AMS 572 REPORT

Understanding Impact of COVID19 on the Profit of the Fortune 100 Companies

Group 5
Zhiwei Lin
Chandrasekhar Thammisetty
Durga Venkatesh Kothapalli

Understanding impact of COVID19 on the profits of the Fortune100 companies

Abstract

As the COVID-19 pandemic has affected all of humanity, we focus on understanding if there is any significant difference in the profits of the Fortune100 companies before the pandemic (2019) and during the recovery phase (2021), and if there is a significant difference, we intend to find the underlying variables which are responsible for the change in profits of these companies during these two time periods. We have chosen multiple variables such as revenue, asset size, market cap, GICS Sector, size of the workforce, CEO sex and the headquarters state political party (all of these are collectively called as 'variables') on which the profit of a company depends and with the help of this study we intend to check whether if any of these variables are significantly contributing to profit. Also, we intend to study the interaction terms between time and all of the above-mentioned variables so as to understand if the profit was impacted adversely or favorably with the passage of time.

Introduction

With the sudden impact of COVID-19 throughout the world, most companies were not prepared to face the slump in sales, which in turn resulted in reduced profits for the year 2020. Companies from different sectors reacted in different ways and most of them adapted to the new normal codes of social distancing and virtual work environments, but still some sectors such as pharma and consumer staples benefited while the hospitality and transportation sectors were adversely affected. By 2021, the first doses of vaccines were made available to the vulnerable sections and slowly the pandemic faded with the rise in vaccination, but did the profits return to the pre pandemic levels is the question we intend to answer.

Sectors which gained significantly from the pandemic:

Pharma & health care sector: As there was a rise in sales of generic medicines, most of the pharma sector instantly benefited from the pandemic. As the vaccines came out, most of the big pharmaceutical companies made huge profits as the pandemic waned.

Logistics: Although the global supply chain had hiccups, most of the last mile delivery services benefited from the pandemic as the delivery services to households increased multifold.

Sectors which performed poorly during the pandemic:

Hospitality sector: As there were many curfews and lockdowns imposed in most of the countries in 2020, mainly the hospitality sector got hit as there were very few people traveling.

Passenger Air Transport: Due to the lockdowns, most of the airlines had near zero occupancy throughout 2020 and the sector only got revived mid-way through 2021.

Our investigation encompasses the study of the first 100 companies in the fortune 500 list of 2019 as this list ranks the top 500 companies by revenue. We intentionally picked the fortune 100 companies as they are presumed to have a very stable stream of revenue.

Using the collection of fortune 100 company's data set (i.e., variables such as revenue, asset size, market cap, GICS Sector, size of the workforce, CEO sex and the headquarters state political party), the hypothesis testing is done to analyze the profit change for pre-covid year 2019 and recovery phase year 2021. This data analysis is used to extract which variable is more responsible or least responsible in effecting the profits. We explored different online sites to extract data and established a data set to commence our research of variable dependency upon profits of fortune 100 companies.

Test 1: We would like to see if the profit levels of the fortune 100 companies during the recovery phase i.e., in 2021 had reached to the pre-pandemic level i.e., in 2019. Let P_{2021}, P_{2019} represent the profits of the fortune 100 list in the years 2021 and 2019 respectively.

Null Hypothesis: $P_{2021} = P_{2019}$

Alternate Hypothesis: $P_{2021} < P_{2019}$

Appropriate data transformations have been made to normalize the data.

Test 2: The second test involves studying the different variables on which the profit is dependent on. Let P be the profit, Y be the Year (0 if 2019 and 1 if 2021), M be the market cap of the company, T be the Type of Company (Public or Private), G be the Global Industry Classification Standard Sector, R be the revenue, E be the Workforce strength, A be the Asset Value of the company, YI be the year in which the company was incorporated, Pol be the political party that won the presidential election in the state which the headquarters of the company is located and Se be the sex of the CEO of the company

$$P = \beta_1 * (Y) + \beta_2 * (M) + \beta_3 * (T) + \beta_4 * (G) + \beta_5 * (R) + \beta_6 * (E) + \beta_7 * (A) + \beta_8 * (YI) + \beta_9 * (Pol) + \beta_{10} * (Se) + \beta_{11} * (M) * (Y) + \beta_{12} * (T) * (Y) + \beta_{13} * (G) * (Y) + \beta_{14} * (R) * (Y) + \beta_{15} * (E) * (Y) + \beta_{16} * (A) * (Y) + \beta_{17} * (YI) * (Y) + \beta_{18} * (Pol) * (Y) + \beta_{19} * (Se) * (Y) + \alpha$$

In the above equation, α is the intercept and $\beta_{12}, \beta_{13}, \ldots\ldots \beta_{19}$ are the interaction terms between the year and the rest of the variables. For example, β_{14} explains the relation between profit and revenue based on time. Here, our main goal is to understand the effect size of each independent variable on the dependent variable (Profit), and this regression model may not be used as a predictor model for other companies.

Data Collection:

In this report, pre-covid period is defined as January 2019 – December 2019, and the recovery period is defined as January 2021- December 2021. The reasoning for these period determinations is that COVID-19 first spread in the U.S.A in January 2019 and COVID-19 vaccine was authorized by FDA for emergency use in December 2020. Therefore, the time before the arrival of COVID-19 virus is pre-covid period and the time after vaccine rollout is recovery period.

With these timelines in mind, we decided to use 2020 Fortune 100 companies as our data set for this project. The reasoning for using 2020 Fortune 100 is that the 2020 ranking is based on 2019 revenues, and companies listed are all large companies in their fields. Therefore, the data set is a sample size of 100 that is representative of the chosen population studied. Moreover, the data set is large enough so the result would be more accurate as it provides smaller margins of error. However, our sample is not chosen by random as we are picking companies in Fortune 100 instead of companies from the population of U.S. large companies. The reasoning for this is due to the limited resources and time we have for this project as there are thousands of large companies out there in the U.S.A and we don't have all the financial data for every company.

For the dataset, it is a sample size of 100 companies and each with 22 variables:

Company name - nominal data, official name of company.

Ticker - nominal data, unique series of letters or numbers that represent a company on the stock market.

Type - nominal data, type of company with possible values: [Public, Private]

Market cap 31/12/2019 - continuous data, total value of all a company's share of stock on 31/12/2019

Market cap 31/12/2021- continuous data, total value of all a company's share of stock on 31/12/2021

GICS (Global Industry Classification Standard) sector - nominal data, a system for categorizing a company by sector. There are 11 possible sectors.

Industry - nominal data, to further classify a company by industry. There are 69 possible industries.

Profit 2019 - continuous data, net profit a company made in 2019 (calendarized).

Profit 2021- continuous data, net profit a company made in 2021 (calendarized).

Rank by profit 2019 - ordinal data, company ranked by profit in 2019.

Rank by profit 2021 - ordinal data, company ranked by profit in 2021.

Revenues 2019- continuous data, revenues a company made in 2019 (calendarized).

Revenues 2021- continuous data, revenues a company made in 2021 (calendarized).

Employees 2019 - continuous data, number of employees a company had in 2019 (calendarized).

Employees 2021 - continuous data, number of employees a company had in 2021(calendarized)

Asset 2019 - continuous data, the amount of assets a company had in 2019 (calendarized).

Asset 2021 - continuous data, the amount of assets a company had in 2021 (calendarized)

CEO sex 2019- nominal data, the sex of 2019 CEO with possible value: [Male, Female]

CEO sex 2021 - nominal data, the sex of 2021 CEO with possible value: [Male, Female]

Year incorporated - discrete data, the founding year of company

HQ state - nominal data, state where the headquarters of company is located

HQ state democratic or republican 2019 - nominal data, HQ state was democratic, or republican based on 2016 presidential election vote. Possible value: [D, R]

HQ state democratic or republican 2021- nominal data, HQ state was democratic, or republican based on 2020 presidential election vote. Possible value: [D, R]

For financial variables like revenues, number of employees, assets and profit, Fortune website have these financial data for all the listed companies; however, these financial data are collected based on their respective fiscal years. Fiscal year is a 12-month period chosen by a company to report financial information. Most companies file their annual report with December 31 as the end of fiscal year, aligning with the calendar year; however, some companies choose a date other than December 31 as the end of fiscal year for their annual report. In order to compare companies' financial data, calendarization is needed to make the company's financials start and end with the same dates.

Equation 1:

$$\frac{(12 - Month) \times Prior\ FY + (Month) \times Current\ FY}{12} = Calendraization,$$

$$where\ Month\ is\ end\ of\ FY(fiscal\ year)$$

For our dataset, we have calendarized the revenue, number of employees, assets and profit by using the formula above for each of 100 companies. For the other variables like ticker, type, market cap 31/12/2019, market cap 31/12/2021, GICS sector, industry, year incorporated, and HQ state, we have collected these variables from various resources such as the company's annual report, analyst calls and financial websites.

Missing Value Mechanism:

When we were collecting the data, we encountered some missing information about the market cap of private companies. Since private companies are not listed on any of the stock exchanges, they don't have official market cap values available. These missing data may decrease our confidence in the accuracy of

results we get in the statistical models. In order to minimize the problem of missingness in the dataset, a proper mechanism for missing value is needed. Identifying the type of missingness in the dataset is a crucial step in deriving an appropriate mechanism for missing values.

In this report, we first assume the missing value is missing completely at random (MCAR). There is no systematic error that makes any value have a higher probability of being missing than other values. Under MCAR assumption, likewise-deletion analysis is an appropriate mechanism to produce reliable and unbiased results. We then assume the missing value is missing not at random (MNAR). For this case, there is systematic error that makes a value have a higher probability of being missing than other values. Under MCAR assumption, an appropriate mechanism is proposed to impute the missing values. The mechanism is by utilizing the median of market cap/asset ratio for public companies to impute the market cap of private companies.

Equation 2:

$$Ratio = \left\{ \frac{MarketCap_{Public\ Company\ i}}{asset_{Public\ company\ i}} \right\}, where\ i = 1,2,3, \dots$$

$$\widehat{MarketCap}_{Private\ company} = Median\ (Ratio) \times asset_{Private\ company}$$

Descriptive statistics:

```
CompanyName           Ticker           MarketCap19         MarketCap21               TypeofCompany
Length:100         Length:100         Min.   :    941.6   Min.   :    572   Private           : 8
Class :character   Class :character   1st Qu.:  37515.0   1st Qu.:  39562   Private to Public: 1
Mode  :character   Mode  :character   Median :  76180.0   Median : 102310   Public            :91
                                      Mean   : 159890.7   Mean   : 234323
                                      3rd Qu.: 205605.0   3rd Qu.: 226885
                                      Max.   :1288000.0   Max.   :2902000
                                      NA's   :9           NA's   :8
                    Sector   RankbyProfit19   RankbyProfit21      Profit19        Profit21
Consumer Staples       :24   Min.   :  1.00   Min.   :  1.00   Min.   :-4979   Min.   :-6520
Financials             :23   1st Qu.: 25.75   1st Qu.: 25.75   1st Qu.: 2451   1st Qu.: 1726
Industrials            :15   Median : 50.50   Median : 50.50   Median : 4313   Median : 6208
Health Care            :14   Mean   : 50.50   Mean   : 50.50   Mean   : 7946   Mean   :10830
Energy                 :11   3rd Qu.: 75.25   3rd Qu.: 75.25   3rd Qu.: 8950   3rd Qu.:13784
Information Technology: 7    Max.   :100.00   Max.   :100.00   Max.   :81417   Max.   :89795
(Other)                : 6
  Revenues19       Revenues21      Employees19       Employees21         Asset19
Min.   : 28941   Min.   : 24634   Min.   :   1779   Min.   :   3023   Min.   :   5992
1st Qu.: 43496   1st Qu.: 48383   1st Qu.:  46000   1st Qu.:  46810   1st Qu.:  50185
Median : 66024   Median : 71007   Median :  90612   Median :  90850   Median : 116685
Mean   : 90712   Mean   :101058   Mean   : 153546   Mean   : 166426   Mean   : 309914
3rd Qu.:110940   3rd Qu.:126238   3rd Qu.: 201250   3rd Qu.: 196900   3rd Qu.: 259756
Max.   :523167   Max.   :571620   Max.   :2200000   Max.   :2300000   Max.   :3503319

   Asset21        CEOsex19 CEOsex21      Year              HQ     HQPolitical19 HQPolitical21
Min.   :   5942   F: 9     F:12     Min.   :1845   New York  :14   D:56          D:65
1st Qu.:  57614   M:91     M:88     1st Qu.:1944   Texas     :14   R:44          R:35
Median : 112635                     Median :1984   California: 9
Mean   : 370840                     Mean   :1971   Illinois  : 9
3rd Qu.: 276532                     3rd Qu.:2000   Ohio      : 6
Max.   :4229166                     Max.   :2021   Virginia  : 5
                                                   (Other)   :43
```

Figure 1. summary table for the dataset using RStudio. It's a sample size of 100 with 22 variables: company name, ticker, market cap 2019, market cap 2021, type of company, sector, rank by profit 2019, rank by profit 2021, profit 2019, profit 2021, revenues 2019, revenues 2021, employees 2019, employees 2021, asset 2019, asset 2021, CEO sex 2019, CEO sex 2019, Year incorporated, Headquarters, HQ political 2019, HQ political 2021. The currency unit is million dollars.

The min, first quartile, median, mean, third quartile and max values for continuous variables are shown in figure 1. There are 9 missing market caps in 2019 and 8 missing market caps in 2021. It's because one of the companies named Albertson INC. launched the IPO, which became a public traded company on Jun 2020. Moreover, it has been observed that the consumer staple sector has most Fortune 100 companies, followed by the financial sector. And most Fortune 100 companies are headquartered in New York and Texas, followed by California and Illinois. It has also been noticed that number of female CEOs increase from 2019 to 2021.

Paired T test under assumption of MCAR:

9 of 100 companies in the dataset have missing market cap values. Assuming that the market cap values are missing completely at random (MCAT), an appropriate option for dealing with these small numbers of missing values is to delete them. This is called likewise-deletion, also known as complete case analysis. This method is simple but quite powerful if only a small part of missingness and missing completely at random. Deleting a few observations does not have much adverse effect on the accuracy of statistical results. The figure below is profit data after the likewise-deletion.

group	count	min	Q1	median	Q3	max	mean	sd	skewness	kurtosis
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Profit Precovid (in milloin)	91	-4979	2672.000	4783.25	9637	81417	8492.689	11986.48	3.617134	19.59547
Profit Recovery (in million)	91	-6520	2554.165	6731.17	14670	89795	11690.603	16216.88	2.667204	11.29646

Figure 2. Profit data after likewise-deletion analysis. Count, minimum, first quartile, median, third quartile, maximum, mean, standard deviation, skewness and kurtosis of profit in pre-covid period and profit in recovery phase are shown above. The figure is generated using RStudio

Paired T test for testing whether companies' profit in recovery phase is less than the companies' profit in pre-covid period at $\alpha = 0.05$.

Let P_{19} be mean profit in pre-covid period and P_{21} be mean profit in the recovery phase.

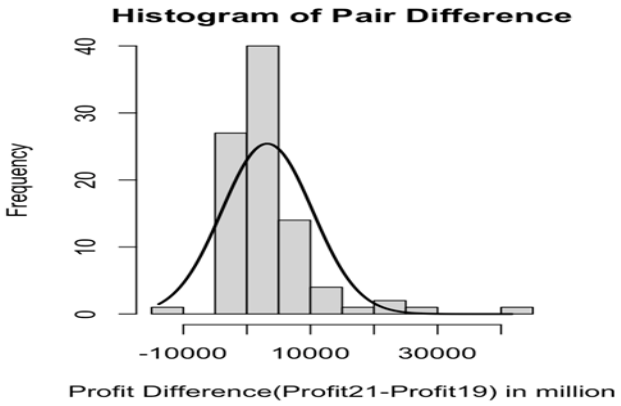
$$H_0: P_{21} - P_{19} = 0, \quad H_a: P_{21} - P_{19} < 0$$

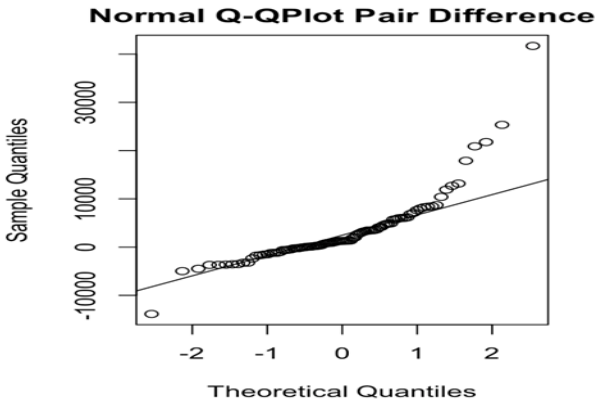
There are three assumptions needed to be satisfied in order to perform paired T-test:

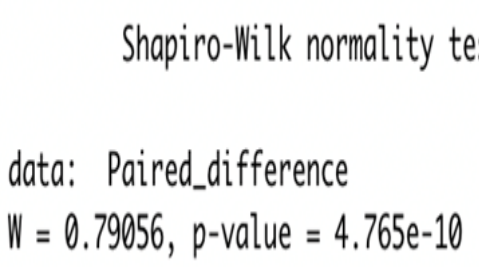
1. Independence: the observations are independent from others
2. Normality: The distribution of difference should be approximately normal.
3. No extreme outliers: No extreme outliers in the difference

For the independence assumption, observations were collected independently from others. Hence this assumption is satisfied.

For normality and no extreme outlier assumptions, QQ plot and Shapiro-Wilk test were used to determine the normality and boxplot was used to detect the extreme outliers.







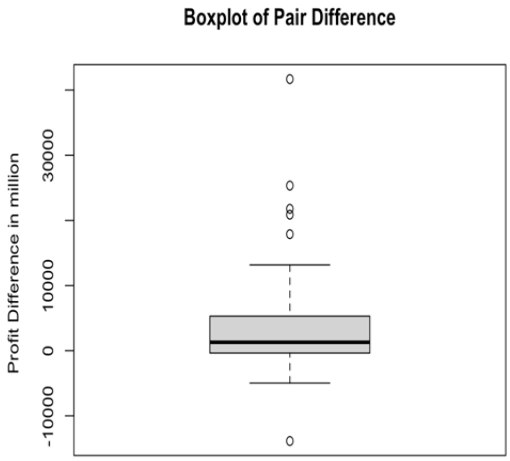


Figure 3-6. Figure 3 on upper left is histogram of paired difference overlaying with normal curve. Figure 4 on the upper right is the QQ plot of paired difference. Figure 5 on lower left is the output of Shapiro-wilk normality test on paired difference. Figure 6 on the lower right is the boxplot of pair difference. Figures are generated using RStudio.

Figures 3 and 4 above clearly show that the distribution of paired difference data is not normally distributed. It is a right-skewed distribution with a longer tail at the right end of the curve. The result of the Shapiro-Wilk normality test in figure 5 is consistent with the non-normality result as p-value is less than 0.05. Moreover, the boxplot in figure 6 shows there are some outliers in the data. Therefore, the assumption of normality and no extreme outlier are violated.

In order to meet the assumptions of paired T-test, data transformation and removal of outliers are needed to make data more closely to the normal distribution and remove extreme outliers from the data. For data transformation, we decided to use log function:

$$Log(x + b)\ , where\ x\ is\ profit\ \ and\ b\ is\ a\ constant$$

The reason for adding b is that the profit value of some companies is negative; therefore, b should be greater than the absolute value of min(x) in order to shift all profits to positive values. the minimum profit value is -6520 (figure 2); therefore, 7000 is chosen for b.

For outlier removal, we decided to use IQR method:

$$Interval: (Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$$
$$where\ Q1\ is\ first\ quartile\ and\ Q3\ is\ third\ \ quartile\ and\ IQR\ is\ Q3 - Q1$$

Note: first quartile and third quartile for profit 2019 and profit 2021 are listed on the figure 2.
Any value not within the interval is considered an outlier and will be removed from the data. As a result, 11 observations are removed and 80 observations remaining.

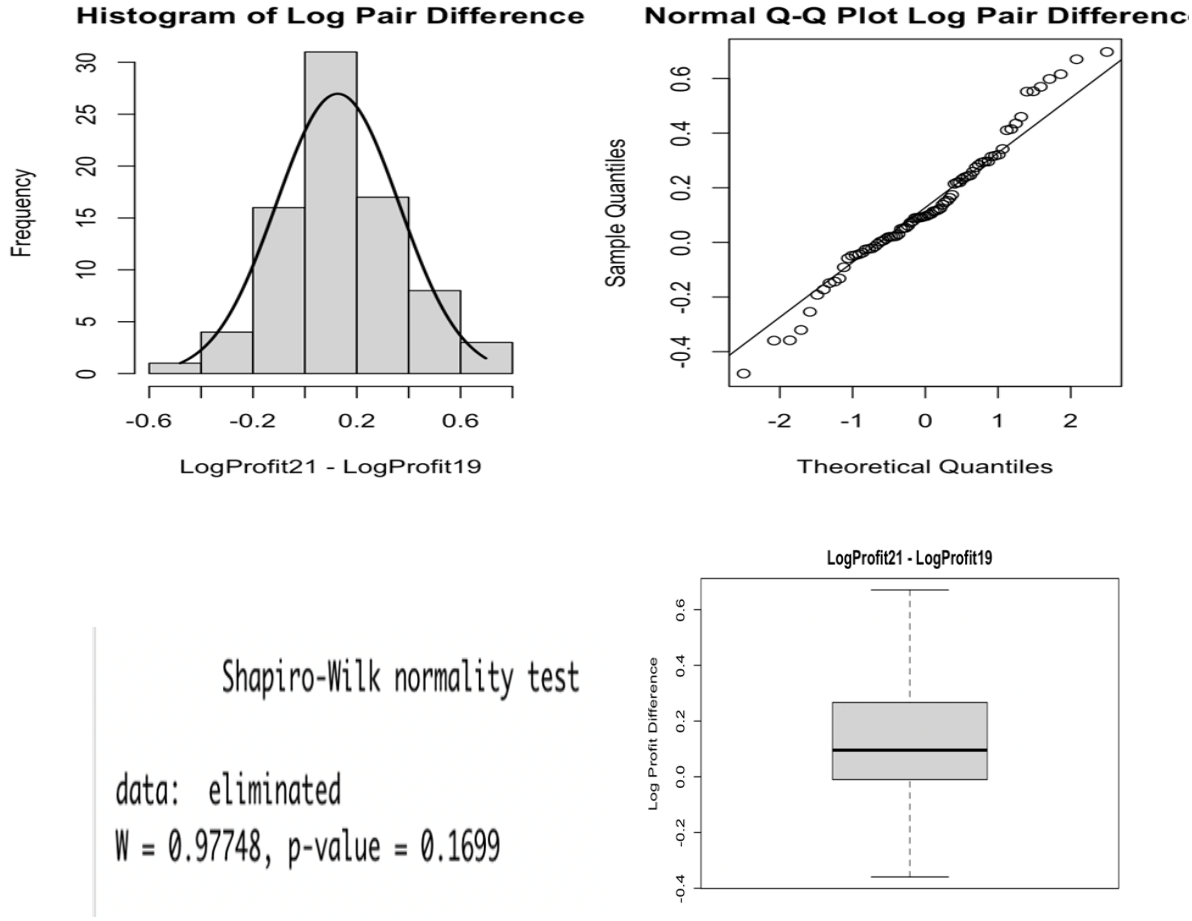


Figure 6-9. Figure 6 on upper left is histogram of log paired difference after log function transformation and removal of outliers. Figure 7 on upper right is QQ plot of log paired difference after log function transformation and removal of outliers. Figure 8 is the output of the Shapiro-Wilk normality test for paired difference data. Figure 9 is a boxplot of log paired difference after log function transformation and removal of outliers. The figures are generated using RStudio.

As can be seen in figure 6, the log paired difference data appear to be closer to a normal distribution, looking like a bell curve with the midpoint having the highest frequency and falling symmetrically around that point. In figure 7, the QQ plot shows that the data points roughly form a straight line. The result of the Shapiro-Wilk normality test is consistent with the normality results as p-value is greater than 0.05.

Moreover, the boxplot in figure 9 shows there are no outliers after the removal of outliers. Therefore, the normality and no extreme outlier assumptions are finally satisfied.

Result of paired T test:

$$H_0: P_{21} - P_{19} = 0, \quad H_a: P_{21} - P_{19} < 0$$

$$t = \frac{d - 0}{s/\sqrt{n}}$$

```
        One Sample t-test

data:  eliminated
t = 4.7811, df = 79, p-value = 1
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
      -Inf 0.1705388
sample estimates:
mean of x
0.1265017
```

The t-value is 4.7811 and the corresponding p-value is 1. P-value is greater than 0.05; hence, we fail to reject null hypothesis. Therefore, 2021 profit is not significantly less than 2019 profit. This result suggests there is increased economic activity in the recovery phase. With the help of vaccination, people can safely return to work, shop and travel.

Figure 10. the output of T-test in RStudio.

Power calculation:

```
     Paired t test power calculation

              n = 80
              d = 0.5345432
      sig.level = 0.05
          power = 0.0000000000853384
    alternative = less

NOTE: n is number of *pairs*
```

Even though we fail to reject null hypothesis, we still want to know the probability of rejecting null hypothesis when it is false (power). As can be seen from the figure 11, power is an extremely small value, which is consistent with the result of failing to reject null hypothesis.

Figure 11. Output of power calculation in RStudio

Paired T test under assumption of MNAR

Missing values we have on the market cap should be considered missing not at random (MNAR) value because whether a row has a missing value on the market cap variable depends on the type of company it is. Deleting rows with missing values would lose all the information about the private companies, in consequence, it might lead to a biased statistical result. Therefore, one way to impute the market cap of private companies is by using median of market cap/asset ratio imputation method (Equation 2)

group	count	min	Q1	median	Q3	max	mean	sd	skewness	kurtosis
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Profit Precovid (in milloin)	100	-4979	2450.575	4313.0	8950.50	81417	7945.944	11572.94	3.778036	21.19372
Profit Recovery (in million)	100	-6520	1725.875	6207.5	13784.32	89795	10829.665	15711.73	2.808098	12.25852

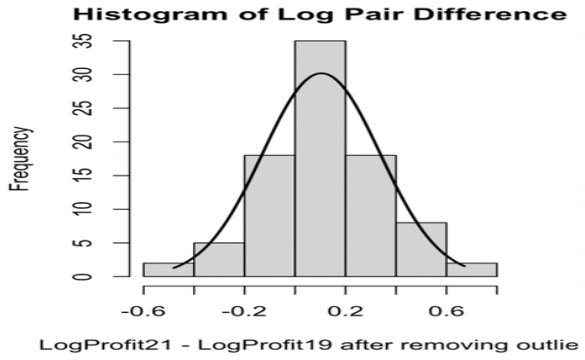
Figure 12. Profit data after median of market cap/asset ratio imputation. Count, minimum, first quartile, median, third quartile, maximum, mean, standard deviation, skewness, and kurtosis of profit in pre-covid period and profit in recovery phase are shown above. The figure is generated using RStudio.

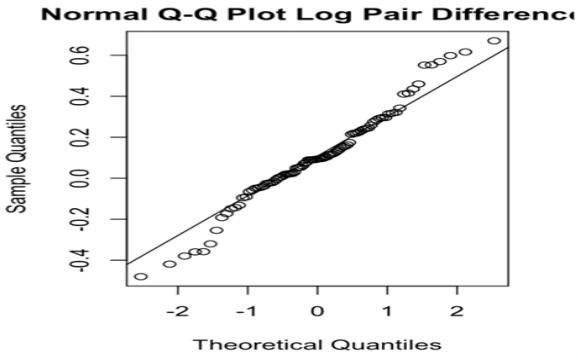
Compared to likely-wise deletion, the imputation method has slightly different Q1, median, Q3, mean, standard deviation, skewness and kurtosis values. As one can see, using different methods to handle the missing values can have different impacts on the variability of the distribution. Therefore, picking an appropriate missing value method is crucial for any statistical study on a dataset with missingness.

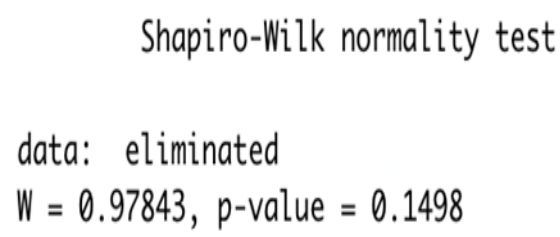
Paired T test for testing whether companies' profit in recovery phase is less than the companies' profit in pre-covid period at $\alpha = 0.05$.

$$H_0: P_{21} - P_{19} = 0, \quad H_a: P_{21} - P_{19} < 0$$

For three assumptions in paired T-test, independence is satisfied because each observation is independent from others. Data transformation and removal of outliers are needed in order to satisfy the normality and no extreme outlier assumptions. The same log transformation and removal of outlier methods are used.







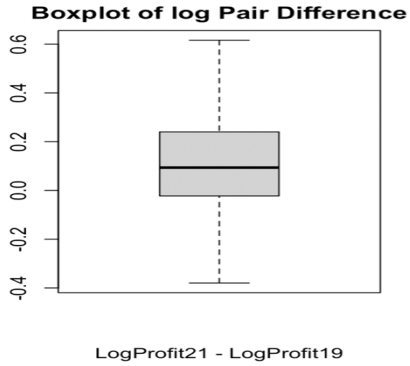


Figure 13-16. Figure 13 on upper left is histogram of log pair difference after log function transformation and removal of outliers. Figure 14 on upper right is QQ plot of log paired difference after log function transformation and removal of outliers. Figure 15 is the output of the Shapiro-Wilk normality test for paired difference data. Figure 16 is a boxplot of log paired difference after log function transformation and removal of outliers.

After log transformation and removal of outliers, the distribution appears to be closer to normal and 12 outliers are removed from the data. Hence, normality and no extreme outlier assumptions are satisfied.

Result of paired T-test:

$$H_0: P_{21} - P_{19} = 0, \quad H_a: P_{21} - P_{19} < 0$$

$$t = \frac{d - 0}{s/\sqrt{n}}$$

```
        One Sample t-test

data:  eliminated
t = 4.1979, df = 87, p-value = 1
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
      -Inf 0.1453212
sample estimates:
mean of x
0.1040949
```

The t-value is 4.1979 and the corresponding p-value is 1. P-value is greater than 0.05, we fail to reject null hypothesis. Therefore, the profit in 2021 is not significantly less than profit in 2019.

Figure 17. output of paired T test in RStudio.

Power calculation:

```
     Paired t test power calculation

              n = 88
              d = 0.4474977
      sig.level = 0.05
          power = 3.111256e-09
    alternative = less

NOTE: n is number of *pairs*
```

Power value is an extremely small, which is consistent with the result of failing to reject null hypothesis

Figure 18. output of power calculation in RStudio.

Overall, the financial performance of large U.S. companies does not appear to differ significantly between the pre-pandemic and recovery phases in both missing values mechanisms. One possible explanation for this is that the company's performance can be categorized as winning, mix-performing, and struggling during the pandemic. Some industries have been hit extremely hard by Covid-19, for example, airlines, entertainment, and fashion retail. As social activity has been significantly reduced during the pandemic, companies in these industries have experienced financial crises, some of which have been permanently closed. Despite this, there are still some winning industries in the pandemic, for example, pharmacy, e-commerce, and remote working software. Companies in these industries have benefited from changes in people's behavior during the pandemic, so profits for these companies have significantly increased. Therefore, the overall financial performance remains unchanged.

Multiple Linear Regression

We have chosen multiple linear regression to study the relationship between profit and the rest of independent variables with the passage of time. The following assumptions were made while taking these points into account:

1) Linear Relationship: It is generally accepted that the higher the revenue, the greater the profit for a company because of the economies of scale. Although there have been examples in history which show that diseconomies of scale is also possible for a company (Example: General Motors), in a general sense higher revenues trickle down to higher profits. In a similar fashion when we plotted Profit vs Asset Size, Profit vs Market Cap, Profit vs Employee size, they all are expected to exhibit a linear relationship. The Scatter plot between profit and all other variables in the data set we have chosen is shown below

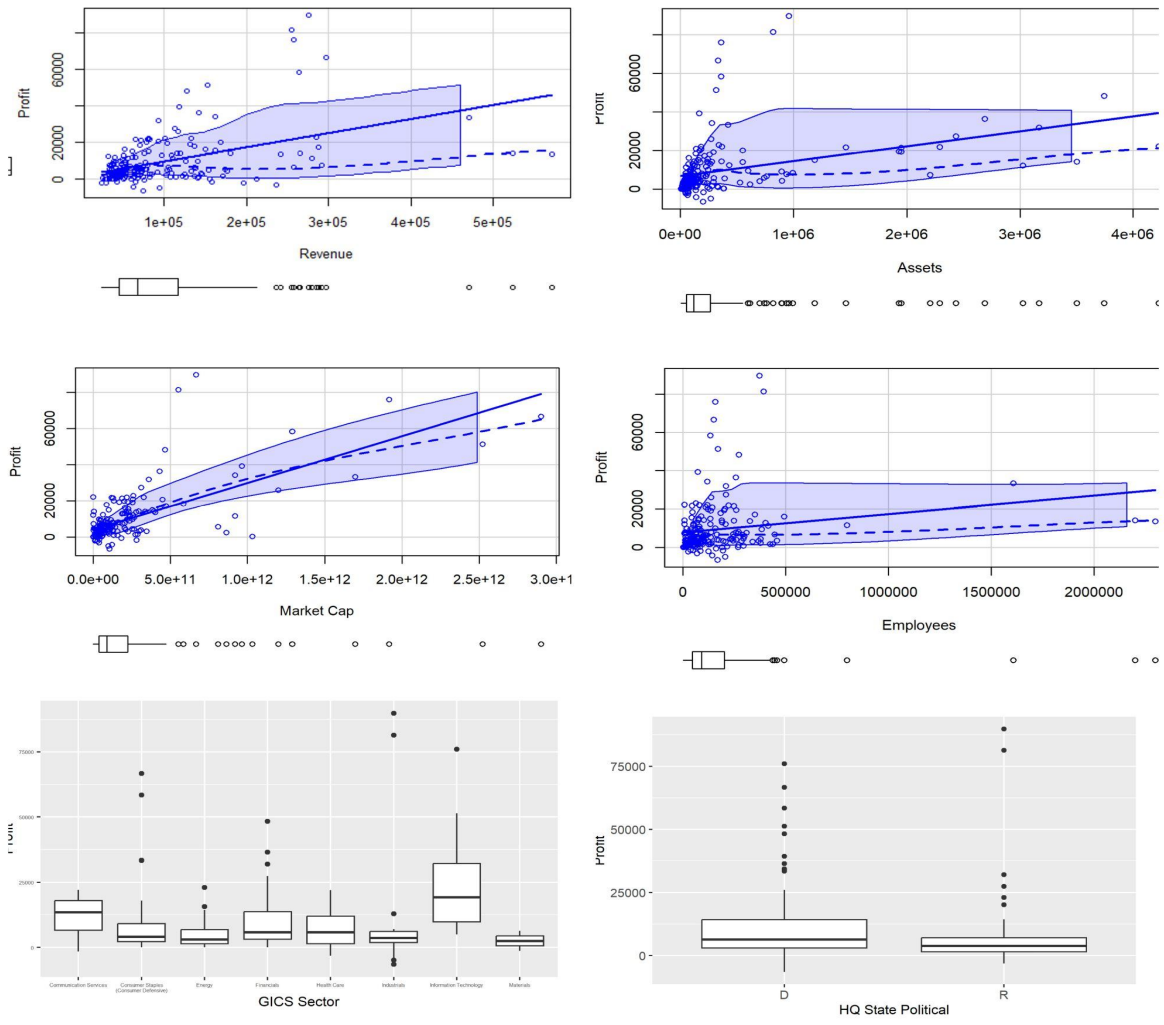


Figure 19-24, first four are scatter plots of Profit vs. Revenues, Assets, Market Cap or Employees, respectively. Last two are boxplots for each sector and HQ state political. Generated using RStudio.

As the above graphs are analyzed it is found that the Employee Size and Year Incorporated has a weak correlation with profit (i.e., correlation coefficient of 182 and 0.133 respectively). Whereas the Asset size and Revenue of these companies have a relatively strong correlation coefficient of 0.378 and 0.456 with profit. Market Cap has a stronger correlation coefficient of 0.69. The above values and graphs show that there is a linear relationship between all of the above-mentioned variables and profit. Although a transformation of independent variables might result in a better correlation; we are sticking to the non-transformed data for the simplicity of this study.

2) Multi-Collinearity: To check the multi collinearity, we have checked the VIF (Variance Inflation Factor) values for all of the independent variables without the interaction terms and they are as follows

	GVIF	Df	GVIF^(1/(2*Df))
Year	1.026640	1	1.013233
`Market Cap`	1.702186	1	1.304679
`Type of Company`	1.825355	1	1.351057
`GICS Sector`	4.012352	7	1.104333
Revenue	1.600070	1	1.264939
Assets	1.852257	1	1.360977
`Year Incorporated`	1.451155	1	1.204639
`HQ State Political`	1.303172	1	1.141565
`CEO Sex`	1.096795	1	1.047280

Figure 25. output of VIF test in RStudio for independent variables

As all of the variables have a VIF values less than 5, the variables do not show multi-collinearity.

3) Independence: A Durbin Watson test is performed to check if the residuals show autocorrelation. The results of this test state that there is a significant autocorrelation between the residuals. The autocorrelation is present as the data is taken from two years (2019 and 2021) for the same companies and the residuals are expected to have some amount of autocorrelation, this can be corrected by adding a lag variable. The study of the lag variable is beyond the scope of this analysis, for the simplicity of this study the autocorrelation is assumed to be negligible.

4) Homoscedasticity: The residuals vs predicted values graph is shown below and it can be seen that there is a minor amount of heteroscedasticity present in the multiple linear regression. This heteroscedasticity problem can be ignored as it is infinitesimal.

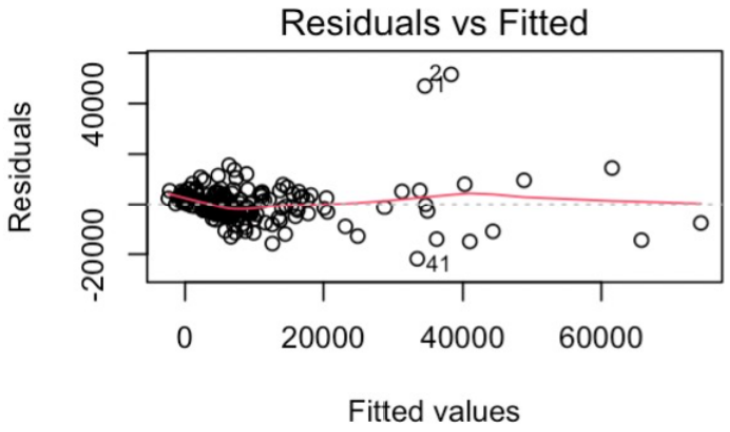


Figure 26. a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. Generated using RStudio

5) Multivariate Normality: A Q-Q plot between Standardized residuals and theoretical quantiles is shown below, the points on the plot roughly follow the diagonal straight line, hence the multivariate normality assumption is met.

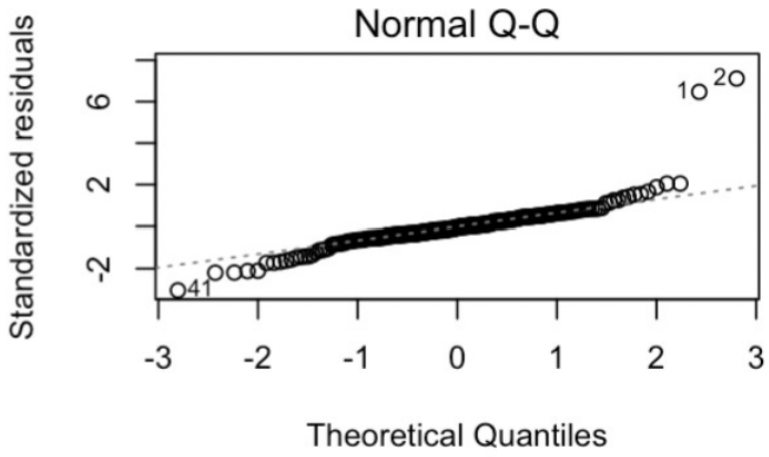


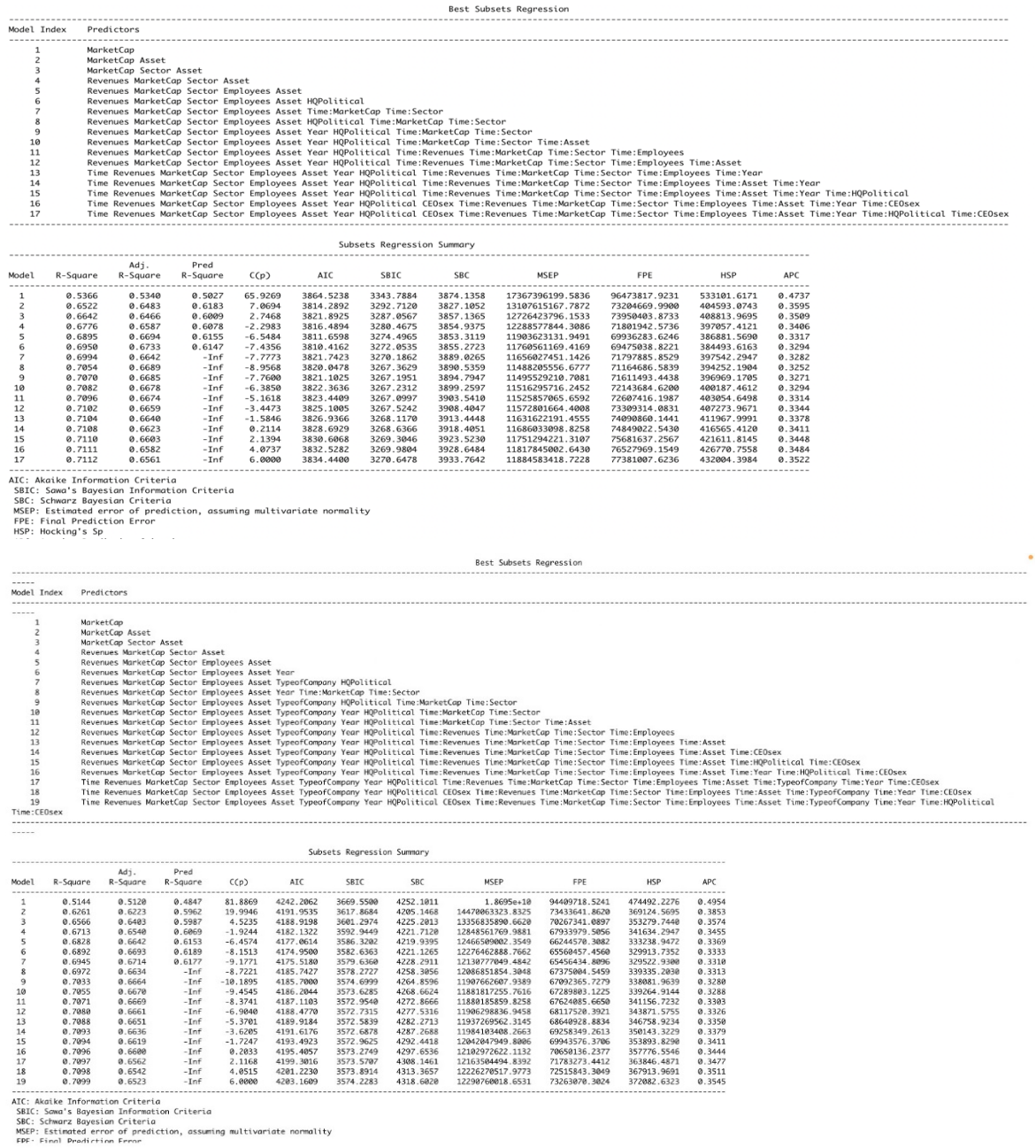
Figure 27. QQ plot of residuals. Generated using RStudio.

6) Higher order Interaction Terms: As our study has interaction between time and various other independent variables, we are narrowing down our interaction terms to two (for example: time & revenue, time & market cap) and interaction terms >3 (for example: time, revenue and market cap) are ignored for the simplicity of the study.

Variable Selection for regression:

As we are conducting a multiple variable regression to understand the dependence of profit of the Fortune100 on different independent variables, firstly we wanted to understand the independent variable

selection for both MCAR and MNAR methods. The following figures show the analysis of selecting different variables to model the profit,



Best Subsets Regression

Model Index	Predictors
1	MarketCap
2	MarketCap Asset
3	MarketCap Sector Asset
4	Revenues MarketCap Sector Asset
5	Revenues MarketCap Sector Employees Asset
6	Revenues MarketCap Sector Employees Asset HQPolitical
7	Revenues MarketCap Sector Employees Asset Time:MarketCap Time:Sector
8	Revenues MarketCap Sector Employees Asset HQPolitical Time:MarketCap Time:Sector
9	Revenues MarketCap Sector Employees Asset Year HQPolitical Time:MarketCap Time:Sector
10	Revenues MarketCap Sector Employees Asset Year HQPolitical Time:MarketCap Time:Sector Time:Asset
11	Revenues MarketCap Sector Employees Asset Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees
12	Revenues MarketCap Sector Employees Asset Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset
13	Time Revenues MarketCap Sector Employees Asset Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Year
14	Time Revenues MarketCap Sector Employees Asset Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:Year
15	Time Revenues MarketCap Sector Employees Asset Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:Year Time:HQPolitical
16	Time Revenues MarketCap Sector Employees Asset Year HQPolitical CEOsex Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:Year Time:CEOsex
17	Time Revenues MarketCap Sector Employees Asset Year HQPolitical CEOsex Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:Year Time:HQPolitical Time:CEOsex

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.5366	0.5340	0.5027	65.9269	3864.5238	3343.7884	3874.1358	17367396199.5836	96473817.9231	533101.6171	0.4737
2	0.6522	0.6483	0.6183	7.0694	3814.2892	3292.7120	3827.1052	13107615167.7872	73204669.9900	404593.0743	0.3595
3	0.6642	0.6466	0.6009	2.7468	3821.8925	3287.0567	3857.1365	12726423796.1533	73950403.8733	408813.9695	0.3509
4	0.6776	0.6587	0.6078	-2.2983	3816.4894	3280.4675	3854.9375	12288577844.3086	71801942.5736	397057.4121	0.3406
5	0.6895	0.6694	0.6155	-6.5484	3811.6598	3274.4965	3853.3119	11903623131.9491	69936283.6246	386881.5690	0.3317
6	0.6950	0.6733	0.6147	-7.4356	3810.4162	3272.0535	3855.2723	11760561169.4169	69475038.8221	384493.6163	0.3294
7	0.6994	0.6642	-Inf	-7.7773	3821.7423	3270.1862	3889.0265	11656027451.1426	71797885.8529	397542.2947	0.3282
8	0.7054	0.6689	-Inf	-8.9568	3820.0478	3267.3629	3890.5359	11488205556.6777	71164686.5839	394252.1904	0.3252
9	0.7070	0.6685	-Inf	-7.7600	3821.1025	3267.1951	3894.7947	11495529210.7081	71611493.4438	396969.1705	0.3271
10	0.7082	0.6678	-Inf	-6.3850	3822.3636	3267.2312	3899.2597	11516295716.2452	72143684.6200	400187.4612	0.3294
11	0.7096	0.6674	-Inf	-5.1618	3823.4409	3267.0997	3903.5410	11525857065.6592	72607416.1987	403054.6498	0.3314
12	0.7102	0.6659	-Inf	-3.4473	3825.1005	3267.5242	3908.4047	11572801664.4008	73309314.0831	407273.9671	0.3344
13	0.7104	0.6640	-Inf	-1.5846	3826.9366	3268.1170	3913.4448	11631622191.4555	74090860.1441	411967.9991	0.3378
14	0.7108	0.6623	-Inf	0.2114	3828.6929	3268.6366	3918.4051	11686033098.8258	74849022.5430	416565.4120	0.3411
15	0.7110	0.6603	-Inf	2.1394	3830.6068	3269.3046	3923.5230	11751294221.3107	75681637.2567	421611.8145	0.3448
16	0.7111	0.6582	-Inf	4.0737	3832.5282	3269.9804	3928.6484	11817845002.6430	76527969.1549	426770.7558	0.3484
17	0.7112	0.6561	-Inf	6.0000	3834.4400	3270.6478	3933.7642	11884583418.7228	77381007.6236	432004.3984	0.3522

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp

Best Subsets Regression

Model Index	Predictors
1	MarketCap
2	MarketCap Asset
3	MarketCap Sector Asset
4	Revenues MarketCap Sector Asset
5	Revenues MarketCap Sector Employees Asset
6	Revenues MarketCap Sector Employees Asset Year
7	Revenues MarketCap Sector Employees Asset TypeofCompany HQPolitical
8	Revenues MarketCap Sector Employees Asset Year Time:MarketCap Time:Sector
9	Revenues MarketCap Sector Employees Asset TypeofCompany HQPolitical Time:MarketCap Time:Sector
10	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:MarketCap Time:Sector
11	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:MarketCap Time:Sector Time:Asset
12	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees
13	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset
14	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:CEOsex
15	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:HQPolitical Time:CEOsex
16	Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:Year Time:HQPolitical Time:CEOsex
17	Time Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:TypeofCompany Time:Year Time:CEOsex
18	Time Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical CEOsex Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:TypeofCompany Time:Year Time:CEOsex
19	Time Revenues MarketCap Sector Employees Asset TypeofCompany Year HQPolitical CEOsex Time:Revenues Time:MarketCap Time:Sector Time:Employees Time:Asset Time:TypeofCompany Time:Year Time:HQPolitical Time:CEOsex

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.5144	0.5120	0.4847	81.8869	4242.2062	3669.5500	4252.1011	1.8695e+10	94409718.5241	474492.2276	0.4954
2	0.6261	0.6223	0.5962	19.9946	4191.9535	3617.8684	4205.1468	14478063323.8325	73433641.8620	369124.5695	0.3853
3	0.6566	0.6483	0.5987	4.5235	4188.9198	3601.2974	4225.2013	13356835890.6620	70267341.0897	353279.7440	0.3574
4	0.6713	0.6540	0.6069	-1.9244	4182.1322	3592.9449	4221.7120	12848561769.9881	67933979.5056	341634.2947	0.3455
5	0.6828	0.6642	0.6153	-6.4574	4177.0614	3586.3202	4219.9395	12466589002.3549	66244570.3082	333238.9472	0.3369
6	0.6892	0.6693	0.6189	-8.1513	4174.9500	3582.6363	4221.1265	12276462888.7662	65560457.4560	329913.7352	0.3333
7	0.6945	0.6714	0.6177	-9.1771	4175.5180	3579.6360	4228.2911	12130777049.4842	65456434.8096	329522.9300	0.3310
8	0.6972	0.6634	-Inf	-8.7221	4185.7427	3578.2727	4258.3856	12086851854.3048	67375004.5459	339335.2030	0.3313
9	0.7033	0.6664	-Inf	-10.1895	4185.7000	3574.6999	4264.8596	11907662607.9389	67092365.7279	338081.9639	0.3280
10	0.7055	0.6670	-Inf	-9.4545	4186.2044	3573.6285	4268.6624	11881817255.7616	67289803.1225	339264.9144	0.3288
11	0.7071	0.6669	-Inf	-8.3741	4187.1103	3572.9540	4272.8666	11880185859.8258	67624085.6650	341156.7232	0.3303
12	0.7080	0.6661	-Inf	-6.9040	4188.4770	3572.7315	4277.5316	11906298836.9458	68117520.3921	343871.5755	0.3326
13	0.7088	0.6651	-Inf	-5.3701	4189.9184	3572.5839	4282.2713	11937269562.3145	68640928.8834	346758.9234	0.3350
14	0.7093	0.6636	-Inf	-3.6205	4191.6176	3572.6878	4287.2688	11984103408.2663	69258349.2613	350143.3229	0.3379
15	0.7094	0.6619	-Inf	-1.7247	4193.4923	3572.9625	4292.4418	12042047949.8006	69943576.3706	353893.8290	0.3411
16	0.7096	0.6600	-Inf	0.2033	4195.4057	3573.2749	4297.6536	12102972622.1132	70650136.2377	357776.5546	0.3444
17	0.7097	0.6562	-Inf	2.1168	4199.3016	3573.5707	4308.1461	12163504494.8392	71783273.4412	363846.4871	0.3477
18	0.7098	0.6542	-Inf	4.0515	4201.2230	3573.8914	4313.3657	12226270517.9773	72515843.3049	367913.9691	0.3511
19	0.7099	0.6523	-Inf	6.0000	4203.1609	3574.2283	4318.6020	12290760018.6531	73263070.3024	372082.6323	0.3545

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error

Figures 28,29: The above figures show how the various methods of selecting the independent variables changes R-Square and Adjusted R-Square under the assumptions of MCAR and MNAR respectively.

For MCAR variable selection analysis, as the R-Square is highest when we consider all the independent variables along with the interaction terms, we will consider Model 17 in the above figure for further analysis. Although Adjusted R-Square is higher for Model 6, we choose to ignore this as it is only marginally higher than the adjusted R-Square of Model 17. In a similar fashion for the MNAR model, we have chosen Model 19 where the R-Square is highest.

As the assumptions required for a multiple variable linear regression have been satisfied, regression analysis has been performed between the dependent variable 'Profit' and the rest of the variables for both MCAR and MNAR cases.

Regression Analysis under assumption of MCAR:

It can be seen that the independent variables which significantly explain the dependent variable (Profit) are Marketcap, Asset size and the interaction between time and Marketcap. All the other independent variables are not found to be significant in explaining Profit. The R squared for the model is 0.7112, this implies that the model explains 71.12% of the variation in the dependent variable.

Note: The type of company variable was removed from MCAR regression because likewise deletion delete all non-public company info.

```
Call:
lm(formula = Profit ~ Time + Revenues + Revenues * Time + MarketCap +
    MarketCap * Time + Sector + Sector * Time + Employees + Employees *
    Time + Asset + Asset * Time + Year + Year * Time + HQPolitical +
    HQPolitical * Time + CEOsex + CEOsex * Time, data = regdata)

Residuals:
   Min     1Q Median     3Q    Max
-22559  -2768   -554   2772  52563

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -5.413e+04  5.540e+04  -0.977  0.33014
Time                            3.920e+04  7.972e+04   0.492  0.62367
Revenues                        3.014e-02  2.135e-02   1.412  0.16010
MarketCap                       3.376e-02  5.229e-03   6.456 1.37e-09 ***
SectorConsumer Staples         -2.521e+03  4.340e+03  -0.581  0.56225
SectorEnergy                   -4.075e+03  4.875e+03  -0.836  0.40456
SectorFinancials               -2.143e+03  4.842e+03  -0.442  0.65879
SectorHealth Care              -2.758e+03  4.531e+03  -0.609  0.54359
SectorIndustrials               1.863e+03  4.508e+03   0.413  0.68000
SectorInformation Technology   -3.028e+03  5.329e+03  -0.568  0.57069
SectorMaterials                -5.357e+03  9.263e+03  -0.578  0.56390
Employees                      -6.044e-03  5.758e-03  -1.050  0.29548
Asset                           6.249e-03  2.115e-03   2.955  0.00363 **
Year                            2.790e+01  2.758e+01   1.012  0.31337
HQPoliticalR                    2.940e+03  2.124e+03   1.384  0.16831
CEOsexM                        -9.991e+02  3.156e+03  -0.317  0.75201
Time:Revenues                   2.943e-02  2.848e-02   1.033  0.30316
Time:MarketCap                 -1.342e-02  5.852e-03  -2.294  0.02317 *
Time:SectorConsumer Staples     4.200e+02  6.211e+03   0.068  0.94618
Time:SectorEnergy               2.092e+02  6.859e+03   0.031  0.97570
Time:SectorFinancials           1.206e+02  6.812e+03   0.018  0.98589
Time:SectorHealth Care         -2.350e+03  6.418e+03  -0.366  0.71478
Time:SectorIndustrials         -8.260e-01  6.415e+03   0.000  0.99990
Time:SectorInformation Technology  8.619e+03  7.503e+03   1.149  0.25243
Time:SectorMaterials            6.563e+03  1.308e+04   0.502  0.61667
Time:Employees                 -6.457e-03  7.732e-03  -0.835  0.40498
Time:Asset                      1.154e-03  2.638e-03   0.437  0.66248
Time:Year                      -2.031e+01  3.952e+01  -0.514  0.60808
Time:HQPoliticalR              -8.140e+02  2.998e+03  -0.272  0.78637
Time:CEOsexM                    1.567e+03  4.264e+03   0.367  0.71384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8391 on 152 degrees of freedom
Multiple R-squared:  0.7112,    Adjusted R-squared:  0.6561
F-statistic: 12.91 on 29 and 152 DF,  p-value: < 2.2e-16
```

Figure 30. Output of Linear model. Dependent variable is Profit, and independent variable are Time (2019, 2021), Revenues, market cap, sector, employees, asset, year incorporation, CEO sex, and interaction between these variables with time. Generated using RStudio.

Regression Analysis under assumption of MNAR:

The results are similar to MCAR case where Marketcap, Asset size and the interaction between time and Marketcap are found to be significant in explaining Profit. All the other independent variables are not found to be significant in explaining Profit. The R squared for the model is 0.7099, this implies that the model explains 70.99% of the variation in the dependent variable.

```
Call:
lm(formula = Profit ~ Time + Revenues + Revenues * Time + MarketCap +
    MarketCap * Time + Sector + Sector * Time + Employees + Employees *
    Time + Asset + Asset * Time + TypeofCompany + TypeofCompany *
    Time + Year + Year * Time + HQPolitical + HQPolitical * Time +
    CEOsex + CEOsex * Time, data = regdata)

Residuals:
   Min     1Q Median     3Q    Max
-21874  -2874   -268   2816  52155

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -5.825e+04  4.959e+04  -1.175   0.2418
Time                               2.334e+04  7.131e+04   0.327   0.7439
Revenues                           3.193e-02  2.054e-02   1.555   0.1219
MarketCap                          3.213e-02  4.982e-03   6.449 1.18e-09 ***
SectorConsumer Staples            -2.217e+03  4.185e+03  -0.530   0.5969
SectorEnergy                      -4.304e+03  4.725e+03  -0.911   0.3637
SectorFinancials                  -2.735e+03  4.621e+03  -0.592   0.5548
SectorHealth Care                 -2.909e+03  4.400e+03  -0.661   0.5094
SectorIndustrials                  1.835e+03  4.378e+03   0.419   0.6756
SectorInformation Technology      -2.443e+03  5.168e+03  -0.473   0.6371
SectorMaterials                   -5.505e+03  9.007e+03  -0.611   0.5419
Employees                         -6.227e-03  5.581e-03  -1.116   0.2662
Asset                              6.389e-03  1.992e-03   3.208   0.0016 **
TypeofCompanyPrivate to Public     2.963e+02  9.685e+03   0.031   0.9756
TypeofCompanyPublic                3.109e+03  4.327e+03   0.718   0.4735
Year                               2.836e+01  2.562e+01   1.107   0.2699
HQPoliticalR                       3.011e+03  1.983e+03   1.518   0.1308
CEOsexM                           -7.760e+02  3.065e+03  -0.253   0.8005
Time:Revenues                      2.726e-02  2.741e-02   0.994   0.3214
Time:MarketCap                    -1.228e-02  5.582e-03  -2.201   0.0291 *
Time:SectorConsumer Staples        9.213e+02  5.980e+03   0.154   0.8777
Time:SectorEnergy                  4.210e+02  6.651e+03   0.063   0.9496
Time:SectorFinancials              1.335e+02  6.510e+03   0.021   0.9837
Time:SectorHealth Care            -2.004e+03  6.235e+03  -0.321   0.7483
Time:SectorIndustrials             3.607e+02  6.225e+03   0.058   0.9539
Time:SectorInformation Technology  8.578e+03  7.272e+03   1.180   0.2399
Time:SectorMaterials               6.583e+03  1.272e+04   0.517   0.6056
Time:Employees                    -6.273e-03  7.495e-03  -0.837   0.4038
Time:Asset                         1.271e-03  2.490e-03   0.511   0.6103
Time:TypeofCompanyPrivate to Public  2.311e+03  1.371e+04   0.169   0.8663
Time:TypeofCompanyPublic           1.710e+03  6.087e+03   0.281   0.7791
Time:Year                         -1.357e+01  3.669e+01  -0.370   0.7120
Time:HQPoliticalR                 -6.356e+02  2.801e+03  -0.227   0.8207
Time:CEOsexM                       2.136e+03  4.051e+03   0.527   0.5987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8161 on 166 degrees of freedom
Multiple R-squared:  0.7099,    Adjusted R-squared:  0.6523
F-statistic: 12.31 on 33 and 166 DF,  p-value: < 2.2e-16
```

Figure 31. Output of Linear model. Dependent variable is Profit, and independent variable are Time (2019, 2021), Revenues, market cap, sector, employees, asset, year incorporation, CEO sex, and interaction between these variables with time. Generated using RStudio.

Interpretation of significant independent variables for both MCAR and MNAR case:

Market Cap: It can be observed that there is a positive relationship between profit and market cap.

Market Cap = Number of public shares * Share Price (As on 31st December for public companies)

The same value is estimated for private companies as mentioned previously in data collection. There is a positive relationship between Market cap and Profit, the companies with larger market capitalizations are likely to make larger profits. The market cap of a company encapsulates the expectations of the public about that company, and as our study confirms, the expectations are in line with the reported profits.

Asset Size: There is a positive relation between asset size and profit, this indicates that the asset heavy companies have performed well in both the years 2019 and 2021, i.e., being asset heavy helped during pre-covid period and during the recovery period.

Time-Market cap: This is the only significant interaction term and this interaction term is negative, this implies that market cap negatively impacted profit in 2021 when compared to 2019. A higher market cap means that there are higher expectations for that company. The profits in the recovery year 2021 fell short of expectations as the market cap blew up more than what was expected. This could be because of many reasons, one of which could be attributed to excessive COVID relief handouts issued by the government, and some studies point out that much of what was provided to the public in the form of COVID social welfare eventually ended up in the stock market. This could also be because of excessive optimism regarding companies with higher market caps and as the investment amounts increased massively in the larger companies, the profits in these companies weren't able to cope up as much. This variable should be studied further to understand if this trend would continue in 2022.

All the other variables are insignificant and further studies must be conducted to understand the relationship between these variables and profit.

Revenue: Although revenue 'the top line' transforms into the profit 'the bottom line', it is not a significant factor contributing to the model. This can only mean one thing, the way in which the costs are managed by the companies play a significant role in determining the profit.

Sector: It is generally understood that certain sectors perform well in some seasons, but as our study shows much of each sector did not contribute to the profits. This result is inconclusive as during the pandemic and also during normal times some sectors tend to outperform other sectors, for example consumer staples sector performs in a stable way even during recessions and booms in the market. For this reason, much deeper analysis is required to draw a firm conclusion regarding the sector.

Workforce Size (Employees): As our study suggests, there isn't a significant contribution made by the workforce size to the bottom line, again this may vary depending on the sector of the industry. Industries such as consumer staples in general require more personnel than IT companies. A deeper analysis must be made sector wise to draw a conclusion for each sector.

Type of company: As our data set contains only 8 private companies and 92 public companies it is harder to draw a conclusion basing solely on this data, a larger sample of private companies must be studied against public companies to understand the effect of the type on the profit.

Year Incorporated: Most of the companies in the dataset have been incorporated after 18th century, and it is generally understood that the older the company the more resilient it is (Lindy effect), while this is true it can be seen that there is no significant impact made to the bottom line based on when the company is incorporated.

Head-quarters location - Presidential election winning party: As we included this variable, we just wanted to study if there is a business-friendly political party based on the headquarter state of the company. The political party which won in the presidential election didn't actually matter in determining the profit for these companies, but it should be understood that this may not be true for smaller companies. This effect must be checked for the smaller companies to understand if the republican or democratic party is creating favorable conditions for business.

CEO Sex: We just wanted to understand if female leadership actually mattered in contributing to the bottom line. As our data consists of the Fortune 100 companies, only 9 and 12 companies had female CEOs in 2019 and 2021 respectively. To understand if the sex of the CEO really matters, we should conduct a deeper study with a larger sample size of female CEO led companies.

Rest of the interaction terms: It can be seen that all other interaction terms except time-market cap interaction term are insignificant, this means that there is no significant interaction between pre-covid and recovery-phase for all the other variables for the Fortune100.

Conclusion:

As we wanted to understand if COVID19 affected the profit levels of the Fortune100, we understood that there isn't any significant difference between the profit levels in 2019 (pre-COVID) and 2021 (Recovery year). This shows that the profit levels of the Fortune100 reached pre COVID levels by the end of 2021.

The second question posed in our research is whether profits of these companies during pre-covid and recovery phase are dependent on a few independent variables, and we found out that market cap, asset size and time-market cap interaction variables are significant contributors to profit. While market cap and asset size are positively correlated with the profits, time-market cap is negatively correlated with profit. This shows that only market cap negatively affected the profit growth during pre-COVID and recovery periods, this could be a result of additional capital which had flown into capital markets because of quantitative easing and under performance of a few Fortune100 companies, but the absolute value of coefficient of time-market cap is an extremely small this implies that this relation may not hold for the future years. Apart from this, having a higher market capital and higher asset size definitely helps a company attain higher profits.

All the other independent variables such as CEO sex, winning political party in the state where the headquarters of a company is located, year incorporated, sector and their respective interaction terms with time variable are insignificant and further study must be conducted with a larger sample size to understand the effects of these variables on profits.

References:

1. “Fortune 500.” Fortune, Fortune, 24 May 2022, https://fortune.com/fortune500/.
2. Edgar: Company Filings.” SEC Emblem, 22 July 2022, https://www.sec.gov/edgar/searchedgar/companysearch.
3. “Paired Samples T-Test in R.” STHDA, http://www.sthda.com/english/wiki/paired-samples-t-test-in-r.
4. Zach. “The Three Assumptions Made in a Paired T-Test.” Statology, 3 Jan. 2022,
5. Zach. “The Five Assumptions of Multiple Linear Regression.” Statology, 16 Nov. 2021, https://www.statology.org/multiple-linear-regression-assumptions/.
6. “Statistics by Jim.” Statistics By Jim, 17 Sept. 2021, http://www.statisticsbyjim.com/.
7. Variable Selection Methods, https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html.
8. Kassambara. “Multicollinearity Essentials and VIF in R.” STHDA, 11 Mar. 2018, http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/.
9. “Olsrr.” RDocumentation, https://www.rdocumentation.org/packages/olsrr/versions/0.5.3.
10. “University of Virginia Library Research Data Services + Sciences.” Research Data Services + Sciences, https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/.
11. *Missing-Data Imputation - Department of Statistics*. http://www.stat.columbia.edu/~gelman/arm/missing.pdf.
12. Christoph Hanck, Martin Arnold. “Introduction to Econometrics with R.” Introduction to Econometrics with R, 6 Oct. 2021, http://www.econometrics-with-r.org/.
13. Getting Started in Fixed/Random Effects Models Using R/Rstudio. https://www.princeton.edu/~otorres/Panel101R.pdf.
14. *Cross Validated*, http://www.stats.stackexchange.com/.
15. *Medium*, https://medium.com/.