S&P 500 Price Change Predictors

Andy Barnes
Ryan Leveille
Tong Sun
Jared Turner

21 November, 2018

Project Scope

S&P 500

The Standard & Poor's 500, often abbreviated as the S&P 500, or just the S&P, is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the New York Stock Exchange (NYSE) or Nasdaq, Inc (NASDAQ). The NYSE and NASDAQ are the American stock exchanges where the stocks that make up the S&P 500 are traded. Market capitalization (market cap) is the market value of a publicly traded company's outstanding shares. Market capitalization is equal to the share price multiplied by the number of shares outstanding. As outstanding stock is bought and sold in public markets, capitalization could be used as an indicator of public opinion of a company's net worth and is a determining factor in some forms of stock valuation. The reason that we chose to use the S&P 500 for our companies is that it is a very broad and diversified index that includes companies from many different industries. The S&P 500 is one of the most commonly followed equity indices, and many consider it one of the best representations of the U.S. stock market, and a bellwether for the U.S. economy. The share prices for each company within the S&P 500 change daily (during trading days) based on trading interest. A share price is the price of a single share of a company, derivative or other financial asset. The share price is the highest amount someone is willing to pay for the stock, or the lowest amount that it can be bought for.

Time Frame

Our project is going to analyze the companies within the S&P 500 over a 7 year period. This period is going to be from 1/01/2011-12/31/2017.

Approach:

Due to the volatility of the stock market and how the data can sometimes be very misleading depending upon the time frame which you choose to analyze the data from, we are going to approach this project with the intention of minimizing any major outliers for our response variable (share price).

For example, if we were to randomly select some date in 2011 and observe the share price for Microsoft it could be $50 per share. Likewise if we took another random date in 2011 to observe the price of Microsoft it could be $25 per share, perhaps even just a week before or after the price was at $50 per share. Furthermore, when we analyze the response variable in 2017 it could be $50 or $25, depending on the time frame you select for the observation.

Therefore, the way that we will make sure that we do not misrepresent our response variable and by keeping the project within the scope of our class we will be taking the average of the share price for each company within the S&P 500 for the years 2011 and 2017. The 5 years in between (2012, 2013, 2014, 2015, 2016) will allow plenty of time for our response variable to change based on our predictor variables.

Another reason why we want to calculate the average share price for each company by year is because our predictor variables will be yearly as well.

The S&P 500 is based on market cap and some companies will grow in size or fall in size so we will be using companies that were in the S&P 500 starting on January 1, 2011, and are still in the S&P 500 on December 31, 2017. As of now, we have 476 companies out of 500 that are still within the S&P 500 to analyze.

Response Variable

`Price.ch`: Our response variable is going to be the share price of each company within the S&P 500. We are going to take the average share price for 2011 and the average share price for 2017 and then calculate the percent difference for each company.

Predictor Variables

We have multiple variables that we are going to use as our predictor variables. These variables have been chosen based on what what we think investors use to analyze a company before making an investment. Also, some other potential variables' data is incomplete or missing so we choose variables that were best for the scope of the project.

Our goal with this project is to find the predictor variables that have the biggest impact on our response variable (share price).

All of our variables are going to be using their percent change in yearly unit during our time period of 2011-2017. We will start with the predictor variables' 2011 observations and end with the predictor variables' 2017 observations. We chose this time frame because it gave us our yearly data of 2011, a 5 year time frame, then our yearly data of 2017.

1. `EPS`: Earnings per share (EPS) is the portion of a company's profit allocated to each share of common stock. Earnings per share serves as an indicator of a company's profitability.
2. `Div`: A dividend is a payment made by a corporation to its shareholders, usually as a distribution of profits. When a corporation earns a profit or surplus, the corporation is able to re-invest the profit in the business and pay a proportion of the profit as a dividend to shareholders.
3. `BV`: Book value per share indicates the book value (or accounting value) of each share of stock. Book value is a company's net asset value, which is calculated by total assets minus intangible assets and liabilities. An easy way to think of book value per share is what is the expected value of the company.
4. `RoA`: The return on assets shows the percentage of how profitable a company's assets are in generating revenue.
5. `RoE`: Return on equity is a measure of the profitability of a business in relation to the book value of shareholder equity, also known as net assets or assets minus liabilities. ROE is a measure of how well a company uses investments to generate earnings growth.
6. `RoIC`: Return on capital (ROC), or return on invested capital(ROIC), is a ratio used in finance, valuation and accounting, as a measure of the profitability and value-creating potential of companies after taking into account the amount of initial capital invested.
7. `DE`: Debt/Equity (D/E) Ratio, calculated by dividing a company's total liabilities by its stockholders' equity, is a debt ratio used to measure a company's financial leverage. The D/E ratio indicates how much debt a company is using to finance its assets relative to the value of shareholders' equity.
8. `Rev`: Revenue is the income that a business has from its normal business activities, usually from the sale of goods and services to customers.

The Data

Our data looks at the 476 stocks in in the S&P 500 today with data dating back to 2011. Each variable is given as the percent change in yearly value or average value from 2011 to 2017. The first 6 rows of data are as follows

Co.	Price.ch	EPS	Div	BV	RoA	RoE	RoIC	DE	Rev
A	1.0041	-0.2632	NA	0.1531	-0.2187	-0.4386	-0.3129	-0.1778	-0.3240
AAL	5.3662	-1.2464	NA	-1.2150	-1.4623	NA	NA	NA	0.7602
AAP	0.9557	0.2564	0.0000	3.0918	-0.4973	-0.6409	-0.6045	-0.3673	0.5193
AAPL	1.8952	1.3316	NA	1.1927	-0.4876	-0.1152	-0.5161	NA	1.1177
ABC	1.2260	-0.3543	2.3953	0.0665	-0.7792	-0.2842	-0.5969	3.8824	0.9091
ABMD	8.9755	-4.6562	NA	2.4388	-2.1889	-2.1493	-2.1050	NA	3.4059

from this we can obtain the following correlation matrix.

	Price.ch	EPS	Div	BV	RoA	RoE	RoIC	DE	Rev
Price.ch	1.0000	0.2264	0.3533	0.2650	0.1495	0.1613	0.3862	0.0228	0.3079
EPS	0.2264	1.0000	0.1746	0.1341	0.7674	0.6148	0.3577	0.0005	0.1163
Div	0.3533	0.1746	1.0000	0.0954	0.0734	0.0707	0.3140	-0.0344	0.0239
BV	0.2650	0.1341	0.0954	1.0000	0.0716	-0.0973	0.0659	-0.1470	0.4739
RoA	0.1495	0.7674	0.0734	0.0716	1.0000	0.7680	0.4130	-0.0146	0.0475
RoE	0.1613	0.6148	0.0707	-0.0973	0.7680	1.0000	0.4423	0.1272	0.0109
RoIC	0.3862	0.3577	0.3140	0.0659	0.4130	0.4423	1.0000	-0.0242	0.1823
DE	0.0228	0.0005	-0.0344	-0.1470	-0.0146	0.1272	-0.0242	1.0000	0.0008
Rev	0.3079	0.1163	0.0239	0.4739	0.0475	0.0109	0.1823	0.0008	1.0000

We see that there is strong correlation between `RoA` and `EPS`, there is some correlation between `Rev` and `BV`.

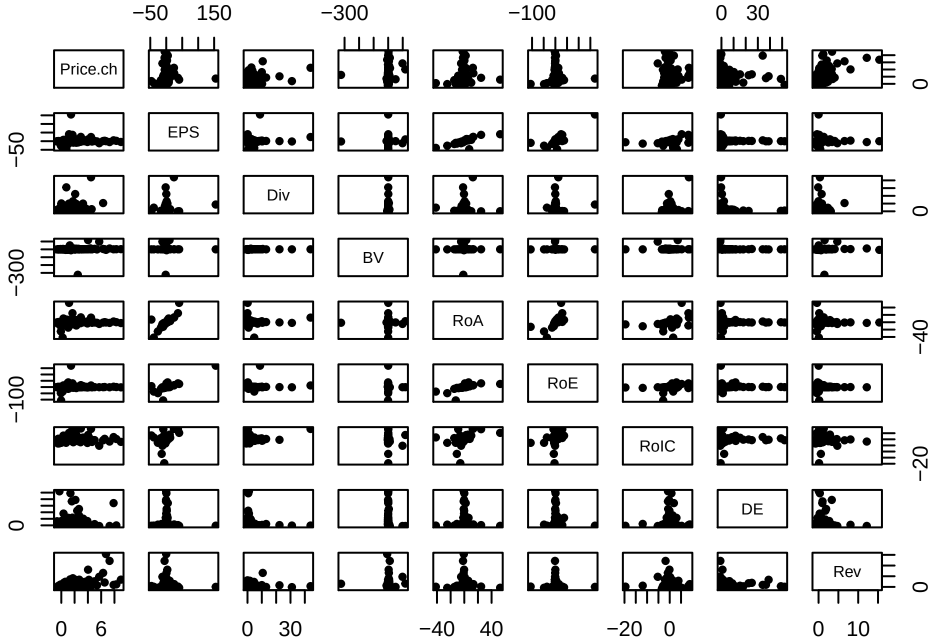

Figure 1: Scatterplot of raw data.

Why we chose this data

We chose this project due to the vast amounts of data that the stock market encompasses. The stock market is a very broad topic and this project can be applied to many different situations.

With our project we hope to find which predictor variables affect the response variable (share price) the most. We are excited to see what our models come up with. The findings will be interesting as there are many different theories as to what is the most important predictor variables for a companies share price.

Data preprocessing

Before formulating our model, we cleaned our data by filling in missing values. For the predictor variable `Div`, missing values were assumed to be 0, this is because dividends other than 0 are usually reported and recorded. For all other predictors, missing values were filled in with the mean among all observations for that data. After our data was cleaned we have the following correlation matrix and plot.

	Price.ch	EPS	Div	BV	RoA	RoE	RoIC	DE	Rev
Price.ch	1.0000	0.0819	0.1026	0.0631	0.0791	0.0643	0.0990	0.0889	0.5314
EPS	0.0819	1.0000	0.1582	0.0134	0.4554	0.8153	0.2446	-0.0128	0.0047
Div	0.1026	0.1582	1.0000	0.0021	0.0579	0.1160	0.1787	-0.0542	-0.0504
BV	0.0631	0.0134	0.0021	1.0000	0.0131	-0.0046	-0.0092	-0.0201	0.0463
RoA	0.0791	0.4554	0.0579	0.0131	1.0000	0.4084	0.4246	-0.0116	-0.0004
RoE	0.0643	0.8153	0.1160	-0.0046	0.4084	1.0000	0.2076	0.0245	-0.0096
RoIC	0.0990	0.2446	0.1787	-0.0092	0.4246	0.2076	1.0000	0.0082	-0.0141
DE	0.0889	-0.0128	-0.0542	-0.0201	-0.0116	0.0245	0.0082	1.0000	0.0404
Rev	0.5314	0.0047	-0.0504	0.0463	-0.0004	-0.0096	-0.0141	0.0404	1.0000

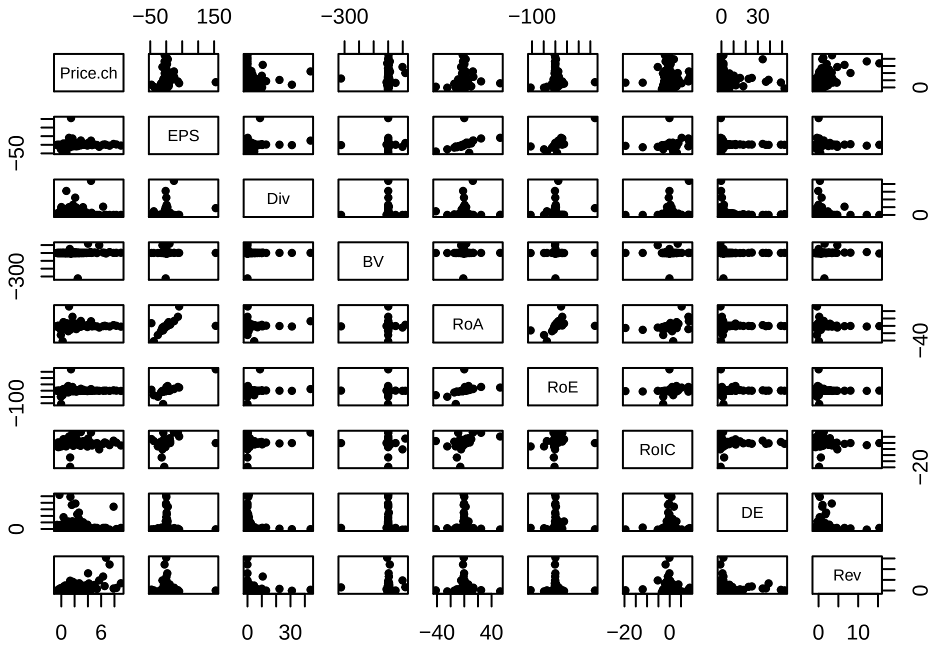

Figure 2: Scatterplot of clean data.

By looking at the scatterplot, there are potential linear relationships between `RoA` and both `EPS` and `RoE`. The correlation matrix and heat map confirm that `RoE` and `EPS` are highly correlated.

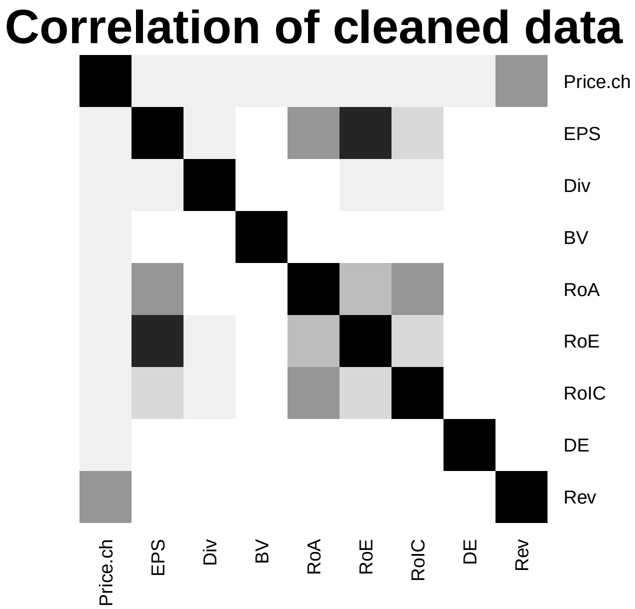

Figure 3: Heat map of correlation matrix for cleaned data, darker shades correspond to greater correlation.

Our Model

We will formulate a simple linear model, where our response, `Price.ch`, is a linear combination of our predictors

$$\widehat{\text{Price.ch}}_i = \beta_0 + \beta_1 \text{EPS}_i + \beta_2 \text{Div}_i + \beta_3 \text{BV}_i + \beta_4 \text{RoA}_i + \beta_5 \text{RoE}_i + \beta_6 \text{RoIC}_i + \beta_7 \text{DE}_i + \beta_8 \text{Rev}_i + \epsilon_i$$

where $\epsilon_i \sim^{\text{iid}} N(0, \sigma^2)$. Where σ^2 is the population variance. Therefore we make the following assumptions for our model

1. Observations are indepent indentically distributed (iid), as is error, ϵ_i.
2. Our response is normally distributed with mean 0 and variance σ^2

After formulating our model, we fit our data to this model using R. The summary is as follows.

```
##
## Call:
## lm(formula = Price.ch ~ ., data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6073 -0.6401 -0.0953  0.3887  6.6643
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.9349505  0.0599953  15.584  < 2e-16 ***
## EPS         0.0037591  0.0101287   0.371  0.71070
## Div          0.0465288  0.0159247   2.922  0.00365 **
## BV           0.0030159  0.0029233   1.032  0.30276
## RoA          0.0097342  0.0147320   0.661  0.50910
## RoE          0.0009637  0.0083038   0.116  0.90766
## RoIC         0.0520675  0.0345008   1.509  0.13193
## DE           0.0190515  0.0098566   1.933  0.05386 .
## Rev          0.5475882  0.0394537  13.879  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.063 on 467 degrees of freedom
## Multiple R-squared:  0.3157, Adjusted R-squared:  0.3039
## F-statistic: 26.93 on 8 and 467 DF,  p-value: < 2.2e-16
```

Therefore our fitted model, with point estimates, of the parameters $\beta_0, ..., \beta_8$, is given by

$$\widehat{\text{Price.ch}}_i = 0.935 + 0.004\text{EPS}_i + 0.047\text{Div}_i + 0.003\text{BV}_i + 0.009\text{RoA}_i + 0.001\text{RoE}_i + 0.052\text{RoIC}_i + 0.019\text{DE}_i + 0.548\text{Rev}_i + \epsilon_i$$

Immediately we see that with such a small p-value, 2.2×10^{-16}, our model is significant. We also see that there are two predictors displaying initial significance, `Div` and `Rev`. However, in addition to our model having an R-squared of 0.3156597, indicating that only about 30% of variation in our response is explained by our predictors, the following diagnostic plots challenge our assumptions.

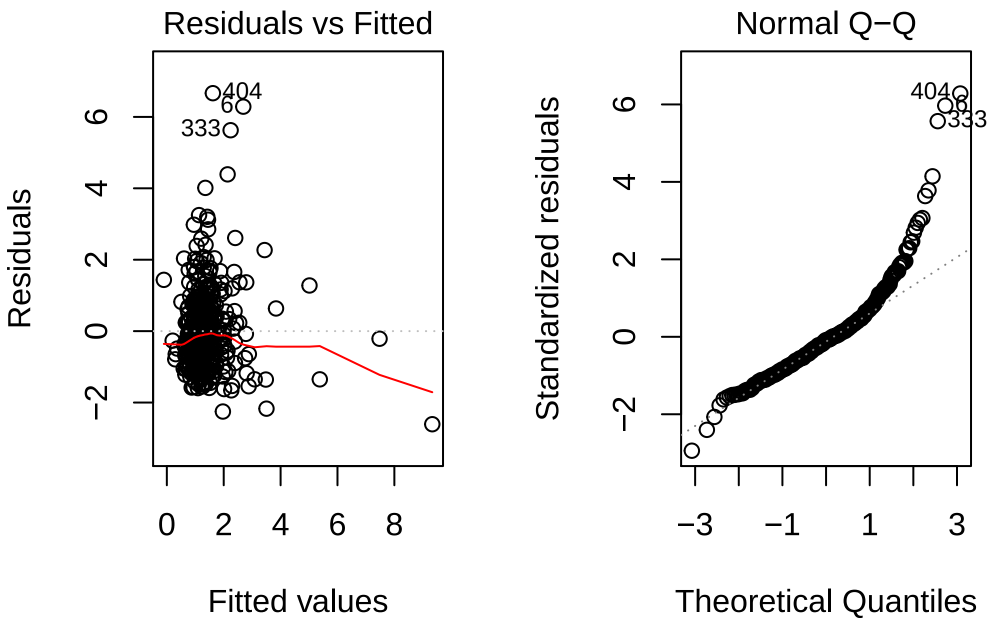

Figure 4: Errors vs fitted values (Left) challenges iid error terms. While QQ Plot (Right) challenges the normality of our response.

Additionally we can see that the histogram of our response values is skewed right, further challenging the normality of our response and indicating outliers.

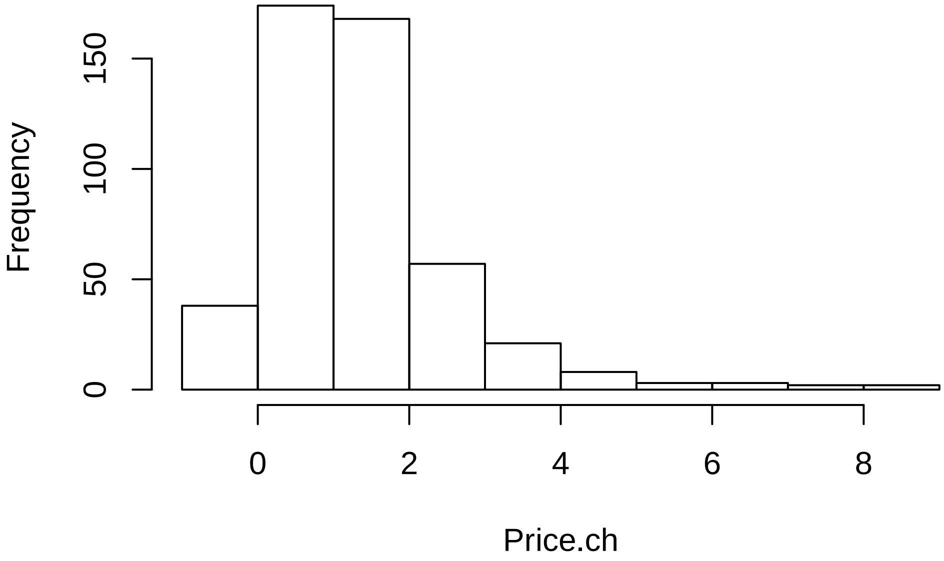

Figure 5: Histogram of response, ‘Price.ch‘, skewness suggests outliers.

Checking for outliers.

In hopes of improving our model, we will set out to find a small subset of outliers we can throw out to improve our model. Our initial goal was to raise our R-squared value to at least 0.5 with the hopes that this would help the distribution of our errors as well as the normality of our response. We wrote a script to this automatically and were able to raise our R-squared above 0.5 by dropping 29 observations. However, the question of whether this affects the validity of our model is a perfectly legitimate one.

Justification for dropping outliers

It is generally not a good idea to drop a significant number of outliers without justification. However, given our knowledge of the market, there are two primary reasons that we believe dropping these outliers do not affect the objective of our model:

1. We are only dropping 6.09% of our data set, which is generally considered acceptable.
2. The stock market is very volatile and is littered with extreme cases. Our predictors are very basic and most extreme cases will likely be correlated with other factors outside the scope of this project.
3. We are studying the SP500 as a subset of popular stocks, but it also gives us insight into the population of all stocks. The SP500 is likely not representative of the population of all stocks. A random subset of 500 stocks from all stocks is more likely to contain a smaller percentage of these extreme observations. In the worst case dropping a small subset of outliers may actually help correct for this, making our sample more representative of the population of all stocks.

Our objective is exploratory in nature and we are not expecting to attain any predictive power from this model. We are only interested in which predictors are significant for most cases. We feel that this model, even with the outliers dropped, is still realistic in this regard.

Final model

We now look assess this subset of our original data, with the outliers withdrawn. We fit this data to our original model using R to obtain the following summary.

```
##
## Call:
## lm(formula = Price.ch ~ ., data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68073 -0.50643 -0.01055  0.43476  1.86390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.772770   0.041222  18.746  < 2e-16 ***
## EPS         0.001594   0.006780   0.235  0.81419
## Div         0.071576   0.011987   5.971 4.87e-09 ***
## BV          0.024859   0.004387   5.666 2.65e-08 ***
## RoA         0.005809   0.010041   0.579  0.56319
## RoE         0.001393   0.005549   0.251  0.80196
## RoIC        0.096121   0.029307   3.280  0.00112 **
## DE          0.006478   0.006929   0.935  0.35035
## Rev         0.477464   0.027011  17.677  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7075 on 438 degrees of freedom
## Multiple R-squared:  0.5061, Adjusted R-squared:  0.4971
## F-statistic:  56.1 on 8 and 438 DF,  p-value: < 2.2e-16
```

We now have an R-squared of 0.5061, indicating that approximately 51% of the variation in our response can be explained by our predictors. Additionally the p-value of our model remained the same, indicating significance of our overall model. However, we now show four initially significant predictors, `Div`, `BV`, `RoIC`, and `Rev`. After generating the same diagnostic plots as before we see improve, albeit not perfection, which is hardly ever to be expected from real world data.

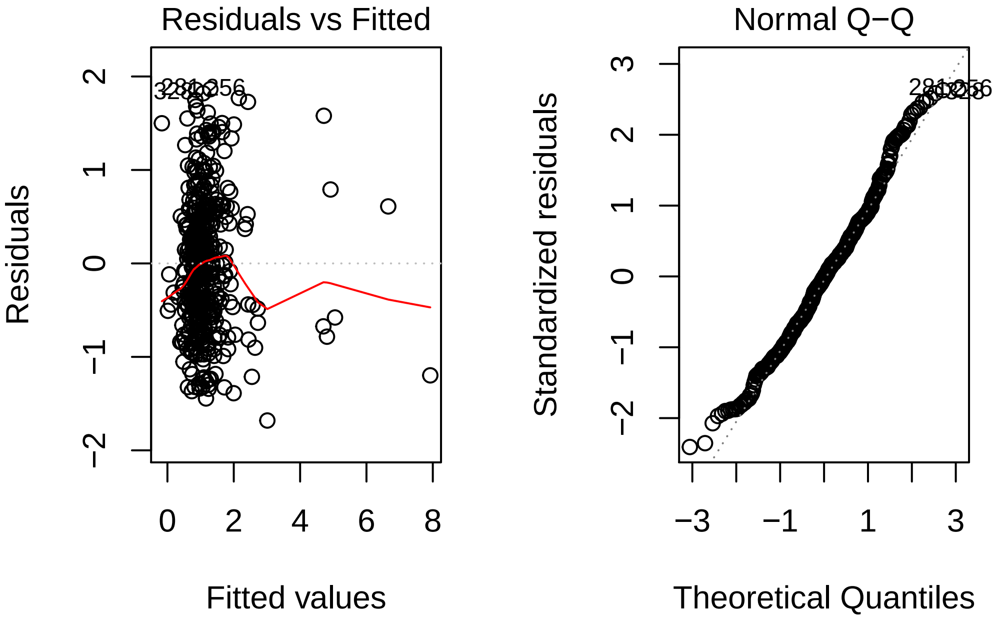

Figure 6: Both the plot of errors vs fitted values (Left) and QQ Plot (Right) show improvement.

However, with these improvements we felt comfortable moving forward with our model and all of its assumptions.

Since not all predictors exihibited significance, we proceeded to reduce our model using R's built in `step` function. This function minimizes an information criteron AIC, which in effect improves model fit and reduces complexity. The output summary of this process and the resulting reduced model is as follows.

```
## Start:  AIC=-300.39
## Price.ch ~ EPS + Div + BV + RoA + RoE + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS      AIC
## - EPS   1     0.028 219.29 -302.338
## - RoE   1     0.032 219.29 -302.330
## - RoA   1     0.168 219.43 -302.053
## - DE    1     0.438 219.70 -301.504
## <none>              219.26 -300.395
## - RoIC  1     5.385 224.65 -291.549
## - BV    1    16.072 235.33 -270.775
## - Div   1    17.850 237.11 -267.411
## - Rev   1   156.423 375.68  -61.693
##
## Step:  AIC=-302.34
## Price.ch ~ Div + BV + RoA + RoE + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS      AIC
## - RoA   1     0.204 219.49 -303.923
## - RoE   1     0.233 219.52 -303.863
## - DE    1     0.427 219.72 -303.468
## <none>              219.29 -302.338
## - RoIC  1     5.396 224.69 -293.472
## - BV    1    16.089 235.38 -272.689
## - Div   1    18.289 237.58 -268.530
## - Rev   1   156.654 375.94  -63.385
##
## Step:  AIC=-303.92
## Price.ch ~ Div + BV + RoE + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS      AIC
## - DE    1     0.413 219.91 -305.082
## - RoE   1     0.469 219.96 -304.968
## <none>              219.49 -303.923
## - RoIC  1     7.697 227.19 -290.517
## - BV    1    16.112 235.61 -274.259
## - Div   1    18.103 237.60 -270.498
## - Rev   1   156.812 376.31  -64.954
##
## Step:  AIC=-305.08
## Price.ch ~ Div + BV + RoE + RoIC + Rev
##
##        Df Sum of Sq    RSS      AIC
## - RoE   1     0.498 220.40 -306.072
## <none>              219.91 -305.082
## - RoIC  1     7.610 227.52 -291.876
## - BV    1    15.878 235.78 -275.919
## - Div   1    17.882 237.79 -272.136
## - Rev   1   157.926 377.83  -65.144
##
## Step:  AIC=-306.07
## Price.ch ~ Div + BV + RoIC + Rev
##
##        Df Sum of Sq    RSS      AIC
## <none>              220.40 -306.072
## - RoIC  1     8.868 229.27 -290.440
## - BV    1    15.868 236.27 -276.995
## - Div   1    18.504 238.91 -272.037
## - Rev   1   157.922 378.33  -66.561
```

```
##
## Call:
## lm(formula = Price.ch ~ Div + BV + RoIC + Rev, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68773 -0.51486 -0.00468  0.42489  1.86630
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.784838   0.039516  19.861  < 2e-16 ***
## Div         0.071806   0.011788   6.092 2.43e-09 ***
## BV          0.024662   0.004372   5.641 3.02e-08 ***
## RoIC        0.109141   0.025881   4.217 3.00e-05 ***
## Rev         0.478997   0.026916  17.796  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7062 on 442 degrees of freedom
## Multiple R-squared:  0.5035, Adjusted R-squared:  0.499
## F-statistic: 112.1 on 4 and 442 DF,  p-value: < 2.2e-16
```

Our reduced model is thus given by

$$\widehat{\text{Price.ch}}_i = \beta_0 + \beta_1 \text{Div}_i + \beta_2 \text{BV}_i + \beta_3 \text{RoIC}_i + \beta_4 \text{Rev}_i + \epsilon_i$$

where $\epsilon_i \sim^{\text{iid}} N(0, \sigma^2)$. Our reduced model carries the same assumptions as our original model. The model with the point estimations of our parameters is given by

$$\widehat{\text{Price.ch}}_i = 0.785 + 0.072 \text{Div}_i + 0.025 \text{BV}_i + 0.109 \text{RoIC}_i + 0.479 \text{Rev}_i + \epsilon_i.$$