STAT 3333 Project

Group 1: Ramadharsh Vanchinathan, Jade Gee, Isabel Delacruz, Gabriela Ixcolin

12/07/2020

```r
library(tidyverse)
library(readxl)
library(dplyr)
library(testequavar)
library(coin)
library(ggplot2)
library(gridExtra)
library(lemon)
```

Introduction, Problem, Purpose

The data set consists of NASDAQ stock market data from December 7, 2015 to December 7, 2020 for key companies. We chose 4 prominent technology companies, Apple (APPL), Microsoft (MSFT), Alphabet (GOOGL), and Amazon (AMZN). They will be compared to industry leaders in other sectors, such as Pfizer (PFE) for health care and Goldman Sachs (GS) in the financial industry.

We would like to determine if there has been a discernible difference in the growth of the value in tech companies compared to companies in other industries. We will be using Permutation Testing, Regression, and Hypothesis Testing to determine any differences.

As Apple, Microsoft, Google, and Amazon have been vying for the title of "most valuable company" for years now, the purpose of this analysis is to determine the following:

> What makes these companies, and the tech industry, different from other companies and industries?

The Data

We are using data from the official NASDAQ website with each company having thousands of days of market data. Each day for each company has 6 variables: `Date`, `Close`, `Volume`, `Open`, `High` and `Low`. We will be using the `High` value because that will show us the change between daily peaks in the stock price. The change in the `High` values will be used to determine the growth rate of each company.

The use of statistics is crucial for determining any meaningful results, as we are trying to deduce the following:

- Are our findings statistically significant?
- Are the differences or similarities between the different companies and different industries significant?

The Hypothesis

The null hypothesis is that tech companies do not grow at a significantly different rate than other industries; while the research hypothesis is that tech companies grow significantly faster than other industries.

$$H_0 : \text{Tech Sector} = \text{Other Sectors}$$

$$H_A : \text{Tech Sector} > \text{Other Sectors}$$

This project is important because as electronics and their manufacturers become ever more pervasive in our societies, homes and lives, their wealth and influence will grow as well. This concept may also be applied to other industries and companies to determine whether it is keeping pace with the rest of the market, and inform financial decisions.

Result and Discussion

Importing and Cleaning the Data

First, we load in all of our data from their .xls files.

```r
# Load in and look at data
AAPL <- read_excel("AAPL.xls")
MSFT <- read_excel("MSFT.xls")
GOOGL <- read_excel("GOOGL.xls")
AMZN <- read_excel("AMZN.xls")
CVX <- read_excel("CVX.xls")
XOM <- read_excel("XOM.xls")
PFE <- read_excel("PFE.xls")
JNJ <- read_excel("JNJ.xls")
GS <- read_excel("GS.xls")
JPM <- read_excel("JPM.xls")
NOC <- read_excel("NOC.xls")
GE <- read_excel("GE.xls")
```

We load in the data, and extract the 'High' column from every companies' stock data as this is the statistic we are interested in. We then put all the data into dataframes based on the sector they belong to.

```r
# Load in and look at data
tech <- full_join(AAPL, MSFT, by="Date")
tech2 <- full_join(GOOGL, AMZN, by="Date")
tech_industries <- full_join(tech, tech2, by="Date") %>% select('Date',contains('High')) %>%
  rename('AAPL' = 'High.x.x','MSFT' = 'High.y.x', 'GOOGL' = 'High.x.y', 'AMZN' = 'High.y.y')

oil_gas <- full_join(CVX, XOM, by="Date")  %>% select('Date',contains('High')) %>%
  rename('CVX' = 'High.x','XOM' = 'High.y')

healthcare <- full_join(PFE, JNJ, by="Date") %>% select('Date',contains('High')) %>%
  rename('PFE' = 'High.x','JNJ' = 'High.y')

finance <- full_join(GS, JPM, by="Date") %>% select('Date',contains('High')) %>%
  rename('GS' = 'High.x','JPM' = 'High.y')

industrial <- full_join(NOC, GE, by="Date")%>% select('Date',contains('High')) %>%
  rename('NOC' = 'High.x','GE' = 'High.y')

head(tech_industries)
```

Date	AAPL	MSFT	GOOGL	AMZN
2015-12-07	29.965	56.1	781	679.99
2015-12-14	28.2	56.79	781.59	682.5
2015-12-21	27.25	55.96	771.9	669.9
2015-12-28	27.3575	56.85	798.69	696.44

Date	AAPL	MSFT	GOOGL	AMZN
2016-01-04	26.4625	55.39	769.2	657.715
2016-01-11	25.2975	54.07	753	625.99

```
head(oil_gas)
```

Date	CVX	XOM
2015-12-07	90.46	77.5
2015-12-14	94	79.61
2015-12-21	93.95	80.27
2015-12-28	92.58	80.08
2016-01-04	90.11	78.14
2016-01-11	86.17	79.92

```
head(healthcare)
```

Date	PFE	JNJ
2015-12-07	32.1393	103.49
2015-12-14	32.0806	105.49
2015-12-21	32.0024	103.91
2015-12-28	32.1686	104.34
2016-01-04	31.5186	101.81
2016-01-11	30.7903	99.47

```
head(industrial)
```

Date	NOC	GE
2015-12-07	188.92	30.8299
2015-12-14	191	31.129
2015-12-21	190.98	30.8997
2015-12-28	191.865	31.3881
2016-01-04	193.2	30.7402
2016-01-11	189.47	29.8232

The Rate of Change of Stock Values

We used the `change` function to determine the rate of change of stock prices from one datapoint to the next.

```
change <- function(table){
  newtab <- numeric(length(table))
  for(i in 1:length(table)-1){
    newtab[i] <- table[i+1] - table[i]
  }
  return(newtab)
}
```

Then we constructed dataframes of these changes values. We created one DF that contains the tech companies, and one that contains the other sectors.

```r
aapl <- change(as.numeric(tech_industries$AAPL))
aapl <- aapl[-length(aapl)]
msft <- change(as.numeric(tech_industries$MSFT))
msft <- msft[-length(msft)]
googl <- change(as.numeric(tech_industries$GOOGL))
googl <- googl[-length(googl)]
amzn <- change(as.numeric(tech_industries$AMZN))
amzn <- amzn[-length(amzn)]

cvx <- change(as.numeric(oil_gas$CVX))
cvx <- cvx[-length(cvx)]
xom <- change(as.numeric(oil_gas$XOM))
xom <- xom[-length(xom)]
Oil <- c(cvx,xom)

pfe <- change(as.numeric(healthcare$PFE))
pfe <- pfe[-length(pfe)]
jnj <- change(as.numeric(healthcare$JNJ))
jnj <- jnj[-length(jnj)]
Health <- c(pfe,jnj)

gs <- change(as.numeric(finance$GS))
gs <- gs[-length(gs)]
jpm <- change(as.numeric(finance$JPM))
jpm <- jpm[-length(jpm)]
Finance <- c(gs,jpm)

noc <- change(as.numeric(industrial$NOC))
noc <- noc[-length(noc)]
ge <- change(as.numeric(industrial$GE))
ge <- ge[-length(ge)]
Industry <- c(noc,ge)

tech_sector <- data.frame(aapl,msft,googl,amzn)

sectors <- data.frame(Oil,Health,Finance,Industry)

head(tech_sector)
```

aapl	msft	googl	amzn
-1.7650	0.69	0.59	2.510
-0.9500	-0.83	-9.69	-12.600
0.1075	0.89	26.79	26.540
-0.8950	-1.46	-29.49	-38.725
-1.1650	-1.32	-16.20	-31.725
0.0675	-1.74	-4.44	-25.890

```r
head(sectors)
```

Oil	Health	Finance	Industry
3.54	-0.0587	-2.23	2.080
-0.05	-0.0782	-4.72	-0.020

Oil	Health	Finance	Industry
-1.37	0.1662	1.01	0.885
-2.47	-0.6500	-6.68	1.335
-3.94	-0.7283	-9.32	-3.730
-0.98	-0.2443	-9.42	-2.190

Permutation Testing

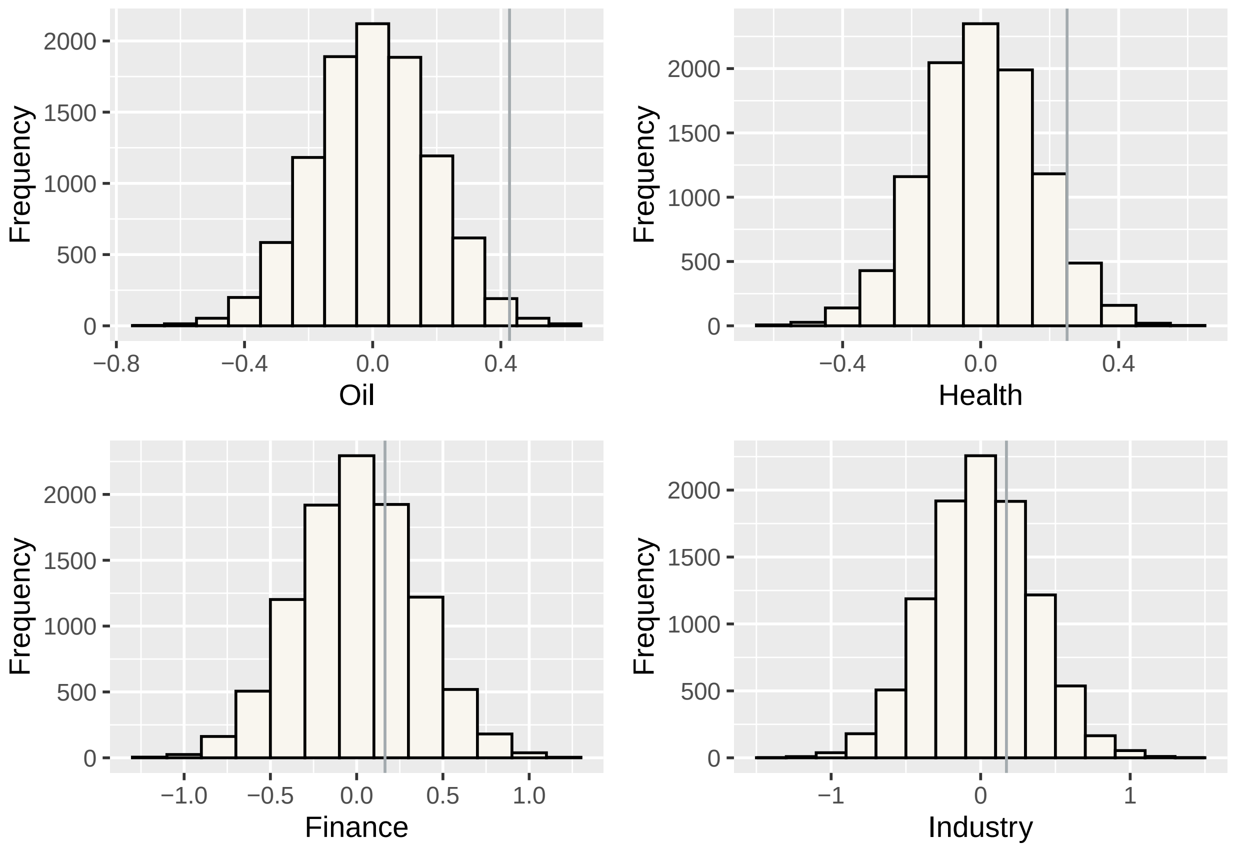
We did the permutation tests comparing each tech company to the other sectors. We did this to determine if each tech company had greater growth than each of the other sectors.

$$H_0 : \mu(\text{Other Sectors}) = \mu(\text{Tech Companies})$$

$$H_A : \mu(\text{Other Sectors}) < \mu(\text{Tech Companies})$$

```r
N <- 10^4-1

observed_oil <- mean(tech_sector$aapl) - mean(sectors$Oil)
observed_health <- mean(tech_sector$aapl) - mean(sectors$Health)
observed_finance <- mean(tech_sector$aapl) - mean(sectors$Finance)
observed_industry <- mean(tech_sector$aapl) - mean(sectors$Industry)



test_oil <- c(tech_sector$aapl, sectors$Oil)
o_len <- length(test_oil)

test_health <- c(tech_sector$aapl, sectors$Health)
h_len <- length(test_health)

test_finance <- c(tech_sector$aapl, sectors$Finance)
f_len <- length(test_finance)

test_industry <- c(tech_sector$aapl, sectors$Industry)
i_len <- length(test_industry)

a <- length(tech_sector$aapl)


perm_Oil <- numeric(N)
perm_Health <- numeric(N)
perm_Finance <- numeric(N)
perm_Industry <- numeric(N)


for(i in 1:N){
  o <- sample(o_len, size = a, replace = FALSE)
  perm_Oil[i] <- mean(test_oil[o]) - mean(test_oil[-o])

  h <- sample(h_len, size = a, replace = FALSE)
  perm_Health[i] <- mean(test_health[o]) - mean(test_health[-o])

  f <- sample(f_len, size = a, replace = FALSE)
  perm_Finance[i] <- mean(test_finance[o]) - mean(test_finance[-o])

  ind <- sample(i_len, size = a, replace = FALSE)
  perm_Industry[i] <- mean(test_industry[o]) - mean(test_industry[-o])
}

perm_aapl <- data.frame(perm_Oil,perm_Health, perm_Finance, perm_Industry)
```

```
require(gridExtra)
po <- ggplot(perm_aapl, aes(x = perm_Oil)) +
  geom_histogram(binwidth = .1, fill = '#F9F6EF', col = 'black') +
  ylab("Frequency") + xlab('Oil') +
  geom_vline(xintercept=observed_oil, col = '#A3AAAE')

ph <- ggplot(perm_aapl, aes(x = perm_Health)) +
  geom_histogram(binwidth = .1, fill = '#F9F6EF', col = 'black') +
  ylab("Frequency") + xlab('Health') +
  geom_vline(xintercept=observed_health, col = '#A3AAAE')

pf <- ggplot(perm_aapl, aes(x = perm_Finance)) +
  geom_histogram(binwidth = .2, fill = '#F9F6EF', col = 'black') +
  ylab("Frequency") + xlab('Finance') +
  geom_vline(xintercept=observed_finance, col = '#A3AAAE')

pi <- ggplot(perm_aapl, aes(x = perm_Industry)) +
  geom_histogram(binwidth = .2, fill = '#F9F6EF', col = 'black') +
  ylab("Frequency") + xlab('Industry') +
  geom_vline(xintercept=observed_industry, col = '#A3AAAE')
grid.arrange(po,ph,pf,pi,ncol = 2, nrow = 2)
```



```
# Oil
(sum(perm_Oil >= observed_oil) + 1)/(N+1)

#Health
(sum(perm_Health >= observed_health) + 1)/(N+1)

#Finance
(sum(perm_Finance >= observed_finance) + 1)/(N+1)

#Industry
(sum(perm_Industry >= observed_industry) + 1)/(N+1)
```

```
## [1] 0.01
## [1] 0.0667
## [1] 0.3195
## [1] 0.3153
```

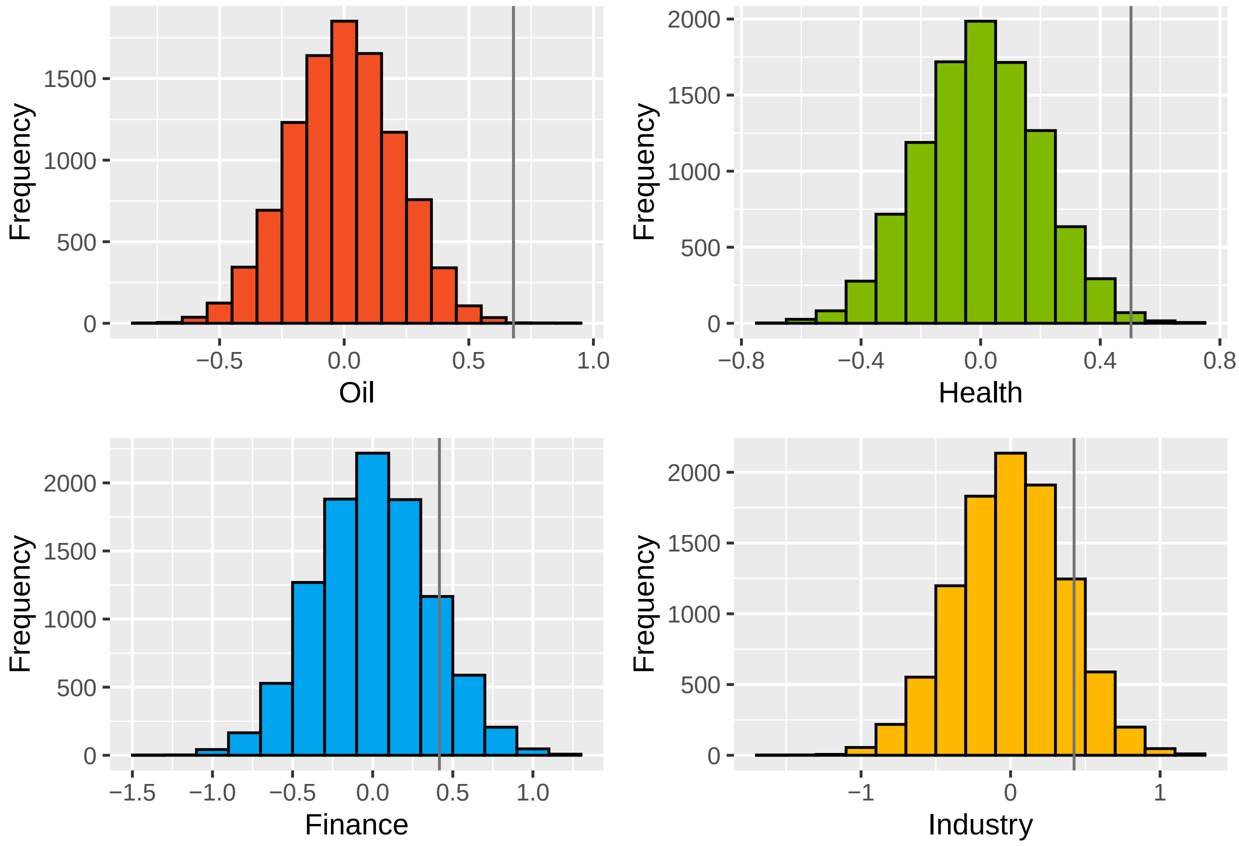
Based on the p-values above, we can say with a 95% significance, AAPL grew faster than only the oil sector over the five year period.

```
N <- 10^4-1

observed_oil <- mean(tech_sector$msft) - mean(sectors$Oil)
observed_health <- mean(tech_sector$msft) - mean(sectors$Health)
observed_finance <- mean(tech_sector$msft) - mean(sectors$Finance)
observed_industry <- mean(tech_sector$msft) - mean(sectors$Industry)
```

```r
test_oil <- c(tech_sector$msft, sectors$Oil)
o_len <- length(test_oil)

test_health <- c(tech_sector$msft, sectors$Health)
h_len <- length(test_health)

test_finance <- c(tech_sector$msft, sectors$Finance)
f_len <- length(test_finance)

test_industry <- c(tech_sector$msft, sectors$Industry)
i_len <- length(test_industry)

a <- length(tech_sector$msft)


perm_Oil <- numeric(N)
perm_Health <- numeric(N)
perm_Finance <- numeric(N)
perm_Industry <- numeric(N)


for(i in 1:N){
  o <- sample(o_len, size = a, replace = FALSE)
  perm_Oil[i] <- mean(test_oil[o]) - mean(test_oil[-o])

  h <- sample(h_len, size = a, replace = FALSE)
  perm_Health[i] <- mean(test_health[o]) - mean(test_health[-o])

  f <- sample(f_len, size = a, replace = FALSE)
  perm_Finance[i] <- mean(test_finance[o]) - mean(test_finance[-o])

  ind <- sample(i_len, size = a, replace = FALSE)
  perm_Industry[i] <- mean(test_industry[o]) - mean(test_industry[-o])
}

perm_msft <- data.frame(perm_Oil,perm_Health, perm_Finance, perm_Industry)
```

```r
require(gridExtra)
po <- ggplot(perm_msft, aes(x = perm_Oil)) +
  geom_histogram(binwidth = .1, fill = '#F25022', col = 'black') +
  ylab("Frequency") + xlab('Oil') +
  geom_vline(xintercept=observed_oil, col = '#737373')

ph <- ggplot(perm_msft, aes(x = perm_Health)) +
  geom_histogram(binwidth = .1, fill = '#7FBA00', col = 'black') +
  ylab("Frequency") + xlab('Health') +
  geom_vline(xintercept=observed_health, col = '#737373')

pf <- ggplot(perm_msft, aes(x = perm_Finance)) +
  geom_histogram(binwidth = .2, fill = '#00A4EF', col = 'black') +
  ylab("Frequency") + xlab('Finance') +
  geom_vline(xintercept=observed_finance, col = '#737373')
```

```r
pi <- ggplot(perm_msft, aes(x = perm_Industry)) +
  geom_histogram(binwidth = .2, fill = '#FFB900', col = 'black') +
  ylab("Frequency") + xlab('Industry') +
  geom_vline(xintercept=observed_industry, col = '#737373')
grid.arrange(po,ph,pf,pi,ncol = 2, nrow = 2)
```



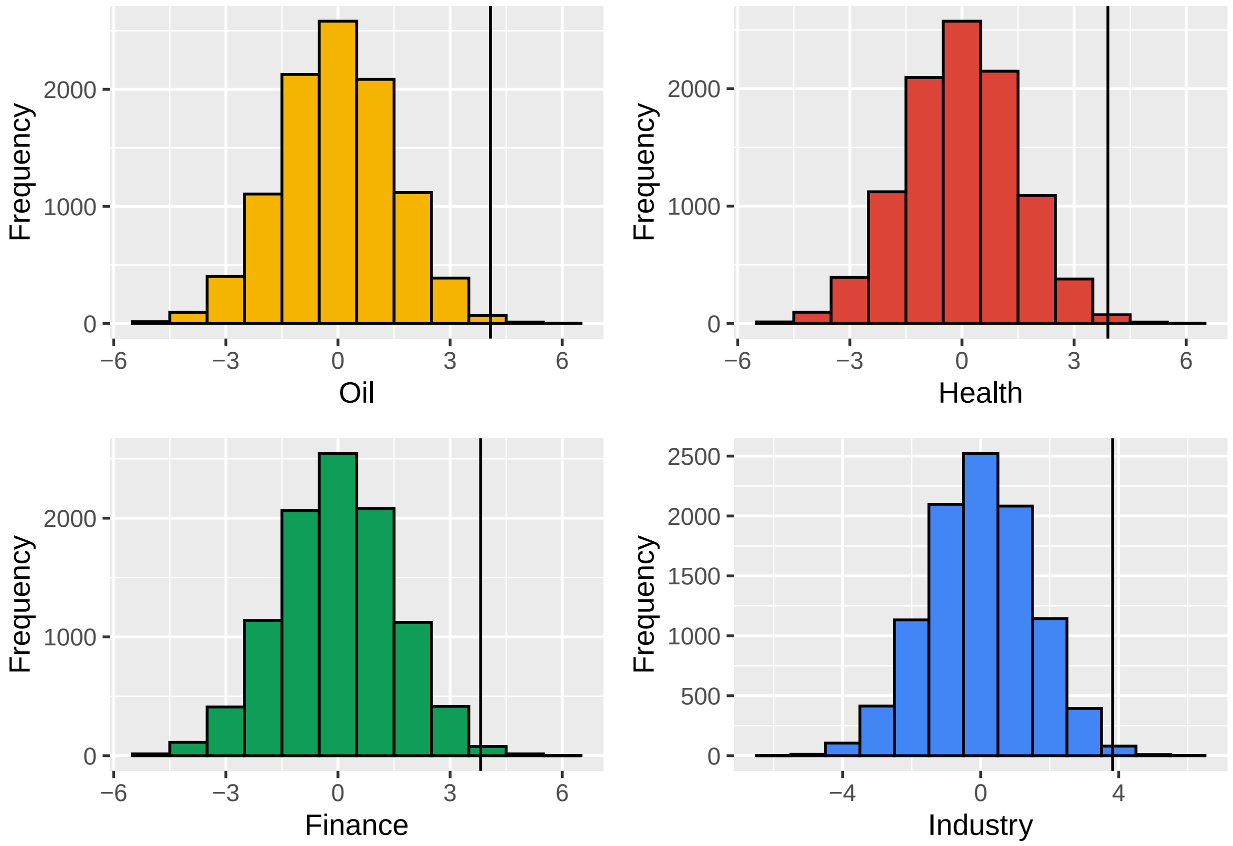
```r
# Oil
(sum(perm_Oil >= observed_oil) + 1)/(N+1)

#Health
(sum(perm_Health >= observed_health) + 1)/(N+1)

#Finance
(sum(perm_Finance >= observed_finance) + 1)/(N+1)

#Industry
(sum(perm_Industry >= observed_industry) + 1)/(N+1)
```

```
## [1] 5e-04
## [1] 0.0046
## [1] 0.126
## [1] 0.1245
```

Based on the p-values above, we can say with a 95% significance, MSFT grew faster than the oil and health care sectors over the five year period.

```r
#
N <- 10^4-1

observed_oil <- mean(tech_sector$googl) - mean(sectors$Oil)
observed_health <- mean(tech_sector$googl) - mean(sectors$Health)
observed_finance <- mean(tech_sector$googl) - mean(sectors$Finance)
observed_industry <- mean(tech_sector$googl) - mean(sectors$Industry)



test_oil <- c(tech_sector$googl, sectors$Oil)
o_len <- length(test_oil)

test_health <- c(tech_sector$googl, sectors$Health)
h_len <- length(test_health)

test_finance <- c(tech_sector$googl, sectors$Finance)
f_len <- length(test_finance)

test_industry <- c(tech_sector$googl, sectors$Industry)
i_len <- length(test_industry)

a <- length(tech_sector$googl)


perm_Oil <- numeric(N)
perm_Health <- numeric(N)
perm_Finance <- numeric(N)
perm_Industry <- numeric(N)


for(i in 1:N){
  o <- sample(o_len, size = a, replace = FALSE)
  perm_Oil[i] <- mean(test_oil[o]) - mean(test_oil[-o])

  h <- sample(h_len, size = a, replace = FALSE)
  perm_Health[i] <- mean(test_health[o]) - mean(test_health[-o])

  f <- sample(f_len, size = a, replace = FALSE)
  perm_Finance[i] <- mean(test_finance[o]) - mean(test_finance[-o])

  ind <- sample(i_len, size = a, replace = FALSE)
  perm_Industry[i] <- mean(test_industry[o]) - mean(test_industry[-o])
}

perm_googl <- data.frame(perm_Oil,perm_Health, perm_Finance, perm_Industry)
```

```r
require(gridExtra)
po <- ggplot(perm_googl, aes(x = perm_Oil)) +
  geom_histogram(binwidth = 1, fill = '#F4B400', col = 'black') +
  ylab("Frequency") + xlab('Oil') +
  geom_vline(xintercept=observed_oil)

ph <- ggplot(perm_googl, aes(x = perm_Health)) +
  geom_histogram(binwidth = 1, fill = '#DB4437', col = 'black') +
  ylab("Frequency") + xlab('Health') +
  geom_vline(xintercept=observed_health)

pf <- ggplot(perm_googl, aes(x = perm_Finance)) +
  geom_histogram(binwidth = 1, fill = '#0F9D58', col = 'black') +
  ylab("Frequency") + xlab('Finance') +
  geom_vline(xintercept=observed_finance)

pi <- ggplot(perm_googl, aes(x = perm_Industry)) +
  geom_histogram(binwidth = 1, fill = '#4285F4', col = 'black') +
  ylab("Frequency") + xlab('Industry') +
  geom_vline(xintercept=observed_industry)
grid.arrange(po,ph,pf,pi,ncol = 2, nrow = 2)
```



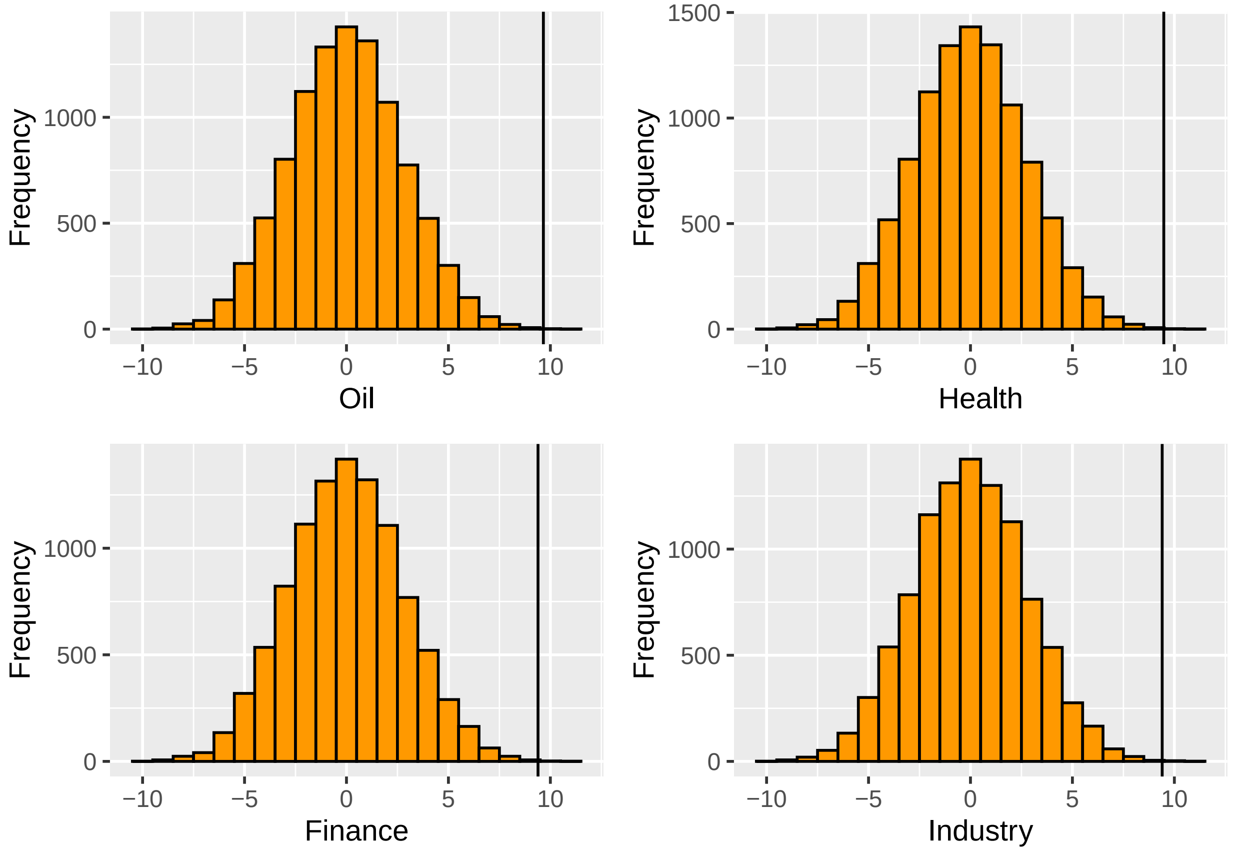
```
# Oil
(sum(perm_Oil >= observed_oil) + 1)/(N+1)

#Health
(sum(perm_Health >= observed_health) + 1)/(N+1)

#Finance
(sum(perm_Finance >= observed_finance) + 1)/(N+1)

#Industry
(sum(perm_Industry >= observed_industry) + 1)/(N+1)
```

```
## [1] 0.0027
## [1] 0.0035
## [1] 0.005
## [1] 0.0052
```

Based on the p-values above, we can say with a 95% significance, GOOGL grew faster than all the other sectors over the five year period.

```r
#
N <- 10^4-1

observed_oil <- mean(tech_sector$amzn) - mean(sectors$Oil)
observed_health <- mean(tech_sector$amzn) - mean(sectors$Health)
observed_finance <- mean(tech_sector$amzn) - mean(sectors$Finance)
observed_industry <- mean(tech_sector$amzn) - mean(sectors$Industry)



test_oil <- c(tech_sector$amzn, sectors$Oil)
o_len <- length(test_oil)

test_health <- c(tech_sector$amzn, sectors$Health)
h_len <- length(test_health)

test_finance <- c(tech_sector$amzn, sectors$Finance)
f_len <- length(test_finance)

test_industry <- c(tech_sector$amzn, sectors$Industry)
i_len <- length(test_industry)

a <- length(tech_sector$amzn)


perm_Oil <- numeric(N)
perm_Health <- numeric(N)
perm_Finance <- numeric(N)
perm_Industry <- numeric(N)


for(i in 1:N){
  o <- sample(o_len, size = a, replace = FALSE)
  perm_Oil[i] <- mean(test_oil[o]) - mean(test_oil[-o])

  h <- sample(h_len, size = a, replace = FALSE)
  perm_Health[i] <- mean(test_health[o]) - mean(test_health[-o])

  f <- sample(f_len, size = a, replace = FALSE)
  perm_Finance[i] <- mean(test_finance[o]) - mean(test_finance[-o])

  ind <- sample(i_len, size = a, replace = FALSE)
  perm_Industry[i] <- mean(test_industry[o]) - mean(test_industry[-o])
}

perm_amzn <- data.frame(perm_Oil,perm_Health, perm_Finance, perm_Industry)
```

```r
require(gridExtra)
po <- ggplot(perm_amzn, aes(x = perm_Oil)) +
  geom_histogram(binwidth = 1, fill = '#FF9900', col = '#000000') +
  ylab("Frequency") + xlab('Oil') +
  geom_vline(xintercept=observed_oil)

ph <- ggplot(perm_amzn, aes(x = perm_Health)) +
  geom_histogram(binwidth = 1, fill = '#FF9900', col = '#000000') +
  ylab("Frequency") + xlab('Health') +
  geom_vline(xintercept=observed_health)

pf <- ggplot(perm_amzn, aes(x = perm_Finance)) +
  geom_histogram(binwidth = 1, fill = '#FF9900', col = '#000000') +
  ylab("Frequency") + xlab('Finance') +
  geom_vline(xintercept=observed_finance)

pi <- ggplot(perm_amzn, aes(x = perm_Industry)) +
  geom_histogram(binwidth = 1, fill = '#FF9900', col = '#000000') +
  ylab("Frequency") + xlab('Industry') +
  geom_vline(xintercept=observed_industry)
grid.arrange(po,ph,pf,pi,ncol = 2, nrow = 2)
```



```r
# Oil
(sum(perm_Oil >= observed_oil) + 1)/(N+1)

#Health
```

```
(sum(perm_Health >= observed_health) + 1)/(N+1)

#Finance
(sum(perm_Finance >= observed_finance) + 1)/(N+1)

#Industry
(sum(perm_Industry >= observed_industry) + 1)/(N+1)
```

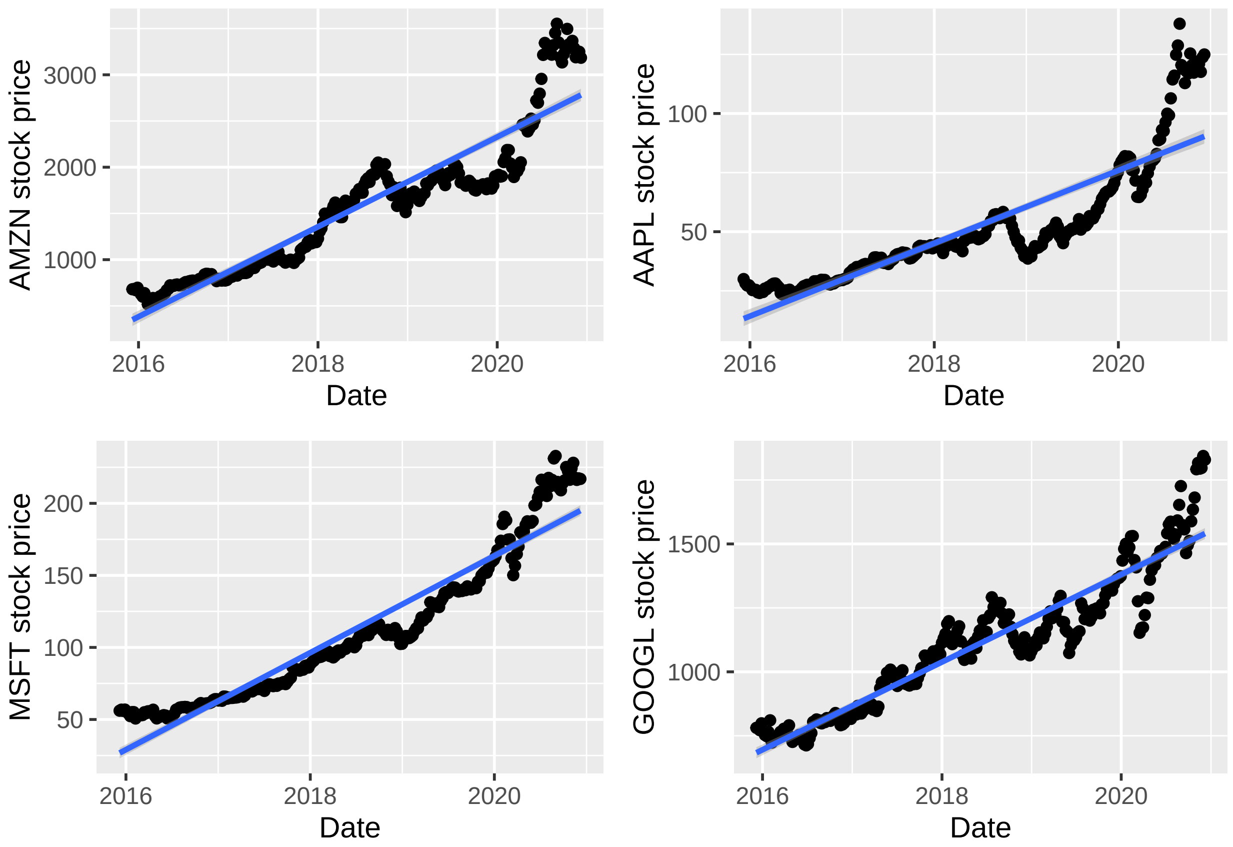
```
## [1] 4e-04
## [1] 4e-04
## [1] 4e-04
## [1] 5e-04
```

Based on the p-values above, we can say with a 95% significance, AMZN grew faster than all the other sectors over the five year period.

Regression

We are finding the linear regressions of each company and each sector to determine their rates of change. This way we will be able to corroborate the results from the permutation testing with the original data. Plotting the data from each tech company we are studying and plotting their linear regressions will help demonstrate this.

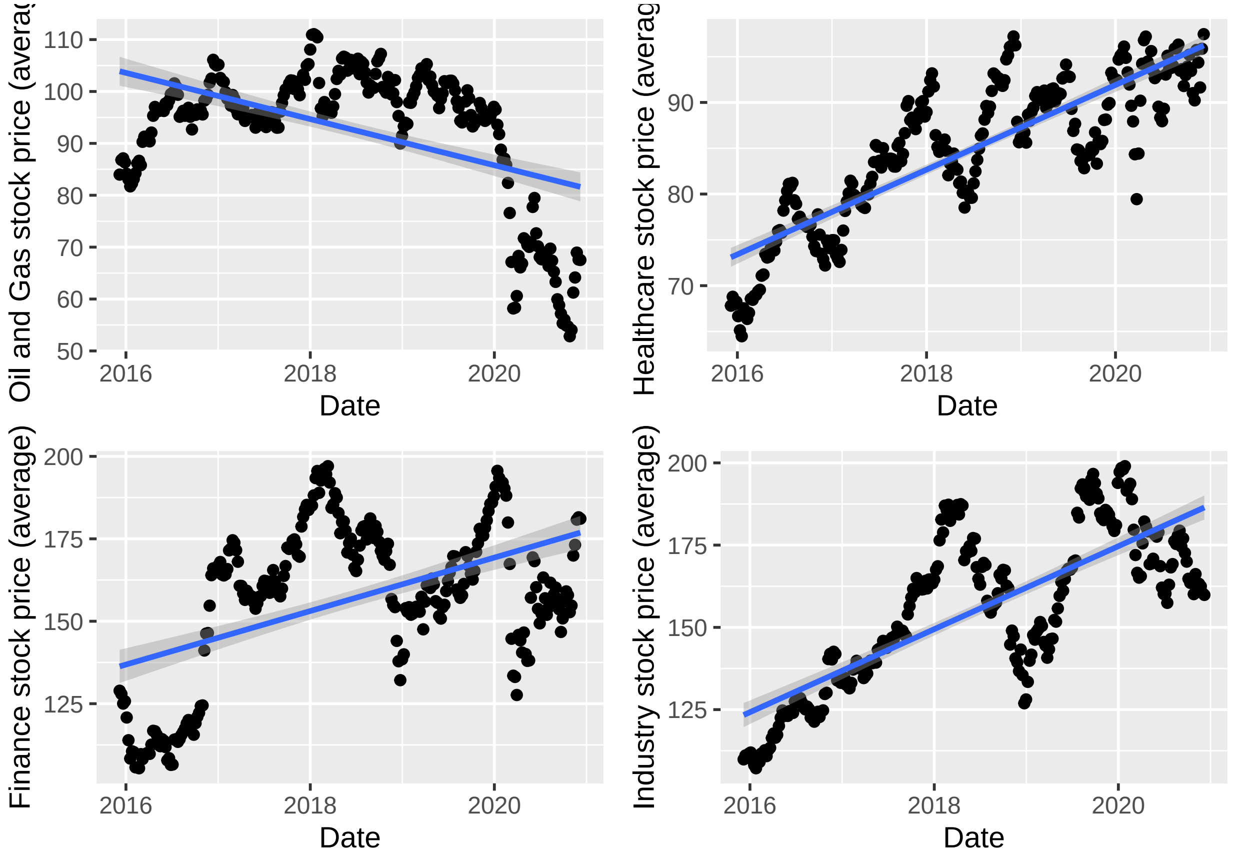
```
require(gridExtra)
amzn_plot <- ggplot(data = tech_industries, aes(x = as.Date(Date), y = as.numeric(AMZN) )) +
  geom_point() + xlab('Date') + ylab('AMZN stock price') + geom_smooth(method='lm', formula= y~x)

aapl_plot <- ggplot(data = tech_industries, aes(x = as.Date(Date), y = as.numeric(AAPL) )) +
  geom_point() + xlab('Date') + ylab('AAPL stock price') + geom_smooth(method='lm', formula= y~x)

msft_plot <- ggplot(data = tech_industries, aes(x = as.Date(Date), y = as.numeric(MSFT) )) +
  geom_point() + xlab('Date') + ylab('MSFT stock price') + geom_smooth(method='lm', formula= y~x)

googl_plot <- ggplot(data = tech_industries, aes(x = as.Date(Date), y = as.numeric(GOOGL) )) +
  geom_point() + xlab('Date') + ylab('GOOGL stock price') +  geom_smooth(method='lm', formula= y~x)

grid.arrange(amzn_plot,aapl_plot,msft_plot,googl_plot,ncol=2, nrow=2)
```



Then we do the same with the other sectors, averaging the values from both to provide a more comprehensive look at the sectors as a whole.

```
# Averaging the sectors
oil2 <- oil_gas %>% mutate(CVX,XOM,avg = ((as.numeric(CVX) + as.numeric(XOM))/2))

health2 <- healthcare %>% mutate(PFE,JNJ,avg = ((as.numeric(PFE) + as.numeric(JNJ))/2))

finance2 <- finance %>% mutate(GS,JPM,avg = ((as.numeric(GS) + as.numeric(JPM))/2))

ind2 <- industrial %>% mutate(NOC,GE,avg = ((as.numeric(NOC) + as.numeric(GE))/2))

# Graphing

require(gridExtra)

oil_plot <- ggplot(data = oil2, aes(x = as.Date(Date), y = as.numeric(avg) )) + geom_point() +
  xlab('Date') + ylab('Oil and Gas stock price (average)') + geom_smooth(method='lm', formula= y~x)

health_plot <- ggplot(data = health2, aes(x = as.Date(Date), y = as.numeric(avg) )) + geom_point() +
  xlab('Date') + ylab('Healthcare stock price (average)') + geom_smooth(method='lm', formula= y~x)

finance_plot <- ggplot(data = finance2, aes(x = as.Date(Date), y = as.numeric(avg) )) + geom_point() +
  xlab('Date') + ylab('Finance stock price (average)') + geom_smooth(method='lm', formula= y~x)

ind_plot <- ggplot(data = ind2, aes(x = as.Date(Date), y = as.numeric(avg) )) + geom_point() +
  xlab('Date') + ylab('Industry stock price (average)') + geom_smooth(method='lm', formula= y~x)
```

```
grid.arrange(oil_plot,health_plot,finance_plot,ind_plot,ncol=2, nrow=2)
```

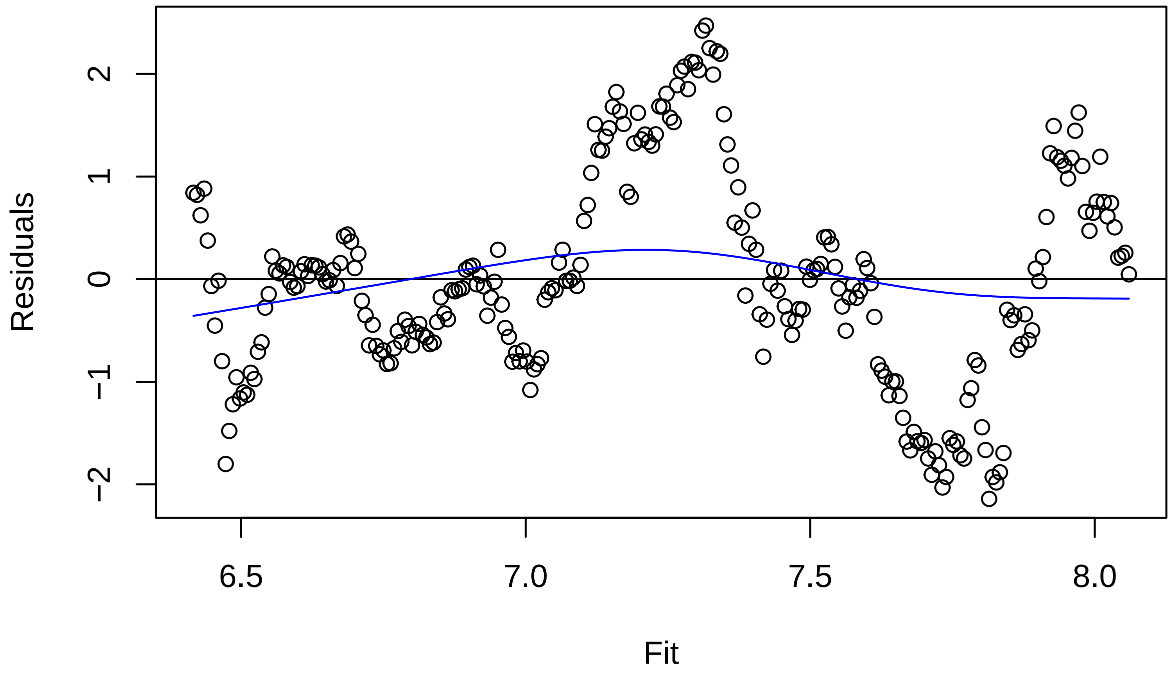


The next step is to find the regression of each tech company and each sector to determine their growth rate from the data itself. We took the log linear regression of High and Date to better show the linear relationship between the two variables.

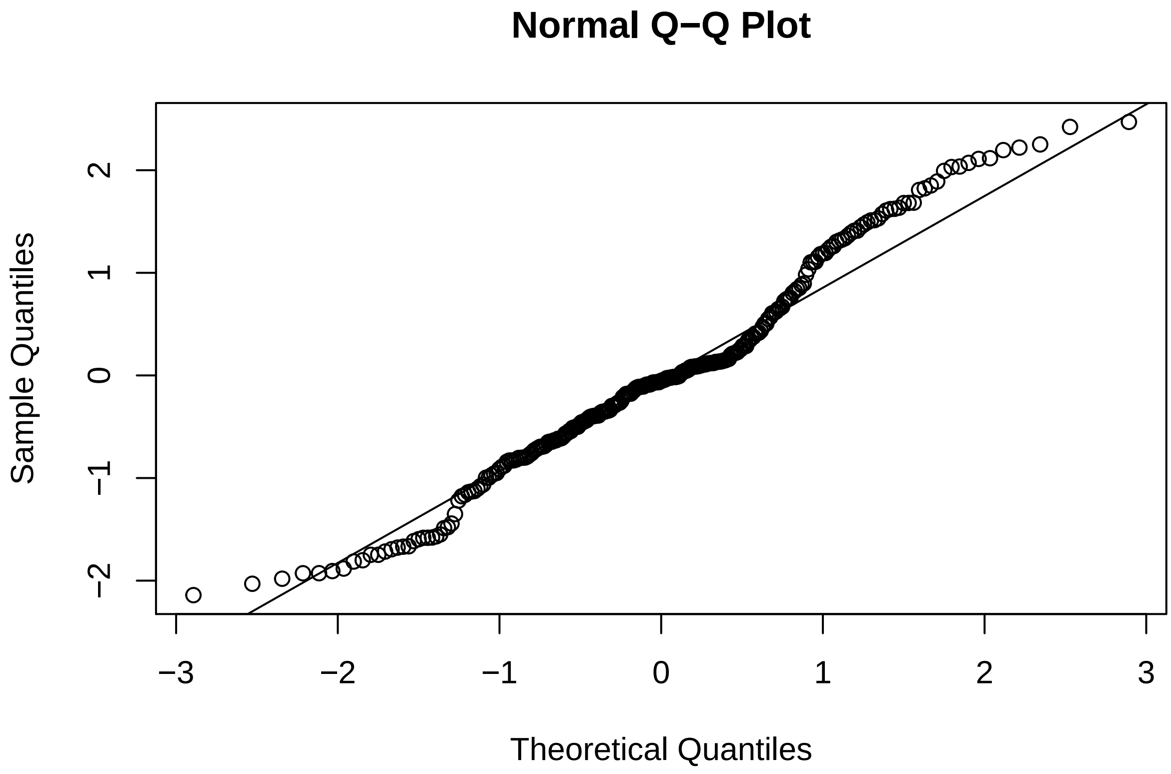
```
# Find the regression values Amazon

reg_AMZN <- AMZN %>% select(Date, High) %>%
  mutate(Date = as.Date(AMZN$Date), High = as.numeric(AMZN$High))

amzn.lm <- lm(log(High) ~ Date, data = reg_AMZN)
summary(amzn.lm)
```

```
##
## Call:
## lm(formula = log(High) ~ Date, data = reg_AMZN)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26780 -0.08106 -0.00670  0.07126  0.30898
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.677e+00  2.613e-01  -33.21   <2e-16 ***
## Date         8.997e-04  1.476e-05   60.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1265 on 260 degrees of freedom
## Multiple R-squared:  0.9346, Adjusted R-squared:  0.9343
## F-statistic:  3713 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(amzn.lm$fit, rstudent(amzn.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(amzn.lm$fit, rstudent(amzn.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(amzn.lm))
qqline(rstudent(amzn.lm))
```



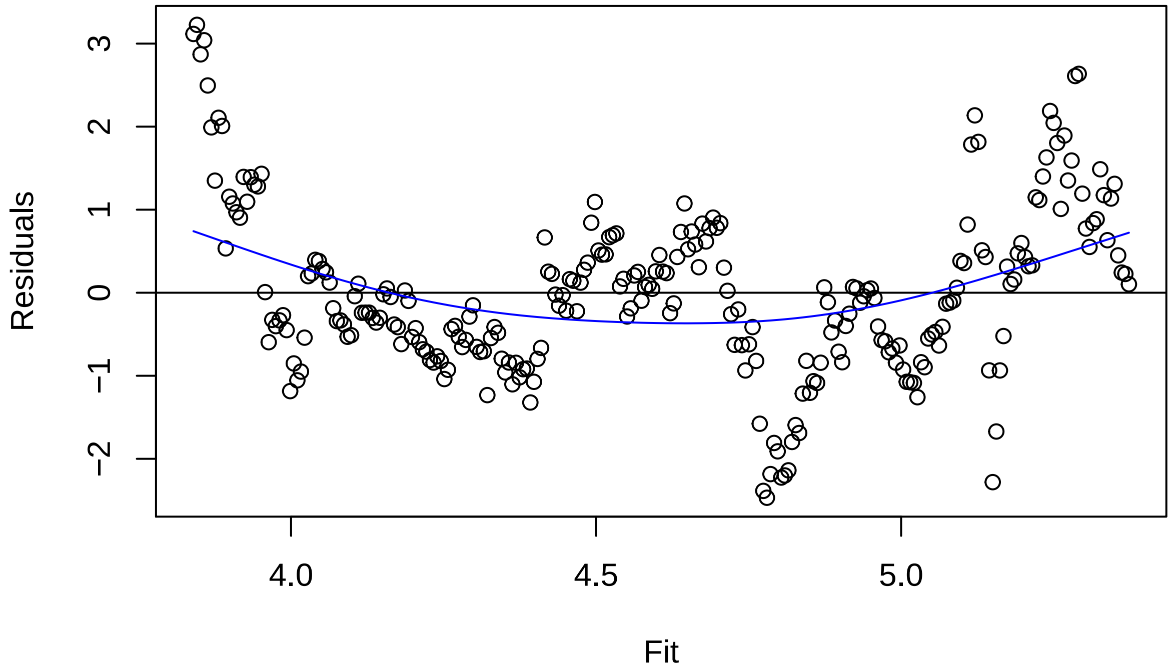
Based on the slope and intercept p-values acquired from the linear regression of AMZN, we can conclude that they are both significantly not zero. The linear regression has a slope of 8.997e-04 and an intercept of -8.677. The residual standard error is 0.1265 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

```
# Find the regression values Microsoft

reg_MSFT <- MSFT %>% select(Date, High) %>%
  mutate(Date = as.Date(MSFT$Date), High = as.numeric(MSFT$High))

msft.lm <- lm(log(High) ~ Date, data = reg_MSFT)
summary(msft.lm)
```

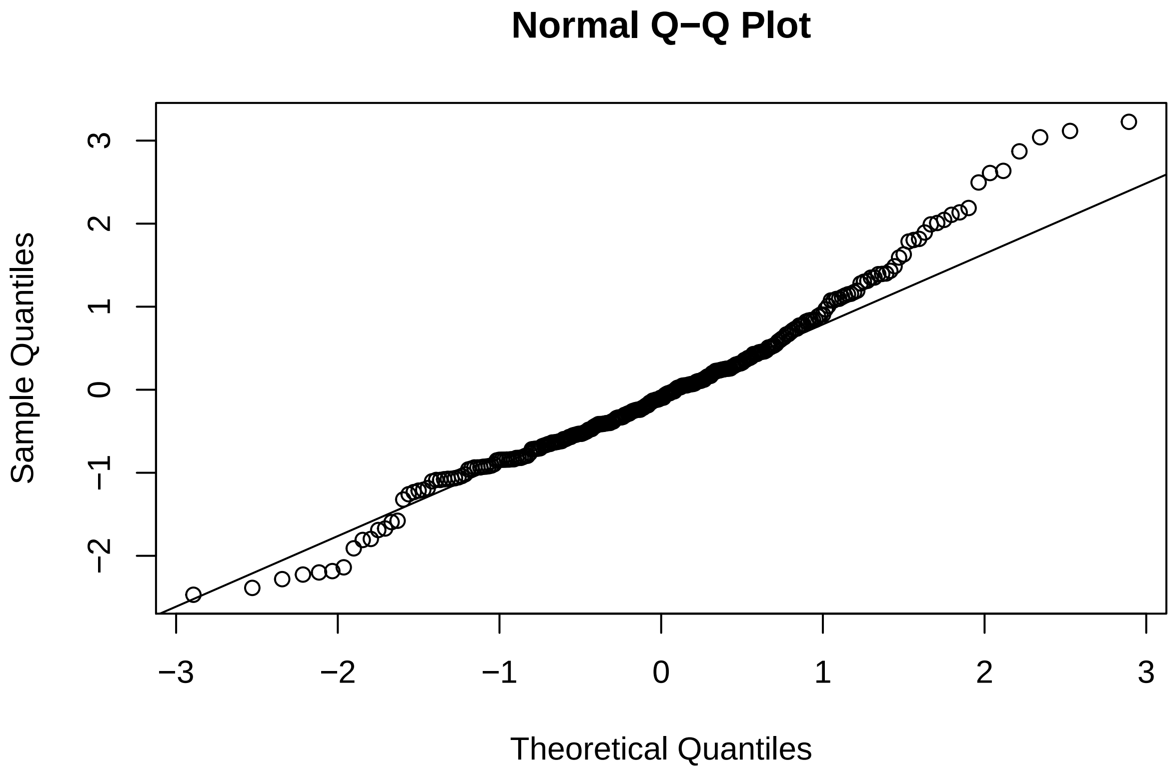
```
##
## Call:
## lm(formula = log(High) ~ Date, data = reg_MSFT)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.150065 -0.039001 -0.005983  0.031323  0.193486
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.024e+01  1.270e-01   -80.6   <2e-16 ***
## Date         8.393e-04  7.179e-06   116.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.06152 on 260 degrees of freedom
## Multiple R-squared:  0.9813, Adjusted R-squared:  0.9813
## F-statistic: 1.367e+04 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(msft.lm$fit, rstudent(msft.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(msft.lm$fit, rstudent(msft.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(msft.lm))
qqline(rstudent(msft.lm))
```



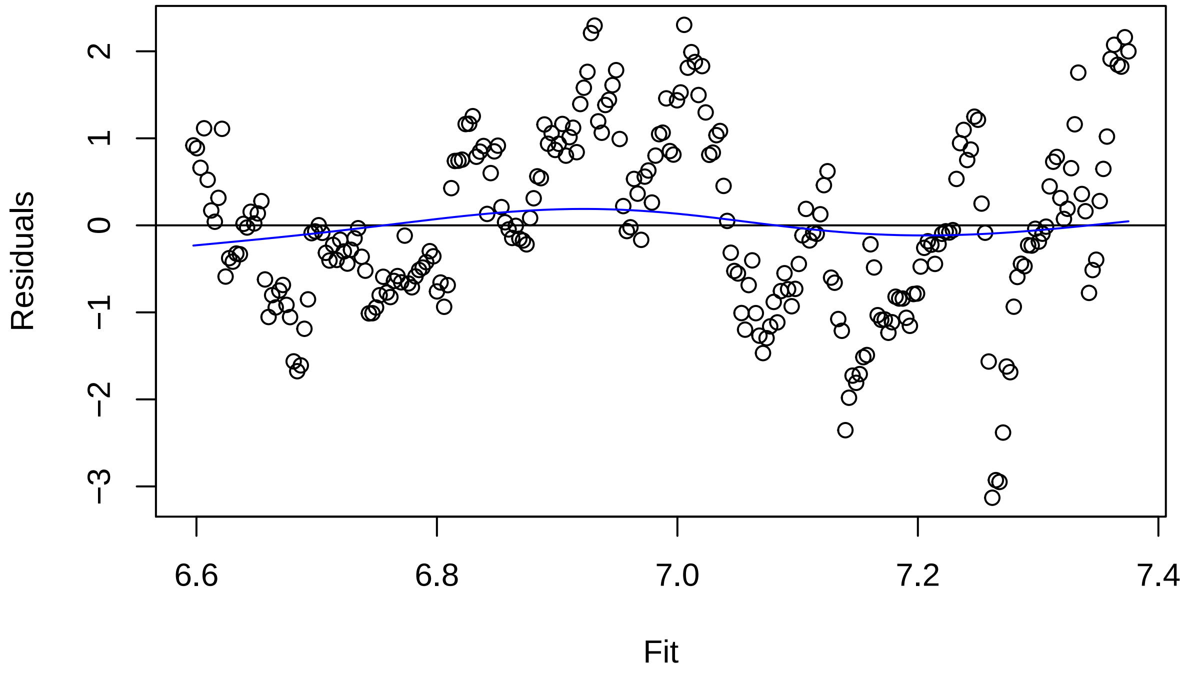
Based on the slope and intercept p-values acquired from the linear regression of MSFT, we can conclude that they are both significantly not zero. The linear regression has a slope of 8.393e-04 and an intercept of -1.024e+01. The residual standard error is 0.06152 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

```
# Find the regression values Google

reg_GOOGL <- GOOGL %>% select(Date, High) %>%
  mutate(Date = as.Date(GOOGL$Date), High = as.numeric(GOOGL$High))

GOOGL.lm <- lm(log(High) ~ Date, data = reg_GOOGL)
summary(GOOGL.lm)
```

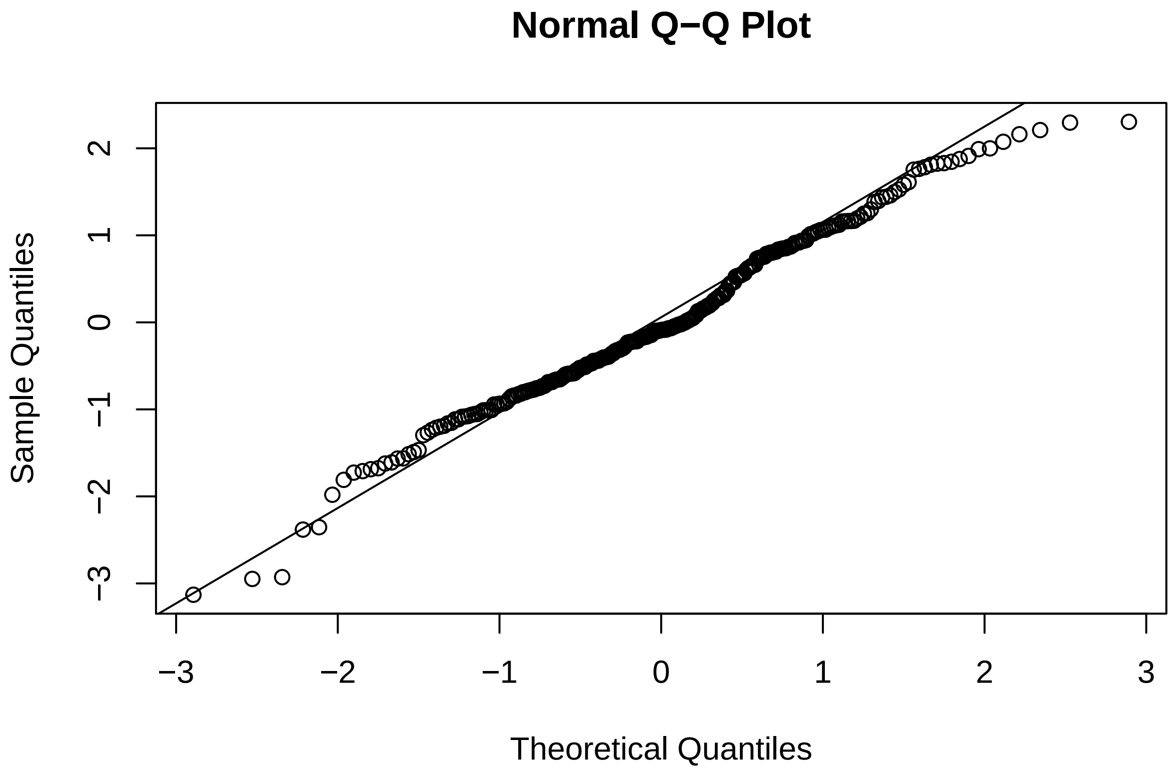
```
##
## Call:
## lm(formula = log(High) ~ Date, data = reg_GOOGL)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.212100 -0.046937 -0.006105  0.055107  0.157877
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.435e-01  1.430e-01  -3.802 0.000179 ***
## Date         4.257e-04  8.078e-06  52.694  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.06922 on 260 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.914
## F-statistic:  2777 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(GOOGL.lm$fit, rstudent(GOOGL.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(GOOGL.lm$fit, rstudent(GOOGL.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(GOOGL.lm))
qqline(rstudent(GOOGL.lm))
```



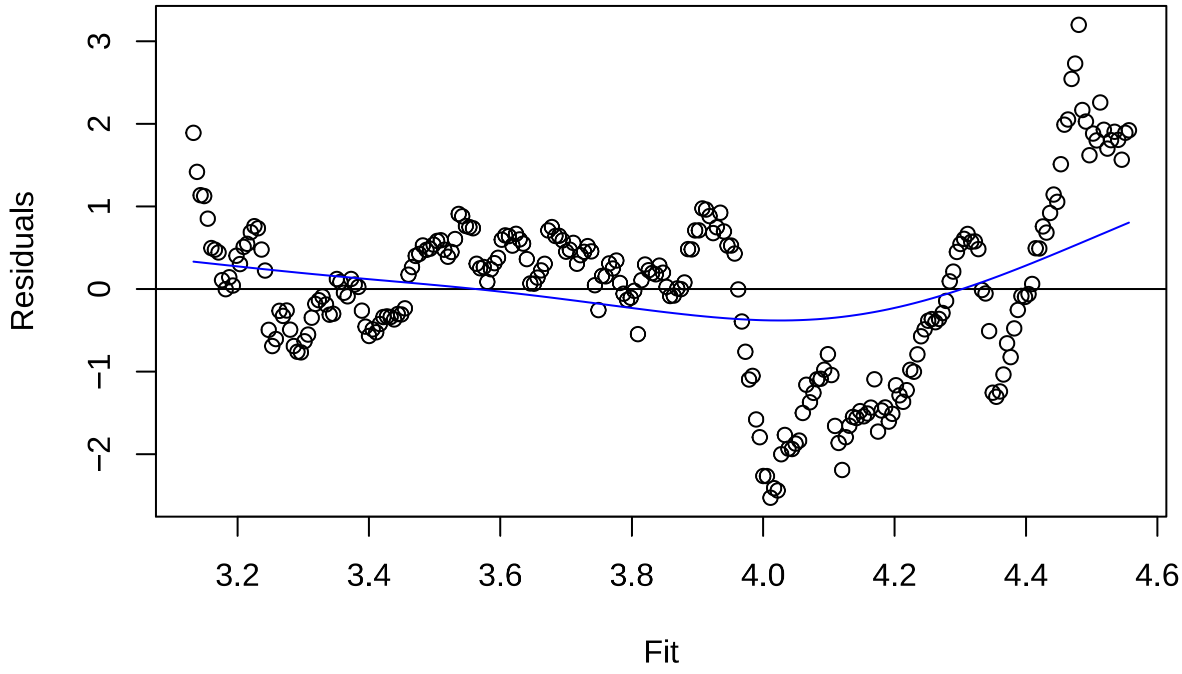
Based on the slope and intercept p-values acquired from the linear regression of GOOGL, we can conclude that they are both significantly not zero. The linear regression has a slope of 4.257e-04 and an intercept of -5.435e-01. The residual standard error is 0.06922 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

```
# Find the regression values Apple

reg_AAPL <- AAPL %>% select(Date, High) %>%
  mutate(Date = as.Date(AAPL$Date), High = as.numeric(AAPL$High))

AAPL.lm <- lm(log(High) ~ Date, data = reg_AAPL)
summary(AAPL.lm)
```

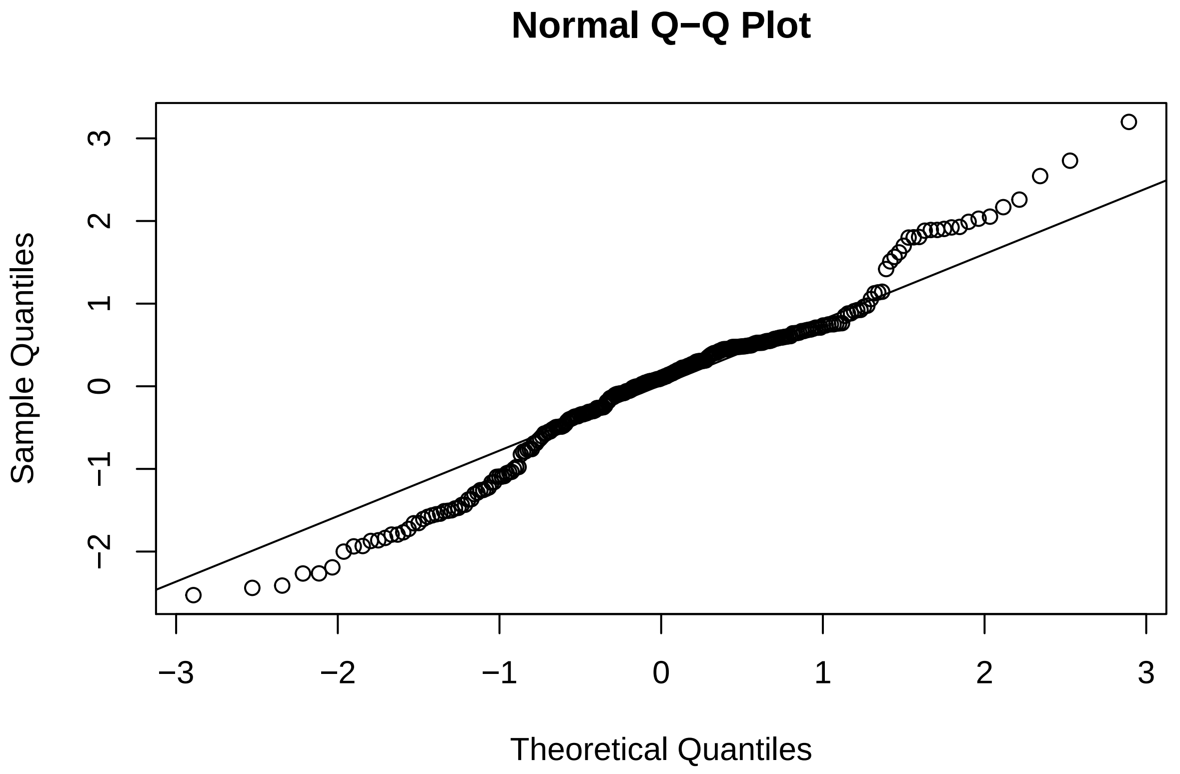
```
##
## Call:
## lm(formula = log(High) ~ Date, data = reg_AAPL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35706 -0.07423  0.01425  0.07840  0.44681
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.9407915  0.2954600  -33.65   <2e-16 ***
## Date         0.0007793  0.0000167   46.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1431 on 260 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.893
## F-statistic:  2179 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(AAPL.lm$fit, rstudent(AAPL.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(AAPL.lm$fit, rstudent(AAPL.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(AAPL.lm))
qqline(rstudent(AAPL.lm))
```

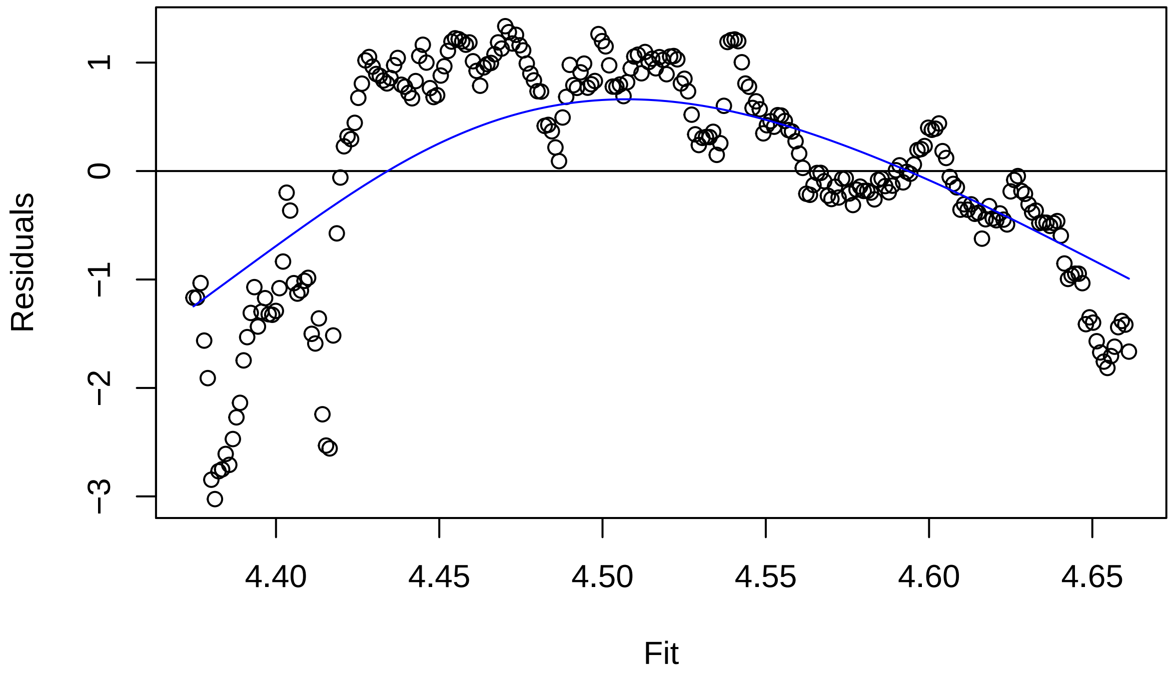


Based on the slope and intercept p-values acquired from the linear regression of AAPL, we can conclude that they are both significantly not zero. The linear regression has a slope of 8.393e-04 and an intercept of -1.024e+01. The residual standard error is 0.06152 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

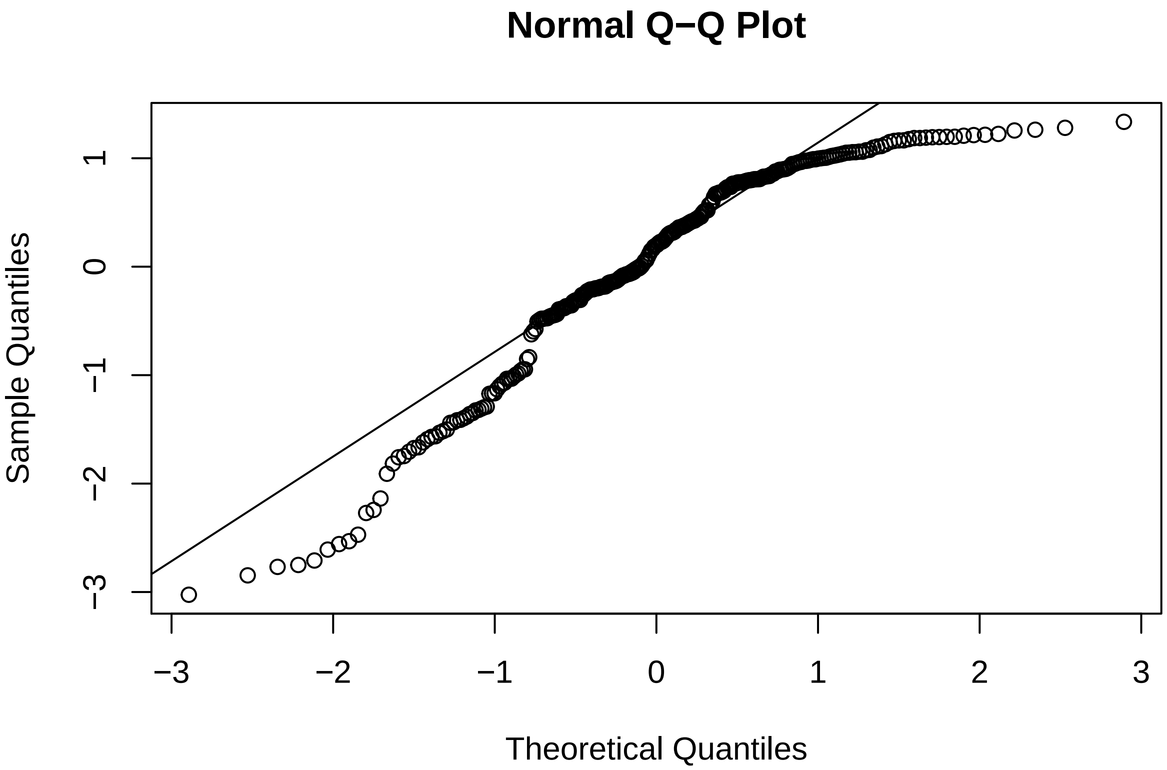
```r
reg_oil <- oil2 %>% select(Date, avg) %>% mutate(Date = as.Date(oil2$Date))

oil.lm <- lm(log(avg) ~ Date, data = reg_oil)
summary(oil.lm)
```

```
##
## Call:
## lm(formula = log(avg) ~ Date, data = reg_oil)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41422 -0.06574  0.02768  0.11607  0.18632
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.292e+00  2.893e-01  25.210   <2e-16 ***
## Date        -1.568e-04  1.634e-05  -9.595   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1401 on 260 degrees of freedom
## Multiple R-squared:  0.2615, Adjusted R-squared:  0.2586
## F-statistic: 92.06 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(oil.lm$fit, rstudent(oil.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(oil.lm$fit, rstudent(oil.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(oil.lm))
qqline(rstudent(oil.lm))
```

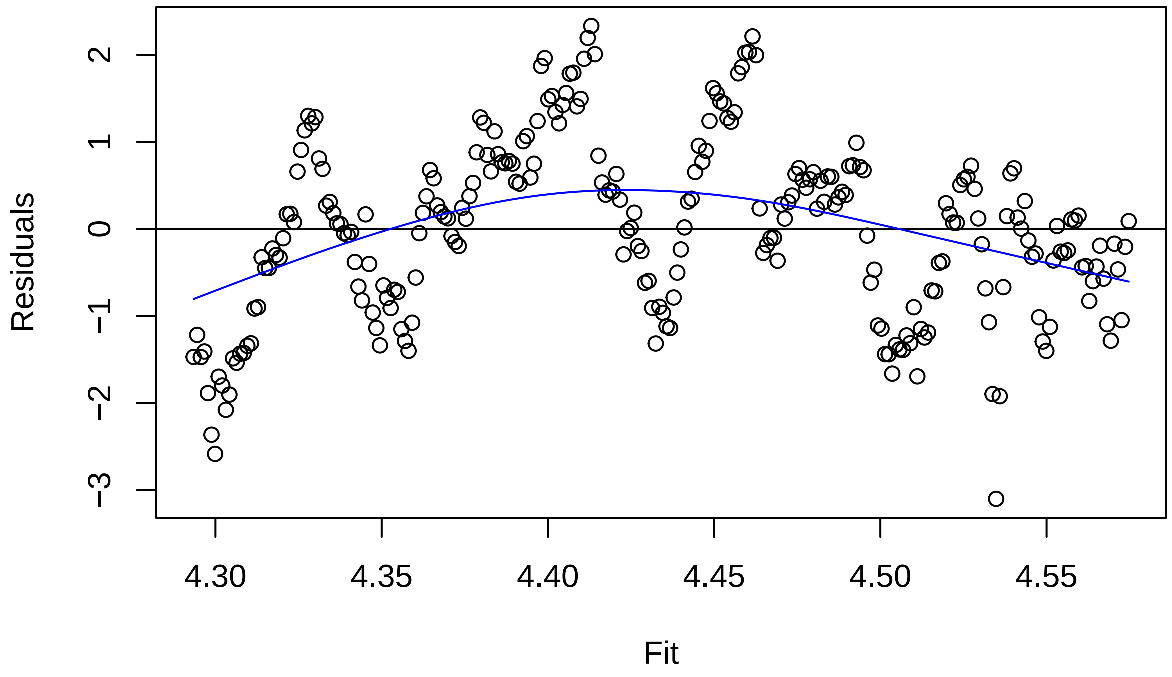


Based on the slope and intercept p-values acquired from the linear regression of the Oil and Gas sector, we can conclude that they are both significantly not zero. The linear regression has a slope of -1.568e-04 and an intercept of 7.292. The residual standard error is 0.1401 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

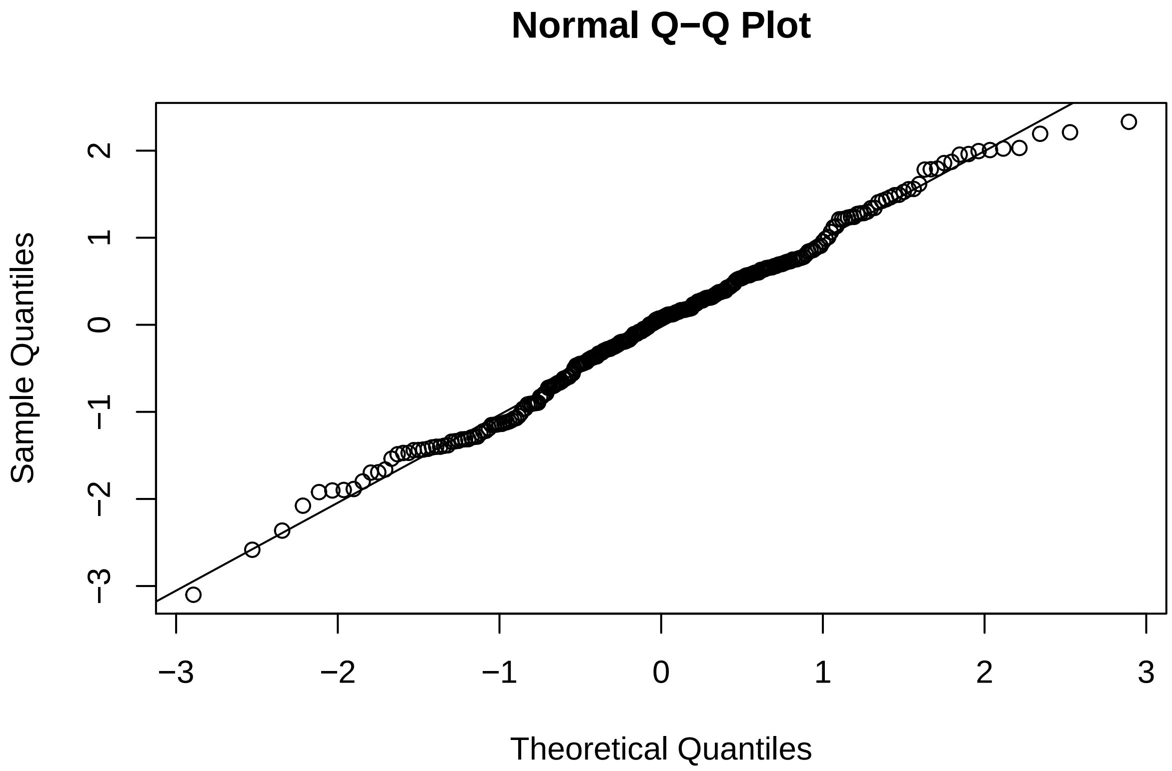
```r
reg_health <- health2 %>% select(Date, avg) %>% mutate(Date = as.Date(health2$Date))

health.lm <- lm(log(avg) ~ Date, data = reg_health)
summary(health.lm)
```

```
##
## Call:
## lm(formula = log(avg) ~ Date, data = reg_health)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.159776 -0.037004  0.003972  0.034471  0.121419
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.711e+00  1.087e-01   15.74   <2e-16 ***
## Date        1.540e-04  6.143e-06   25.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05264 on 260 degrees of freedom
## Multiple R-squared:  0.7073, Adjusted R-squared:  0.7061
## F-statistic: 628.2 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(health.lm$fit, rstudent(health.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(health.lm$fit, rstudent(health.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(health.lm))
qqline(rstudent(health.lm))
```

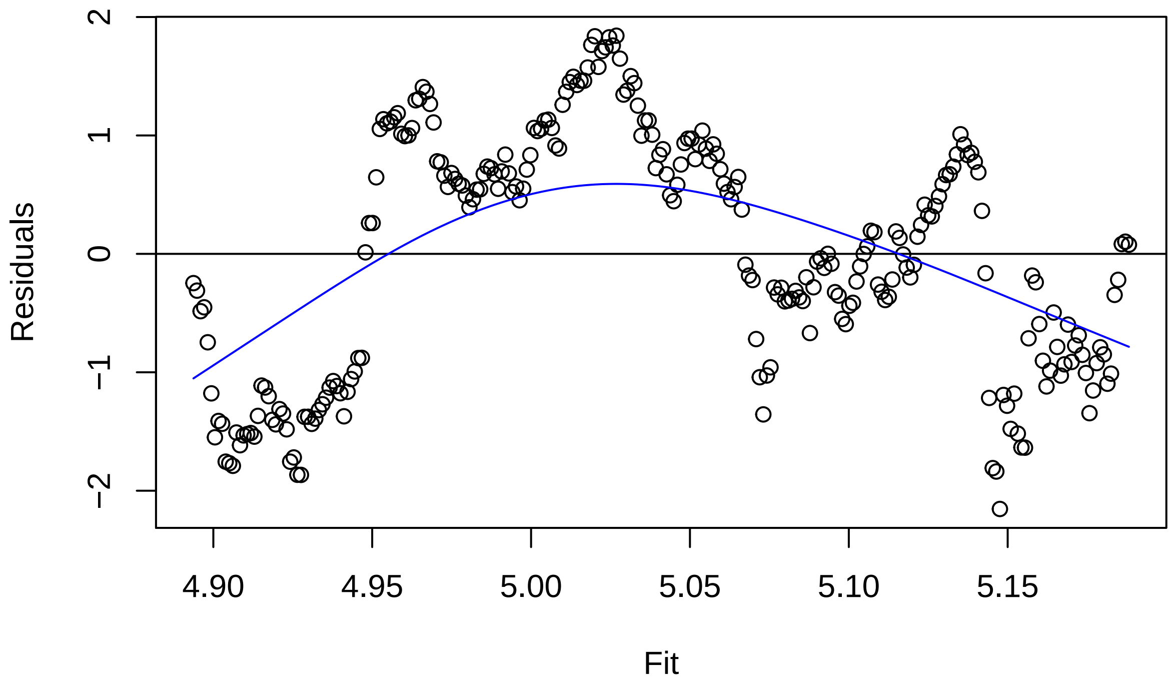


Based on the slope and intercept p-values acquired from the linear regression of the Healthcare sector, we can conclude that they are both significantly not zero. The linear regression has a slope of 1.540e-04 and an intercept of 1.711. The residual standard error is 0.05264 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

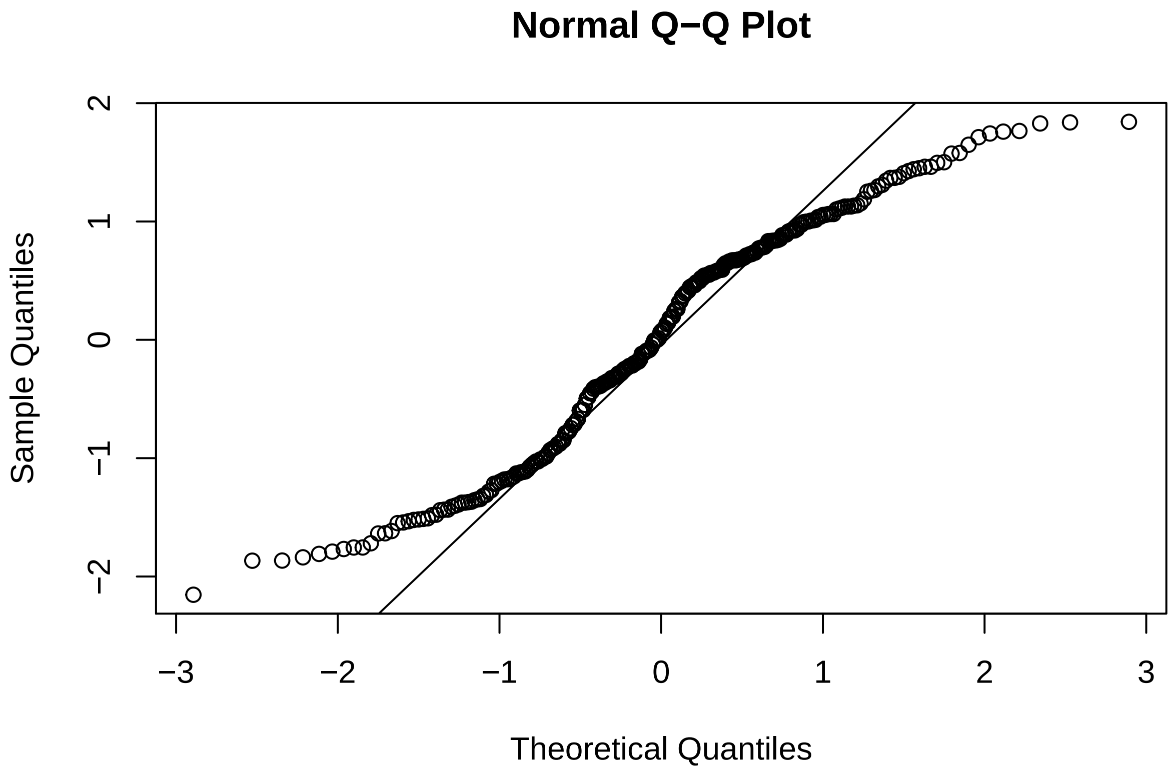
```r
reg_finance <- finance2 %>% select(Date, avg) %>% mutate(Date = as.Date(finance2$Date))

finance.lm <- lm(log(avg) ~ Date, data = reg_finance)
summary(finance.lm)
```

```
##
## Call:
## lm(formula = log(avg) ~ Date, data = reg_finance)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.298380 -0.128116  0.009854  0.116627  0.256509
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.190e+00  2.895e-01   7.566 6.67e-13 ***
## Date        1.612e-04  1.636e-05   9.853  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1402 on 260 degrees of freedom
## Multiple R-squared:  0.2719, Adjusted R-squared:  0.2691
## F-statistic: 97.07 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(finance.lm$fit, rstudent(finance.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(finance.lm$fit, rstudent(finance.lm), df = 3), col = "blue")
```



```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(finance.lm))
qqline(rstudent(finance.lm))
```

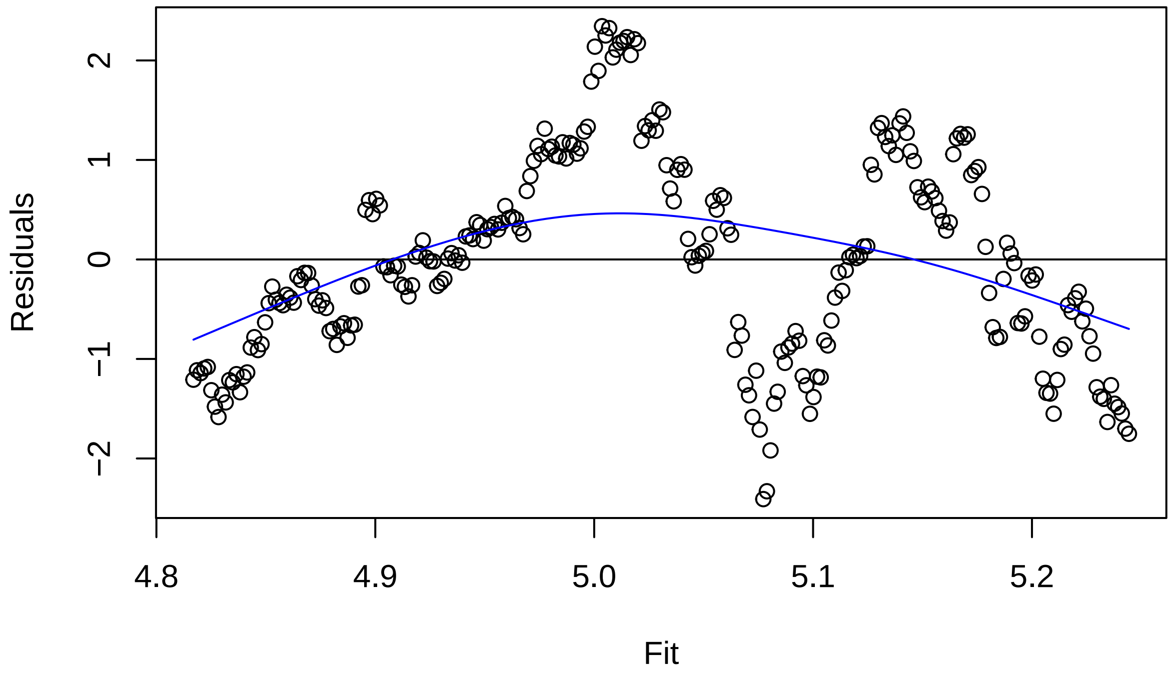


Based on the slope and intercept p-values acquired from the linear regression of the Financial sector, we can conclude that they are both significantly not zero. The linear regression has a slope of 1.612e-04 and an intercept of 2.190. The residual standard error is 0.1402 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

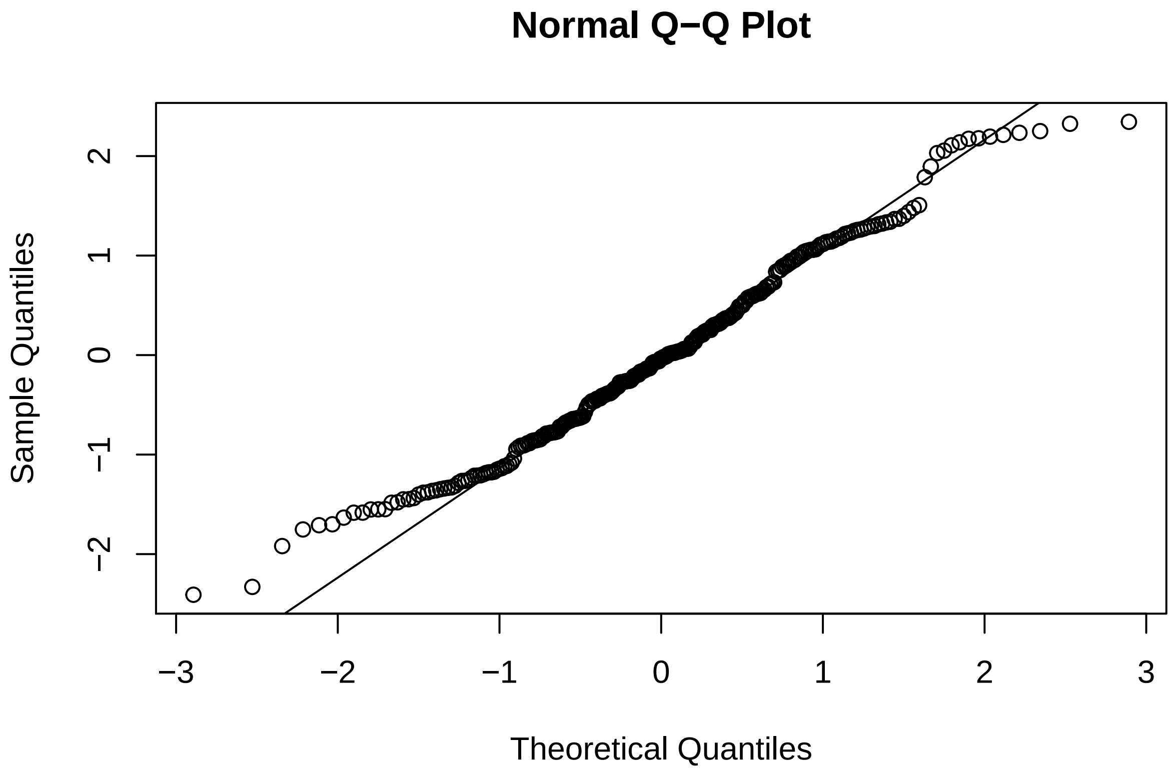
```
reg_ind <- ind2 %>% select(Date, avg) %>% mutate(Date = as.Date(ind2$Date))

ind.lm <- lm(log(avg) ~ Date, data = reg_ind)
summary(ind.lm)
```

```
##
## Call:
## lm(formula = log(avg) ~ Date, data = reg_ind)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.233653 -0.076001 -0.003392  0.069212  0.227578
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.927e-01  2.026e-01   4.405 1.55e-05 ***
## Date        2.339e-04  1.145e-05  20.430  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09812 on 260 degrees of freedom
## Multiple R-squared:  0.6162, Adjusted R-squared:  0.6147
## F-statistic: 417.4 on 1 and 260 DF,  p-value: < 2.2e-16
```

```
#Diagnostics:Independence of Residuals? constant variance?
plot(ind.lm$fit, rstudent(ind.lm), xlab = "Fit", ylab = "Residuals")
abline(h = 0)
lines(smooth.spline(ind.lm$fit, rstudent(ind.lm), df = 3), col = "blue")
```



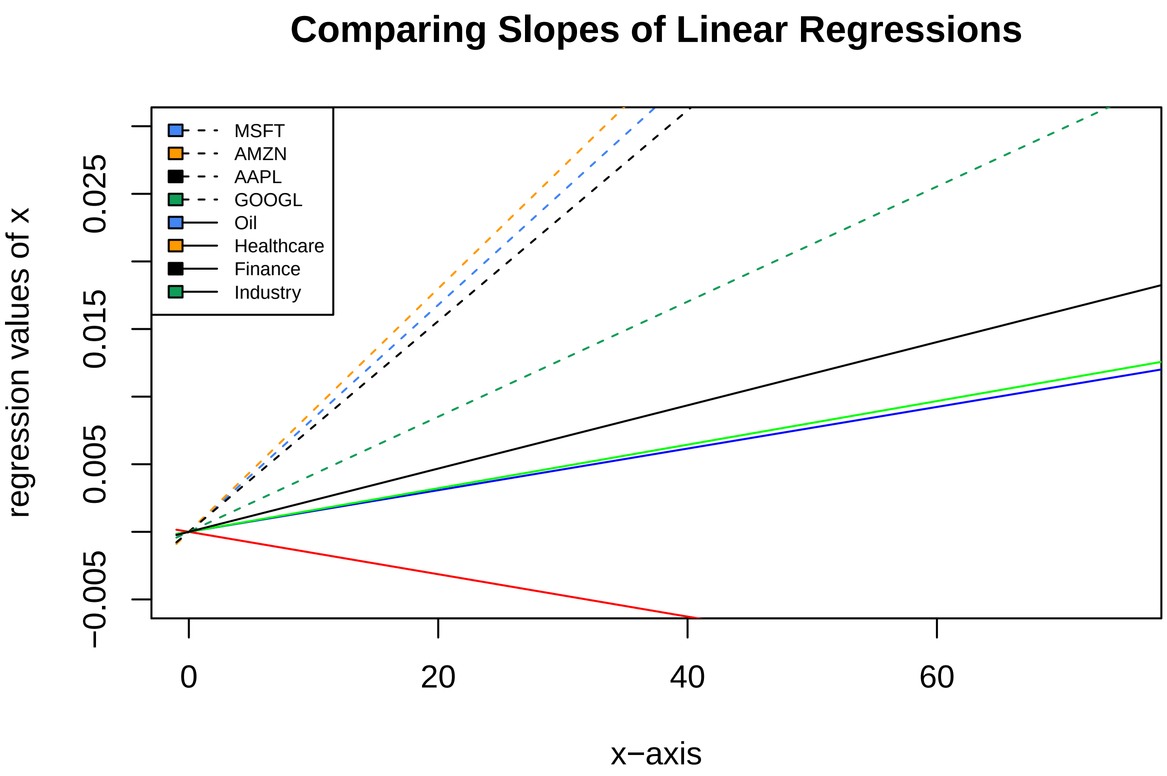
```
#Diagnostics: Normality of residuals?
qqnorm(rstudent(ind.lm))
qqline(rstudent(ind.lm))
```



Based on the slope and intercept p-values acquired from the linear regression of the Industrial sector, we can conclude that they are both significantly not zero. The linear regression has a slope of 2.339e-04 and an intercept of 8.927e-01. The residual standard error is 0.09812 on 260 degrees of freedom, and the normality graph shows that the data is acceptably normal. Thus the regression model is a good fit.

Then to visualize the rates of change, we plotted all the slopes, without the intercepts, to show the different slopes of each regression we have taken so far.

```r
# tech regressions
amzn_y <- function(x){return(amzn.lm$coefficients[2]*x)}
aapl_y <- function(x){return(AAPL.lm$coefficients[2]*x)}
msft_y <- function(x){return(msft.lm$coefficients[2]*x)}
googl_y <- function(x){return(GOOGL.lm$coefficients[2]*x)}
# sector regressions
oil_y <- function(x){return(oil.lm$coefficients[2]*x)}
health_y <- function(x){return(health.lm$coefficients[2]*x)}
finance_y <- function(x){return(finance.lm$coefficients[2]*x)}
ind_y <- function(x){return(ind.lm$coefficients[2]*x)}

# plotting
plot(seq(-1,100),msft_y(seq(-1,100)),type = "l",lty=2, col = '#4285F4', xlab = 'x-axis',
     ylab = 'regression values of x', ylim = c(-.005 , .03), xlim = c(0,75))
lines(seq(-1,100),amzn_y(seq(-1,100)), col = '#FF9900',lty=2)
lines(seq(-1,100),aapl_y(seq(-1,100)), col = 'black',lty=2)
lines(seq(-1,100),googl_y(seq(-1,100)), col = '#0F9D58',lty=2)

lines(seq(-1,100),oil_y(seq(-1,100)), col = 'red',lty=1)
lines(seq(-1,100),health_y(seq(-1,100)), col = 'blue',lty=1)
lines(seq(-1,100),finance_y(seq(-1,100)), col = 'green',lty=1)
lines(seq(-1,100),ind_y(seq(-1,100)), col = 'black',lty=1)

title(main="Comparing Slopes of Linear Regressions")
legend(-3,.0314, legend = c('MSFT','AMZN','AAPL','GOOGL','Oil','Healthcare','Finance','Industry'),
       c('#4285F4','#FF9900','black','#0F9D58'), lty=c(2,2,2,2,1,1,1,1), cex = .6 )
```



This plot illustrates that every tech company has a greater slope, or rate of growth, than every other sector. This helps confirm our conclusion that we reject the null hypothesis, implying that the tech companies we studied grew at a faster rate than companies in other sectors.

Conclusion

After analyzing the data set consisting of NASDAQ market data from December 7, 2015 to December 7, 2020, we can say with confidence that our research hypothesis is evident in statistically showing that stock growth of consumer technology companies grows faster than other sectors. We were able to determine this with permutation testing. Microsoft (MSFT) grew faster than oil and health care, Apple (APPL) grew faster than oil, Alphabet(GOOGL) and Amazon(AMZN) grew faster than every other sector. Taking the linear regression of all the companies and all the sectors, we could conclude that all the tech companies grew faster than the other sectors. Therefore, we can reject our null hypothesis: Tech companies do not grow at a significantly higher rate than other industries. Our findings were statistically significant, supporting the claim that Apple, Microsoft, Google, and Amazon are the “fastest growing companies.” Our analysis helps determine the rapid integration of these corporations and their products into everyday life. By comparing them to the other major sectors in the stock market, they help us identify these companies and the sector as a whole, to have massive potential for growth in the future. It also helps inform investment strategies and technological trends.

Reference

“Alphabet Inc. Class A Common Stock (GOOGL) Historical Data.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/googl/historical.

“AMZN.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/amzn.

“Apple Inc. Common Stock (AAPL) Historical Data.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/aapl/historical.

“CVX.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/cvx.

“Exxon Mobil Corporation Common Stock (XOM) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/xom/advanced-charting.

“General Electric Company Common Stock (GE) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/ge/advanced-charting.

“Goldman Sachs Group, Inc. (The) Common Stock (GS) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/gs/advanced-charting.

“Johnson & Johnson Common Stock (JNJ) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/jnj/advanced-charting.

“JP Morgan Chase & Co. Common Stock (JPM) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/jpm/advanced-charting.

“Microsoft Corporation Common Stock (MSFT) Historical Data.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/msft/historical.

“Northrop Grumman Corporation Common Stock (NOC) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/noc/advanced-charting.

“Pfizer, Inc. Common Stock (PFE) Advanced Charting.” Nasdaq, 7 Dec. 2020, www.nasdaq.com/market-activity/stocks/pfe/advanced-charting.