Forecasting Dow Jones Industrial Average Monthly Excess Return

Zheyuan Fan

07/04/2020

Introduction

In this report, I am interested in forecasting the monthly excess return of March and April 2020 as accurately as possible using different time series models. After identifying the data as smooth time series, I used two time series models, ARIMA (parameter determination by pacf and acf) and holt-winter model. Then I computed the AIC values, and showed Q-Q plots to check whether the residuals are normally distributed. I also used box.test to check whether the residuals were white noise. Finally, based on the above, the best model was confirmed as the ARIMA(2, 0, 2) model, thus I forecasted the monthly excess return of March and April based on this model. The exact procedure is described below.

Procedures

First, load the data into R and generate dependent variable, the monthly excess return.

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

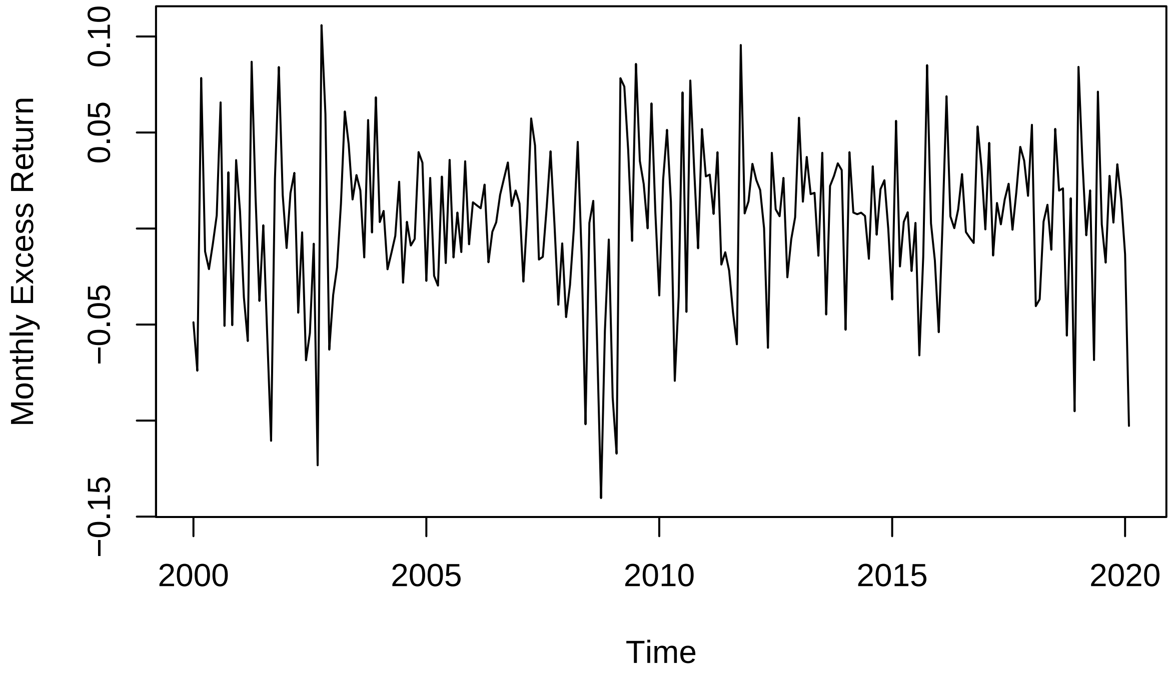
```
data <- read.csv("/Users/6ixlegend/Desktop/STA457/^DJI.csv")
data$MER <-(data$Close-data$Open)/data$Open
data <- data[,c(8)]
```

Define the time series data using the ts() function and generate the ACF and PACF plots.

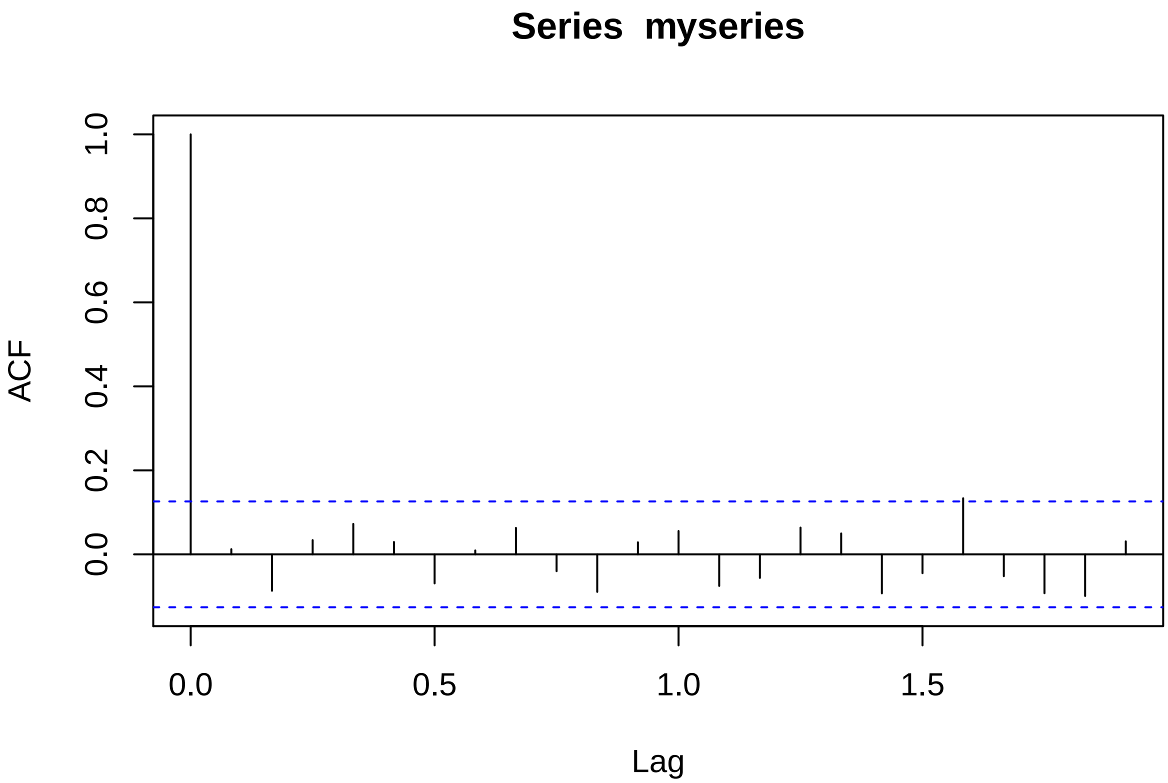
```
myseries <- ts(data,frequency =12, start = c(2000,1),end = c(2020,2))
str(myseries)
```

```
##  Time-Series [1:242] from 2000 to 2020: -0.0488 -0.074 0.0784 -0.0119 -0.0211 ...
```

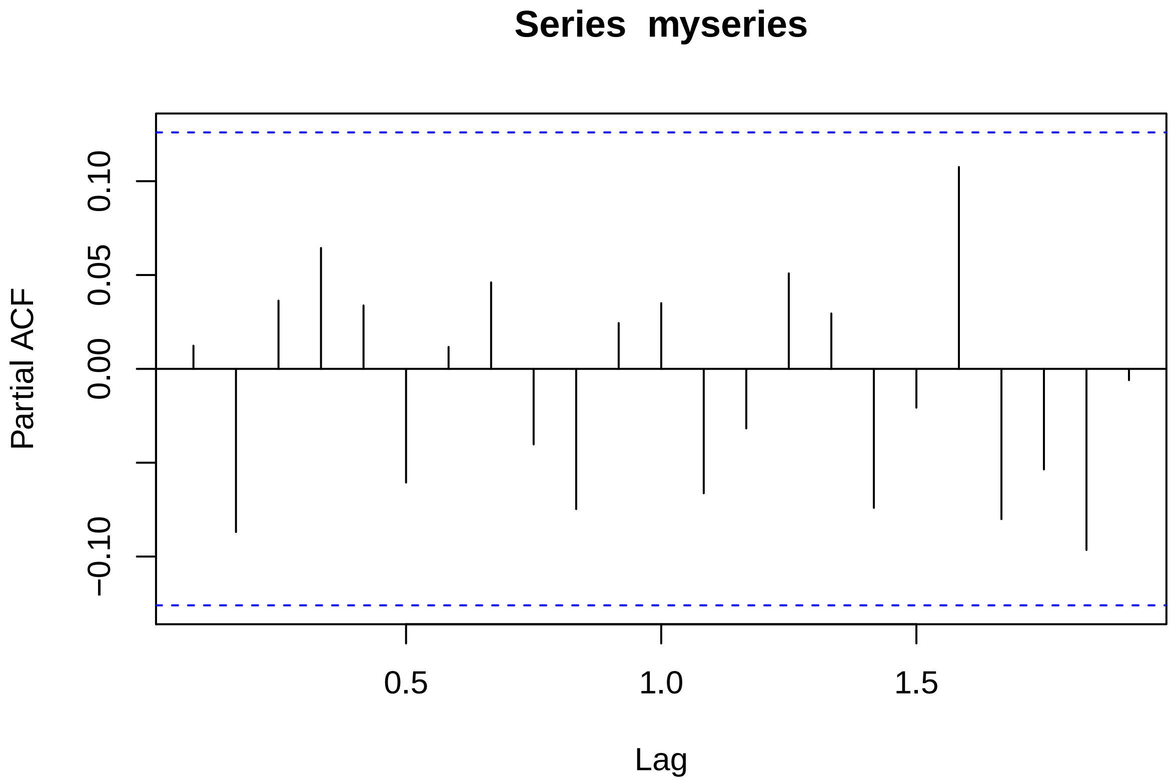
```
plot(myseries,type="l",xlab="Time",ylab="Monthly Excess Return")
```



```r
acf(myseries)
```



```r
pacf(myseries)
```



Then confirm that the data is actually stationary, the result below shows that no differences are required to make the time series stationary.

```r
ndiffs(myseries)
```

```
## [1] 0
```

Now fit two ARIMA models, one with parameters (2, 0, 1) and the other with parameters (2, 0, 2). We choose such parameters because we see that the sample ACF cuts off after lag 1.5. (Note that ARIMA(p, 0, q) is actully ARMA(p, q) because no differencing processing is required, thus I am essentially fitting ARMA(2, 1) model). We see AIC for model1 is -852.23 and AIC for model2 is -857.72. model2 has a smaller AIC so it can be considered as a reasonably better model.

```r
model1<-arima(myseries,order=c(2,0,1),method="ML")
model1
```

```
##
## Call:
## arima(x = myseries, order = c(2, 0, 1), method = "ML")
##
## Coefficients:
##           ar1      ar2     ma1  intercept
##       -0.1729  -0.0903  0.1867     0.0040
## s.e.   0.5128   0.0673  0.5126     0.0025
##
## sigma^2 estimated as 0.00166:  log likelihood = 431.11,  aic = -852.23
```

```r
model2<-arima(myseries,order=c(2,0,2),method="ML")
model2
```

```
##
## Call:
## arima(x = myseries, order = c(2, 0, 2), method = "ML")
##
## Coefficients:
##           ar1      ar2     ma1     ma2  intercept
##       -0.1157  -0.9506  0.1235  0.8886     0.0040
## s.e.   0.0589   0.0445  0.0889  0.0641     0.0025
##
## sigma^2 estimated as 0.001607:  log likelihood = 434.86,  aic = -857.72
```

These results below shows the accuracies of the two model.

```r
accuracy(model1)
```

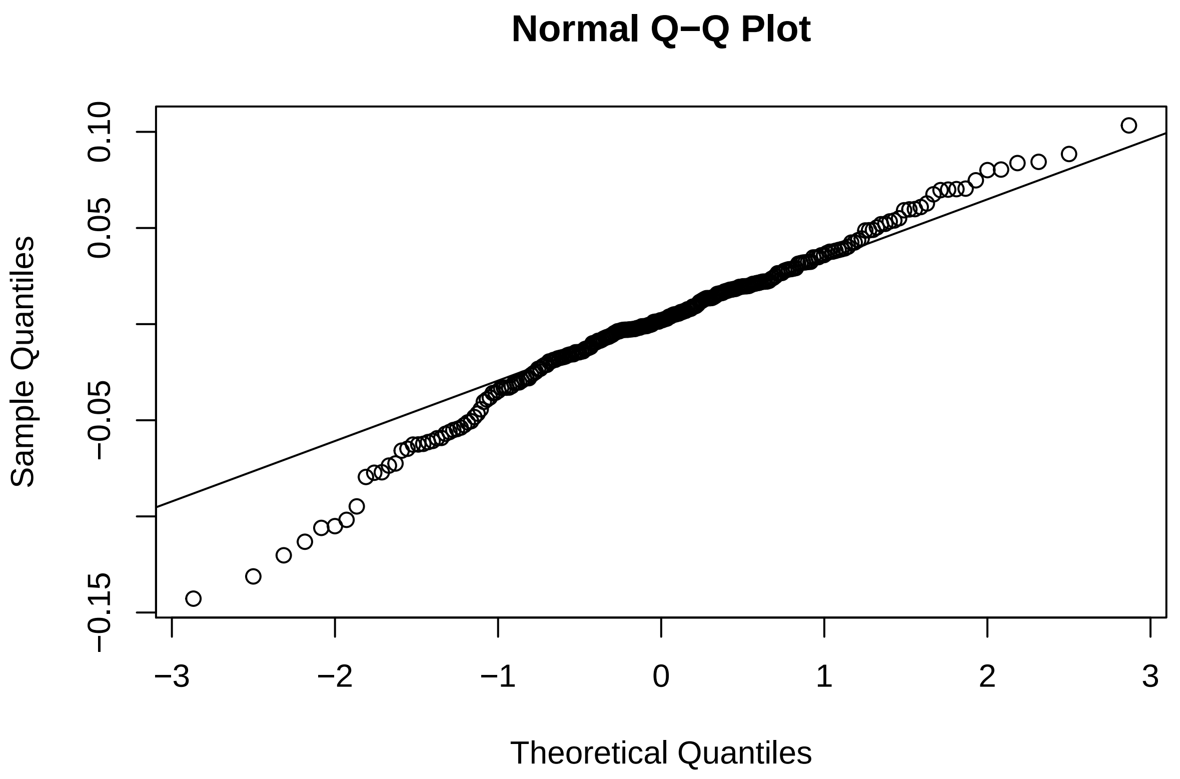
```
##                         ME       RMSE       MAE      MPE     MAPE      MASE
## Training set -4.129783e-05 0.04074424 0.0305443 111.1082 155.2355 0.7011763
##                     ACF1
## Training set 0.001942939
```

```r
accuracy(model2)
```

```
##                         ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -8.528295e-05 0.04008995 0.03036243 165.2475 185.5155 0.6970012
##                     ACF1
## Training set 0.008461902
```

Now testing whether the residuals of the model are white noise by using the command Box.test(), we see that some residuals are not normally distributed, but most of them are on the straight line. The p-value of the Ljung-Box Test is 0.9757, which is larger than 0.05. Fxing significance level at 5%, we fail to reject the null hypothesis that the model residuals are white noise. So we can say that the residuals are white noise.

```r
qqnorm(model1$residuals)
qqline(model1$residuals)
```

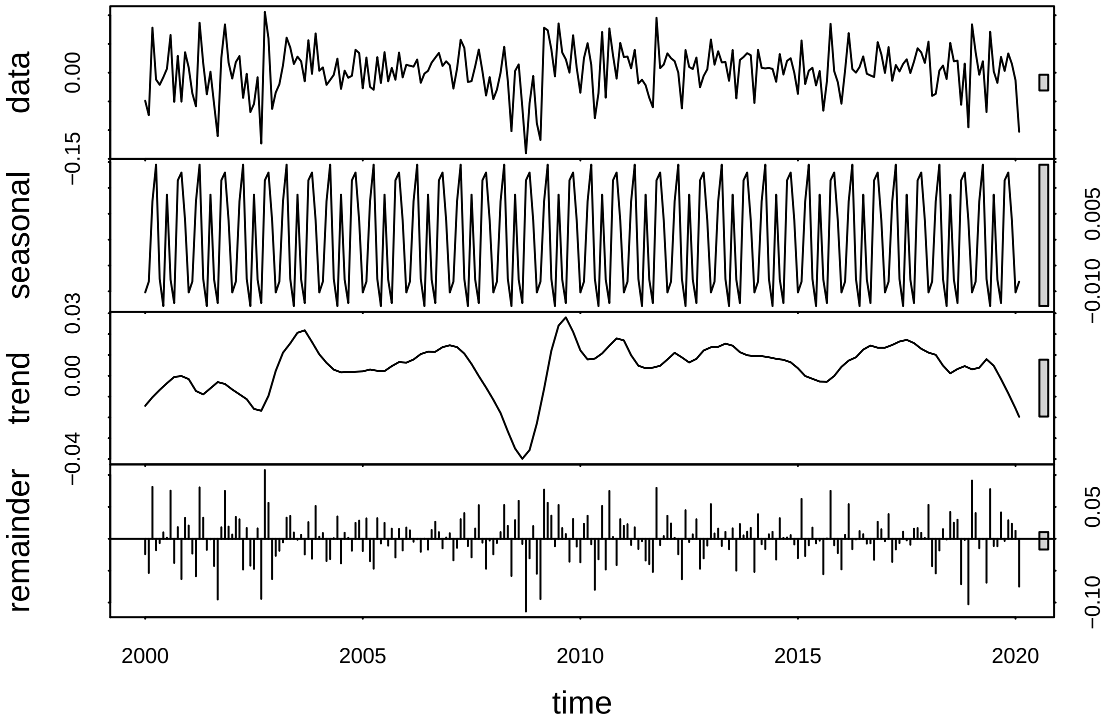


```
Box.test(model1$residuals,type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  model1$residuals
## X-squared = 0.00092493, df = 1, p-value = 0.9757
```

Now fit another model Holt-Winters. Here, stl( ) function decomposes the time series into seasonal, trend and irregular components using loess. The Holt-Winters model Holtfit is constructed by the ets() function. We see the model has an AIC value of -192.1588.

```
holt<-stl(myseries,s.window="period")
plot(holt)
```



```r
holtfit <- ets(myseries, model = "AAA")
holtfit
```

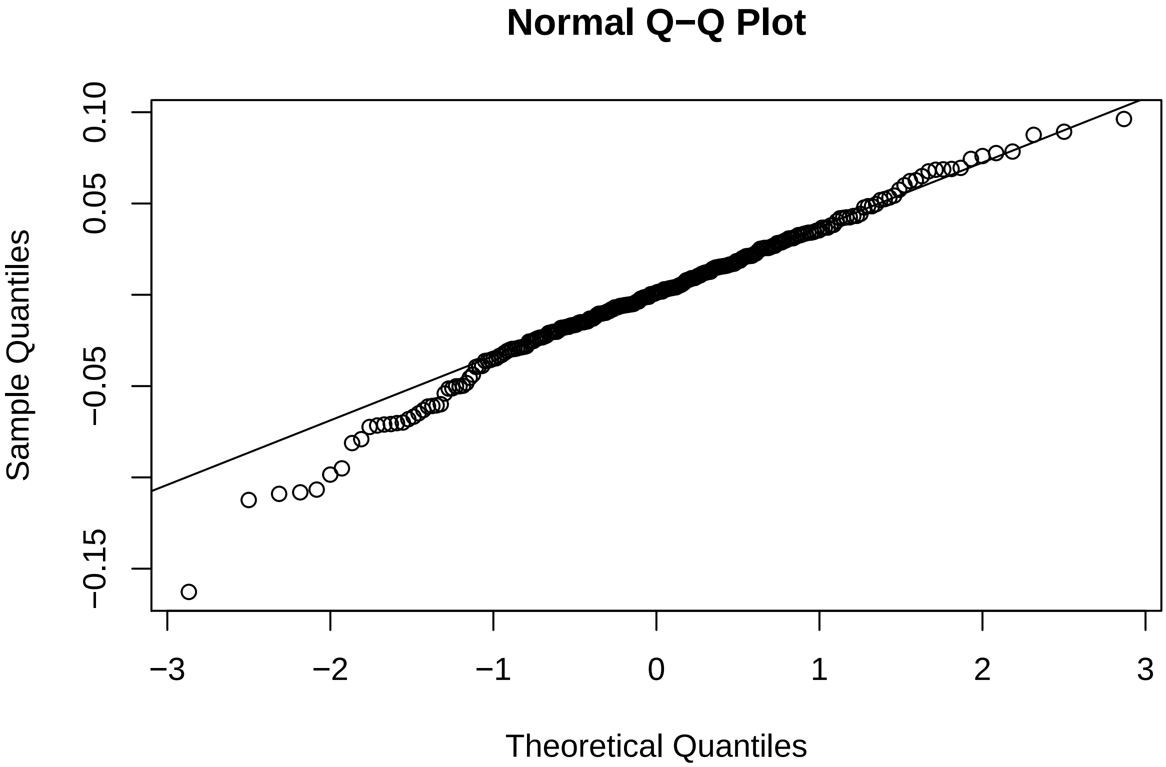
```
## ETS(A,A,A)
##
## Call:
##  ets(y = myseries, model = "AAA")
##
##   Smoothing parameters:
##     alpha = 2e-04
##     beta  = 1e-04
##     gamma = 0.0337
##
##   Initial states:
##     l = 0.0018
##     b = 0
##     s = 0.0035 0.0142 0.0147 -0.0133 0.0014 0.0061
##            -0.0105 -0.008 0.017 0.0078 -0.0077 -0.0252
##
##   sigma:  0.0417
##
##       AIC      AICc       BIC
## -192.1588 -189.4266 -132.8468
```

```r
accuracy(holtfit)
```

```
##                         ME       RMSE        MAE       MPE     MAPE      MASE
## Training set -0.0007848178 0.04028649 0.03069148 0.1527364 325.3258 0.7083102
##                     ACF1
## Training set 0.003581663
```

Now testing whether the residuals of the model are white noise by using the command Box.test(), we see that most points are on the straight line, which is a good indication of normally distuibuted residuals. Meanwhile, the Ljung-Box Test gives a p-value of 0.9553, which is larger than 0.05, Fxing significance level at 5%, we fail to reject the null hypothesis that the model residuals are white noise. So we can say that the residuals are white noise.

```
qqnorm(holtfit$residuals)
qqline(holtfit$residuals)
```



```
Box.test(holtfit$residuals,type="Ljung-Box")
```

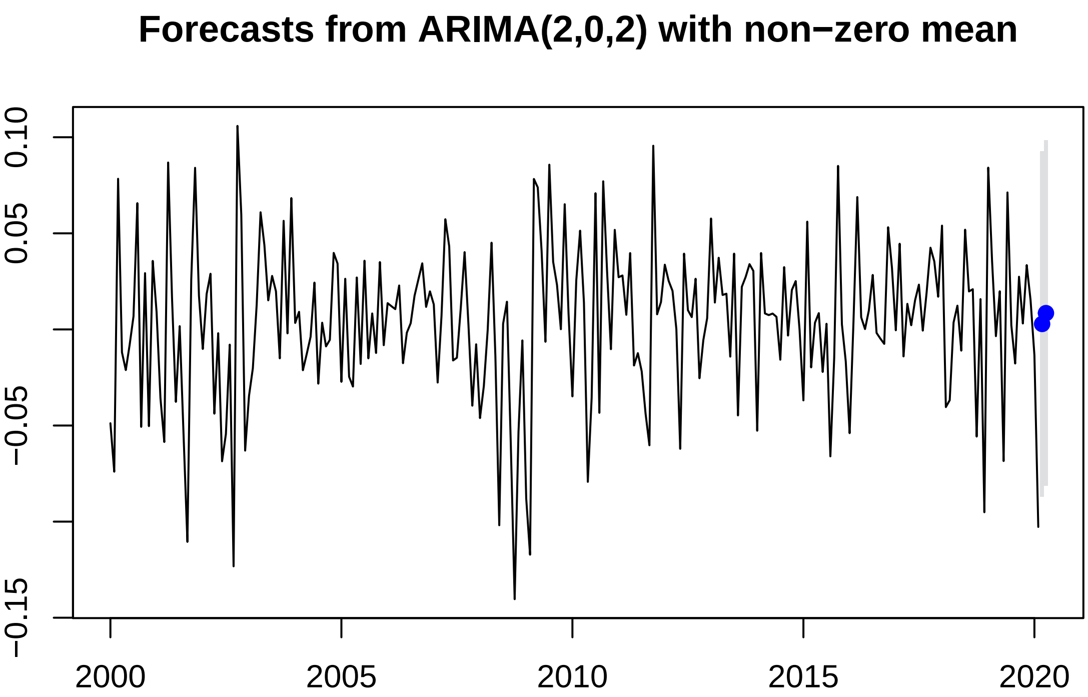
```
##
##  Box-Ljung test
##
## data:  holtfit$residuals
## X-squared = 0.0031431, df = 1, p-value = 0.9553
```

Finally, based on the above analysis, the ARIMA(2, 0, 2) model seems to be a better fit, because it has the lowest AIC. Therefore, I forecast the monthly excess return of March and April 2020 based on this model. The predicted point estimates are shown as two blue points in the plot, their corresponding confidence intervals are included as well. Based on this prediction, the predicted excess return for March 2020 is 0.002845780, with a 95% confidence interval of (-0.08701194, 0.09270351), the predicted excess return for April 2020 is 0.008486569, with a 95% confidence interval of (-0.08137392, 0.09834705).

```
p1<-forecast(model2,h=2,level=c(97.5))
p1
```

```
##          Point Forecast     Lo 97.5    Hi 97.5
## Mar 2020    0.002845780 -0.08701194 0.09270351
## Apr 2020    0.008486569 -0.08137392 0.09834705
```

```
plot(p1)
```



# Appendix The data was downloarded at finance.yahoo.com .
“Dow Jones Industrial Average (^DJI) Historical Data.” Yahoo! Finance, Yahoo!, 6 Apr. 2020, finance.yahoo.com/quote/$^{\text{DJI/history?p=}}$DJI.