AMS 572 Fall 2022 Group 2 Project

Stony Brook University- Dec 1, 2022

Project By:

Aakash Doshi

Raj Shah

Efthimios Vlahos

Professor:

Pei fen Kuan

Introduction

It can very well be said that the main purpose of any statistical study is to attempt to give us, objectively, as much confidence as mathematically possible in our decision-making process between two opposing scenarios in the form of a null hypothesis and an alternate hypothesis. Across virtually all industries, statistics and probability theory gives us guidelines on how to approach many difficult questions that can be answered with a high degree of certainty. In this group project, we investigate two hypotheses with their associated alternative hypotheses using various statistical methodologies that give us a guideline to make reasonable decisions based on two datasets involving the New York Stock Exchange (NYSE).

Portfolio Managers have long emphasized and have been highly overweight on IT sector stocks in the last decade. It has been commonly perceived that technological advancements have led the massive growth in the IT sector. Even retail investors have been attracted to fancy tech products and have been investing on a large scale in IT-based companies. In comparison to this, Utilities being an essential consumer product have been delivering consistent returns over the years. As financial experts say, IT stocks are the hot stocks, and superior returns are guaranteed. But is this true?

Thereby we have undertaken this study whereby we are checking if IT companies have higher returns as compared to the Utility sector companies.

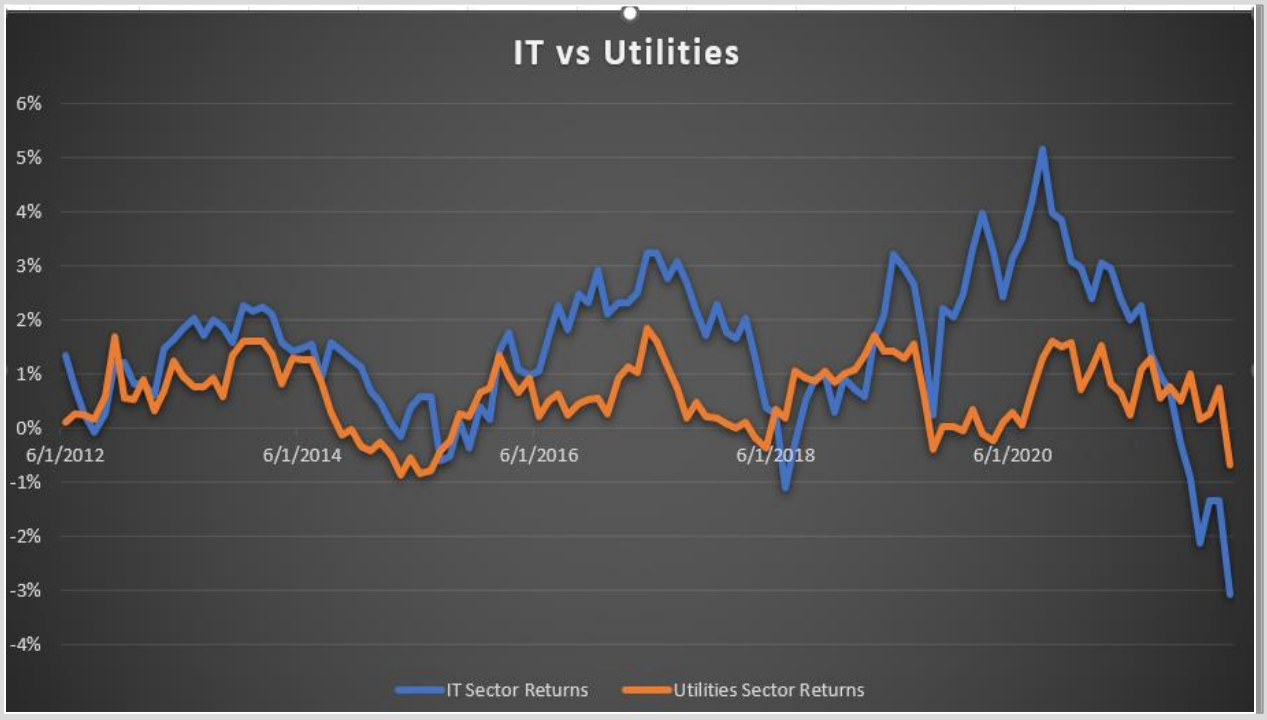


Figure 1: Chart showing IT vs Utilities sector returns ranging from 2012-2020 [1]

The data set is stored as a Microsoft Excel CSV file. In this project, we will use the R language and environment to do statistical computing and graphics work. Before doing hypotheses testing, we need to import the data set file into R and installing of the below mentioned packages:

1) H(0) (Null-Hypothesis)= The IT sector has similar returns with the utilities sector

V.S.

H(1) (Alternative Hypothesis)= The IT sector has greater returns than the utilities sector

Data For First Hypothesis

The NYSE .csv contains information about all the stocks listed on the New York Stock exchange (well over the required 50 samples as required by project guidelines) in which, by itself, there are over 79 features/variables for each ticker/example that includes various economic parameters about that company/row in the data set. As for the securities and prices .csv, they contain a general description of each company with the division of the sector (containing 8 variables) and raw, as in daily, prices of the tickers listed in the NYSE. The NYSE dataset consists of financial information on the entire list of stocks trading on the exchange. We have thereby created two populations consisting of an entire list of IT sector companies and Utilities sector companies respectively. We will then randomly sample the data from the population for the hypothesis testing. With a 5% margin of error, we will be sampling 57 companies from the IT sector and 27 companies from the Utilities sector.

List of the variables:

Sr No	Variable Name	Brief
1	Symbol	Ticker/code of the company which is used as an identifier on all the financial sector data portals
2	Stock Name	name of the company listed on the NYSE
3	Sector	the sector to which the company belongs
4	Sub-industry	sub-industry is a one-step further bifurcation of the sector
5	Close_2016	closing price of the last trading day of 2016
6	Close_2015	closing price of the last trading day of 2015
7	returns	Year-on-Year (Y-o-Y returns of the stock)

Data Exploration and Results of First Hypothesis

```
> mean(final_IT$returns)
[1] 17.49189
> mean(final_Utilities$returns)
[1] 12.36635
```

This quick and simple builtin R function gave us the average return from all the stocks in the IT and Utilities Sector which was 17.49189 % and 12.3665 %. As a preliminary result, we see that the IT sector has a higher percentage of returns in the year spanning 2015-2016 but still cannot conclude this definitively since we have not conducted any of the standard procedures to be able to feel confident about this claim.

After looking at the mean return of both sectors, it is natural to see if there are any stocks in either of the sectors that skew the data to the right or left with respect to returns. We can get a general idea of this visually by looking at the boxplots below:

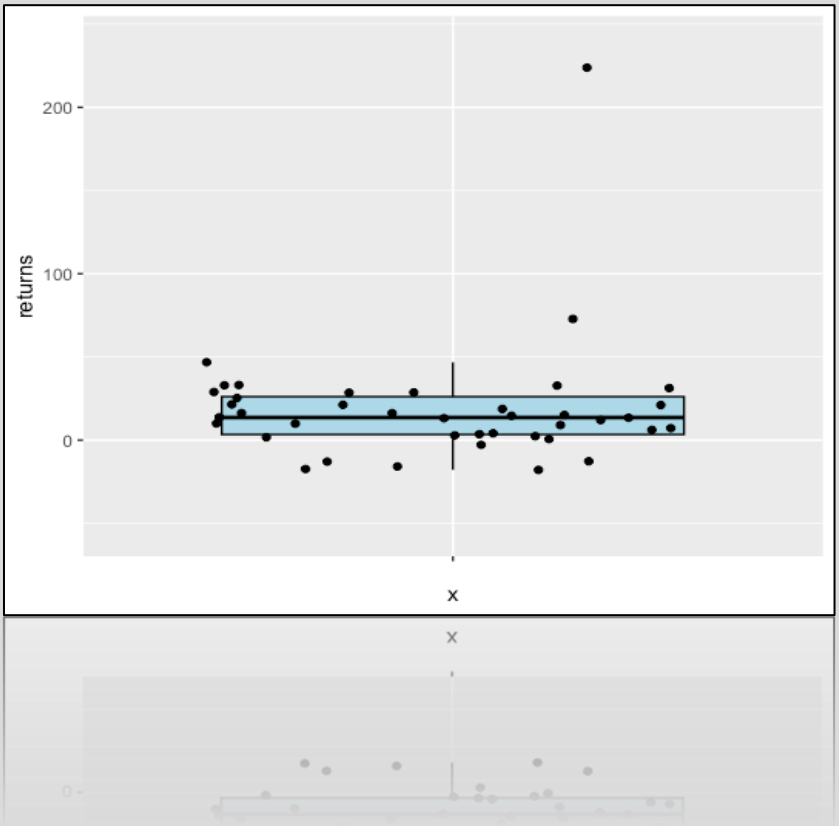


Figure 2: Boxplot of IT Sample (with outliers)

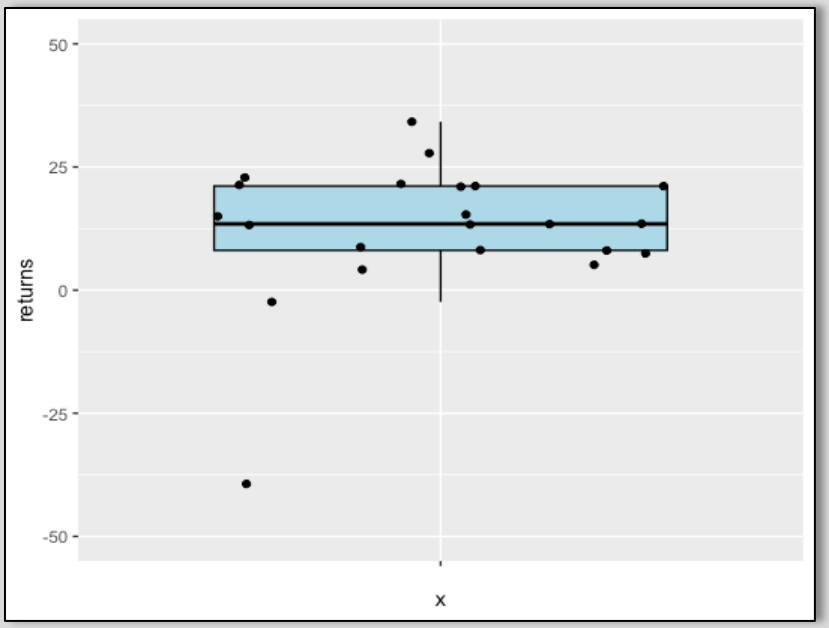


Figure 3: Boxplot of Utilities Sample (with outliers)

We can see visually from the two boxplots above that the IT sector sample has one stock outperforming by a very large percentage and the Utilities sector has the same but instead of outperforming it is underperforming.

The above graphs seem to suggest that the sample from the IT sector is positively skewed while the sample from the Utilities sector is negatively skewed: this suggests that, given the sample data available to us, also considering the fact that for the Utilities sector there are less than thirty examples, we cannot assume normality in the data and a separate test other than the two sample t-test is needed to be performed in order to reject or fail to reject the null hypothesis.

To further support our claim, we conducted Shapiro-Wilkinson tests on both samples as well:

```
> shapiro.test(sampleIT$returns)
Shapiro-Wilk normality test
data:  sampleIT$returns
W = 0.65004, p-value = 1.615e-08
```

```
> shapiro.test(sampleUtilities$returns)
Shapiro-Wilk normality test
data:  sampleUtilities$returns
W = 0.82135, p-value = 0.001101
```

Since we cannot use the parametric two sample t-test to objectively decide whether to reject or fail to reject the null hypothesis, we resort to the non-parametric Wilcoxon-test hypothesis test:

```
> wilcox.test(sampleIT$returns,sampleUtilities$returns,alternative = 'greater')

        Wilcoxon rank sum exact test
data:  sampleIT$returns and sampleUtilities$returns
W = 465, p-value = 0.3603
alternative hypothesis: true location shift is greater than 0
```

The Wilcoxon indicates to us that we fail to reject the null hypothesis: we cannot definitively say without a reasonable doubt that the IT sector performs better than the Utilities sector. The graphs, as well as the Wilcoxon test, are telling us that the outliers in the IT sector and Utilities sector are the main culprits for the false assumption that the IT sector has better returns than the Utilities sector, which turns out to be mainly attributed to the positively skewed data for the IT sector and the negatively skewed data for the Utilities sector.

Missing Values Completely At Random (MCAR)

We are hereby simulating the hypothesis by artificially seeding the missing values in the dataset. Considering we are working on the returns of the individual stocks; we hereby assume that values are missing randomly from the population meaning there is no correlation in the missingness of the data.

We have artificially seeded the missing values (NA) using the “missForest” library with 20% of the data in variables Sub Industry, Close_2016, and Close_2015 being replaced with NA values.

This 20% missing data is re-seeded using mice imputation considering we have categorical data in the variable Sub Industry. We have used the statistical function polynomial regression for iterating and imputing categorical data and “pmm“ function for iterating and imputing numerical data.

```
imput_it <- mice(it_data1,m=5,method=c("polyreg","pmm","pmm"),maxit = 20)

imput_uti <- mice(uti_data1,m=5,method=c("polyreg","pmm","pmm"),maxit = 20)
```

Performing the hypothesis test on MCAR data, the parameters are unbiased and thereby the conclusion of the hypothesis remains the same though the power of the test might decrease due to deficiency in the design of the system resulting in MCAR data.

Missing Values Not At Random (MNAR)

In this case, missingness is not random as previously explained. And so, missingness depends on the observed data that is the missingness is related to events or factors which are not measured by the researcher.

In the financial industry, we commonly observe survivorship bias whereby the companies that have performed poorly or have been bankrupt, do not form part of the dataset. Also, the NYSE has a certain set of criteria whereby only those companies meeting these criteria are listed in the NYSE and others do not form part of the exchange. Below is the list of stringent criteria for a rebalancing of the list of stocks on the exchange:

1. The aggregate market value of publicly held shares must be at least US$40 million for IPO companies, or US$100 million for companies seeking to list their existing securities or to transfer to NYSE.
2. Have at least 400 holders of 100 shares or more and an average monthly trading volume of at least 100,000 shares for the most recent six months.
3. Have at least 2,200 total shareholders and an average monthly trading volume of at least 100,000 shares for the most recent six months.
4. Have at least 500 total shareholders, with an average monthly trading volume of at least 1 million shares for the most recent 12 months

Additionally, there are private companies like new startups that are not listed on any of the exchanges and the data for such a set of companies is not freely available on any of the platforms.

Dealing with data that has missing values, not due to randomness is a complex deal. We need to develop complex models to deal with missingness that is not random effectively. We now have certain data companies like MSCI., that are developing indexes in such a way whereby both the issues i.e., survivorship bias as well as private sector company data have been looked into.

```
> wilcox.test(final_final_it$returns,final_final_uti$returns,alternative = 'greater')

        Wilcoxon rank sum test with continuity correction

data:  final_final_it$returns and final_final_uti$returns
W = 24174, p-value = 0.3934

alternative hypothesis: true location shift is greater than 0
```

Conclusion of First Hypothesis

After conducting the Wilcox test, we conclude we don't have sufficient evidence to reject the null hypothesis. The higher returns from the IT sector turn out to be a result of the few outliers that exist in the data. Thus, we cannot accept the alternative hypothesis that the IT sector does not have higher returns than the utilities sector. And so, the industry perception that the IT sector has been clocking higher returns needs to be looked into further detail and thereby make wise investment decisions.

II. Data Exploration and Results of Second Hypothesis

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g., how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. 2. The value of the dependent variable at a certain value of the independent variables (e.g., the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Assumptions of Multiple Linear Regression:

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods and there are no hidden relationships among variables.
3. In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.
4. Normality: The data follows a normal distribution.
5. Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

Formula for Multiple Linear Regression

$$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \beta 3 * x3 + \beta 4 * x4 + \beta 5 * x5 + \beta 6 * x6$$

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The t statistic of the overall model.
- The associated p value (how likely it is that the t statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the t statistic and p value for each regression coefficient in the model.

Hypothesis II

H0 : IT sector YOY Returns dependent on fundamental factors including Investments(X1), Net Income(X2), Earnings per Share(X3), Net Cash Flow(X4), Long Term Debt(X5), Research and Development(X6).

Ha : IT sector YOY Returns are independent of fundamental factors including Investments(X1), Net Income(X2), Earnings per Share(X3), Net Cash Flow(X4), Long Term Debt(X5), Research and Development(X6).

Formula for this hypothesis:

H0 : X1,X2,X3,X4,X5,X6 = 0

Ha : X1,X2,X3,X4,X5,X6 ≠ 0

Now let's start the hypothesis which is basically using multiple linear regression to predict which of the following six factors are returns in IT sector dependent on. What we start by doing is gathering all the IT sector stocks from the list and listing all the parameters or variables in it. By doing so we are segregating IT stocks from the list.

Secondly, compute the IT returns of one year(2015) for all the stocks and then add the column of YOY Returns and compute them.

Why we have used Multiple Linear Regression Model :

It is important to know that a multiple linear regression model consist of predator variable or dependent variable and it shows the dependency on many other independent variables which helps to find correlation between them.
Also multiple linear regression model is used to estimate relationship between two or more independent variable and one dependent variables.
We also needed to find hypothesis testing using multiple linear regression which is finding driving force in IT sector returns from given six variables.

```
install.packages(c("Hmisc","dplyr","ggplot2","mice","Amelia","missForest","random","car"))
```

Now while creating a model we need to test and train the dataset accordingly. So we have trained 80% of our data and tested on the remaining 20% of it. So we randomly generated values from the dataset using runif() function and implementing training and testing methodologies on it.

```
#runif() used to randomly generate values from my_data
A1 = runif(506)
#Sorting the data
A2 = order(A1)
#Testing and training my_data:
train_data = my_data[A2[1:350],]
test_data = my_data[A2[351:506],]
```

Now the data is tested and trained we just need to run the multiple linear regression model using lm() function. Using summary function to get desired output:

```
Call:
lm(formula = YOYReturns ~ Investments + NetIncome + EarningsPerShare +
    NetCashFlow + Long.TermDebt + ResearchandDevelopment, data = train_data
1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29106 -0.10298 -0.00151  0.07941  0.43231

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.650e-01  3.406e-02   4.843 2.74e-05 ***
Investments              2.653e-11  1.366e-11   1.943   0.0604 .
NetIncome                2.272e-11  1.607e-11   1.413   0.1667
EarningsPerShare         6.652e-03  1.191e-02   0.559   0.5800
NetCashFlow              4.266e-11  2.605e-11   1.638   0.1107
Long.TermDebt           -1.683e-11  7.588e-12  -2.218   0.0334 *
ResearchandDevelopment   5.076e-11  2.034e-11   2.495   0.0176 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1638 on 34 degrees of freedom
  (309 observations deleted due to missingness)
Multiple R-squared:  0.2403,    Adjusted R-squared:  0.1062
F-statistic: 1.792 on 6 and 34 DF,  p-value: 0.1302
```

From the above it can be clearly seen that intercept here is the risk free rate(R)f which is 0.0165(β0) which is 1.65% is the minimum returns anyone gets when the buy any stock. Below formula shows that except long term debt all are positively correlated.

y=0.0165+ 0.00000000002653*x1+ 0.00000000002272*x2+* 0.006652x3+ 0.00000000004266*x4- 0.00000000001683*x5+ 0.00000000005076*x6

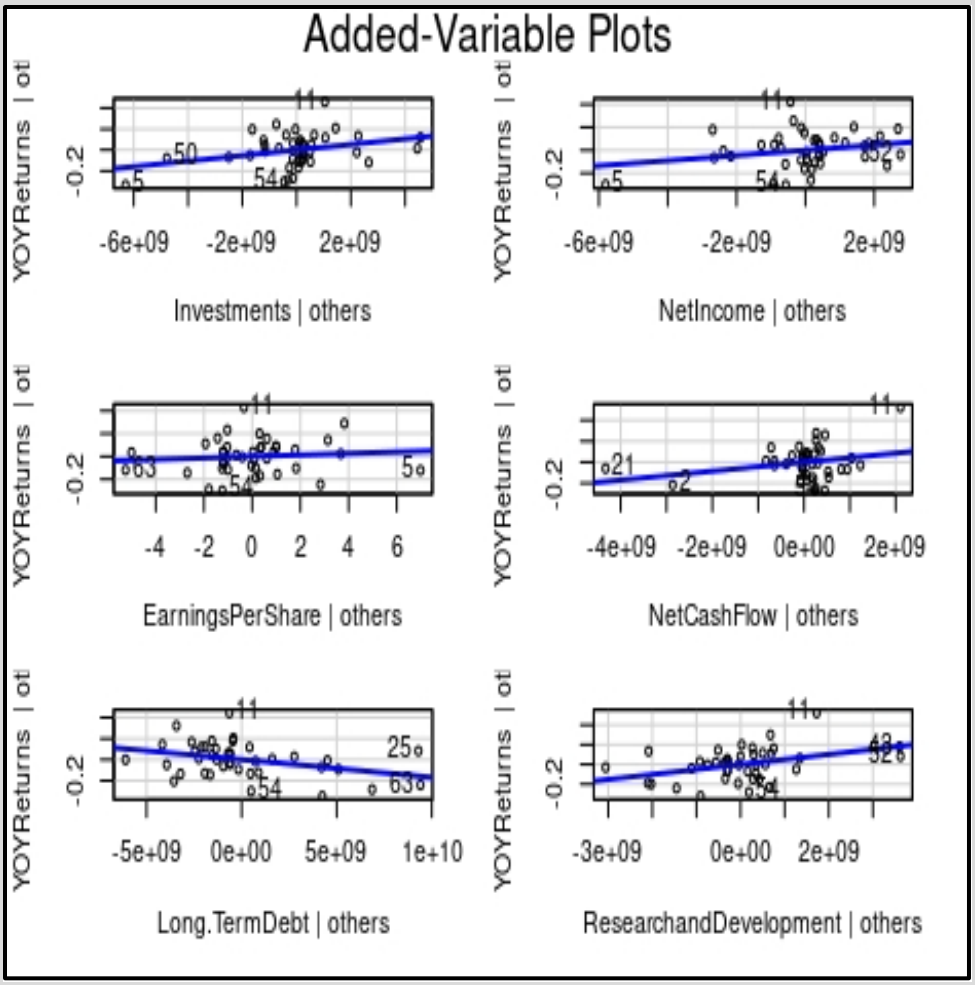
If the p value for a variable is less than your significance level your sample data provide enough evidence to reject the null hypothesis for the entire population Your data favor the hypothesis that there is a non-zero correlation. Changes in the independent variable are associated with changes in the dependent variable at the population level. This variable is statistically significant and probably a worthwhile addition to your regression model.

On the other hand, when a p value in regression is greater than the significance level it indicates there is insufficient evidence in your sample to conclude that a non-zero correlation exists.

While looking at above summary at an level of significance of 0.05 it can be seen that two variables i.e Long.Term Debt and Research and Development are not significant in terms of generating returns to it sector.

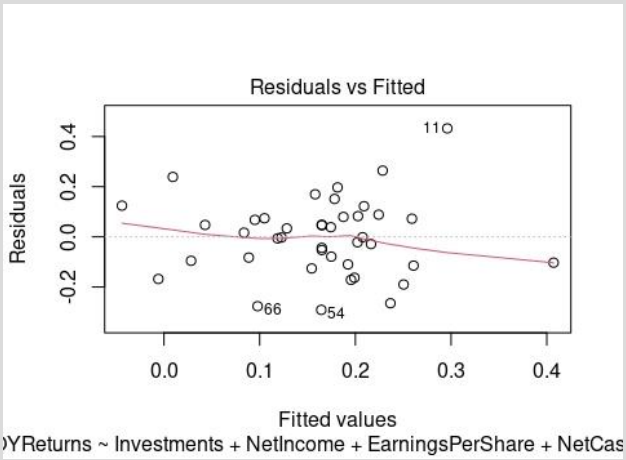
Using confint() function with level of significance of 5%:

```
< confint(model,level=0.95)
```



Now finding the residuals for the model :

```
#Finding residual of the model:
model.resid <- resid(model)
```



Effect of Missing Values

In most datasets, there might be missing values either because it wasn't entered or due to some error. Replacing these missing values with another value is known as Data Imputation. There are several ways of imputation. Common ones include replacing with average, minimum, or maximum value in that column/feature. Different datasets and features will require one type of imputation method. For example, considering a dataset of sales performance of a company, if the feature loss has missing values then it would be more logical to replace a minimum value.

Missing values completely at random:

MCAR introduce artificial missing completely at random values in a given complete data set. Missing values are multivariate and have generic pattern. We used the given dataset to used to predict MCAR effect on the result.

We used the impute function to randomly get the generated numbers for all the six variables which we used and then we run the model.

From the model re-run, what we see is by infatuating randomly generated values in the data we actually just have one significantly correlated values which is net income which is closest to p value and hence resulting in this composition.

Missing value not at random:

In this we need to fill missing values with not randomly but with strategically placing NA or '0' in the dataset and then trying to figure out if the solution remains the same or not

Using prodNA function to seed 20% dataset with NA values and simultaneously removing '0's from the dataset and then run the model.

After infatuating missing values with NA we get different results and Net Income and investments are two significantly correlated factors which are given significance.

So, based on missing values factors we can see the significant variables changes when we use different missing values techniques. So based on missing values we are not able to predict which are the significant variables which determine returns in IT sectors.

So, as hypothesized IT returns are not dependent on mentioned six factors but on various other factors as well.

Conclusion of Second Hypothesis

From the above hypothesis we deduce that we reject the null hypothesis which states that IT sector YOY Returns dependent on fundamental factors including Investments(X1), Net Income(X2), Earnings per Share(X3), Net Cash Flow(X4), Long Term Debt(X5), Research and Development(X6). We used several methods like multiple linear regression model to predict variables affecting the IT YOY returns.When using multiple linear regression we found that two factors are significant but when we did further analysis based on effects of missing values. We deduce that when infactuated with missing values our results are not the same anymore and this must be further researched. While comparing values we found that earnings per share is the driving force which remains the same across all models but which cannot be said true for other variables. So, we reject the null hypothesis.

REFERENCES

[1]https://stockcharts.com/articles/mindfulinvestor/2022/03/utilities-vs-technology-which-838.html

[2] https://www.kaggle.com/datasets/dgawlik/nyse?select=fundamentals.csv

[3] https://builtin.com/data-science/boxplot

[4] https://www.statisticshowto.com/shapiro-wilk-test/

{5] https://resourcehub.bakermckenzie.com/en/resources/cross-border-listings-handbook/north-america/new-york-stock-exchange/topics/principal-listing-and-maintenance-requirements-and-procedures

[6]https://journals.lww.com/epidem/Fulltext/2011/03000/Sensitivity_Analysis_When_Data_Are_Missing.25.aspx

[7] https://www.geeksforgeeks.org/how-to-impute-missing-values-in-r/

[8] https://stats.stackexchange.com/questions/197192/to-remove-more-than-predictors-in-lm-function-in-r

[9] http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/