Predicting Stock Movement with Company’s Financial Data

Revenue Growth and Stock Returns

Kit Yi Wong (s3970390)

Last updated: 22 May, 2024

Introduction

In recent years, the stock market has been experiencing a surge. It seems like there is a phenomenon that emotions drive market behavior instead of a company's fundamental performance. Some research also indicates that the stock markets may not be truly efficient. Marc G.(2005) study showed the level of security market traders' investment may not be directly associated with the information efficiency of stock price. This is because managers have discretion in investment choices by giving the appropriate incentives to traders.

Indeed, the fundamental analysis does provide significant information about future stock returns, especially when considering the long-term trends of a stock movement. In an efficient market, the market price is expected to reflect the estimates of the intrinsic value of a securities. If discrepancies exist between the market price and the true value of a stock, the market will gradually realign itself toward its calculated intrinsic worth. In examining a company's fundamental performance, one of the critical variables is the revenue growth rate. Aloke G.(2005) study demonstrated that revenue-growth firms exhibit higher earnings ability than cost-reduction firms. Narasimhan J. (2022) study revealed a significant correlation between stock price and the magnitude of revenue surprises on the earnings announcement dates. The correlation effect will be weakened after the earnings announcement dates but a weak correlation still exists.

Problem Statement

Did the behavior of the stock market has changed? Is assessing a company's fundamental information still an effective method for predicting stock performance? This project will try to verify this question with a hypothesis test on stock return and revenue growth rate. Referring to Chris B. (2022) from McKinsey Research, an extra five percentage points of revenue per year correlates with an additional three to four percentage points of total shareholder returns.[1] This project will use the latest stock market information to conduct a hypothesis test on the statement above.

To simplify the hypothesis statement, this project will examine the relationship between Revenue Growth and Stock Return rates, to determine if stocks exhibit an average of 0.7 Stock Return to Revenue Growth rate. (McKinsey Research Suggested: Stock Return/Revenue Growth Rate = 0.035/0.05 = 0.7)

Data

This project sourced the S&P500 Companies dataset and the S&P500 Stocks dataset from Kaggle (Larxel, 2024)[5] with a data snapshot from 17 May 2024. The S&P 500 Companies dataset contains the company's basic information and the S&P500 Stocks dataset contains the stock price of stocks listed in the S&P500 index. This project combined 2 files into 1 dataset to obtain the revenue growth rate and year-to-date (YTD) return for stocks from the S&P500 index. The S&P 500 Index is a stock index that measures the performance of 500 leading publicly traded companies in the United States.[6] The stocks in the S&P500 account for approximately 75% of all publicly traded stocks in US stock market.[7] It provides a fair representation of the overall US stock market performance.

- Data Retrieval
  - *Load the S&P 500 Stocks and S&P 500 Companies datasets from csv files*
  - *Calculate the % of return per stock for the YTD period (17 May 2023 to 17 May 2024)*
    YTD return = ((Current value - Beginning value) / Beginning value) [8]
  - *Merge the S&P 500 Companies dataset and S&P 500 Stocks dataset with filtering fields required in this project*
  - *Dataset Fields: Stock Symbol [Symbol], Company Short Name[Shortname], Company Sector[Sector], Company Industry[Industry], Revenue Growth Rate[Revenuegrowth], YTD Return Rate[YTD Return])*

Data (Cont.)

```r
# Import S&P500 companies and stock price csv file
comp_df <- data.frame(read.csv( "Data/sp500_companies_20240517.csv", header=TRUE))
stock_df <- data.frame(read.csv( "Data/sp500_stocks_20240517.csv", header=TRUE))
# Filter to take the YTD stock price snapshot
stock_df_2024 <- stock_df %>% filter(Date == "2024-05-17")
stock_df_2023 <- stock_df %>% filter(Date == "2023-05-17")
# Merge the YTD stock price snapshots for YTD return calculation
stock_df_YTD = merge(stock_df_2023, stock_df_2024, by="Symbol")
# YTD return calculation
stock_df_YTD <- mutate(stock_df_YTD,
                       Return = ((stock_df_YTD$Adj.Close.y - stock_df_YTD$Adj.Close.x)
                       /stock_df_YTD$Adj.Close.x))
# Merge the S&P Companies Dataset with the S&P Stock YTD Return Dataset
df = merge(subset(comp_df, select = c(Symbol, Shortname, Sector, Industry, Revenuegrowth)),
           subset(stock_df_YTD, select = c(Symbol, Return)), by="Symbol")
# preview dataset size
dim(df)
```

```
## [1] 503   6
```

Descriptive Statistics and Visualisation

Missing Values: Missing values are identified from revenue growth and YTD return fields. For revenue growth, missing values have been filled back with data from Google Finance. [9] For YTD return, it was discovered that 3 stocks (Symbol: SOLV, GEV, VLTO) were found not listed in the S&P500 index one year ago. Consequently, YTD returns could not be calculated for these stocks, leading to the removal of them from the dataset.[10][11]

```
# check missing values - Revenuegrowth
df[(is.na(df$Revenuegrowth) | df$Revenuegrowth==""), 'Symbol' ]
```

```
## [1] "WDC"
```

```
# fix issue cases with Google Finance data [19]
df[(df$Symbol=="WDC"), 'Revenuegrowth'] = -0.034
# check missing values - Return
df[(is.na(df$Return) | df$Return==""), 'Symbol']
```

```
## [1] "GEV"  "SOLV" "VLTO"
```

```
# fix issue cases by remove it from the dataset. Stocks not listed in S&P500 1yr ago.[10][11]
df = df[!df$Symbol %in% c("SOLV","GEV","VLTO") ,]
```

Decsriptive Statistics and Visualisation (Cont.)

Outliners: Use Z-Scores to identify outliners. Stocks with a z-score > than 3 on the return to revenue growth rate were removed.

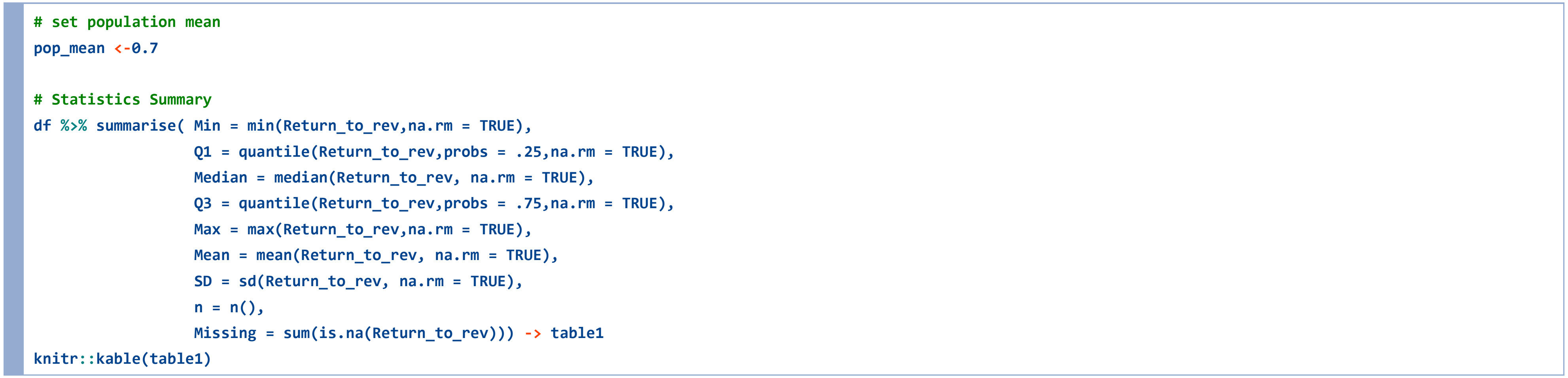
```r
# Construct Return to Revenue Growth Rate column
df <- mutate(df, Return_to_rev = (df$Return / df$Revenuegrowth))

# identify outliners - use Z-Scores
z.scores <- df[complete.cases(df$Return_to_rev),'Return_to_rev'] %>% scores(type ="z")
df[which(abs(z.scores) > 3),'Symbol']
```

```
## [1] "AMAT" "BLDR" "CAT"  "CHTR" "GNRC" "MMM"  "NXPI" "PH"   "PKG"
```

```r
# fix outliners
df = df[!df$Symbol %in% c("AMAT","BLDR","CAT", "CHTR", "GM", "MMC", "NWSA", "PGR", "PHM") ,]
```

Decsriptive Statistics and Visualisation (Cont.)

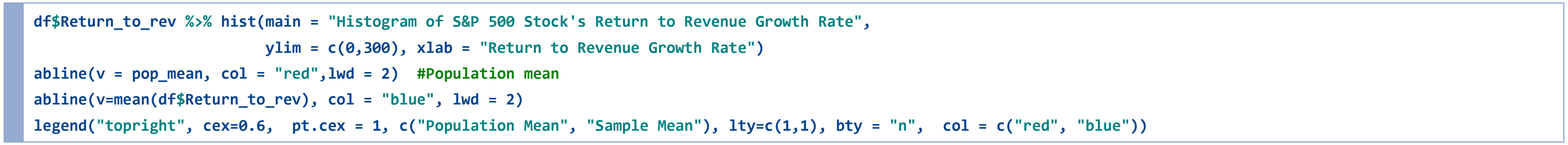
According to the Statistics Summary, I know the Return to Revenue Growth rate sample mean($\overline{x}$) is 1.95. How unusual is that assuming the population mean(μ) of Return to Revenue Growth rate is 0.7?



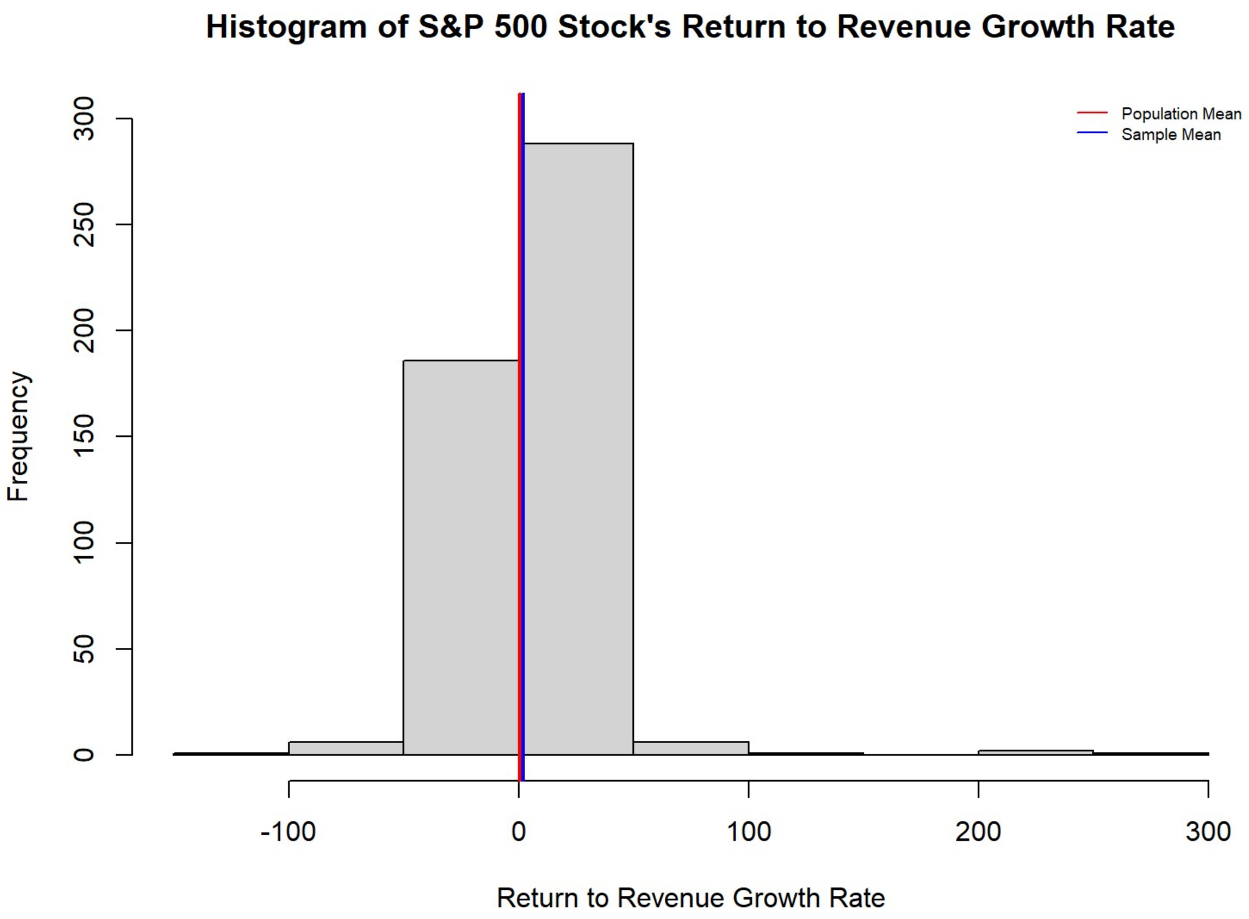
```r
# set population mean
pop_mean <-0.7

# Statistics Summary
df %>% summarise( Min = min(Return_to_rev,na.rm = TRUE),
                  Q1 = quantile(Return_to_rev,probs = .25,na.rm = TRUE),
                  Median = median(Return_to_rev, na.rm = TRUE),
                  Q3 = quantile(Return_to_rev,probs = .75,na.rm = TRUE),
                  Max = max(Return_to_rev,na.rm = TRUE),
                  Mean = mean(Return_to_rev, na.rm = TRUE),
                  SD = sd(Return_to_rev, na.rm = TRUE),
                  n = n(),
                  Missing = sum(is.na(Return_to_rev))) -> table1
knitr::kable(table1)
```

Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
-117.8518	-1.523982	1.0272	3.672476	281.6642	1.954044	24.97035	491	0

Decsriptive Statistics and Visualisation - Histogram

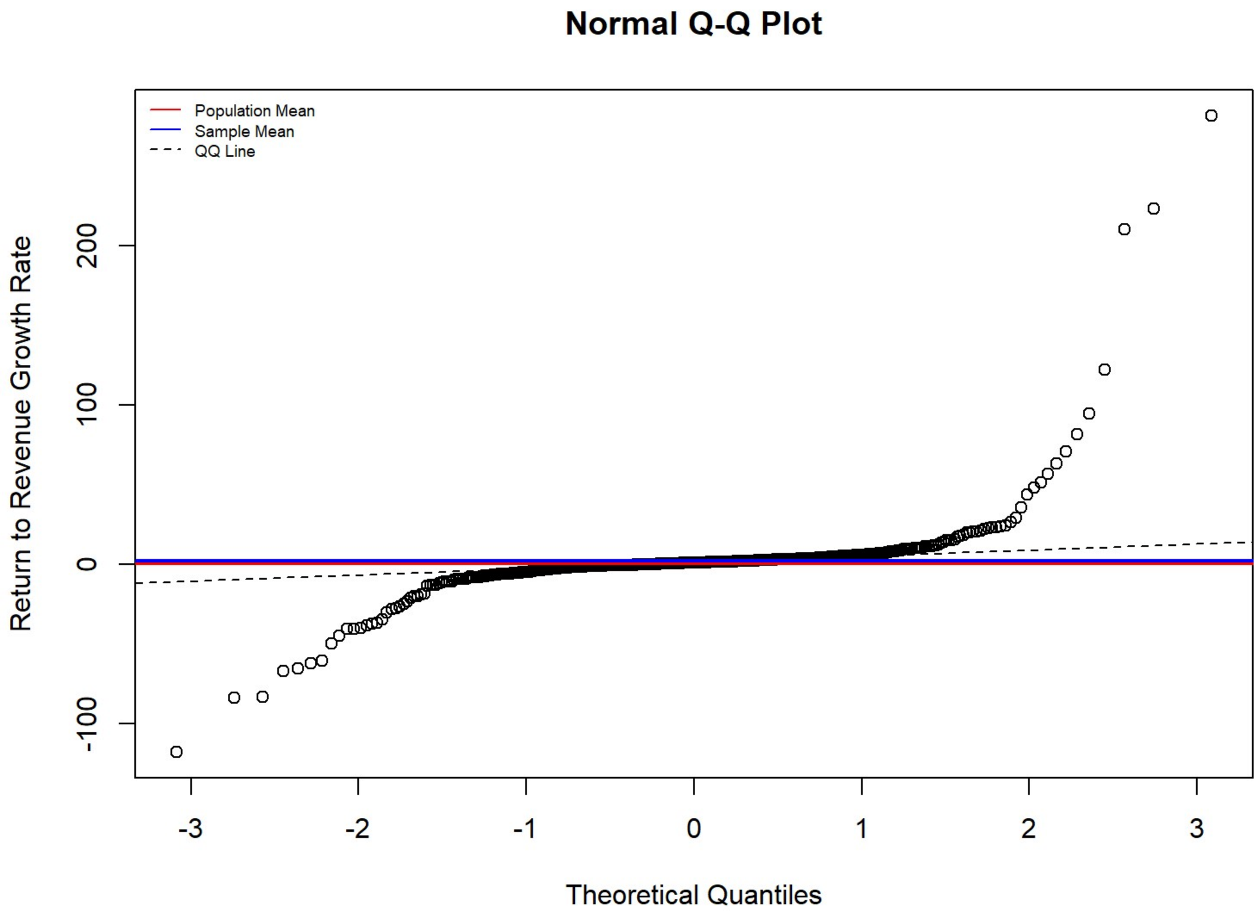


```
df$Return_to_rev %>% hist(main = "Histogram of S&P 500 Stock's Return to Revenue Growth Rate",
                          ylim = c(0,300), xlab = "Return to Revenue Growth Rate")
abline(v = pop_mean, col = "red",lwd = 2)  #Population mean
abline(v=mean(df$Return_to_rev), col = "blue", lwd = 2)
legend("topright", cex=0.6,  pt.cex = 1, c("Population Mean", "Sample Mean"), lty=c(1,1), bty = "n",  col = c("red", "blue"))
```



Decsriptive Statistics and Visualisation - QQplot

```
df$Return_to_rev %>% qqnorm(ylab = "Return to Revenue Growth Rate")
qqline(df$Return_to_rev, col = "black", lwd=1,lty=2)
abline(h = pop_mean, col = "red", lwd = 2) #Population mean
abline(h=mean(df$Return_to_rev), col = "blue", lwd = 2)
legend("topleft", cex=0.6,  pt.cex = 1, c("Population Mean", "Sample Mean", "QQ Line"), lty=c(1,1,2), bty = "n",  col = c("red", "blue", "black"))
```



Hypothesis Testing and Confidence Interval

State the Null and Alternate hypothesis:
A one-sample t-test was conducted to test the hypothesis because we have a known population mean, a sample mean, and a unknown population standard deviation. The sample size is larger than 30, so I can assume the sampling distribution will be normally distributed. The Significance level, representing the likelihood of a usual occurrence, for the test was set at 5%.

$$H_0 \; : \; \mu_{=}0.7$$

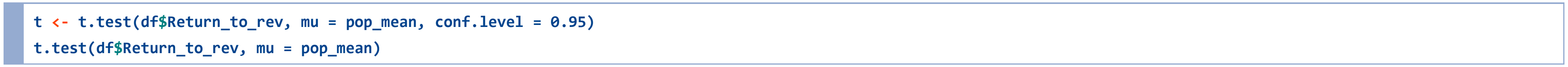
$$H_a \; : \; \mu_{\neq}0.7$$

Null hypothesis: The population mean of the Return to Revenue Growth rate is 0.7
Alternate hypothesis: The population mean of the Return to Revenue Growth rate is not 0.7

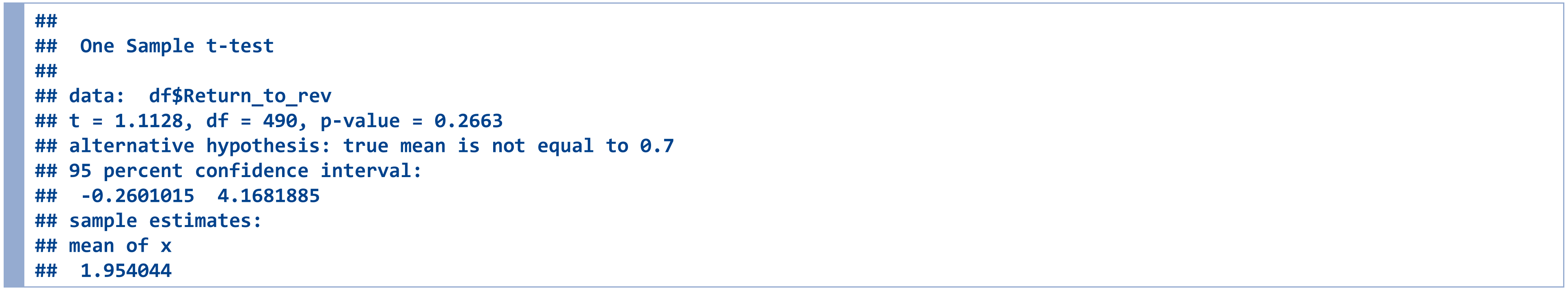
```r
# set population mean
pop_mean <-0.7
```

Hypothesis Testing and Confidence Interval (Cont.)

Decision: Since the p-value is gather than our significance level α = 0.05, I do not reject the null hypothesis. Other than that, the 95% CI of the estimated population mean [-0.3918217, 5.2181602], which does capture μ = 0.7. The results of the one-sample t-test were therefore not statistically significant. In summary, the average S&P500 YTD Return to Revenue Growth rate was not significantly different from the stated rate 0.7



```
t <- t.test(df$Return_to_rev, mu = pop_mean, conf.level = 0.95)
t.test(df$Return_to_rev, mu = pop_mean)
```



```
##
##  One Sample t-test
##
## data:  df$Return_to_rev
## t = 1.1128, df = 490, p-value = 0.2663
## alternative hypothesis: true mean is not equal to 0.7
## 95 percent confidence interval:
##  -0.2601015  4.1681885
## sample estimates:
## mean of x
##  1.954044
```

Regression analysis (Cont.)

Hypotheses for the overall linear regression model:
Next, this project want to understand does the S&P500 YTD Return to Revenue Growth rate fits the linear regression model for stocks from a specific sector. Stocks from the Technology sector will be used for this test, below are the hypotheses.

Null hypothesis: The data does not fit the linear regression model. Alternate hypothesis: The data fits the linear regression model.

Decision Rules: Reject null hypothesis If p value < 0.05 (α significance level). Test will be statistically significant if reject null hypothesis.

The sample sizes of stocks from the Technology and Utilities sectors exceed 30.

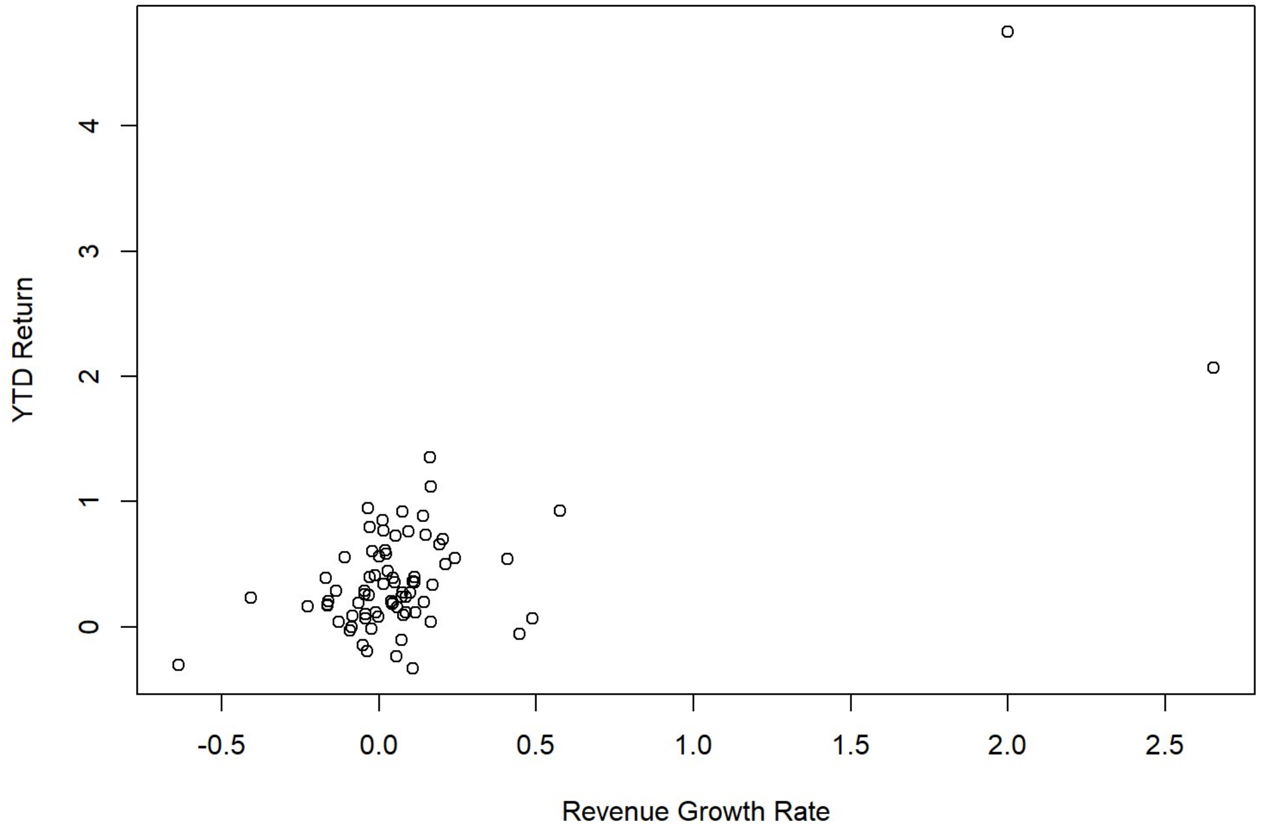
```
dim(df[(df$Sector=="Technology"), ])
```

```
## [1] 75  7
```

Regression analysis (Cont.)

From the YTD Return against Revenue Growth Rate Scatter Plot, there appears to be a linear positive relationship for stocks from the Technology sector.

```
plot(Return ~ Revenuegrowth, data = df[(df$Sector=="Technology"), ],
     xlab = "Revenue Growth Rate",
     ylab = "YTD Return")
```



Regression analysis (Cont.)

Is data fit the linear regression model statistically significant?

```
# Fit a linear regression model to the data and assess the fit
model1 <- lm(Return ~ Revenuegrowth, data = df[(df$Sector=="Technology"), ])
model1 %>% summary()
```

```
## 
## Call:
## lm(formula = Return ~ Revenuegrowth, data = df[(df$Sector == 
##     "Technology"), ])
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.22202 -0.20389 -0.03019  0.20589  2.19916 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.31074    0.05272   5.895 1.08e-07 ***
## Revenuegrowth  1.12202    0.12520   8.962 2.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4444 on 73 degrees of freedom
## Multiple R-squared:  0.5239, Adjusted R-squared:  0.5173 
## F-statistic: 80.32 on 1 and 73 DF,  p-value: 2.195e-13
```

```
model1 %>% confint()
```

```
##                   2.5 %    97.5 %
## (Intercept)   0.2056751 0.4157989
## Revenuegrowth 0.8725000 1.3715404
```

Regression analysis (Cont.)

Is data fit the linear regression model statistically significant?

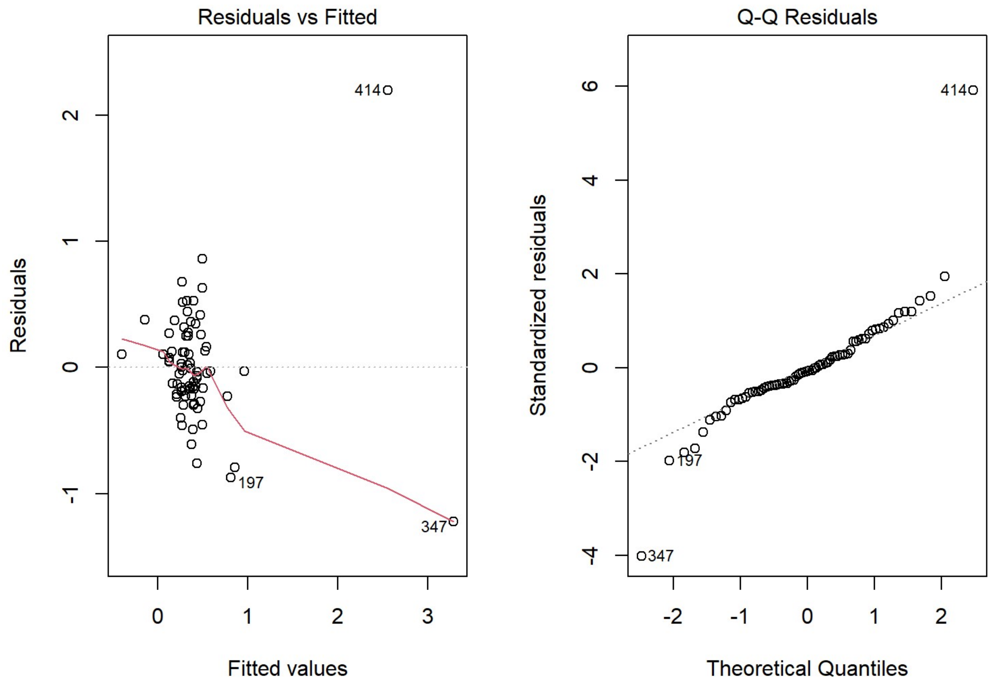
- The p value for the F-test is really small, F(1, 73) = 80.32, $p < 0.001$.
- The linear model was statistically significant.
- The estimated average YTD Return when Revenue Growth Rate = 0 was 0.31074.
- The intercept of the regression was statistically significant, a = 0.31074, $p < 0.001$, 95% CI (0.21, 0.42).
- For every one unit increase in Revenue Growth Rate, YTD Return rate was estimated to increase on average by 1.12.
- The slope of the regression was statistically significant, b = 1.12202, $p < 0.001$, 95% CI (0.87, 1.37).

Regression analysis (Cont.)

Check the main assumptions for linear regression:

- Independence: Independence was assumed as each stock came from a different company.
- Linearity: The Residual vs. Fitted scatter plot suggested a linear relationship. The trend line is considered flat. Variability on y axis is constant across the range of values on the x asix. No sign of heteroscedasticity.
- Normality of residuals: The residuals fall close to the QQline. The normal Q-Q plot didn't indicate obvious evidence of departures from normality.

```r
par(mfrow=c(1,2))
model1 %>% plot(which = 1)
model1 %>% plot(which = 2)
```



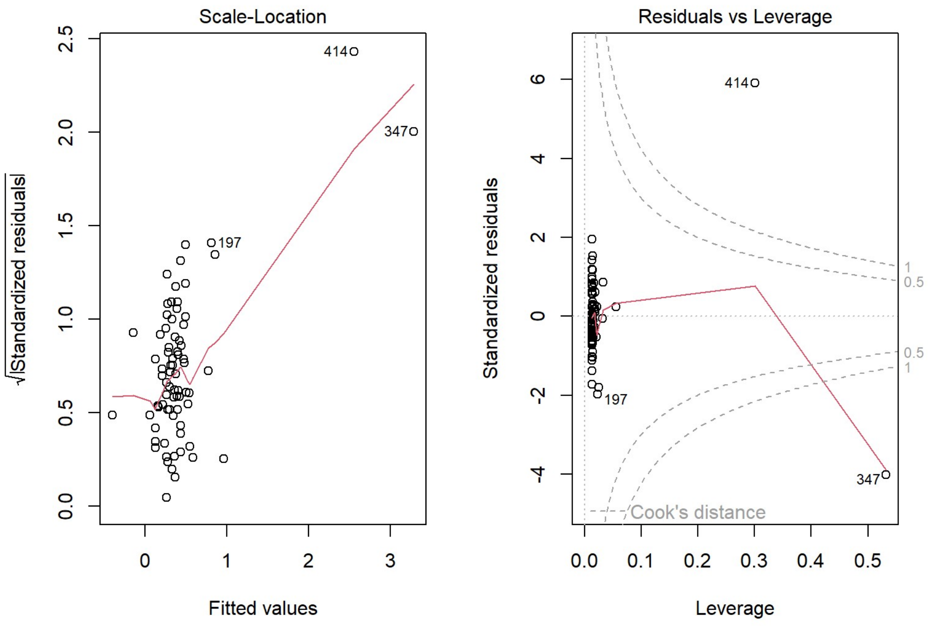
Regression analysis

Check the main assumptions for linear regression

- Influential cases: The Residuals vs Leverage scatter plot found 2 values fall within band 1. There appeared to be 2 influential cases that could have a disproportional impact on the fit of the regression model. These 2 outliners should be removed.

- Homoscedasticity: If excluded the outliner points identified from the Residuals vs Leverage plot, the variance in residuals is consistent across fitted values in the Scale-Location plot. The polt presents a Homoscedasticity outlook.

```
par(mfrow=c(1,2))
model1 %>% plot(which = 3)
model1 %>% plot(which = 5)
```



Discussion

- From the Hypotheses Testing, this project found the average S&P500 YTD Return to Revenue Growth rate was not significantly different from the stated rate, captured from McKinsey Research, 0.7.
- From the Hypotheses Testing of does the S&P500 YTD Return to Revenue Growth rate fits the linear regression model for stocks from a Technology sector, this project found the linear model was statistically significant.
- The hypothesis testing offers a clear and straightforward method for concluding our hypotheses. It is user-friendly and easy to understand. However, a drawback of hypothesis testing is its restricted to test only one parameter at a time. Typically, stock analysis requires the consideration of multiple variables. If I want to prove another fundamental variable relationship to stock return, a separate hypothesis test would be required, which can be time-consuming.
- If this project can be extended, I aim to examine the relationship of YTD Return to Revenue Growth across various sectors, like the Utilities sector, to assess whether the data exhibits different behaviors from different sectors.
- This project has demonstrated that, on average, the revenue growth rate from the company's fundamental information on average has exhibited a positive linear relationship with the mid-term year-to-date (YTD) Return. Despite the presence of emotional buying activity observed in the Technology sector this year, the data behavior of the YTD Return to Revenue Growth rate remains valid. Therefore, investors should not overlook the importance of the company's fundamental information.

References

[1] Chris B., Rebecca D., Nicholas N., Tido R. (2022). Empirical research reveals what it takes to generate value-creating growth today. Mckinsey & Company. Retrieved May 20, 2024 from https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/the-ten-rules-of-growth

[2] Aloke G., Zhaoyang G., Prem C. Jain (2005). Sustained Earnings and Revenue Growth, Earnings Quality, and Earnings Response Coefficients. Springer Science. Retrieved May 21, 2024 from https://link.springer.com/article/10.1007/s11142-004-6339-3

[3] Marc G., Timothy K., David W (2005). Emotions can drive market behavior in a few short-lived situations. But fundamentals still rule. Mckinsey & Company. Retrieved May 21, 2024 from https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/do-fundamentalsor-emotionsdrive-the-stock-market

[4] Narasimhan J. (2022). Revenue Growth and Stock Returns. S&P Global. Retrieved May 21, 2024 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=314962 [5] Larxel (2004). S&P 500 Stocks (daily updated). Kaggle. Retrieved May 20, 2024 from https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks/data

[6] Will K (2023). S&P 500 Index: What It's for and Why It's Important in Investing. Investopedia. Retrieved May 20, 2024 from https://www.investopedia.com/terms/s/sp500.asp

[7] Paula P. (2021). What Are the S&P 500, the Nasdaq, and the Dow?. The Balance Money. Retrieved May 21, 2024 from https://www.thebalancemoney.com/the-sandp-500-nasdaq-dow-jones-what-is-this-stuff-453745#:~:text=The%20S%26P%20500%20tracks%20500%20large%20U.S.%20companies,represent%20roughly%2075%25%20of%20all%20publicly%20traded%20stocks.

[8] Claire B. (2023). How Do I Calculate the Year-to-Date (YTD) Return on My Portfolio?. Investopedia. Retrieved May 20, 2024 from https://www.investopedia.com/ask/answers/060115/how-do-i-calculate-my-yeartodate-ytd-return-my-portfolio.asp

[9] Google Finance(2024). Western Digital Corp. Retrieved May 20, 2024 from https://www.google.com/finance/quote/WDC:NASDAQ

[10] Yahoo Finance. (2024). GE Vernova and Solventum Set to Join S&P 500; Dentsply Sirona to Join S&P MidCap 400; Others to Join S&P SmallCap 600. Retrieved May 20, 2024 from https://finance.yahoo.com/news/ge-vernova-solventum-set-join-221200444.html

[11] Nasdaq(2023). Veralto To Be Added To S&P 500. Retrieved May 20, 2024 from https://www.nasdaq.com/articles/veralto-to-be-added-to-sp-500

[12] Dr.Laleh T(2024). Week 08- Class Worksheet Answers. RMIT - Applied Analytics (MATH1324). Retrieved May 20, 2024 from https://rmit.instructure.com/courses/124444/files/36559873/download?download_frd=1

[13] Dr.Laleh T(2024). Week 11- Class Worksheet Answers.pdf. RMIT - Applied Analytics (MATH1324). Retrieved May 20, 2024 from https://rmit.instructure.com/courses/124444/files/36559844/download?download_frd=1