum")
lines(Sbar$spec ~ Sbar$freq, col = 'black')
lines(Sbar$spec*Sbar$df / qchisq(0.975, df = Sbar$df) ~ Sbar$freq, col = 'black', lty =
'dashed')
lines(Sbar$spec*Sbar$df / qchisq(0.025, df = Sbar$df) ~ Sbar$freq, col = 'black', lty =
'dashed')
lines(power_transfer(ARIMASpec$spec,ARIMASpec$freq) ~ ARIMASpec$freq, col = 'blue')
lines(power_transfer(IMASpec$spec,IMASpec$freq) ~ IMASpec$freq, col = 'red')
legend("topright", legend = c("raw periodogram", "smoothed periodogram",
  "ARMA(2,2) spectrum", "ARMA(4,1) spectrum"),col = c("darkgray", "black", "red", "blue"),
lty = 1)
# ARIMA (2,1,2) is reasonable for the temporal residuals
```

```{r Fit ARIMA model}
log_bac_sp500 <- data.frame("bac_price" = as.vector(log(bac_price)),
          "SP500" = as.vector(log(SP500)),"XLF" = as.vector(log(XLF)))
arima_model <- arima(x = log_bac_sp500$bac_price, order = c(2, 1, 2),
                     xreg = cbind(log_bac_sp500$SP500,log_bac_sp500$XLF))
par(mfrow = c(2,2), mar = c(4,4,2,2))
plot(resid(arima_model))
plot(zlag(resid(arima_model)), resid(arima_model), xlab = "Lag -1", ylab = "Residual")
acf(resid(arima_model))
pacf(resid(arima_model))
arima_model_2 <- arima(x = log_bac_sp500$bac_price, order = c(4, 1, 1),
                     xreg = cbind(log_bac_sp500$SP500,log_bac_sp500$XLF))
par(mfrow = c(2,2), mar = c(4,4,2,2))
plot(resid(arima_model_2))
plot(zlag(resid(arima_model_2)), resid(arima_model_2), xlab = "Lag -1", ylab = "Residual")
acf(resid(arima_model_2))
pacf(resid(arima_model_2))
```

```{r More diagnostics}
LjungBoxPlot(arima_model) # LjungBoxPlot is the only problematic one
AICc(arima_model) # Better AICc and BIC of (2,1,2)
BIC(arima_model)
LjungBoxPlot(arima_model_2)
AICc(arima_model_2)
BIC(arima_model_2)
```

```{r Compare ARIMA and OLS}
summary(ols_bac_sp500)[4]
~~~

~~~
arima_model # Compare ols and arima; similar coefficients but larger Se for Arima
arima_model_2
```

```{r Model Forecast}
# Forecast three months of real data (2603 trading days: 67 test; 2536 training)
bankofamerica <- ts(log(tq_get("BAC",from = '2011-01-01',to = '2021-05-07',get =
'stock.prices')
                    %>%select(adjusted)))
SP500_test<- log(tq_get("SPY",from = '2021-02-01',to = '2021-05-07',get =
'stock.prices')%>%
  select(adjusted))
XLF_test <- log(tq_get("XLF",from = '2021-02-01',to = '2021-05-07',get =
'stock.prices')%>%
  select(adjusted))
prediction <- predict(arima_model, n.ahead = 67,newxreg=cbind(SP500_test,XLF_test))
prediction_2 <- predict(arima_model_2, n.ahead = 67,newxreg=cbind(SP500_test,XLF_test))
BAC_test <- bankofamerica[2537:2603,]
fit1 <-
Arima(BAC_test,xreg=data.matrix(data.frame(SP500_test,XLF_test)),model=arima_model)
onestep1 <- as.vector(fitted(fit1))
onestep1<- ts(onestep1,start=c(2537), end=c(2603))
plot(bankofamerica,xlim = c(2500,2603),ylim = c(3.3,4), col="black")
lines(onestep1,col = "blue")
lines(prediction$pred,col="red")
lines(prediction$pred + 2*prediction$se, lty = "dashed",col="red")
lines(prediction$pred - 2*prediction$se, lty = "dashed",col="red")
abline(v=2536)
legend("bottomright",col=c("black","blue","red"),lty=1,
       legend=c("True BAC Price","One-step Ahead Prediction","ARIMA Forecast and 95% CI"),
cex=0.8)
# Reasonably good forecast across testing data
```

```{r CI}
upper_bond <- prediction$pred + 2*prediction$se
lower_bond <- prediction$pred - 2*prediction$se
compare_95 <- data.frame(BAC_test,upper_bond,lower_bond)
compare_95 %>% mutate(correct  = ifelse(BAC_test > lower_bond , 1, 0)) %>%
summarize(mean(correct))
compare_95 %>% mutate(correct  = ifelse(upper_bond > BAC_test , 1, 0)) %>%
summarize(mean(correct))
# Stay within 95% CI
```

```{r Predict up and down}
BAC_return <-as.vector(diff(BAC_test))/BAC_test
onestep_return1 <- as.vector(diff(onestep1))/onestep1
predication <- data.frame(BAC_return,onestep_return1)
predication %>% mutate(correct = onestep_return1*BAC_return > 0) %>%
summarize(mean(correct))
# 88.06% Correctness
```


## R Codes for Moderna Analysis

```{r Data Retrieving}
## Getting training data from 2020-02-01 to 2021-01-01; After Covid happened.
MRNA<- tq_get('MRNA',
               from = "2020-02-01",
               to = "2021-01-01",
               get = "stock.prices")$adjusted
SPY<- tq_get('SPY',
               from = "2020-02-01",
~~~

~~~
               to = "2021-01-01",
               get = "stock.prices")$adjusted
XPH<- tq_get('XPH',
               from = "2020-02-01",
               to = "2021-01-01",
               get = "stock.prices")$adjusted

data <-data.frame(MRNA,SPY,XPH)
```

### EDA

```{r}
# Time Series, First Difference
par(mfrow=c(1,3))
plot(data$MRNA, type = 'l',main = "Raw MRNA Stock Price")
plot(diff(data$MRNA),type = 'l', main = "First Difference of MRNA Stock Price")
plot(diff(log(data$MRNA)),type = 'l', main = "1E. First Difference of Log (MRNA)", ylab =
" ", xlab = "Time")
```

### OLS Model Fitting and Residual Analysis

```{r fig.align="center", echo = F,fig.width=10, fig.height=2}
MRNA_SPY_XPH_model<- lm(log(MRNA) ~ log(SPY)+log(XPH),data=data)
MRNA_SPY_model <- lm(log(MRNA) ~ log(SPY),data=data)
##summary(MRNA_SPY_model)
MRNA_resid <- ts(resid(MRNA_SPY_model))
par(mfrow=c(1,4))
plot(MRNA_resid, main = "OLS Residual v.s. Time")
plot(diff(MRNA_resid),main = "First Difference of OLS Residual")
acf(diff(MRNA_resid))
pacf(diff(MRNA_resid))
```

### TS Model Fitting and Residual Analysis

```{r fig.align="center", echo = F,fig.width=10, fig.height=2}
##auto.arima(MRNA_resid,stepwise=FALSE,ic='aic',trace=FALSE) ## Using Auto_Arima
arima_model_moderna <- arima(x = data$MRNA, order = c(3, 1, 2),
                     xreg = data$SPY)
par(mfrow = c(1,4))
plot(resid(arima_model_moderna), main="Model Residual v.s. Time")
plot(zlag(resid(arima_model_moderna)), resid(arima_model_moderna), xlab = "Lag -1", ylab =
"Residual", main = "Lag 1 Residuals")
acf(resid(arima_model_moderna))
LjungBoxPlot(arima_model_moderna)
```
### TS Model Spectrum Analysis

```{r fig.align="center", echo = F,fig.width=6, fig.height=3}
resid <- MRNA_resid
resid.arima <- arima(diff(resid), c(3,0,2))
power_transfer <- function(spec,freq){spec/(2 - 2*cos(2*pi*freq))}
I <- spec.pgram((resid), demean = TRUE, detrend = FALSE, plot = FALSE)
Sbar <- spec.pgram((resid), demean = TRUE, detrend = FALSE, plot = FALSE,
                   kernel("modified.daniell", c(5,5)))
ARIMASpec <- ARMAspec(model = list(ar = resid.arima$model$phi, ma =
resid.arima$model$theta,
                                 sigma2 = resid.arima$sigma2), plot = FALSE)
plot(I$spec*exp(-digamma(1)) ~ I$freq, type='l', col = 'darkgray',
     xlab = "frequency", ylab = "spectrum", log  = "y", main ="Figure.6 Moderna: Raw
Periodogram v.s Model Spectrum", cex.main =0.8)
lines(Sbar$spec ~ Sbar$freq, col = 'black')
lines(Sbar$spec*Sbar$df / qchisq(0.975, df = Sbar$df) ~ Sbar$freq, col = 'black', lty =
~~~

~~~
'dashed')
lines(Sbar$spec*Sbar$df / qchisq(0.025, df = Sbar$df) ~ Sbar$freq, col = 'black', lty =
'dashed')
lines(power_transfer(ARIMASpec$spec,ARIMASpec$freq) ~ ARIMASpec$freq, col = 'blue')
legend("topright", legend = c("raw periodogram", "smoothed periodogram", "ARIMA(3,1,2)
spectrum"),
       col = c("darkgray", "black", "blue"), lty = 1,, cex=0.55)
```

### TS Model Prediction

```{r fig.align="center", echo = F}

MRNA <- ts(tq_get("MRNA",from = '2020-02-01',to = '2021-05-07',get = 'stock.prices')
                    %>%select(adjusted))

SP500_test<- tq_get("SPY",from = '2021-02-01',to = '2021-05-07',get =
'stock.prices')$adjusted
prediction <- predict(arima_model_moderna, n.ahead = 67,newxreg=SP500_test)
MRNA_test <- MRNA[252:318]
fit2 <- Arima(MRNA_test,xreg=SP500_test,model=arima_model_moderna)
onestep <- as.vector(fitted(fit2))
onestep<- ts(onestep,start=c(252), end=c(318))

plot(MRNA,xlim = c(220,318), ylim = c(0,300),col="black",
     main = "Forecasting MRNA Stock Price using Regression with ARIMA(3,1,2)",
     ylab = "adjusted price")
lines(onestep,col = "blue")
lines(prediction$pred,col="red")
lines(prediction$pred + 2*prediction$se, lty = "dashed",col="red")
lines(prediction$pred - 2*prediction$se, lty = "dashed",col="red")
abline(v=252)
legend("topleft",col=c("black","blue","red"),lty=1,
       legend=c("True MRNA Price","One-step Ahead Prediction","ARIMA Forecast and 95%
CI"), cex=0.8)
```

### Predication Correction Rate
```{r}
SPY_Test <- ts(tq_get("SPY",from = '2021-02-01',to = '2021-05-01',get = 'stock.prices')
                    %>%select(adjusted))
fit2 <- Arima(SPY_Test,model=arima_model)
onestep <- as.vector(fitted(fit2))
SPY_return <-as.vector(diff(SPY_Test))/SPY_Test
onestep_return <- as.vector(diff(onestep))/onestep
pred_data %>% mutate(Correct = (onestep_return*SPY_return > 0)) %>%
summarize(mean(Correct))
```


## R Codes for Zoom Analysis

```{r load-data}
zoom <- tq_get("ZM",from = '2019-04-18',to = '2021-05-16',get =
'stock.prices') # 450 trading days
```
```{r EDA, fig.align='center',fig.width=6,fig.height=4}
zm_price <- ts(zoom%>%select(adjusted))
summary(zm_price)
par(mfrow=c(2,2))
plot(zm_price,main='raw price')
plot(diff(zm_price),main='Zoom 1st diff')
plot(log(zm_price),main='log price')
plot(diff(log(zm_price)),main='log price 1st diff')
```
~~~

~~~
### SP500
```{r,echo = TRUE, results = 'hide',fig.show = 'hide'}
SP500 <- tq_get("SPY",from = '2019-04-18',to = '2021-05-16',get =
'stock.prices')
SP500_price <- ts(SP500 %>% select(adjusted))
par(mfrow=c(1,2))
plot(SP500_price)
plot(diff(SP500_price))
```
### XLK
```{r,echo = TRUE, results = 'hide',fig.show = 'hide'}
XLK <- tq_get("XLK",from = '2019-04-18',to = '2021-05-16',get =
'stock.prices')
XLK_price <- ts(XLK %>% select(adjusted))
par(mfrow=c(1,2))
plot(XLK_price)
plot(diff(XLK_price))
```
#### Raw Scale
```{r, echo = TRUE, results = 'hide',fig.show = 'hide'}
stock <- data.frame("zm" = as.vector(zm_price),
                    "sp500" = as.vector(SP500_price),
                    "xlk" = as.vector(XLK_price))
stock_ts <- ts(stock)
plot(stock_ts,yax.flip=T)
```
```{r, echo = TRUE, results = 'hide'}
ols_zm <- lm(zm~sp500+xlk,data=stock)
summary(ols_zm)
```
#### Log scale
```{r,fig.align='center',fig.height=3,fig.width=5}
par(mfrow = c(1,1))
stock <- data.frame("zm" = as.vector(log(zm_price)),
                    "sp500" = as.vector(log(SP500_price)),
                    "xlk" = as.vector(log(XLK_price)))
stock_ts <- ts(stock)
plot(stock_ts,yax.flip=T,main='logged stock price')
```
```{r,echo = TRUE, results = 'hide'}
ols_zm <- lm(zm~sp500+xlk,data=stock)
summary(ols_zm)
```
#### Truncate Time
```{r,fig.align='center',fig.height=3,fig.width=5}
stock2 <- stock[c(234:523),]
stock2_ts <- ts(stock2)
plot(stock2_ts,yax.flip=T)
```
### Modeling
#### OLS
```{r}
ols_zm <- lm(zm~sp500 + xlk, data = stock2) # Logged and Truncated
summary(ols_zm)
```
#### OLS Residuals Diagnostics
```{r}
par(mfrow=c(2,2))
plot(resid(ols_zm))
plot(resid(ols_zm),type = 'l')
acf(resid(ols_zm)) # Suggest first difference
pacf(resid(ols_zm)) # AR(2)?
```
#### ARIMA
```{r, echo = TRUE, results = 'hide'}
~~~

~~~
resid <- ts(resid(ols_zm))
auto.arima(resid, stepwise = FALSE, ic = "aicc", trace = FALSE)
#ARIMA(4,1,0)
auto.arima(resid, stepwise = FALSE, ic = "bic", trace = FALSE)
#ARIMA(1,1,1)
auto.arima(resid, stepwise = FALSE, ic = 'aic', trace = FALSE)
#ARIMA(4,1,0)
```
#### ARIMA Residual Diagnostics
```{r,fig.align='center',fig.width=6,fig.height=6}
arima_model <- arima(x = stock2$zm, order = c(4, 1, 0),
                     xreg = cbind(stock2$sp500,stock2$xlk))
par(mfrow = c(3,2))
plot(resid(arima_model), main = 'Residuals')
plot(zlag(resid(arima_model)), resid(arima_model), xlab = "Lag  1", ylab
= "Residual", main='lagged Residuals')
acf(resid(arima_model))
pacf(resid(arima_model))
#par(mfrow = c(1,1))
LjungBoxPlot(arima_model,main='LjungBoxPlot')
```


```{r}
summary(ols_zm)
arima_model
```
#### Periodogram/Spectrum Diagnostics
```{r,fig.align='center',fig.width=6,fig.height=4}
par(mfrow = c(1,1))
log_zm <- stock2$zm
arima_spec <- ARMAspec(model = list(ar = arima_model$model$phi,
                                    ma = arima_model$model$theta,
                                    sigma2 = arima_model$sigma2),
plot = FALSE)
I <- spec.pgram(diff(log_zm),plot = FALSE) # Raw Periodogram (for 1st diff log_zm)
Sbar <- spec.pgram(diff(log_zm),plot=FALSE,
                   kernel('modified.daniell',c(3,3))) # Smoothed
plot(I$spec*exp(-digamma(1)) ~ I$freq, type = 'l', col = 'gray',
     xlab = 'frequency', ylab = 'log spectrum', log = "y") # Logged Raw
Periodogram
lines(Sbar$spec ~ Sbar$freq, col = 'black') # Smoothed Periodogram
lines(Sbar$spec*Sbar$df / qchisq(0.975, df = Sbar$df) ~ Sbar$freq,
      col = 'black', lty = 'dashed') # CI
lines(Sbar$spec*Sbar$df / qchisq(0.025, df = Sbar$df) ~ Sbar$freq,
      col = 'black', lty = 'dashed') # CI
lines(arima_spec$spec ~ arima_spec$freq, col = 'blue')
legend("bottomleft",
       legend = c("raw periodogram",
                  "smoothed periodogram",
                  "ARIMA(4,1,0) spectrum"),
       col = c("gray",
"black",
               "blue"),
       lty = 1)
```
### Prediction
```{r fig.algin = "center"}
arima_model <- arima(x = stock2$zm, order = c(4, 1, 0),
                     xreg = cbind(stock2$sp500,stock2$xlk))
zoom <- log(tq_get("ZM",from = '2019-04-18', to = '2021-05-29', get =
'stock.prices')
               %>% select(adjusted))
zoom <- ts(zoom %>% slice(234:n()))
SP500_test <- log(tq_get("SPY",from =
'2021-05-17',to='2021-05-29',get='stock.prices')
~~~

```
                  %>% select(adjusted))
XLK_test <- log(tq_get("XLK",from =
'2021-05-17',to='2021-05-29',get='stock.prices')
                  %>% select(adjusted))
zoom_test <- zoom[291:300,]
# Direct Prediction
prediction <- predict(arima_model, n.ahead = 10,

newxreg = cbind(SP500_test,XLK_test))
# Onestep ahead prediction
fit1 <- Arima(zoom_test,
              xreg = data.matrix(cbind(SP500_test,XLK_test)),
              model = arima_model)
onestep1 <- as.vector(fitted(fit1))
onestep1 <- ts(onestep1, start = 291, end = 300)
plot(zoom,ylab="log Zoom Stock Price",
     xlim = c(150,300))
abline(v=290)
lines(onestep1, col = 'blue')
lines(prediction$pred,col='red')
lines(prediction$pred + 2*prediction$se, lty = "dashed",col="red")
lines(prediction$pred - 2*prediction$se, lty = "dashed",col="red")
legend("bottomleft",col=c("black","blue","red"),lty=1,
       legend=c("True log Zoom Stock","One-step Ahead Prediction",
"Direct Prediction and 95% CI"),
cex=0.8)
```
#### Accuracy
```{r}
zoom_return <-as.vector(diff(zoom_test))/zoom_test
onestep_return <- as.vector(diff(onestep1))/onestep1
accuracy <- data.frame(zoom_return,onestep_return)
accuracy %>% mutate(correct = onestep_return*zoom_return > 0) %>%
summarize(mean(correct))
```
```
```