Mustafa Abdullayev

Price Prediction Project

AY6050 – Intro to Enterprise Analytics

Term: Winter 2020

Instructor: Roy Wada

03/09/2020

# Introduction

This is Microsoft Word Report accompanying R Script. In my Script, my main aim was to perform both regression and time series analysis for future price prediction. As a data source, I used real World stock prices (Pepsi, Apple, Microsoft, Disney and Home Depot) between January 2018 and March 2020. (Courtesy of Yahoo Finance). For the first two parts, I utilized exponential smoothing analysis and tried find optimal smoothing parameters (alpha and beta). Secondly, I used simple and multiple regression analysis in order to predict future stock prices. In this part, I utilized powerful R built-in functions and graphs, such as bar charts, to dive deeper to observe and verify assumptions of regression models such as normality of residuals. Moreover, Chi-square test was utilized in order to formally prove/disprove my findings. Since I also provided R script with all the codes and comments, I removed some of the codes and comments from my report (such as package loading). Also, I did not incorporate all of my R code to report since it was too lengthy. However, all necessary outputs (Test results and graphs) are added. It is due to keep my report brief, succinct and to the point.

# Problem 1

For the first problem, I tried to use exponential smoothing in order to forecast stock prices for March 2020 (2nd, 3rd,4th,5th, and 6th day of March). I utilized Holt-Winters smoothing and different alpha values in order to find optimal smoothing factor(alpha) for each stock. In plain English, smoothing factor or alpha is the weight we assign to the newest observation ("EXPONENTIAL SMOOTHING", 2001). For instance, setting value of alpha close to 1 means we pay attention to only newest observations. As I expected, almost each stock has a different optimal value for alpha. For Pepsi and Microsoft, the optimal alpha value is 0.35 since I got the smallest error (Pepsi MSE = 6.51 and Microsoft MSE = 18.34). On the other hand, 0.15 is the optimal alpha value for Apple and Home Depot stocks (Apple MSE = 185.54 and Home Depot MSE = 34.39). Finally, for Disney Stocks, using exponential smoothing with alpha = 0.55 is the best choice (Disney MSE = 13.82).

In my opinion, small alpha values, usually, yielded better result because there are wild swings in the recent stock prices due to external risks (i.e. Coronavirus). Since there is no fundamental change in stocks pricing mechanisms and recent observations are just price fluctuations, putting heavy weight in recent observations is not wise.

###################### Problem 1 ###########################################################  
  
PEP <- read.csv("PEP\_Train.csv")  
train\_PEP <- ts(PEP$Close,frequency=21)  
test\_PEP <- read.csv("PEP\_Test.csv")$Close  
  
AAPL <- read.csv("AAPL\_Train.csv")  
train\_AAPL <- ts(AAPL$Close,frequency = 21)  
test\_AAPL <- read.csv("AAPL\_Test.csv")$Close  
  
DIS <- read.csv("DIS\_Train.csv")  
train\_DIS <- ts(DIS$Close,frequency = 21)  
test\_DIS <- read.csv("DIS\_Test.csv")$Close  
  
HD <- read.csv("HD\_Train.csv")  
train\_HD <- ts(HD$Close, frequency = 21)  
test\_HD <- read.csv("HD\_Test.csv")$Close  
  
MSFT <- read.csv("MSFT\_Train.csv")  
train\_MSFT <- ts(MSFT$Close,frequency = 21)  
test\_MSFT <- read.csv("MSFT\_Test.csv")$Close  
  
alpha\_values = c(0.15,0.35,0.55,0.75)  
beta\_values = c(0.15,0.25,0.45,0.85)  
  
  
 ####### Pepsi ###########

for (val in alpha\_values){  
   
 holt <- HoltWinters(train\_PEP,alpha = val,seasonal = "multiplicative")  
 prediction <-forecast(holt,h=5)  
 pred <- prediction$mean  
 actual <- test\_PEP  
   
 print(MSE(pred,actual))  
}

## [1] 20.84253

## [1] 6.510273  
## [1] 12.93786  
## [1] 11.44949

######## Apple ###########

### Same Technique with Pepsi

## [1] 154.2569  
## [1] 185.5413  
## [1] 366.6918  
## [1] 229.9055

######## Microsoft ###########

### Same Technique with Pepsi

## [1] 80.27978  
## [1] 18.33509  
## [1] 32.51709  
## [1] 36.04999

######## Disney ###########

### Same Technique with Pepsi

## [1] 134.0496  
## [1] 31.6091

## [1] 13.8204  
## [1] 15.7706

######## Home Depot ###########

### Same Technique with Pepsi

## [1] 34.39  
## [1] 57.41938

## [1] 131.0011  
## [1] 72.26973

# Problem 2

To continue my time series analysis, I tried to use adjusted exponential smoothing in order to forecast stock prices for March 2020 (2nd, 3rd,4th,5th, and 6th day of March).Again, I utilized Holt-Winters smoothing. But this time, I tried find optimal beta value for my model. For this purpose, I assumed alpha to be constant and equal to 0.75. In plain English, beta used to control the trend calculation. Smaller beta values make model to concentrate on long-term trend whereas models with greater beta values put more emphasis on short-term trend (----------). Here, I found that, except for Apple, setting beta equals to 0.15 yields best results. Interestingly, optimal beta value for Apple stock is found to be 0.85. (Detailed MSE values are in the output part below)

Despite having smallest error for beta = 0.85, beta value of 0.15 was the second best for Apple stock and their difference in terms of MSE was not extremely high (2284 vs 1424 ). This in turn, demonstrate that, since recent trend is not rational and panic selling (due to Coronavirus), we should not put heavy emphasize on recent trade in order to predict future values. Since there is no fundamental change in stocks pricing mechanisms and recent observations are just price fluctuations, we should put more emphasize on long-term trend.

#################################### Problem 2 ###############################  
  
 ####### Pepsi ###########

for (val in beta\_values){  
   
 holt <- HoltWinters(train\_PEP,alpha = 0.75,beta = val,seasonal = "multiplicative")  
 prediction <-forecast(holt,h=5)  
 pred <- prediction$mean  
 actual <- test\_PEP  
   
 print(MSE(actual,pred))  
}

## [1] 124.9577  
## [1] 265.2611  
## [1] 451.4204  
## [1] 515.69

######## Apple ###########

### Same Technique with Pepsi

## [1] 2284.147  
## [1] 3018.845  
## [1] 2387.595  
## [1] 1424.743

######## Microsoft ###########

### Same Technique with Pepsi

## [1] 155.7627  
## [1] 309.396  
## [1] 274.3486  
## [1] 344.046

######## Disney ###########

### Same Technique with Pepsi

## [1] 54.95933  
## [1] 131.1905  
## [1] 144.2038  
## [1] 93.37999

######## Home Depot ###########

### Same Technique with Pepsi

## [1] 782.1339  
## [1] 1567.875  
## [1] 2454.345  
## [1] 3129.399

# Problem 3

In the third part of my analysis, I utilized simple regression analysis in order to predict stock prices for 4th, 5th and 6th of March 2020. Additionally, with the help of R graphs and charts, I tried to validate assumptions of simple regression analysis. For each stock, I choose 2 days lagged stock price as an independent variable in order to construct my regression model. After constructing my simple regression model, I used histogram to observe distribution of residuals. From below graphs, (Figure 1.1, Figure 2.1, Figure 3.1, Figure 4.1 and Figure 5.1) we can observe that residuals are almost normally distributed. Also, there is a negative skewness in all graphs. In order to dive deeper, I constructed probability plots for residuals of 5 regression models for each stock (Figure 1.2 , Figure 2.2, Figure 3.2, Figure 4.2, Figure 5.2 ). Although there is a mismatch in the starting values, we can say that residuals for each stock fit well to normally distributed values. Also, in my opinion, this mismatch in the starting values stem from negative skewness in the probability distribution of residuals. Finally, in order to prove that residuals are normally distributed, I carried out Chi-Square test with 0.05 significance. My hypotheses are as follow :

1. H0 : There is no significant difference in the distribution of residuals and Normal distribution
2. There is a significant difference between distribution of residuals and Normal Distribution.

As a result of these test, for all regression models, my p-values were bigger than 0.05 (between 0.2 and 0.3). So, I do not have enough evidence to reject my null hypotheses. Thus, I can not say that residual values for my simple regression analyses are not normally distributed.

To continue my analysis of residual. I plotted scatter plot , residuals values against time, in order to see if their independency (Figure 1.3, Figure 2.3, Figure 3.3, Figure 4.3, Figure 5.3). From these graphs, we can clearly see that my residual values are not dependent on time. At any given time, my residual values are independent of each other.

Finally, I tried to check homoscedasticity property of residuals. That assumption say that at any given value of observation, variance of residuals should be the same. So, I plotted my residuals against stock prices for each of my stocks (Figure 1.4, Figure 2.4, Figure 3.4, Figure 4.4, Figure 5.4). Except for the last week of stock prices, it is obvious that variance of residuals is same for any given value of stock price. Indeed, due to recent panic in markets (i.e. Coronavirus) there is a greater variance in the residuals of regression analyses for each stock.

After making sure that all of assumptions of simple linear regression is satisfied for my regression models, I tried to predict future stock prices for each stock. Before that, I observed that Coefficient of Determination (R-square) value for my regression models. It turns out that , for every model my Coefficient of Determination is between 0.95 and 1. That means, my regression models can explain the more than 95% of variance in the actual stock prices. Also, Coefficient of correlation between my dependent (2 days lagged stock prices) and dependent values (actual stock prices) are more than 0.95. It demonstrates that, as expected, there is a heavy positive correlation between stock prices and 2 day lagged stock prices (Actual R-squared and correlation values are in the output part.) It is indeed impressive but neither R-square value nor correlation coefficient do not say anything about our prediction power. So, in order to test my models, predicted stock prices for 4th,5th and 6th of March. I got MSE values of 15.96, 71.74,29.96, 7.62 and 109.60 for Pepsi, Apple, Microsoft, Disney and Home Depot, respectively. Indeed, these are good prediction. For Apple and Disney, using simple linear regression with 2 days lagged prices gave the best results. However, for remaining stocks (Pepsi, Microsoft and Home Depot) using simple exponential smoothing is still best choice.

################### Problem 3#####################################  
  
  
############### Pepsi ############  
  
PEP\_lag0 <- Lag(PEP$Close,shift = 0)  
  
PEP\_lag2 <- Lag(PEP$Close,shift = 2)  
  
PEP\_lag <- data.frame(PEP\_lag0,PEP\_lag2)  
names(PEP\_lag) <- c("lag0","lag2")  
  
  
PEP\_test\_lag2 <- Lag(test\_PEP,shift = 2)  
  
PEP\_test\_lag <- data.frame(PEP\_test\_lag2)  
names(PEP\_test\_lag) <- c("lag2")  
  
model <- lm(lag0 ~ lag2,data = PEP\_lag)  
  
## Coefficient of Determination  
summary(model)

## Multiple R-squared: 0.9814, Adjusted R-squared: 0.9814   
## F-statistic: 2.844e+04 on 1 and 539 DF, p-value: < 2.2e-16

## Coefficient of Correlation  
cor\_PEP <-cor(PEP\_lag, use = "pairwise.complete.obs")  
print(round(cor\_PEP,2))

## lag0 lag2  
## lag0 1.00 0.99  
## lag2 0.99 1.00

## Actual Prediction and MSE  
pred\_PEP\_3 <- predict(model,newdata = PEP\_test\_lag)  
pred\_PEP\_3[3:5]

## 3 4 5   
## 137.5179 135.5325 142.2926

mse\_3 <- MSE(pred\_PEP\_3[4:5],test\_PEP[4:5])  
print(mse\_3)

## [1] 15.95939

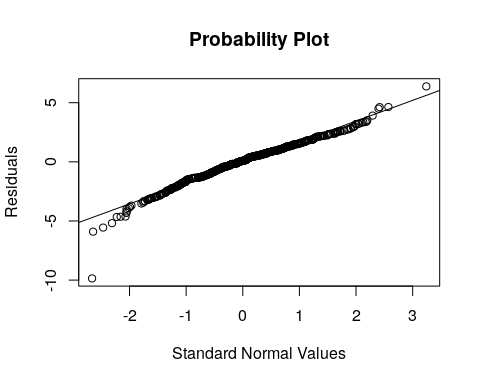
## Histogram and Probability fplot for residuals  
hist(model$residuals, xlab = "bins",ylab = "residuals", main = "Distribution of residuals")

Figure .1 – Distribution of residuals for Pepsi



qqPlot(x=rnorm(length(model$residuals)),y=model$residuals,add.line = TRUE,xlab = "Standard Normal Values",ylab = "Residuals", main = "Probability Plot")

Figure .2 – Probability plot of residuals for Pepsi

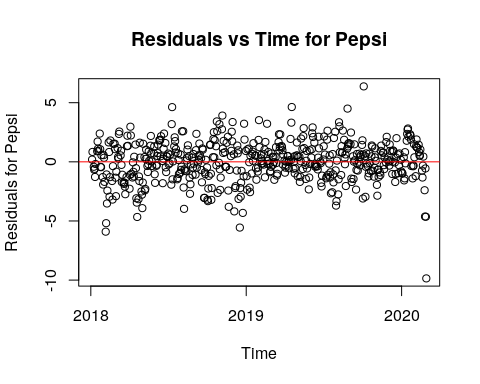


## Chi Square Test  
PEP\_test <- chisq.test(x = model$residuals, y = rnorm(length(model$residuals)))

## X-squared = 292140, df = 291600, p-value = 0.2396

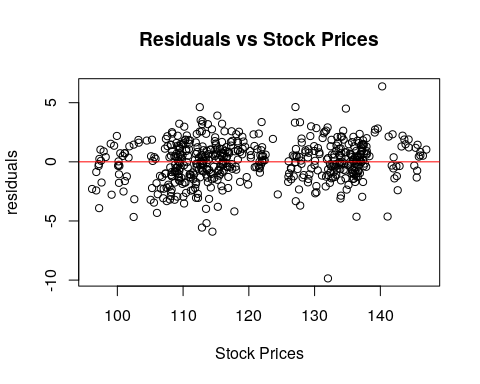
## Residuals vs Time  
plot(x = as.Date(PEP$Date[3:543]),y = model$residuals,main = "Residuals vs Time for Pepsi",xlab = "Time",ylab = "Residuals for Pepsi")  
abline(0,0,col="red")

Figure .3 –Residuals against time for Pepsi



## Residuals vs Stock Price  
plot(x=PEP$Close[3:543],y = model$residuals,xlab = "Stock Prices",ylab = "residuals",main = "Residuals vs Stock Prices")  
abline(0,0,col = "red")

Figure .4 – Homoscedasticity of residuals for Pepsi



############## Apple #############3  
  
## Coefficient of Determination  
summary(AAPL\_model)

## Multiple R-squared: 0.9843, Adjusted R-squared: 0.9843   
## F-statistic: 3.387e+04 on 1 and 539 DF, p-value: < 2.2e-16

## Coefficient of Coreelation  
cor\_AAPL <-cor(AAPL\_lag, use = "pairwise.complete.obs")  
print(round(cor\_AAPL,2))

## lag0 lag2  
## lag0 1.00 0.99  
## lag2 0.99 1.00

## Actual Prediction and MSE  
pred\_AAPL\_3 <- predict(AAPL\_model,newdata = AAPL\_test\_lag)  
pred\_AAPL\_3[3:5]

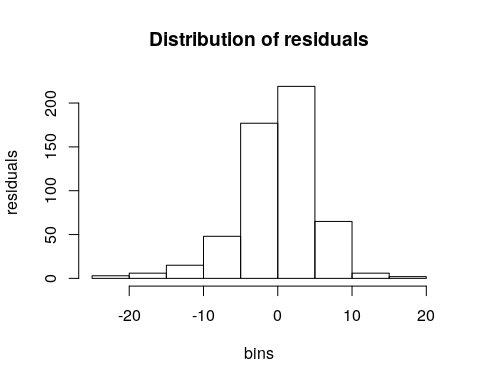
## 3 4 5   
## 298.7784 289.3302 302.6911

mse\_3 <- MSE(pred\_AAPL\_3[3:5],test\_AAPL[3:5])  
mse\_3

## [1] 71.7352

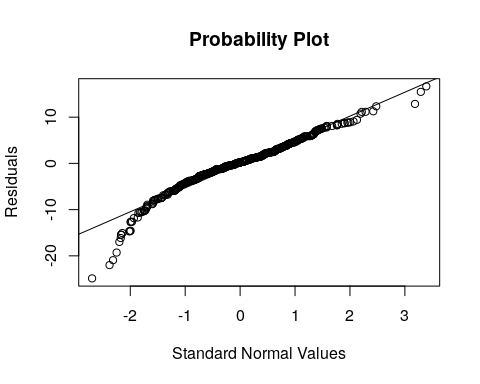
## Histogram and Probability fplot for residuals  
hist(AAPL\_model$residuals,xlab = "bins",ylab = "residuals", main = "Distribution of residuals")

Figure 2.1 – Distribution of residuals for Apple



qqPlot(x=rnorm(length(AAPL\_model$residuals)),y=AAPL\_model$residuals,add.line = TRUE,xlab = "Standard Normal Values",ylab = "Residuals", main = "Probability Plot")

Figure 2.2 – Probability plot of residuals for Apple

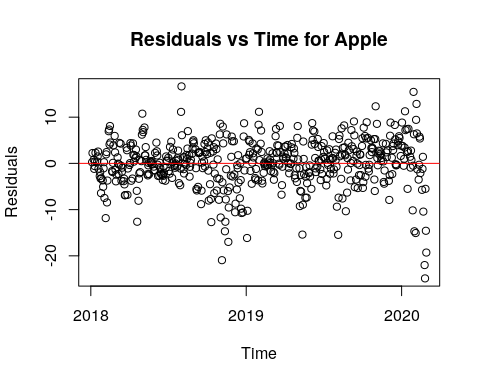


## Chi Square Test  
AAPL\_test <- chisq.test(x = AAPL\_model$residuals, y = rnorm(length(AAPL\_model$residuals)))

## X-squared = 292140, df = 291600, p-value = 0.2396

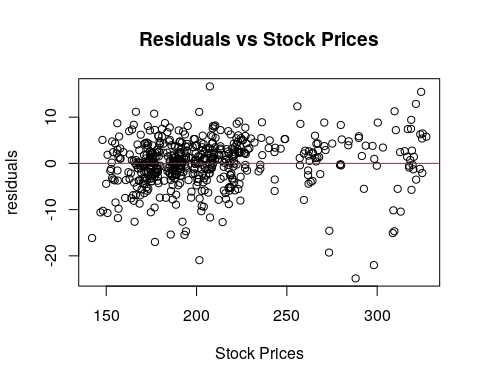
## Residuals vs Time  
plot(x = as.Date(AAPL$Date[3:543]),y = AAPL\_model$residuals,main = "Residuals vs Time for Apple",xlab = "Time",ylab = "Residuals")  
abline(0,0,col="red")

Figure 2.3 –Residuals against time for Apple



## Residuals vs Stock Price  
plot(x=AAPL$Close[3:543],y = AAPL\_model$residuals,xlab = "Stock Prices",ylab = "residuals",main = "Residuals vs Stock Prices")  
abline(0,0,col = "red")

Figure 2.4 – Homoscedasticity of residuals for Apple



############## Microsoft #############3

summary(MSFT\_model)

## Multiple R-squared: 0.9885, Adjusted R-squared: 0.9885   
## F-statistic: 4.636e+04 on 1 and 539 DF, p-value: < 2.2e-16

## Coefficient of Coreelation  
cor\_MSFT <-cor(MSFT\_lag, use = "pairwise.complete.obs")  
print(round(cor\_MSFT,2))

## lag0 lag2  
## lag0 1.00 0.99  
## lag2 0.99 1.00

## Actual Prediction and MSE  
pred\_MSFT\_3 <- predict(MSFT\_model,newdata = MSFT\_test\_lag)  
pred\_MSFT\_3[3:5]

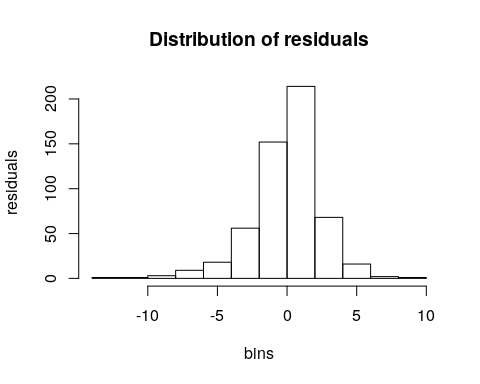
## 3 4 5   
## 172.8463 164.6002 170.6155

mse\_3 <- MSE(pred\_MSFT\_3[3:5],test\_MSFT[3:5])  
mse\_3

## [1] 29.96084

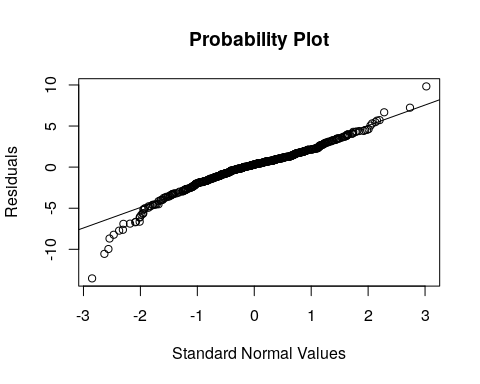
## Histogram and Probability fplot for residuals  
hist(MSFT\_model$residuals, xlab = "bins",ylab = "residuals", main = "Distribution of residuals")

Figure 3.1 – Distribution of residuals for Microsoft



qqPlot(x=rnorm(length(MSFT\_model$residuals)),y=MSFT\_model$residuals,add.line = TRUE,xlab = "Standard Normal Values",ylab = "Residuals", main = "Probability Plot")

Figure 3.2 – Probability plot of residuals for Microsoft

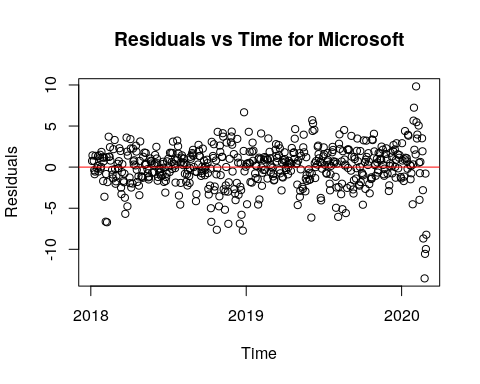


## Chi Square Test  
MSFT\_test <- chisq.test(x = MSFT\_model$residuals, y = rnorm(length(MSFT\_model$residuals)))

## data: MSFT\_model$residuals and rnorm(length(MSFT\_model$residuals))  
## X-squared = 292140, df = 291600, p-value = 0.2396

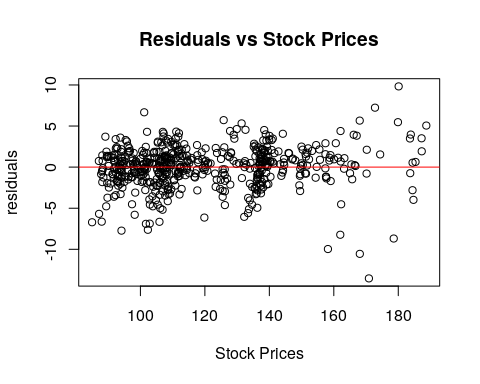
## Residuals vs Time  
plot(x = as.Date(MSFT$Date[3:543]),y = MSFT\_model$residuals,main = "Residuals vs Time for Microsoft",xlab = "Time",ylab = "Residuals")  
abline(0,0,col="red")

Figure3.3 –Residuals against time for Microsoft



## Residuals vs Stock Price  
plot(x=MSFT$Close[3:543],y = MSFT\_model$residuals,xlab = "Stock Prices",ylab = "residuals",main = "Residuals vs Stock Prices")  
abline(0,0,col = "red")

Figure 3.4 – Homoscedasticity of residuals for Microsoft



############## Disney #############3  
  
## Coefficient of Determination  
summary(DIS\_model)

## Multiple R-squared: 0.976, Adjusted R-squared: 0.9759   
## F-statistic: 2.19e+04 on 1 and 539 DF, p-value: < 2.2e-16

## Coefficient of Coreelation  
cor\_DIS <-cor(DIS\_lag, use = "pairwise.complete.obs")  
print(round(cor\_DIS,2))

## lag0 lag2  
## lag0 1.00 0.99  
## lag2 0.99 1.00

## Actual Prediction and MSE  
pred\_DIS\_3 <- predict(DIS\_model,newdata = DIS\_test\_lag)  
pred\_DIS\_3[3:5]

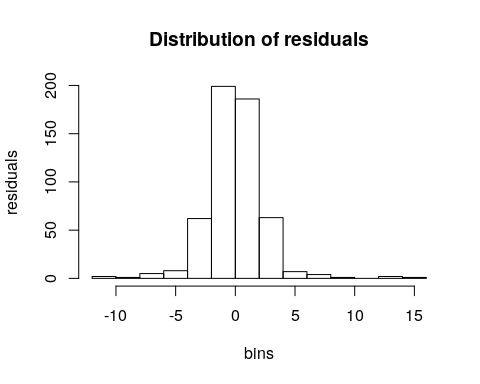
## 3 4 5   
## 120.0157 116.5303 119.2258

mse\_3 <- MSE(pred\_DIS\_3[3:5],test\_DIS[3:5])  
mse\_3

## [1] 7.616939

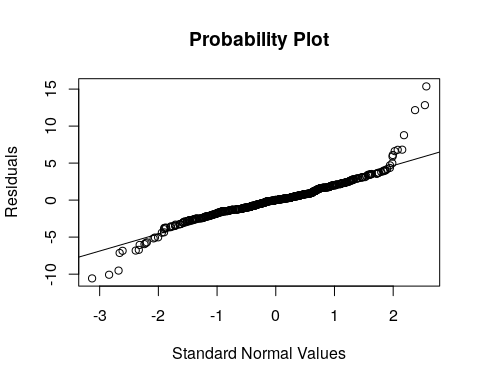
## Histogram and Probability fplot for residuals  
hist(DIS\_model$residuals, xlab = "bins",ylab = "residuals", main = "Distribution of residuals")

Figure 4.1 – Distribution of residuals for Disney



qqPlot(x=rnorm(length(DIS\_model$residuals)),y=DIS\_model$residuals,add.line = TRUE,xlab = "Standard Normal Values",ylab = "Residuals", main = "Probability Plot")

Figure 4.2 – Probability plot of residuals for Disney

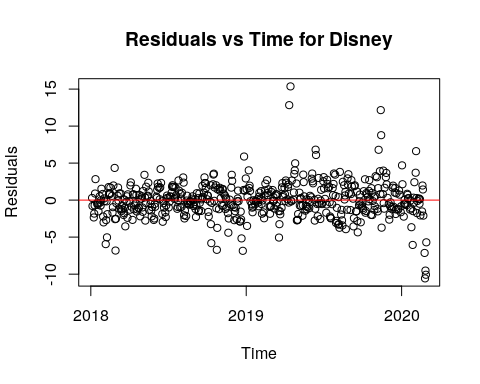


## Chi Square Test  
DIS\_test <- chisq.test(x = DIS\_model$residuals, y = rnorm(length(DIS\_model$residuals)))

## X-squared = 292140, df = 291600, p-value = 0.2396

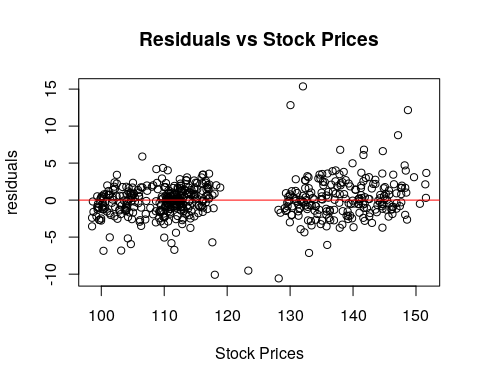
## Residuals vs Time  
plot(x = as.Date(DIS$Date[3:543]),y = DIS\_model$residuals,main = "Residuals vs Time for Disney",xlab = "Time",ylab = "Residuals")  
abline(0,0,col="red")

Figure 4.3 –Residuals against time for Pepsi



## Residuals vs Stock Price  
plot(x=DIS$Close[3:543],y = DIS\_model$residuals,xlab = "Stock Prices",ylab = "residuals",main = "Residuals vs Stock Prices")  
abline(0,0,col = "red")

Figure 4.4 – Homoscedasticity of residuals for Disney



############## Home Depot #############3  
  
  
## Coefficient of Determination  
summary(HD\_model)

## Multiple R-squared: 0.9637, Adjusted R-squared: 0.9637   
## F-statistic: 1.432e+04 on 1 and 539 DF, p-value: < 2.2e-16

## Coefficient of Coreelation  
cor\_HD <-cor(HD\_lag, use = "pairwise.complete.obs")  
print(round(cor\_HD,2))

## lag0 lag2  
## lag0 1.00 0.98  
## lag2 0.98 1.00

## Actual Prediction and MSE  
pred\_HD\_3 <- predict(HD\_model,newdata = HD\_test\_lag)  
pred\_HD\_3[3:5]

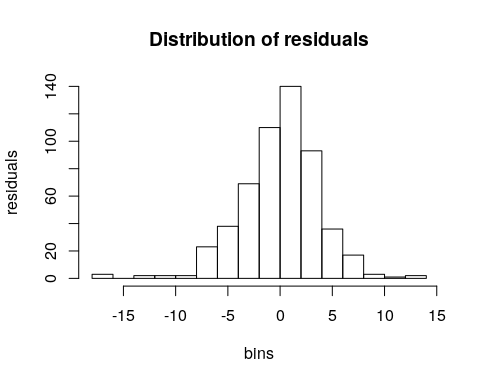
## 3 4 5   
## 229.5633 227.5971 240.5148

mse\_3 <- MSE(pred\_HD\_3[3:5],test\_HD[3:5])  
mse\_3

## [1] 109.5922

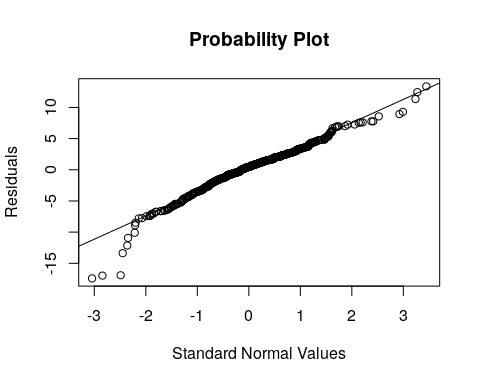
## Histogram and Probability fplot for residuals  
hist(HD\_model$residuals, xlab = "bins",ylab = "residuals", main = "Distribution of residuals")

Figure 5.1 – Distribution of residuals for Home Depot



qqPlot(x=rnorm(length(HD\_model$residuals)),y=HD\_model$residuals,add.line = TRUE,xlab = "Standard Normal Values",ylab = "Residuals", main = "Probability Plot")

Figure 5.2 – Probability plot of residuals for Home Depot

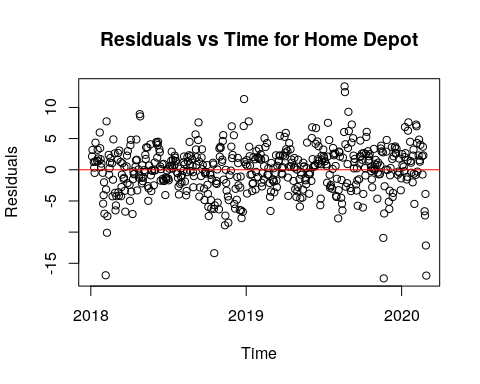


## Chi Square Test  
HD\_test <- chisq.test(x = HD\_model$residuals, y = rnorm(length(HD\_model$residuals)))

## X-squared = 292140, df = 291600, p-value = 0.2396

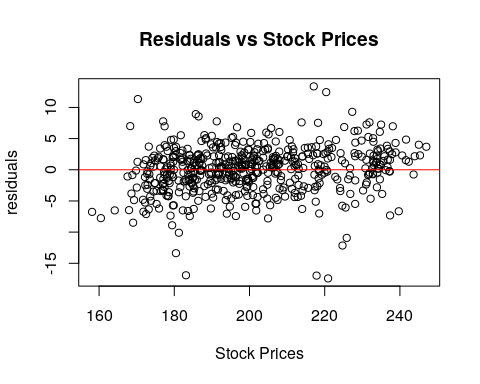
## Residuals vs Time  
plot(x = as.Date(HD$Date[3:543]),y = HD\_model$residuals,main = "Residuals vs Time for Home Depot",xlab = "Time",ylab = "Residuals")  
abline(0,0,col="red")

Figure5.3 –Residuals against time for Home Depot



## Residuals vs Stock Price  
plot(x=HD$Close[3:543],y = HD\_model$residuals,xlab = "Stock Prices",ylab = "residuals",main = "Residuals vs Stock Prices")  
abline(0,0,col = "red")

Figure 5.4 – Homoscedasticity of residuals for Home Depot



# Problem 4

In the final part of my analysis, I tried to predict Pepsi stock prices using multiple linear regression model. In order to that, I used remaining stock prices a independent variables. Similar to problem 3, I checked the assumptions of regression analysis using charts and graphs (Figure 5.1, Figure 5.2, Figure 5.3, Figure 5.4 ). From figure 5.1 and 5.2 , we can see that residuals for our model are normally distributed. In order to prove this formally , I constructed a hypothesis test using Chi-Square test with 0.05 significance. My hypotheses are as follow :

1. H0 : There is no significant difference in the distribution of residuals and Normal distribution
2. There is a significant difference between distribution of residuals and Normal Distribution.

As a result of this test, my p-values was bigger than 0.05 (almost 0.53). So, I do not have enough evidence to reject my null hypotheses. Thus, I cannot say that residual values for my regression analyses are not normally distributed. On the other hand, I plotted residuals against time in order to see their independency (Figure 5.3). But, here, I observed that as time moves, residuals tends to get smaller and there is line pattern. But, since this pattern is not strong, I could not say that residuals are not independent. Moreover, from figure 5.4, we can see that homoscedasticity assumption of linear regression , stating that variance of residuals does not vary for different values, are violated. It is obvious that, as we are getting close to today, variance gets smaller.

After checking assumptions of linear regression, I analyzed the coefficient of correlation and coefficient of determination. First, R-squared value was 0.887. This means my model can explain the 89% of variance in Pepsi stock prices. On the other hand, there was a strong positive correlation between variable of my regression model (precise values in the output part).

It is indeed impressive but neither R-square value nor correlation coefficient do not say anything about our prediction power. So, in order to test my model, I predicted stock prices for ,2nd , 3rd ,4th,5th and 6th of March. I got MSE value of 13.18. Indeed, this is a good prediction. However, for Pepsi , using simple exponential smoothing is still the best choice.

################### Problem 4 ##################################  
  
########### Pepsi ###########  
  
names <- c("PEP","AAPL","MSFT","DIS","HD")  
  
together <- cbind(PEP$Close,AAPL$Close,MSFT$Close,DIS$Close,HD$Close)  
together <- data.frame(together)  
names(together) <- names  
  
together\_test = cbind(test\_AAPL,test\_DIS,test\_HD,test\_MSFT)  
together\_test <- data.frame(together\_test)  
names(together\_test) <- names[2:5]  
  
  
tog\_model <- lm(data=together, PEP ~ AAPL + MSFT + DIS + HD)  
  
## Coefficient of Determination  
summary(tog\_model)

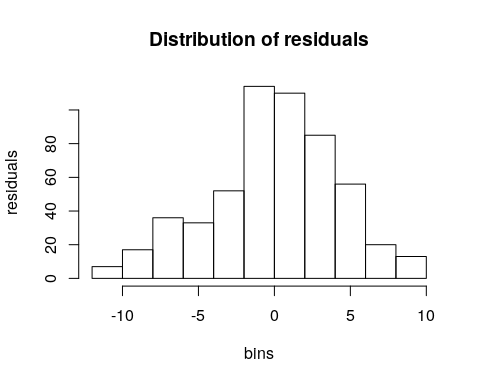
## Multiple R-squared: 0.887, Adjusted R-squared: 0.8861   
## F-statistic: 1055 on 4 and 538 DF, p-value: < 2.2e-16

# Coefficient of Correlation  
cor\_together <- cor(together,use= "pairwise.complete.obs")  
print(round(cor\_together,2))

## PEP AAPL MSFT DIS HD  
## PEP 1.00 0.70 0.90 0.90 0.81  
## AAPL 0.70 1.00 0.87 0.68 0.80  
## MSFT 0.90 0.87 1.00 0.89 0.83  
## DIS 0.90 0.68 0.89 1.00 0.75  
## HD 0.81 0.80 0.83 0.75 1.00

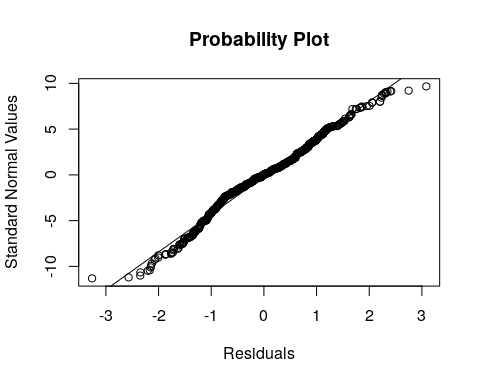
## Histogram and probability plot of resisuals  
hist(tog\_model$residuals, xlab = "bins",ylab = "residuals", main = "Distribution of residuals")

Figure 5.1 – Distribution of residuals for Pepsi



qqPlot(x=rnorm(length(tog\_model$residuals)), y = tog\_model$residuals, xlab = "Residuals",ylab = "Standard Normal Values", main = "Probability Plot",add.line = TRUE)

Figure 5.2 – Probability plot of residuals for Pepsi

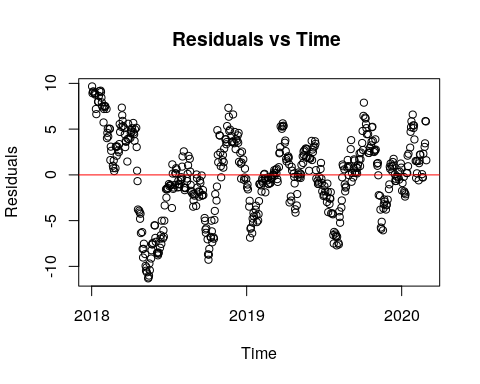


## Chi Square Test  
tog\_test <- chisq.test(x = tog\_model$residuals, y = rnorm(length(tog\_model$residuals)))

## X-squared = 294306, df = 293764, p-value = 0.5214

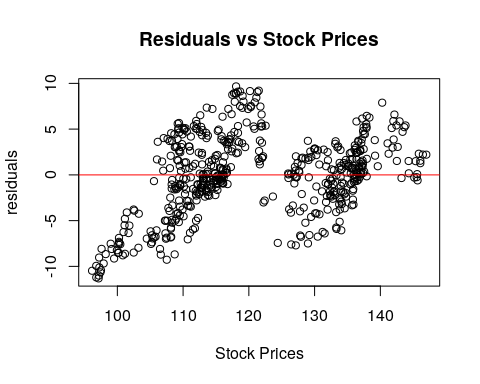
## Residuals vs Time  
plot(x = as.Date(PEP$Date),y = tog\_model$residuals,main = "Residuals vs Time ",xlab = "Time",ylab = "Residuals")  
abline(0,0,col="red")

Figure 5.3 – Residuals over time Pepsi



## Residuals vs Stock Price  
plot(x=together$PEP,y = tog\_model$residuals,xlab = "Stock Prices",ylab = "residuals",main = "Residuals vs Stock Prices")  
abline(0,0,col = "red")

Figure 5.4– Homoscedasticity of residuals for Pepsi



#### Actual Prediction  
pred\_linear <- predict(tog\_model, new = together\_test)  
test\_PEP

## [1] 137.58 135.58 142.39 138.10 137.26

mse\_linear <- MSE(pred\_linear,test\_PEP)  
print(mse\_linear)

## [1] 13.17703

# Conclusion

To conclude, in my analysis, I utilized real world stock prices and different forecasting techniques in order to predict future prices. For different stock, different type of analysis yielded the best results. For 3 of them, namely Pepsi, Microsoft and Home Depot, using simple exponential smoothing with alpha values of 0.35,035 and 0.15 was the optimal model. Having small alpha values make sense, since recent fluctuation in the prices is due to panic selling, thus not fundamentally backed. So, we should not put heavy emphasize on recent observations.

On the other hand, for Apple and Disney, I got most precise forecasting by using simple linear regression analysis against 2-days lagged stock price. Also, there was a strong positive correlation between variables , as expected.

# References :

Exponential Smoothing. (n.d.). Retrieved December 1, 2016, from <https://www.exploreanalytics.com/wiki/index.php?title=Exponential_Smoothing>

EXPONENTIAL SMOOTHING. (2001, June 5). Retrieved from <https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/exposmoo.htm>

Yahoo Finance - Stock Market Live, Quotes, Business & Finance News. (n.d.). Retrieved from <https://finance.yahoo.com/>