Stock Price Forecasting for P&G Stock Price

Steve Yoon 24918013

Min Zhao 37275476

Huanqing Wang 43566330

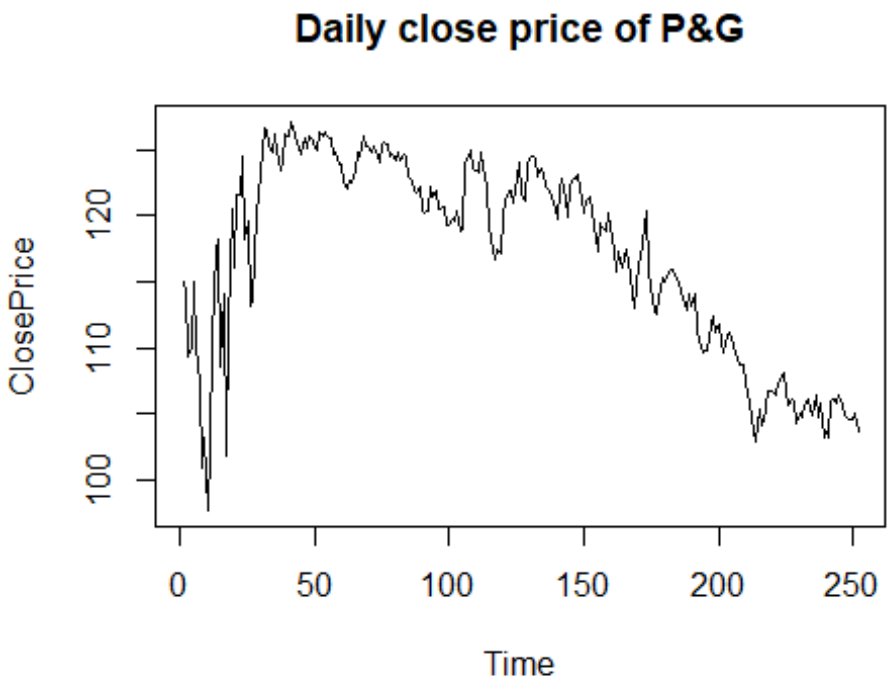
Introduction:

We have recorded P&G stock price time series in order to analyze if there is trend or seasonal effect of data and to test acf and pacf plot of this data. We suspected the data to have a trend because as soon as stock price increase, people will have information about P&G company stock price resulting more people to try to buy company's stock and increase demand. 252 data of stock price from 7th December of 2019 to April 3rd by daily with acf and pacf plot will be given. We will look at those plots to see which model will fit the data. We will take difference to remove any trend and try to fit the model to ARMA model. Also, we will use Ljung-Box test to be more specific on which model best fit. At last, we will use portmanteau test to see if residual is white noise meaning residuals are random and has normal distribution with expectation equal to 0. Furthermore, we will provide predicted value to compare with actual value and see how it will be continuing to act for further prediction.

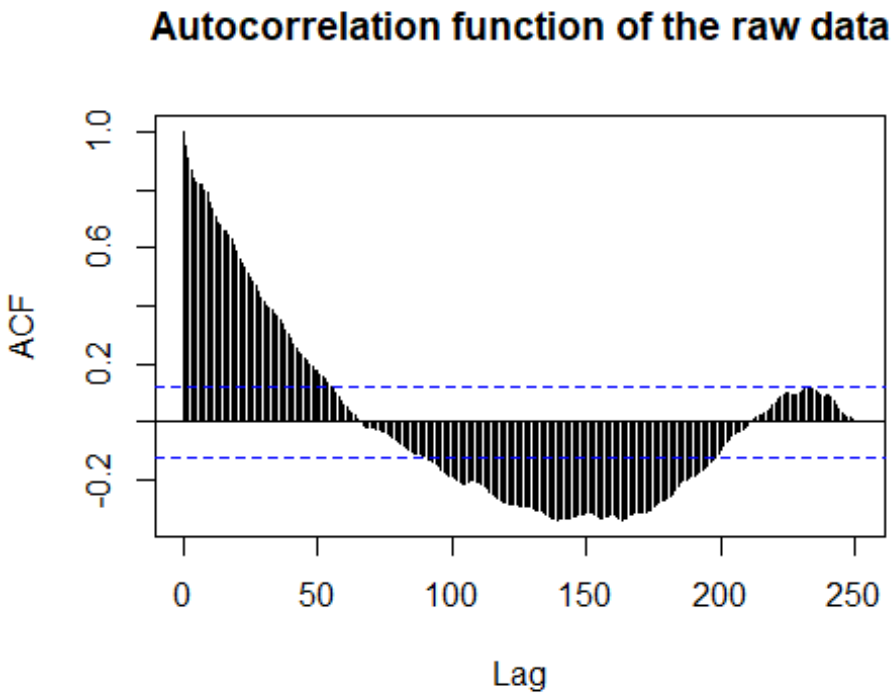
Analysis of Data:

We first plot the data and its acf:

```
data =read.csv("HistoricalQuotes.csv", header = TRUE)
ClosePrice = as.ts(as.numeric(gsub('[$]','',data$Close.Last)))
plot(ClosePrice,main = "Daily close price of P&G")
```

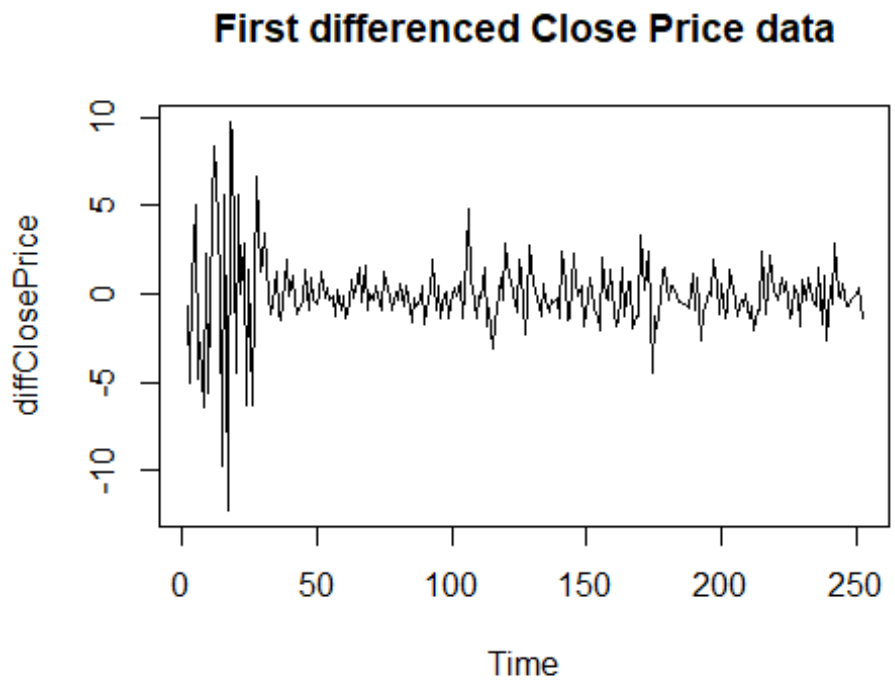



```
acf(ClosePrice,lag.max="252",main ="Autocorrelation function of the raw
 data")
```

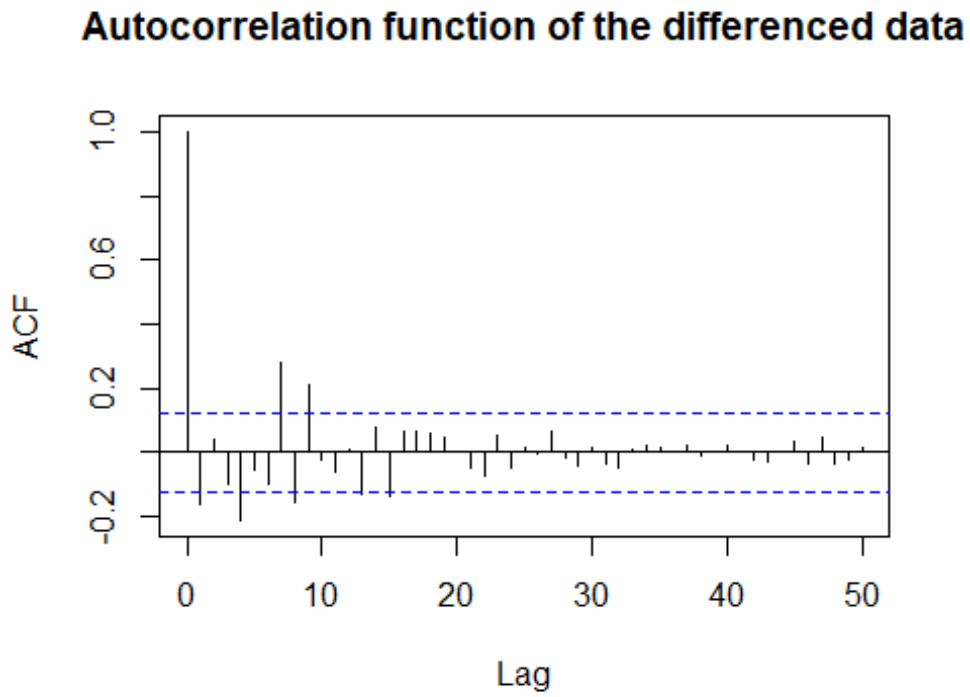


The plot shows that the series has an overall decreasing trend after appproximately the 30^{th} observation. The acf also decreases very slowly, which also suggests that the series has a trend. Hence, we try to difference the data to make it stationary:

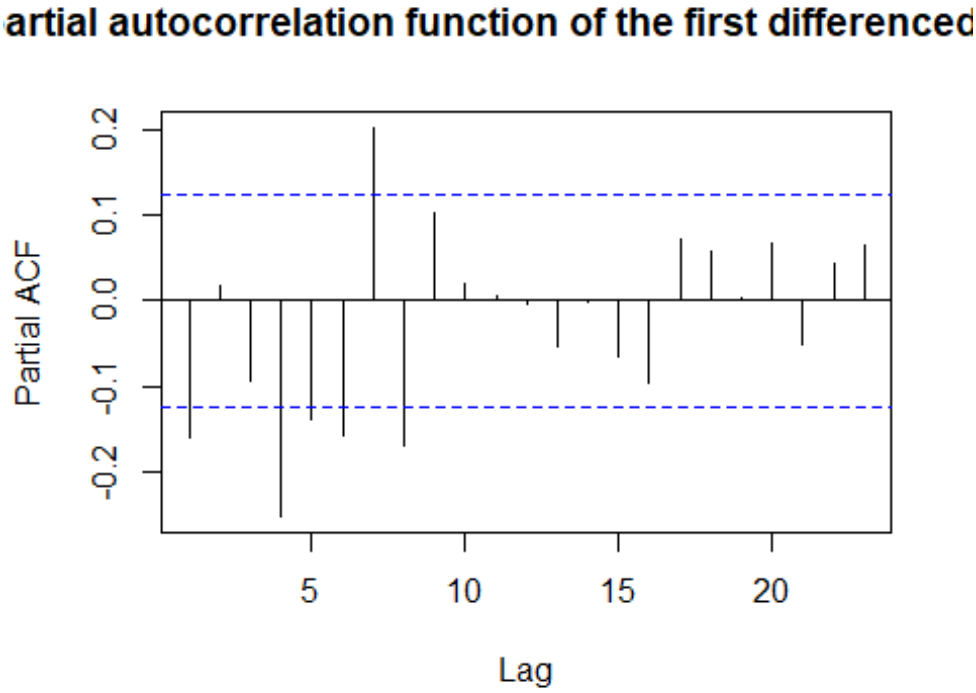
```
diffClosePrice = diff(ClosePrice,lag=1)
plot(diffClosePrice,main = "First differenced Close Price data")
```




```
acf(diffClosePrice,lag.max=50,main ="Autocorrelation function of the di
fferenced data")
```



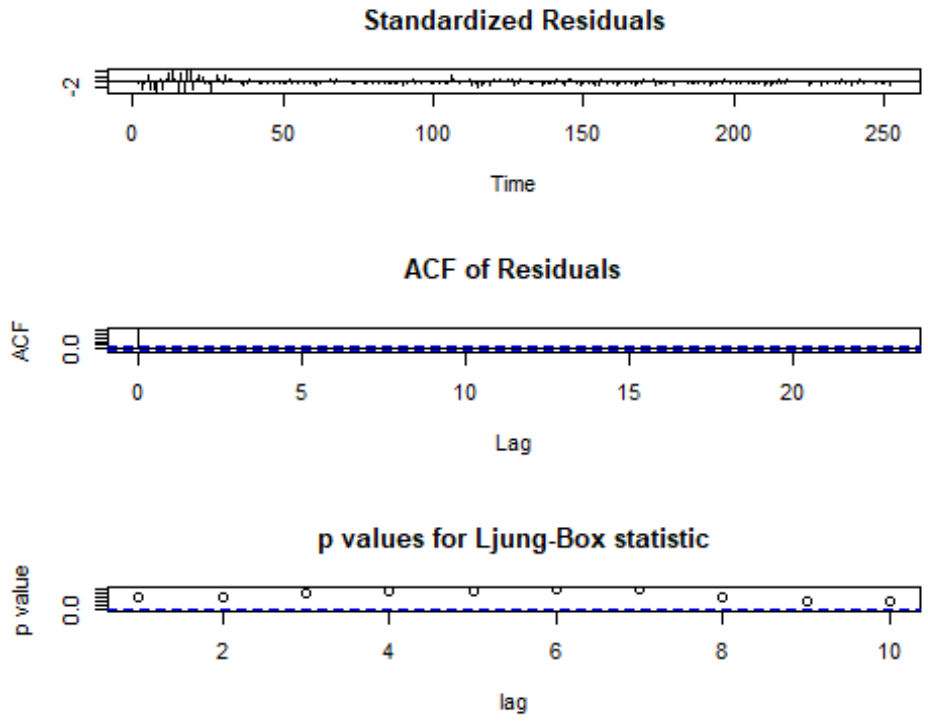

```
pacf(diffClosePrice, main = "partial autocorrelation function of the fi
rst differenced data")
```



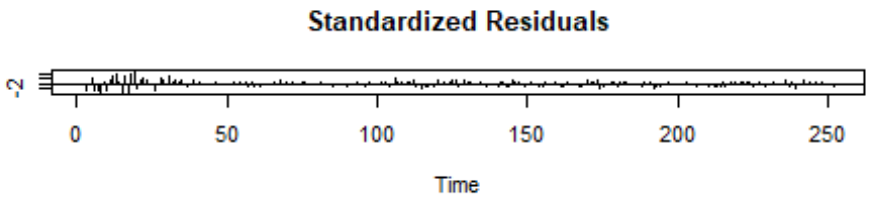
The plot suggests that the new series might be stationary because the mean value is about constant so that there is no trend, and it also does not have any seasonal effect. Nevertheless, there are some extreme values at the beginning of the series. This corresponds to the sudden increase of the stock price in our raw data. Since acf does not decay slowly this time, the new series looks staionary. The acf is significant at lags less than 10 and is the most significant at lag 7. The pacf plot is significant at lags less than 8.

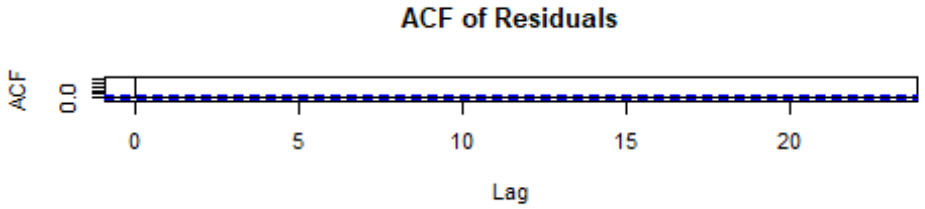
We then decided to fit an ARMA model to our stationary series. We selected three models by trial and error and by looking at the diagnostic plots, plus the principle of parsimony. Since all models with $p + q < 7$ have three or more significant Ljung-Box statistics, we eventually chose three possible models with $p + q = 7$.

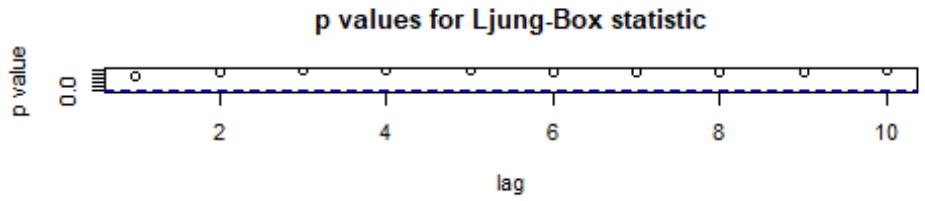
```
ar7=arima(diffClosePrice,order = c(7,0,0))
arma61=arima(diffClosePrice,order = c(6,0,1))
arma16=arima(diffClosePrice,order = c(1,0,6))
tsdiag(ar7)
```



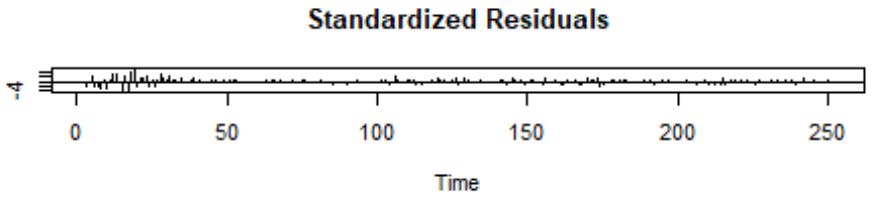

```
tsdiag(arma61)
```

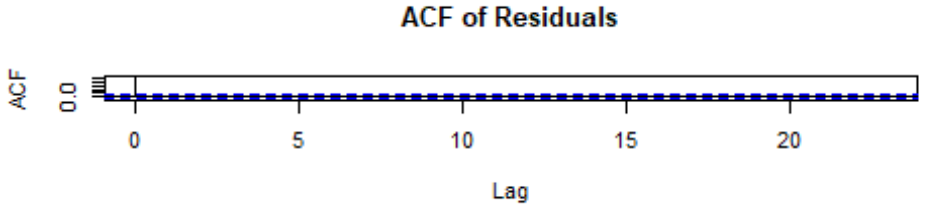


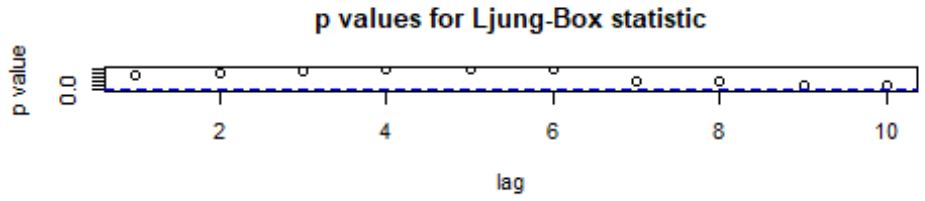





```
tsdiag(arma16)
```







We saw that all the p-values are not significant in these models; most acf's are not significant; and most of the standardized residuals are within the range ± 2. Notice that some residuals at the beginning of the plots are significant, but this corresponds to the sudden increase of our raw data, and therefore poorly fitted values might be expected. Since all these three models fit our data pretty good, we decided to split our data into a training set and a test set to choose the best one.

```
m=18
train<- 1:(length(diffClosePrice)-m)
trainx<- diffClosePrice[train]
testx<- diffClosePrice[-train]
foremodel7 = predict(ar7, m)
foremodel61 = predict(arma61, m)
foremodel16 = predict(arma16, m)
sum((testx - foremodel7$pred)^2)

## [1] 25.53418

sum((testx - foremodel61$pred)^2)

## [1] 24.65403

sum((testx - foremodel16$pred)^2)

## [1] 30.41133
```

We notice that the error in the ARMA(6,1) model is the smallest. Hence, our model is: $X(t) + 0.0393 = -0.8938 * (X(t-1) + 0.0393) - 0.2109 * (X(t-2) + 0.0393) - 0.1711 * (X(t-3) + 0.0393) - 0.3659 * (X(t-4) + 0.0393) - 0.3628 * (X(t-5) + 0.0393) - 0.3066 * (X(t-6) + 0.0393) + 0.7130 * Z(t-1) + Z(t)$ where $X(t)$ represents the difference in the close price of P&G company, and $Z(t)$ is the white noise process with estimated variance 4.09.

```
arma61

##
## Call:
## arima(x = diffClosePrice, order = c(6, 0, 1))
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ar6     ma1
intercept
##       -0.8938  -0.2109  -0.1711  -0.3659  -0.3628  -0.3066  0.7130
  -0.0393
## s.e.   0.0945   0.0806   0.0776   0.0782   0.0817   0.0623  0.0831
   0.0665
##
## sigma^2 estimated as 4.09:  log likelihood = -533.46,  aic = 1084.91
```

At last, we double check that the residuals of our model indeed follow a white noise process by using the portmanteau lack-of-fit test.

```
N = length(arma61$residuals)
M1 = 15
M2 = 25
M3 = 30
rho = acf(arma61$residuals,lag = 50,plot = F)$acf
Q1 = 200*sum(rho[2:(M1+1)]**2)
Q2 = 200*sum(rho[2:(M2+1)]**2)
Q3 = 200*sum(rho[2:(M3+1)]**2)
pchisq(Q1,M1,ncp=0,lower.tail=F)

## [1] 0.9878957

pchisq(Q2,M2,ncp=0,lower.tail=F)

## [1] 0.9938309

pchisq(Q3,M3,ncp=0,lower.tail=F)

## [1] 0.9979189
```

It turns out that none of the p-values is significant, so we do not have enough evidence to reject the null hypothesis that the residuals are not from the white noise process in 5% significance level.

We could also predict next week's (beginning at April 4th) stock prices:

```
pred=predict(arma61,n.ahead = 7)
cumsum(pred$pred)+ClosePrice[length(ClosePrice)]

## [1] 104.0132 103.9113 103.9351 104.1103 104.0868 104.2468 103.8654
```

Conclusion

Our initial analysis shows stock price is non-stationary. It makes sense since today's stock price is highly correlated to the previous day: people tend to buy the stock if the stock price has decreased yesterday, and that might be the reason why we have negative values in the AR coefficients, in particular at lag 1 which is the biggest value in magnitude. The prediction might not be super accurate since recent stock price is very unpredictable due to 2019-corononvirus. But we also find some interesting exceptions worth looking into. There are a number of limitations and shortcomings in our experiment and potential improvements can be made to our data collection and analysis method. Further researches can be done with possible improvements such as more refined search data and more accurate algorithm to compute stock price values.

Reference

https://www.nasdaq.com/market-activity/stocks/pg/historical?fbclid=IwAR0GbNenFN7EkcjzN_mBE3jr954xlDy1FqecYjcmGQAEG8I6zbYVkqAf_rk