Columbia University

Industrial Engineering and Operations Research

Healthcare Stock Market Analysis

by:

Rohan Talati (rst2138), Keshava Dilwali (kd2593) & Akshan Mehta (apm2174)

Date: December 12, 2016

Abstract:

In this project for IEORE4150 course, Introduction to probability and statistics, we have analyzed a group of 10 stocks from the healthcare sector. The main objective of the project is to generate insights for investing in the given stocks by understanding trends and figures plotted using R Studio and presented with Shiny. To start with simple visualization, we observed the log returns of daily closing stock prices followed by linear regression and finally forecasting the future stock values.

1. Data Set

In this project, the data set comprised of daily closing stock prices for ten companies from the Healthcare sector. The companies chosen had similar market capitalization to ensure that the comparison was commensurable. Data was recorded for each trading day from December 1, 2015 to November 30, 2016. All the ten companies in our data set are listed on the New York Stock Exchange (NYSE). They are as follows:

- Johnson & Johnson (JNJ)
- Pfizer Inc. (PFE)
- Novartis AG (NVS)
- Merck & Co., Inc. (MRK)
- UnitedHealth Group Inc. (UNH)
- GlaxoSmithKline (GSK)
- Sanofi SA (SNY)
- AbbVie Inc. (ABBV)
- Abbott Laboratories (ABT)
- Bristol-Myers Squibb Co (BMY)

The stock price data for these companies was obtained from Google Finance. The web link for our application is - https://rinfinity.shinyapps.io/shinyapp/

2. Project Goals

During every U.S. presidential election, the healthcare sector goes through many fluxes creating both opportunities and challenges for investors. It is a topic of great discord between the two major political parties, with each proposing a wide range of regulations on the industry which affect their functioning in the long run. In this project, we choose 10 healthcare related companies and perform statistical analysis on log-returns of their daily closing stock prices during the election period. We use this data to analyze the healthcare sector and forecast future stock behavior. This will provide an investor with insights that can assist in making an informed decision about which stocks to invest in, along with the general market trend.

The healthcare industry comprising of hospital services, medical equipment, pharmaceuticals and biotechnology, is vastly influenced by regulatory and economic policy issues. Even though the demand for these products and services is largely inelastic given its indispensable nature, political rhetoric brings volatility and uncertainty to this market.

Towards this, we examined certain characteristics of each stock taken either one-at-a-time or two-at-a-time. We calculated log-returns of the closing stock prices, and tested if the data was normally distributed. Next, we estimated the confidence intervals for the mean and variance of each stock, and performed a regression of the log-return on time. Lastly, we performed graphical analysis on the stocks to forecast their future values. Subsequently, using two-stocks, we tested the equality of the two population means, and performed a regression of one log-return on another. Consequently, we used this data to forecast the future stock prices.

3. Analysis

3.1 Single Stock Analysis

For each stock, we first examined if log-returns of the daily closing prices were normally distributed by drawing a histogram of the stock data with the normal probability curve, as shown in Figure 1. To further verify our claim, we plotted the data on a q-q normal plot as shown in Figure 2. If the data is normal, then points in the Q-Q normal plot will lie on the straight line. Our stock data at the tail does not show normal behavior. However, it does demonstrate that normality is a reasonably good approximation for the distribution.

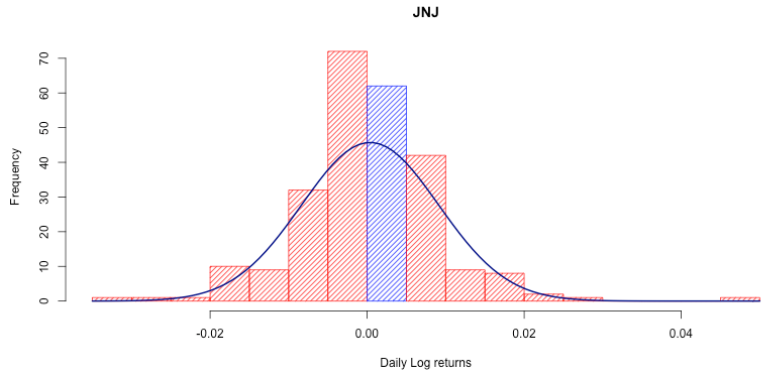

Fig 1: Histogram for JNJ stock data with the normal probability curve

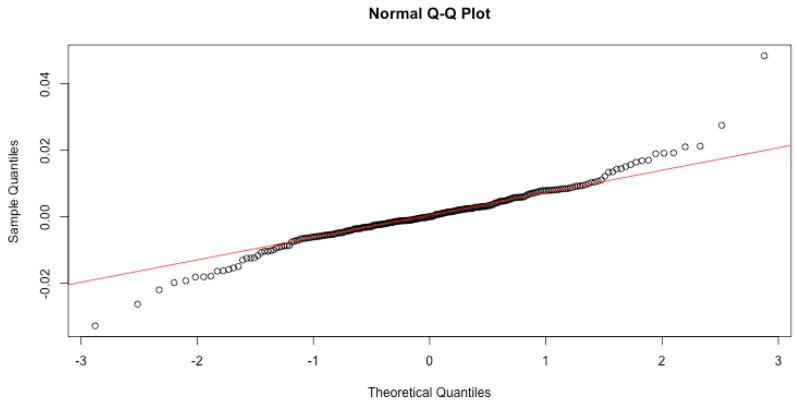

Fig 2: Q-Q Plot for JNJ stock data

We then calculated an approximate confidence interval for the population mean and variance for a user-defined confidence level. At 95% confidence level, the confidence intervals for each stock are presented in Table 1.

Stock	Confidence Interval for Mean (μ)	Confidence Interval for Variance (σ^2)
JNJ	[-0.00071, 0.00146]	[6e-05, 9e-05]
PFE	[-0.00168, 0.00127]	[0.00012, 0.00017]
NVS	[-0.00241, 0.00065]	[0.00013, 0.00018]
MRK	[-0.00115, 0.0022]	[0.00015, 0.00022]
UNH	[-3e-04, 0.00273]	[0.00013, 0.00018]
GSK	[-0.0018, 0.00127]	[0.00013, 0.00018]
SNY	[-0.00214, 0.00146]	[0.00018, 0.00025]
ABBV	[-0.00204, 0.00238]	[0.00027, 0.00038]
ABT	[-0.0026, 0.00133]	[0.00021, 3e-04]
BMY	[-0.00314, 0.00161]	[0.00031, 0.00044]

Table 1: Confidence Intervals for mean and variance at 95% confidence level

We found that stock BMY had the highest volatility (0.0191) and thus the highest risk, while, stock JNJ had the lowest volatility (0.0087), and thus the lowest risk. Next, we performed regression of the stock log-returns on time as shown in Figure 3. It shows us how the stock JNJ varies over a period of the last 250 trading days.

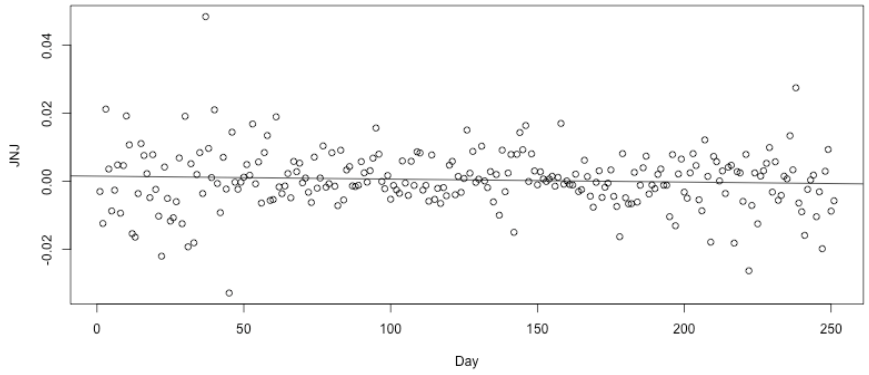

Fig 3: Regression of JNJ stock log returns on time

In this project, we aimed to build a regression model to forecast the future values of each stock. Before forecasting, we need to study the behavior of the independent predictors (in this case, each stock) using different visualization techniques. First, each stock was plotted on a correlation matrix plot to measure linear dependencies between them. A value of +1 or -1 suggests high correlation, while a value of 0 suggests weak relationship between the variables. Figure 4 shows the correlation matrix between

stocks. We can see that stocks BMY and MRK have the least correlation while SNY and GSK have the highest correlation.

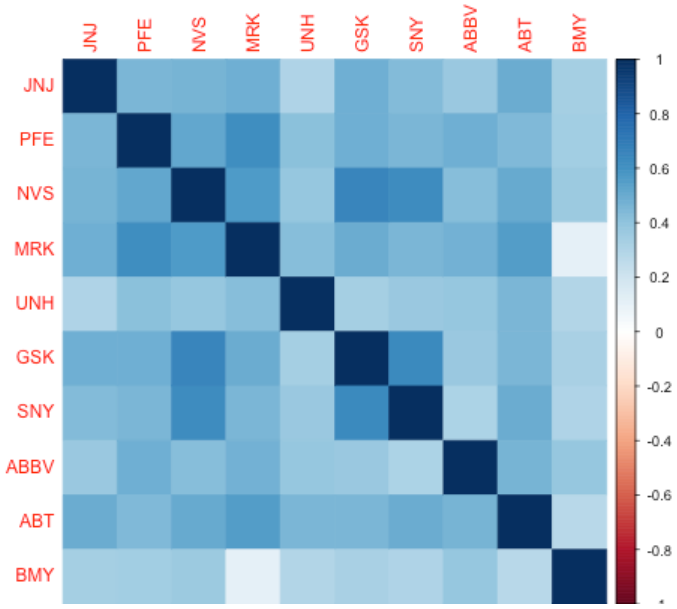

Fig 4: Correlation Matrix between each pair of stock

Next, a box plot was drawn for each stock to find any outlier observations in the data. These outliers can skew the slope of the best-fit line, thereby affecting our prediction. A data point lying outside 1.5 * Interquartile range is said to be an outlier and must be taken into consideration during forecasting. Figure 5 shows the box plot for JNJ along with its outlier data points.

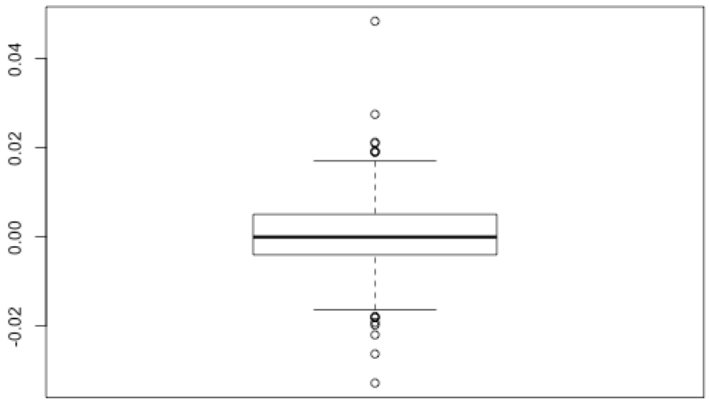

Fig 5: Box plot for JNJ

Finally, a density plot of the stocks was drawn to see the distribution of the predictor variable. Figure 6 shows the density plot for JNJ.

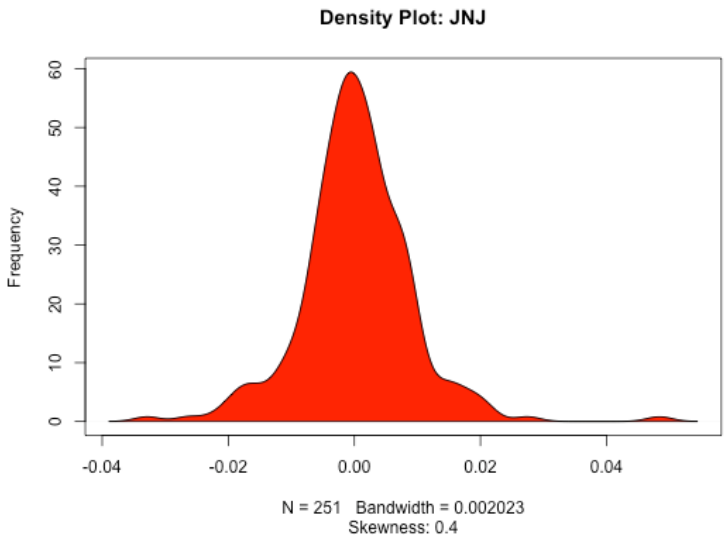

Fig 6: Density plot for JNJ

Using this information, we then proceeded to predict the future stock prices using the given sample data. Figure 7 shows the forecast for JNJ at 95% confidence level using 50 days of sample data.

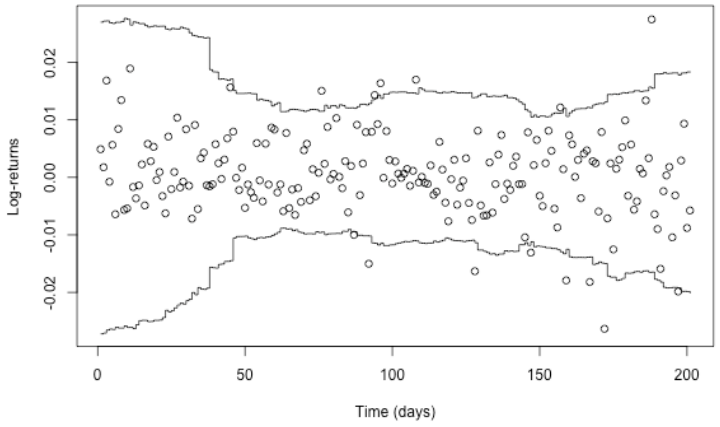

Fig 7: Forecast for JNJ at 95% confidence level

3.2 Two Stock Analysis

For each pair of stocks, we first tested the equality of two population means using t-test. Here we set the null hypothesis H_0 as the event where the two population means were equal. In the case of JNJ and PFE, we saw that the null hypothesis can be rejected for a p-value greater than 0.4072. Next, we performed regression of one log-return on another to quantify the trend of the pair of stocks. Here, slope of the trend line shows which stocks have a positive or negative momentum in comparison to the other stocks. Figure 8 shows the regression plot of JNJ on PFE, while Figure 9 shows its residual plot. This residual plot is used to validate our model as it can help to assess if the observed error is consistent with the stochastic error.

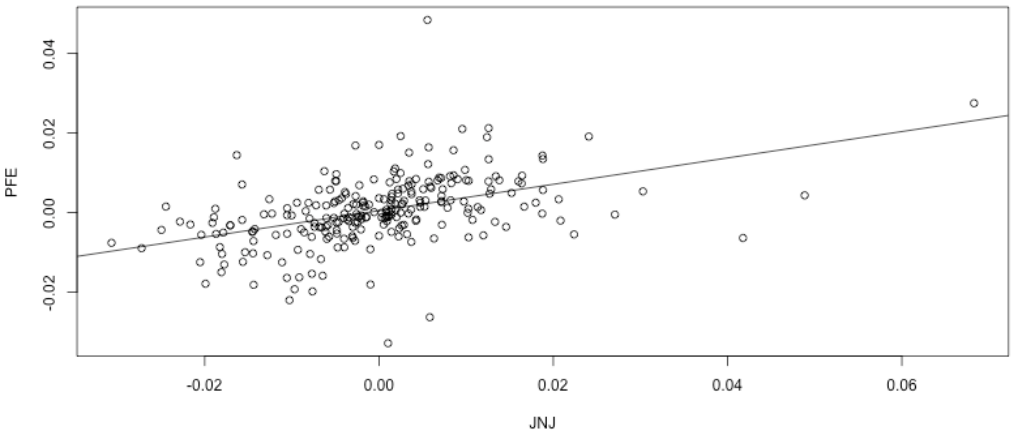

Fig 8: Regression plot for JNJ on PFE

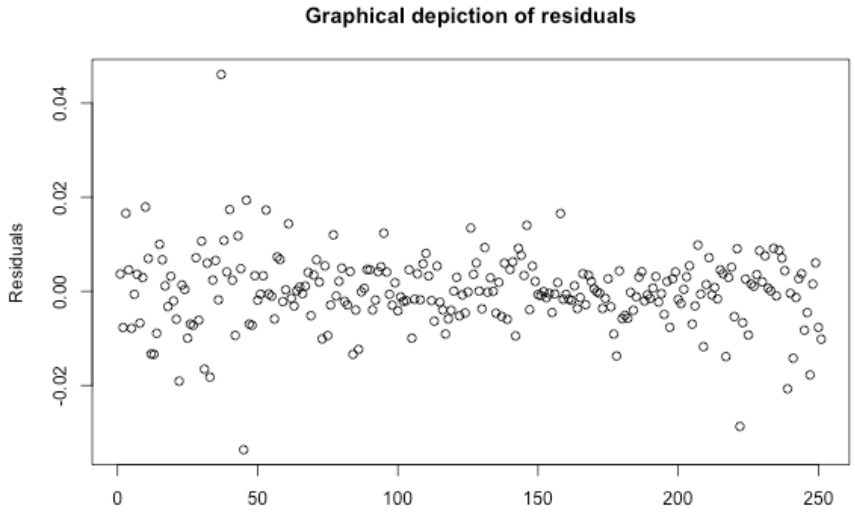

Fig 9: Graphical depiction of residuals for JNJ on PFE

4. Conclusion

In this project, we analyzed daily log-returns of closing prices for healthcare stocks over a period of 1 year. As it was an election year, we expected to see some interesting industry trends affected by political rhetoric around the healthcare sector. In this unpredictable electoral climate, we also wished to predict future stock prices using sample stock data. Towards these goals, using single-stock data we performed tests for normal distribution, estimated the confidence intervals for the mean and variance of the sample data, and performed regression of each stock data on time. Further, we used the available information to create a linear model used for forecasting future trends of each stock. Lastly, by taking a pair of two-stocks at a time, we tested the equality of the two population means, and performed regression for a stock on each other stock.

5. References

I. https://www.thestreet.com/story/13863354/1/4-health-care-stock-picks-for-your-long-term-portfolio.html
II. http://r-statistics.co/Linear-Regression.html