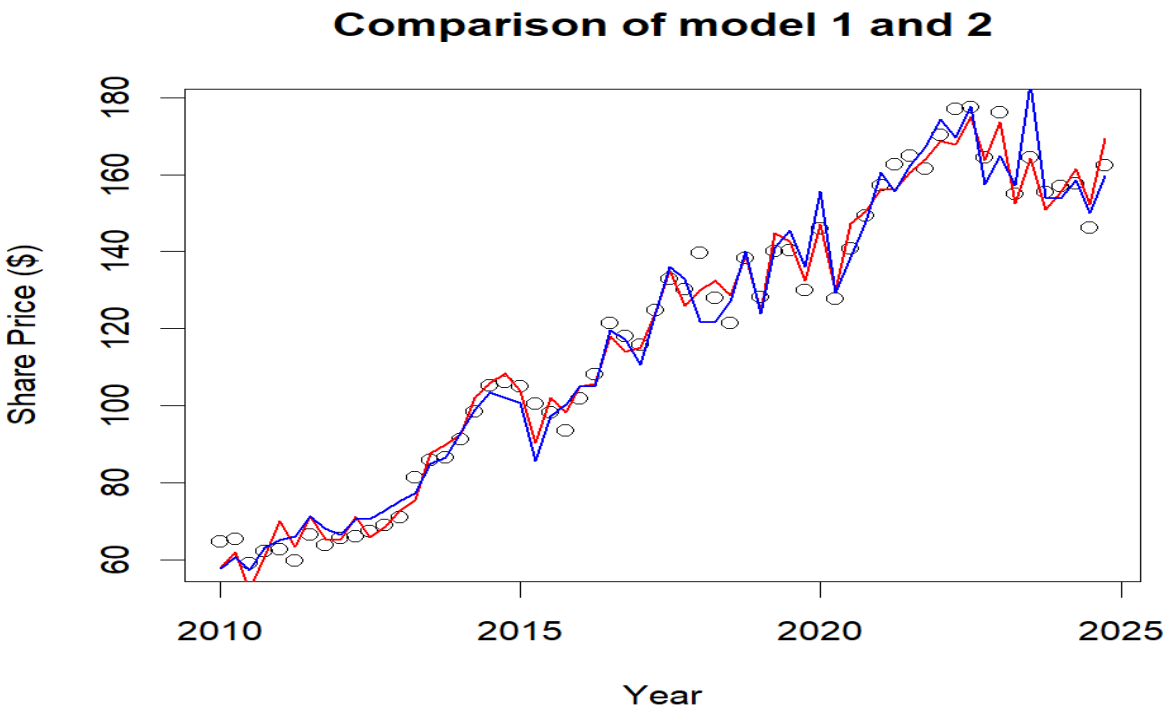
Abhijay Upadhyay

Analyzing and Predicting the Quarterly Share Price of Johnson and Johnson using Macroeconomic, Financial, and Stock Market Data



ABSTRACT:

The purpose of this paper is to predict JNJ's share price using the financial and stock data of various companies as well as macroeconomic factors. We also developed new categorical variables to aid in the prediction of JNJ share price. Two models were made, both achieving an R^2 above 0.975. In the end, we found Shareholder Equity, various parts of BD stock and financial data, and our developed categorical variables to be the best predictors of JNJ's quarterly share price data.

Background Information on JNJ:

The stock which we will be analyzing in this project is Johnson and Johnson, or in its shortened form, JNJ. JNJ is a global health company that deals in medical technology, biotechnology, and pharmaceuticals. It has been around since 1886 and has been publicly traded since 1944. I chose JNJ because, as a business in the industry of health, I feel like its stock wouldn't fluctuate as much as firms in other industries do, since health is a necessity

for everybody. In addition, I found the stock of a long-standing firm would be less susceptible to fluctuations than other firms' stocks. This would make JNJ's stock hopefully follow trends unique from those of other firms. In addition, since JNJ is based in New Jersey, which is where I'm from, I feel like I have context and insights that could allow me to perform better analysis and get better results.

Description of Factors:

For the data, I made 26 columns, all downloaded from either Macrotrends.net or from the Federal Reserve Economic Data (FRED). Data dealing with financial or stock data is from Macrotrends. Data dealing with macroeconomic factors is from FRED. The dependent variable here is Quarterly_Share_Price, which is the share price of JNJ every quarter from 2010 to the end of 2024.

The 25 other variables are:

"Quarter" = The quarter in which the data's observations have been made.

"Time_Index" = The Quarter column turned into an index for easier calculations.

"Quarterly_Revenue" = Revenue in the quarter for JNJ.

"Quarterly_Gross_Profit" = Gross profit in the quarter for JNJ.

"Quarterly_Assets" = Value of the assets JNJ possesses in the quarter.

"Quarterly_Current_Ratio" = Current ratio in the quarter, current assets/current liabilities.

"Quarterly_Quick_Ratio" = Quick ratio in the quarter, (current assets-current inventory)/current liabilities.

"Shareholder_Equity" = Shareholder equity in the quarter.

"Debt_To_Equity_Ratio" = Debt to equity ratio for the quarter.

"TTM_Net_Income" = Trailing twelve months net income, the net income of a quarter a year ago from the current quarter. Can be thought of in terms of Time Series Modelling terms as an autoregressive function with lag 4.

"ROI" = Return on investment in the quarter.

"ROE" = Return on equity in the quarter.

"JNJ_Quarterly_EPS" = Earnings per share in the quarter.

"JNJ_PE_Ratio" = PE ratio in the quarter, Share Price/EPS.

"Merck_Share_Price" = Share price of Merck, a company in a similar industry and location. Wanted to see if stock changes could have been due to market fluctuation and if Merck and JNJ had similar stock trends.

"Merck_PE_Ratio" = PE ratio for Merck in the quarter.

"Merck_Quarterly_EPS" = Earnings per share for Merck in the quarter.

"BD_Share_Price" = Share price of BD, another company in a similar industry and location, for the quarter.

"BD_PE_Ratio" = PE ratio for BD in the quarter.

"BD_Quarterly_EPS" = EPS for BD in the quarter.

"SP_Decrease_Dummy" = Boolean variable for if the stock price has decreased from the previous quarter.

"Negative_EPS_Dummy" = Boolean variable for if the earnings per share for the quarter was negative.

"Healthcare_Spending" = Overall healthcare spending. Wanted to see if stock growth was driven by higher spending in the healthcare industry.

"CPI" = Consumer Price Index, essentially a benchmark/estimate for inflation. Wanted to see if share price growth was driven by inflation.

"Bad_PE_Dummy" = Boolean variable for if the PE ratio of JNJ for the quarter was over 30. Included because a high PE ratio for an established company could indicate that the company is currently being overvalued and that its share price could decrease in value in the future.

The model:

The model itself is a linear multivariable regression model made in R.

The model parameters are as follows:

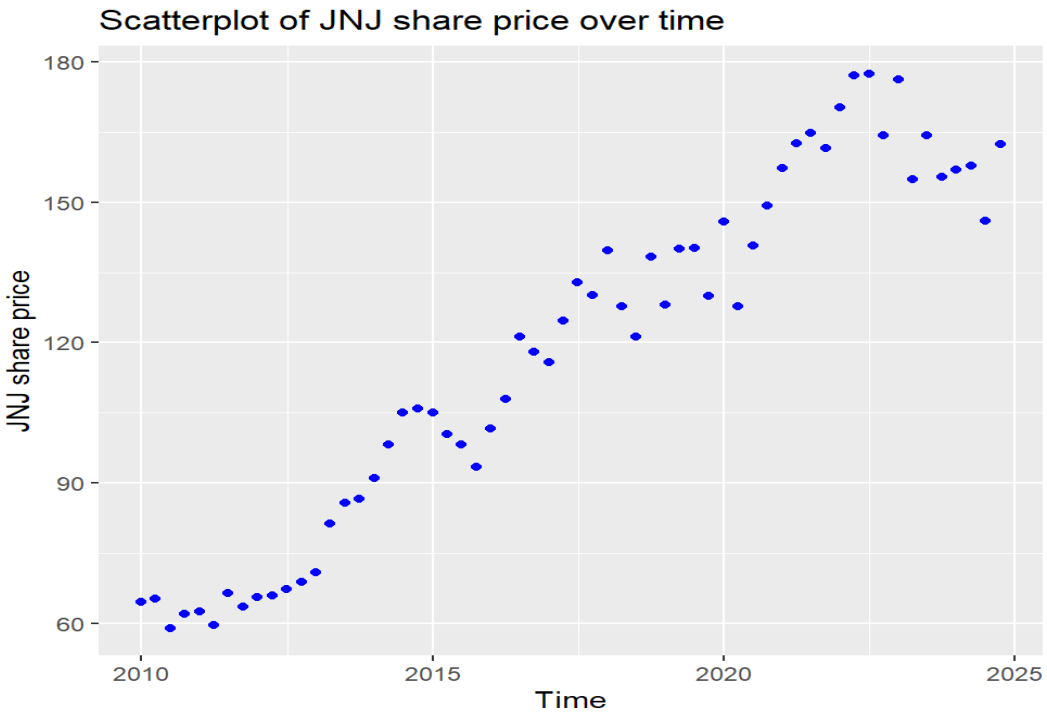
B0 + B1 * x1 + B2 * x2 + … + B24 *x24.

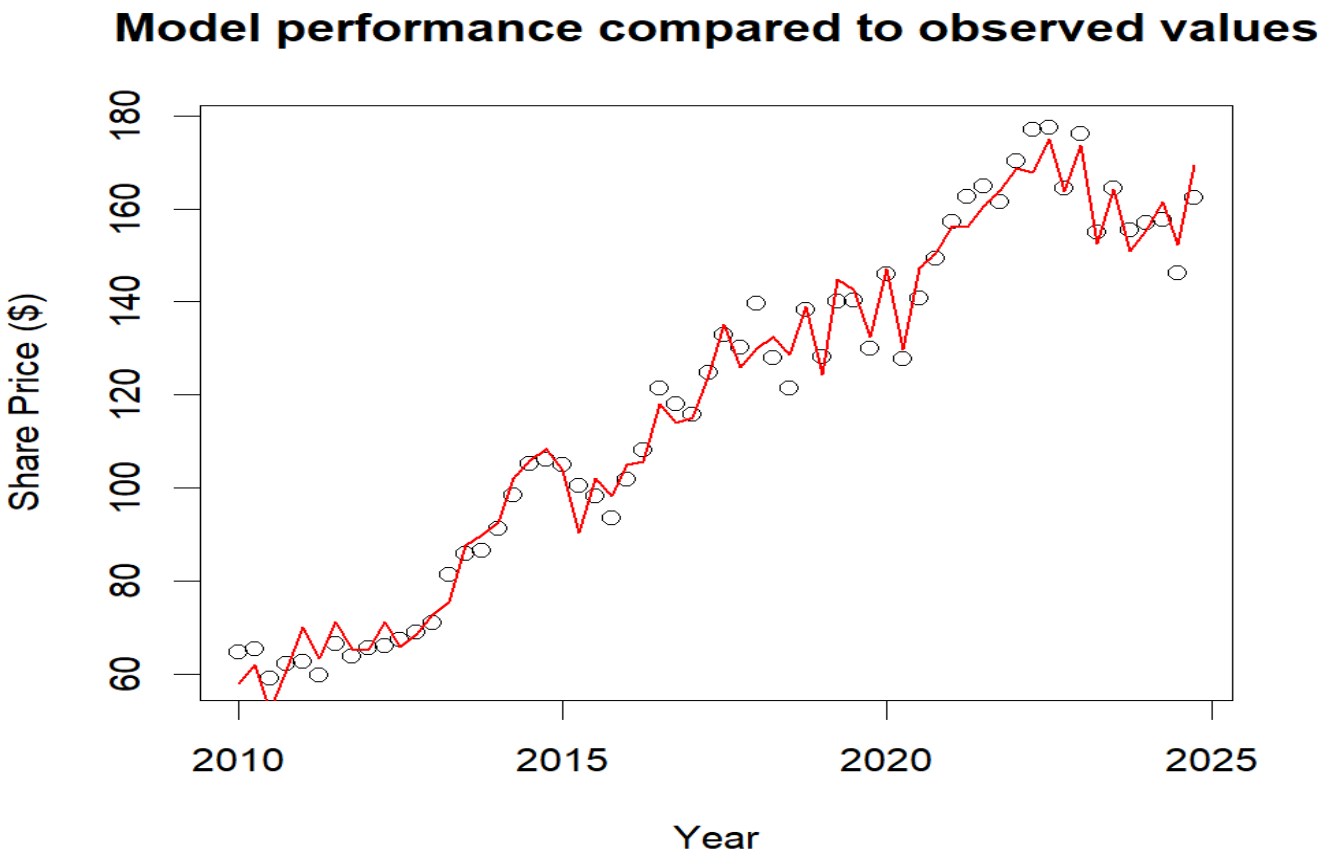
And our model is as follows:

-1.739783e+02 (intercept) + -1.557303e-01 * Time_Index + -2.634603e-0 * Quarterly_Revenue + 4.306976e-03 * Quarterly_Gross_Profit + -2.069066e-02 * Quarterly_Assets + -1.013456e+01 * Quarterly_Current_Ratio + 1.348790e+01 * Quarterly_Quick_Ratio + 1.322770e+00 * Shareholder_Equity + 3.304277e+01 * Debt_To_Equity_Ratio + 1.050314e+00 * TTM_Net_Income + 6.252832e+01 * ROI + -4.233514e+01 * ROE + 6.923931e-01 * JNJ_Quarterly_EPS + 3.763052e-02 * JNJ_PE_Ratio + 1.727485e-02 * Merck_Share_Price + 2.459640e-03 * Merck_PE_Ratio + 1.105520e+00 * Merck_Quarterly_EPS + 3.403821e-01 * BD_Share_Price + -2.572246e-02 * BD_PE_Ratio + -1.332627e-01 * BD_Quarterly_EPS + -7.556970e+00 * SP_Decrease_Dummy + 6.656657e+00 * Negative_EPS_Dummy + -2.814685e-02 * Healthcare_Spending + 6.025647e-01 * CPI + -1.829531e+01 * Bad_PE_Dummy

Our model has an R^2 of 0.9867.

Below we've provided a scatterplot of the observations to give an idea of what the values for share price that we're trying to predict look like. We've also provided a graph that shows our model's predicted values compared to the actual values.


Scatterplot of JNJ share price over time
JNJ share price
180
150
120
90
60
2010
2015
2020
2025
Time

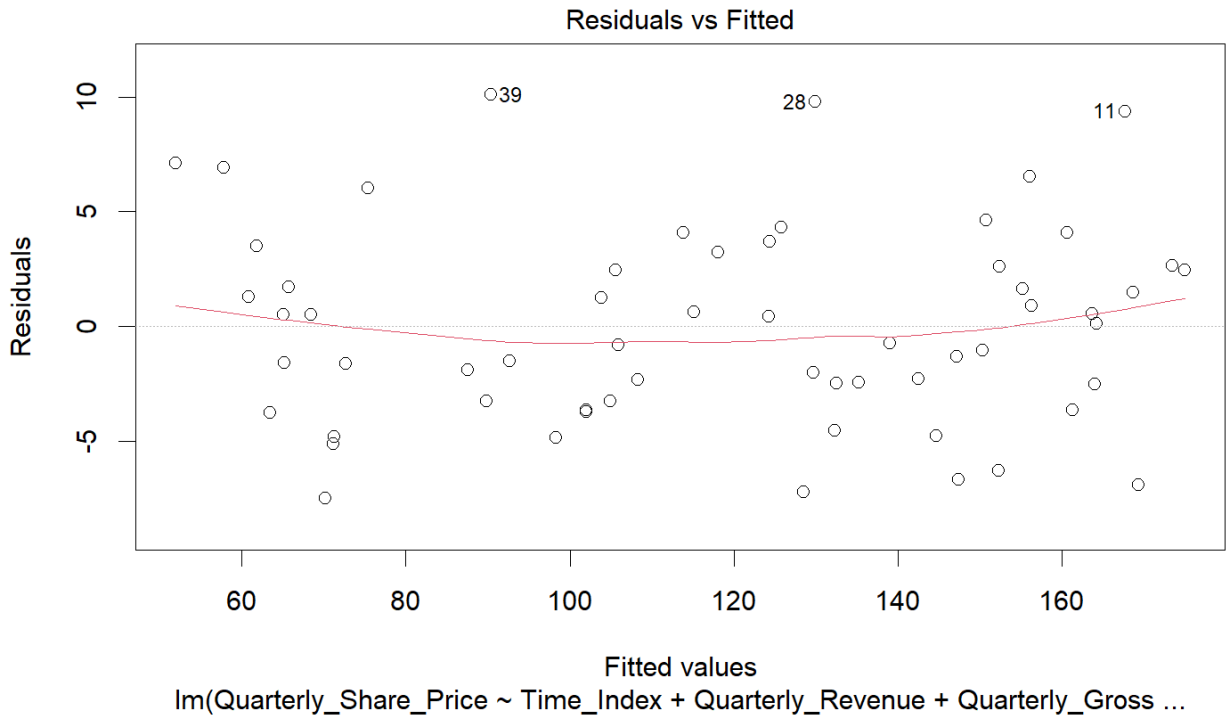

Model performance compared to observed values
Share Price ($)
60
80
100
120
140
160
180
2010
2015
2020
2025
Year

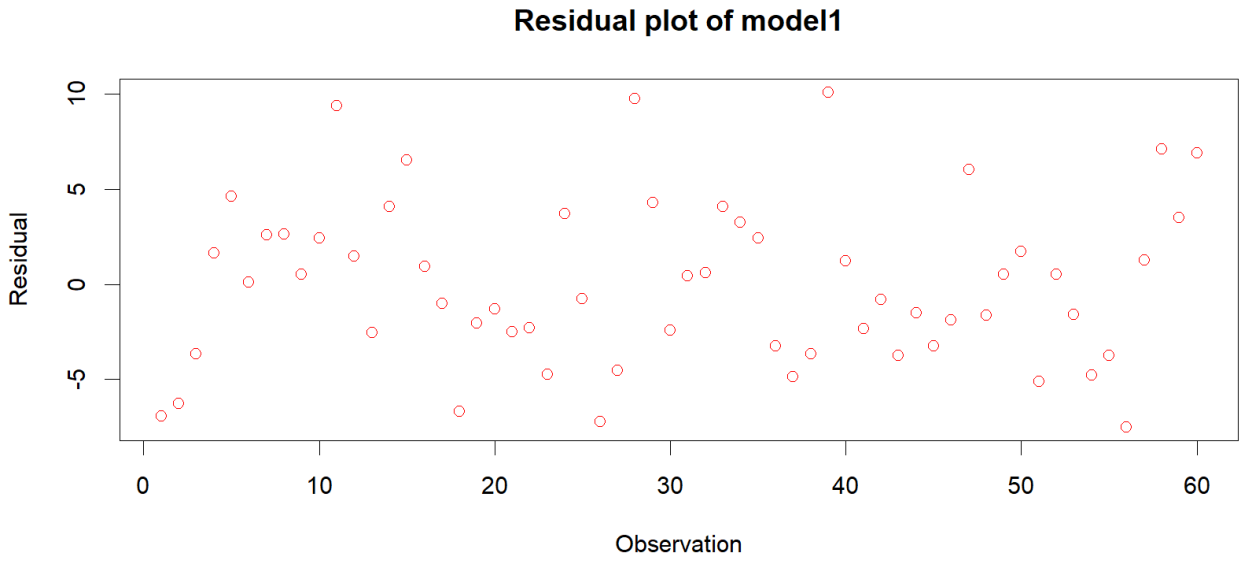
Assumptions of the model and Recommendations in case of violations:

The assumptions of any Ordinary Least Squares Regression model are:

- Linearity
- Independence of error terms
- Homoscedasticity
- No perfect multicollinearity
- Zero mean of errors
- No autocorrelation of errors

For linearity, we're testing if the y variable has a linear relationship with our x variables. The model should have a linear relationship between x and y variables so that it can be optimized to make the most accurate predictions possible. To understand why, think of making a linear model to predict a curved y variable like drawing a curve using a straight line. There will always be some area with errors, and to avoid these errors we can simply turn the curve to a straight line. To check linearity, we look at the regression plot of the model:





Looking at our plots, we see that the general trend of the residuals (the red line) isn't a straight line, but rather that it is a curve. What this implies is that using share price may not be the best course of action. It may be better to transform the share price by finding its square root or share price, or use the natural log of share price, so that a linear relationship between our x and y variables may be achieved, which would allow our model to make better predictions.

No perfect multicollinearity means that our variables are not perfectly correlated to each other. Perfect multicollinearity needs to be avoided because it means spreading the influence of 1 variable over many variables, decreasing the significance of any specific variable and giving the impression that a useful variable may be less important than it actually is. To test for perfect multicollinearity, we use code vif(model1).

```
Time_Index     Quarterly_Revenue
        328.721211          116.480701
Quarterly_Gross_Profit      Quarterly_Assets
        137.297236          32.407498
Quarterly_Current_Ratio  Quarterly_Quick_Ratio
        2805.006794         3077.804558
  Shareholder_Equity   Debt_To_Equity_Ratio
        28.050770           109.699280
     TTM_Net_Income                 ROI
        28.811160           37.600113
               ROE     JNJ_Quarterly_EPS
        11.159110           4.585577
      JNJ_PE_Ratio      Merck_Share_Price
        9.734915            53.110665
     Merck_PE_Ratio    Merck_Quarterly_EPS
        2.844045            2.843253
     BD_Share_Price          BD_PE_Ratio
        46.224504           2.207212
    BD_Quarterly_EPS     SP_Decrease_Dummy
        1.784307            1.900065
  Negative_EPS_Dummy    Healthcare_Spending
        2.839392           85.622510
```

A vif score above 5 signifies problematic multicollinearity. In the case of multicollinearity, we should inspect each variable with a vif score above 5 and remove the ones that are insignificant to our model. If a model has a high multicollinearity, it means that it's similar to another variable, so dropping it wouldn't hurt the model significantly.

For independence of error terms and for no autocorrelation of errors, we're checking if our residuals are random and without pattern, i.e. independent of each other. No residual should be influenced by the residuals before it. Independence and no autocorrelation of residuals are important because if there is a pattern in our residuals, it means there's a trend in the observations we're not accounting for, which means our model is suboptimal and that our variables may not be being used to their full potential. In short, if there's a trend in our residuals, it means that one of our variables may be able to predict it and contribute to the model but isn't currently, and so if we let the trend remain, we may not let the variable reveal its significance, and so we'd drop it for future models, leaving us with worse insights and future models. We can get a general idea of if independence of error terms and no autocorrelation of residuals are present through looking at the residual plot given above. But to go more in depth and in order to properly test for model independence of error terms and no autocorrelation of residuals, we use the Durbin-Watson test.

Durbin-Watson test

data: model1

DW = 1.3763, p-value = 5.383e-05

alternative hypothesis: true autocorrelation is greater than 0.

The hypothesis test here is if true autocorrelation is equal to 0 or if it is greater than 0.

H0: True autocorrelation is not greater than 0.

H1: True autocorrelation is greater than 0.

Our significance level is 0.05, and so if the p-value for the test is below that, we say it's too unlikely for our results to simply be random chance and that the alternative hypothesis is true. Here, our p-value is significantly below 0.05, and so we accept the alternative hypothesis H1 that true autocorrelation is greater than 0. This violates the assumptions of independence and no autocorrelation of errors. In order to address this, we should examine the trend in our residuals and see if there's perhaps some way or variable that can explain it. For now, it may just be that the trend is due to the fact that our x variables don't share a linear relationship with our y variable, leaving our predictions too high in one area and too low in another. It could also be quarterly fluctuations in the JNJ share price.

Homoscedasticity is the assumption that the variance of our residuals is constant. In other words, it means that there's no point in which the residuals are larger on average than any other point. Violating homoscedasticity means having heteroscedasticity, and heteroscedasticity is bad because it can give us a poor view of the model's performance, and in addition, the model can be skewed to focus solely on the region with higher residual variance at the expense of everywhere else. To look for homoscedasticity, we can also refer to the residual plot shown above. A better way to check for homoscedasticity is by using the Breusch-Pagan test.

studentized Breusch-Pagan test

data: model1

BP = 21.871, df = 24, p-value = 0.5869

The Breusch-Pagan test has a null hypothesis of homoscedasticity, and an alternative hypothesis of heteroscedasticity. The p-value of 0.5869 is greater than our significance level of 0.05, and so we fail to reject the null hypothesis, and say that the model has homoscedasticity. If our model had heteroscedasticity, something that could be done is transform the y variable so that it had a lower variance at troublesome points, which could be done by taking the natural logarithm or the square root of our dependent variable.

Zero mean of errors is the assumption that the mean of our residuals is equal to 0. This needs to be the case, because if it isn't, it means our model is predicting a less accurate value on average than it could be. This assumption will always be satisfied by the Ordinary Least Squares method of calculating a linear model, and so if we had a model that didn't satisfy zero mean of errors, we would need to consider using an OLS linear model instead of that current model.

Analysis of Significance of Variables:

Some statistical analysis has already been performed to check the Gauss-Markov assumptions listed above. Now, the next phase of this project is to look at a summary of our model and remove any variables deemed insignificant, as well as try to correct any errors that caused violations of Gauss-Markov assumptions.

As a whole, the slopes of our model are jointly significant, as it has a joint p-value below 2 * 10^-16, which is far lower than the significance level of 0.05.

H0: The slopes of the model are not jointly significant

H1: The slopes of the model are jointly significant

the p-value being below 0.05 means we reject H0 and say the slopes of the model are jointly significant. In addition, our model has an R^2 of 0.9867, which is quite high.

Using the code summary(model), we see that these are the variables that are currently significant to our model, and have a p-value below 0.05:

Shareholder_Equity

BD_Share_Price

SP_Decrease_Dummy

Negative_EPS_Dummy

Bad_PE_Dummy

These are variables that are close to our 0.05 significance level threshold:

Debt_To_Equity_Ratio

ROE

Healthcare_Spending

CPI

BD_PE_Ratio

In terms of hypothesis testing,

H0: the slope of the variable isn't significant, and

H1: the slope of the variable is significant.

Only variables with p-values below 0.05 are considered significant. Therefore, we only consider Shareholder_Equity, BD_Share_Price, SP_Decrease_Dummy, Negative_EPS_Dummy, and Bad_PE_Dummy significant. For every other variable we fail to reject H0, and so we can't say the variable is significant. Every variable besides the ones we specifically mentioned above seem pretty insignificant to our model.

Model Modification:

Given the observations above, these variables and a select few other ones that are close to being significant are the only ones that shall be used in the next model. In addition to that, we've changed the predicted variable from Quarterly_Share_Price to new variable called log_SP, which is the log of the Quarterly_Share_Price. We compared using the log of the share price and actual share price, and the log of the share price yielded a slightly yet noticeably better model.

The variables included in our new model are

Shareholder_Equity, BD_Share_Price, SP_Decrease_Dummy, Negative_EPS_Dummy, Bad_PE_Dummy, Debt_To_Equity_Ratio, ROE, Healthcare_Spending, CPI, BD_PE_Ratio

This new model has the parameters:

Log of share price = B0 + B1 * x1 + B2 * x2 + … + B10 * x10

This new model has the formula:

3.3624353 + Shareholder_Equity * 0.0182811 + Debt_To_Equity_Ratio * 0.2331648 + ROE * 0.0038958 + BD_Share_Price * 0.0036880 + BD_PE_Ratio * -0.0001316 + SP_Decrease_Dummy * -0.0548543 + Negative_EPS_Dummy * 0.0111817 + Healthcare_Spending * 0.0001084 + CPI * -0.0037138 + Bad_PE_Dummy * -0.0501736

The new model has an R^2 of .9782, which is really quite close to the R^2 of the old model considering we dropped more than half the variables.

Some things that stand out to me are that BD seems to be very heavily correlated with JNJ, whereas Merck really wasn't. In addition, the variables we either added or invented because they felt useful all seem to have been significant. This clues us in to the idea that using categorical variables and thinking outside the box may lead to better predictions. At the very least, effort should be taken to ensure that most of the variables are not heavily correlated with one another or cover similar data.

In terms of hypothesis testing for the new data,

H0 is that the slopes of the model are not all jointly significant

H1 is that the slopes of the model are all jointly significant

The p-value of the model is 2.2 * 10^-16, which is far less than 0.05, so we reject the null hypothesis and say the slopes of the model are all jointly significant.

For testing the variables individually, we look at their p-values

H0 is that the slopes for the variables are not individually significant

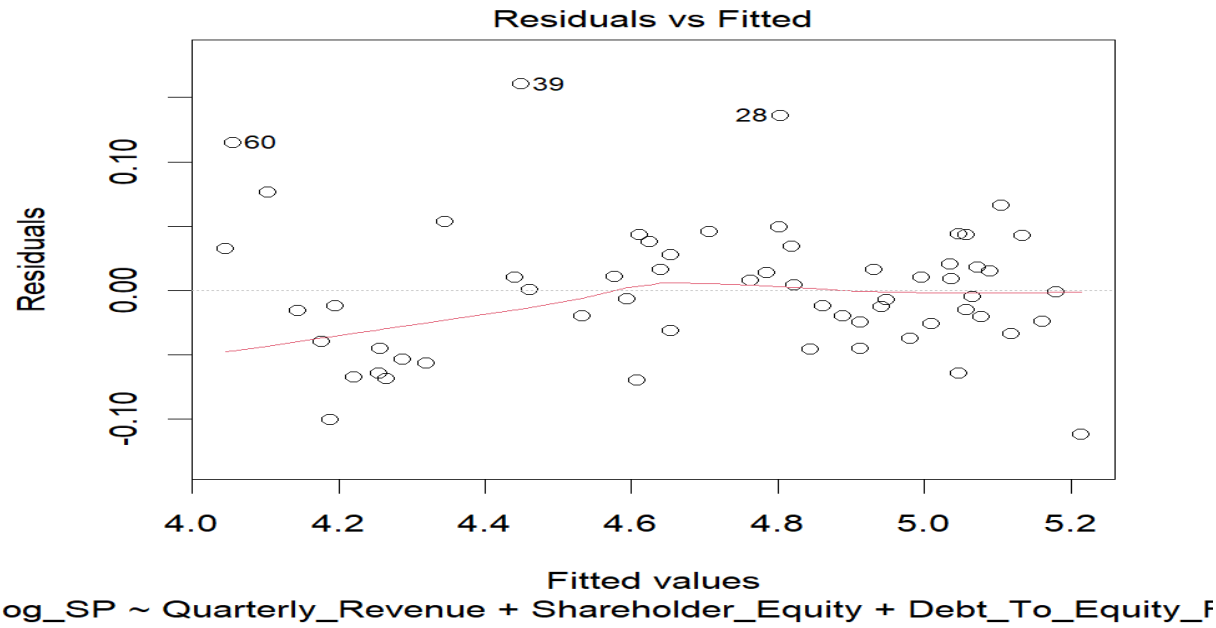
H1 is that the slopes for the variables are individually significant

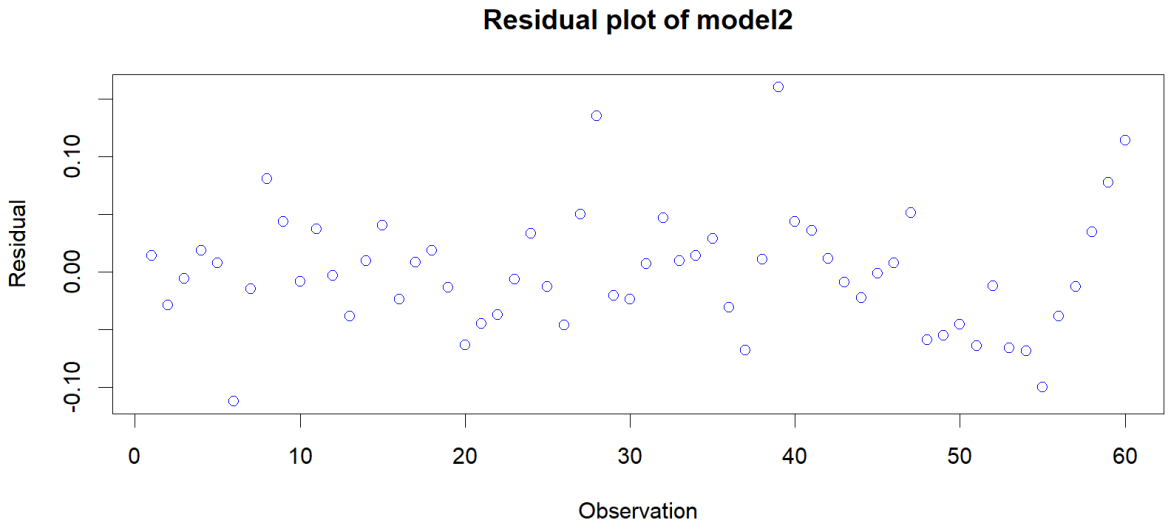
Shareholder_Equity, Debt_To_Equity_Ratio, BD_Share_Price, SP_Decrease_Dummy, and CPI are the only significant variables in this new model. Every other variable has too high of a p-value to be deemed significant.

Now we move on to checking the assumptions of model2. They're the same assumptions as before.

- Linearity
- Independence of error terms
- Homoscedasticity
- No perfect multicollinearity
- Zero mean of errors
- No autocorrelation of errors

For Linearity, we once again generate a residual plot, only using the new model this time.





The trend of the residuals starts off being far lower than expected, but then it seems to converge to 0. This is likely due to the model using the natural log of Quarterly_Share_Price instead of the share price itself. It seems there must be a better way to transform the data than a log transformation. But while this residual plot may look worse, it's important to remember that the scaling is different. This model only seems worse because the residuals are so close to 0 that the trend being off is amplified. In reality, we're quite confident this residual plot is better than the first one. However, it can still be improved by transforming the y variable differently/more.

When testing for multicollinearity we only see 4 variables with concerning vif scores (scores above 5), those being Debt_To_Equity_Ratio, BD_Share_Price, Healthcare_Spending, and CPI. The percentage of problematic variables is far lower than before, where a majority of our variables had high vif scores. However, what can still be done is removing the remaining variables with high vif scores.

To check for independence of error terms and for no autocorrelation of errors, we look at the residual plot above and use the Durbin-Watson test on model 2.

Durbin-Watson test

data: model2

DW = 1.2005, p-value = 3.071e-05

alternative hypothesis: true autocorrelation is greater than 0

Here, H0 is that true autocorrelation is not greater than 0 and H1 is that true autocorrelation is greater than 0. With the p-value below 0.05, we reject the null hypothesis and say that true autocorrelation is greater than 0, which violates our assumptions of independence of error terms and no autocorrelation of errors.

Once again, given the p-value of the test and how it is below 0.05, we reject the null hypothesis and say there's no independence of error terms and autocorrelation in the variables. This is likely due to the transformation we performed not being the most optimal one to perform.

To test for homoscedasticity, we look at the residual plot above and use the Breusch-Pagan test.

studentized Breusch-Pagan test

data: model2

BP = 10.714, df = 11, p-value = 0.4676

The null hypothesis H0 here is that the model has homoscedasticity (consistent variance of residuals), and the alternative hypothesis H1 is that the model has heteroscedasticity (inconsistent variance of residuals).

Once again, we see that, given the high p-value, we fail to reject the null hypothesis of homoscedasticity, and therefore we say the model has homoscedasticity.

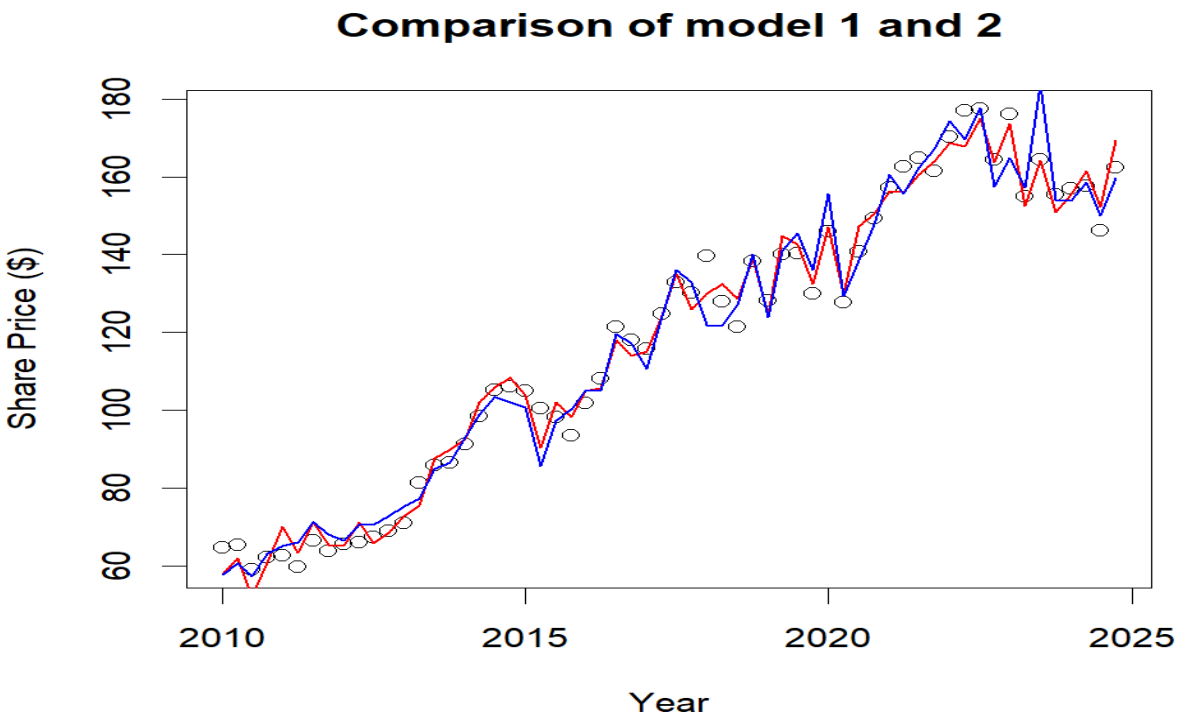
Since we are using the OLS method, our model is guaranteed to have 0 mean error.

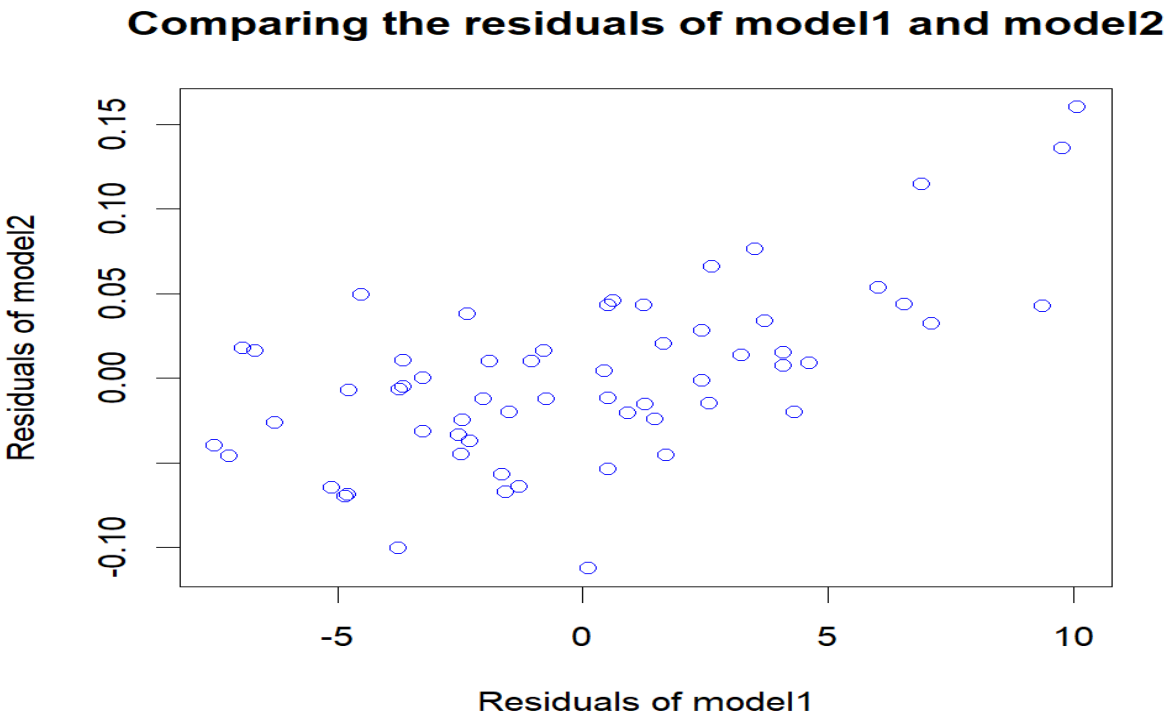
Model Comparison:

Overall, model2 improves on many parts of model1, such as having variables with lower vif scores, likely having a better residual plot, and having a higher percentage of significant variables. In addition to that, model2 does not seem worse than model1 in any significant way. The R^2 scores of the model are also quite close, despite model2 being much simpler.

Looking at the plots below, we can see a couple of things. In the plot directly below, model1 is in red and model2 is in blue. One observation is that model 2 is predicting more extreme values than model 1 is at some points, and less extreme values than model 1 is at other points. Looking at the plot that compares the residuals of the models, we see the spread of residuals for model 1 is wider than the spread of residuals for model 2. As such, that leads me to the conclusion that model 2 is better on average at predicting the share price than model 1 is.

Factoring in how model2 is a simpler model with better variables and that better suits the Gauss-Markov assumptions, while not having any significant drawbacks, and being about as good of a predictor of share price as model1, I say that I find model2 to be the better model.


Comparison of model 1 and 2
Share Price ($)
60
80
100
120
140
160
180
2010
2015
2020
2025
Year


Comparing the residuals of model1 and model2
Residuals of model2
-0.10
0.00
0.05
0.10
0.15
-5
0
5
10
Residuals of model1

Citation list:

For this project I used two websites.

Macrotrends - http://www.macrotrends.net/

Federal Reserve Economic Data - https://fred.stlouisfed.org/