Determinants of Stock Market in DOW 30

Muskan Yadav Manasvinee Amin
Quantitative Methods for Finance, Boston University

June 27, 2022

Abstract

The purpose of this project is to use linear regression and various plot techniques that influence the stock price with multiple linear regression. The accuracy of the final results depends on the quality of the data and its validity. Using mathematical and statistical models to analyze the stock market is important. The data is for 30 DOW companies. To better understand what factors might be influencing the stock prices of these companies, a variety of attributes were examined on a particular trading day. After processing the data with various statistical methods, multiple linear regression was selected to predict the most important factor affecting the stock prices of the DOW 30 companies. This project uses backward step wise regression to select the optimal model. All the factors are taken into the linear regression model. Where each factor is then eliminated one by one, at the end the chosen factor provides the optimal model.

Introduction

Since the multiple linear regression has numerous independent variables, the calculation is complex. Thus, statistical analysis is used in practical applications generally. The primary question of interest, therefore, is whether or not the stock prices are affected by the change in price earnings ratio, earnings per share, Beta, Dividend yield and Market capital. That is, if the stock market is majorly affected by one of the following factors.

In the paper, various statistical methods are used to process the raw data. This paper does not process the data by adjusting and cleaning the outliers of the data since the number of variables used for analysis are only a few. In the project, the outlier might not affect the results, so eliminating the outlier is not considered necessary. The data used in this paper is analyzed in two ways. First, we processed the data and established the hypothesis. Then, we built a regression model to study relationships between the dependent variable (stock price) and a number of price-attributes (independent variables).

Data Description

The data set is imported from Barchart[1]. It contains 5 predictor variables indicating whether or not the stock price changes. To prepare the data for regression the predictor variables are normalized.

Choice of covariates

The independent variables that are regressed against the stock price are chosen on the basis of :

- The price-to-earnings (P/E) ratio relates a company's share price to its earnings per share. A high P/E ratio means that the stock is overvalued.
- Earnings per share (EPS) is the amount of a company's profit allocated to each outstanding share of a company's common stock.
- Beta indicates how volatile a stock's price is in comparison to the overall stock market. A beta greater than 1 indicates a stock's price swings more wildly than the overall market.

- Market capital is arrived at by multiplying the share price by the number of shares outstanding. So rise in stock's price leads to rise in its market capital.
- For dividend yield it is assumed that the dividend is not raised or lowered, the yield will rise when the price of the stock falls.

Figure 1 shows the scatter plot of each of the independent variable plotted against stock price. If a high degree of multi collinearity exists, the measurements should be concentrated around a straight line in the scatter plot. Some outliers are observed but there is no clear pattern followed by the data points plotted. This can be resolved by collecting more data points or by trying to remove some of the covariates from the model.

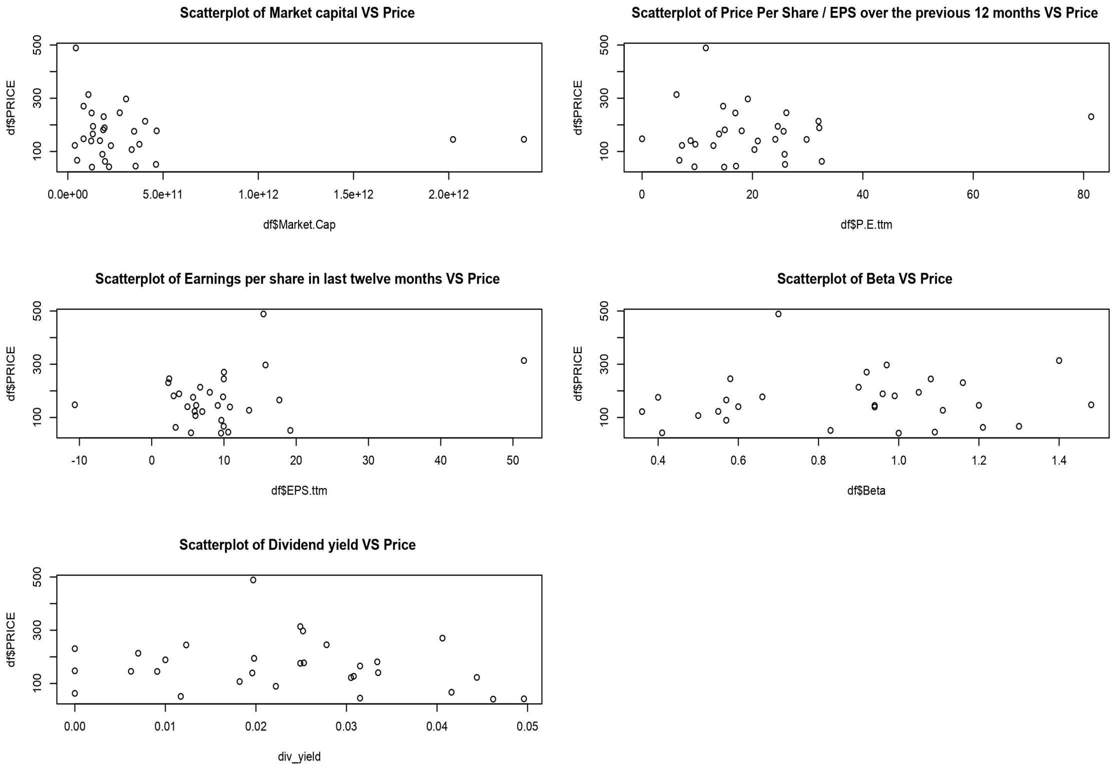


Figure 1. Scatter plots of each of the 5 independent variables

Another scatter plot is seen to observe the data points against the linear line. In Figure 2 it is observed that the data points are scattered away from the best fitted line.

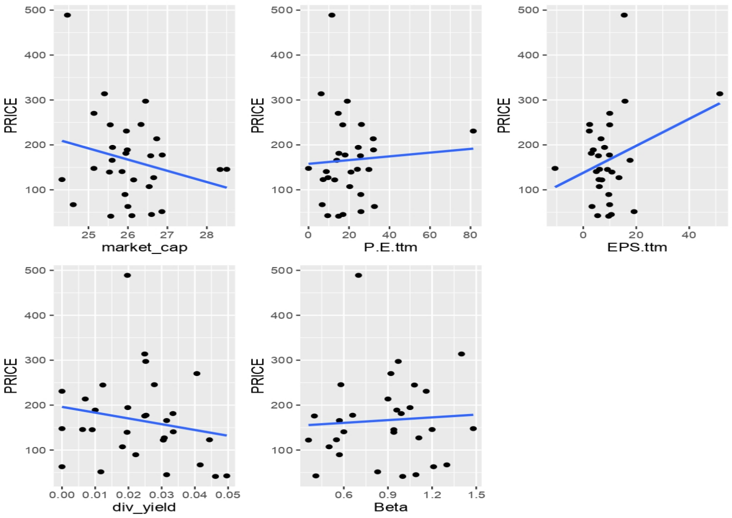


Figure 2. Scatter plots of each of the 5 independent variables with best fitted line

Plotting the Data

Normally distributed data looks similar to a bell shaped curve with one peak and symmetric around the mean. Based on the histograms below none of the independent variables are normally distributed. Histogram are used because they represent a large amount of data and therefore indicate the shape of the distribution. Thus, the histogram of the 5 independent variables were plotted.

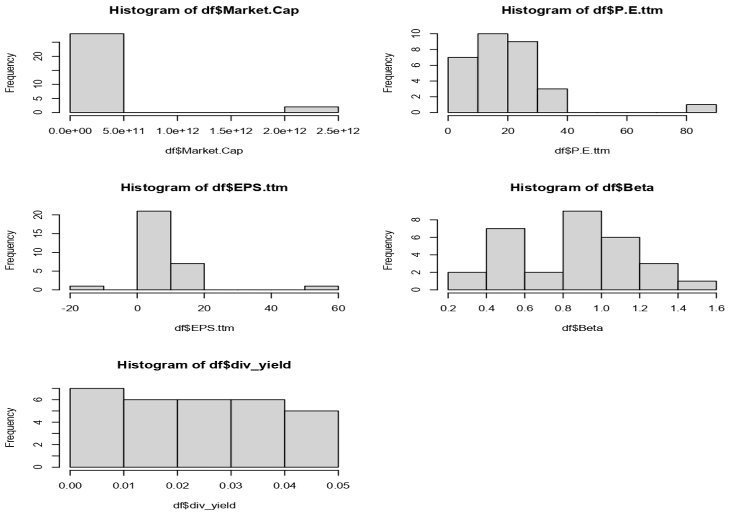


Figure 3. Histograms of each of the 5 independent variables

Figure 3 shows the histograms. Again there are some clear outliers noticed. The data plotted does not seem to follow normal distribution on just observation. An attempt is made to take the log-normal of the data and then plot the histograms for it again

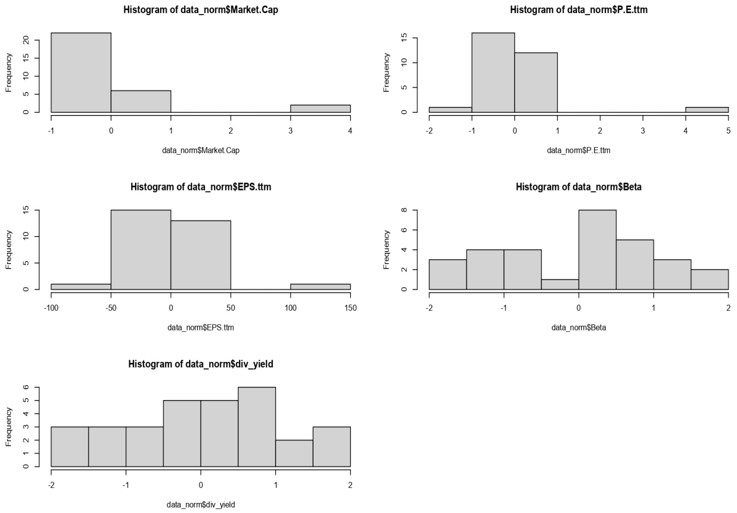


Figure 4. Histograms of each of the 5 independent variables after normalisation

Figure 4 shows the histograms of the five variables after normalising. Since the data seems to follow a bell shaped curve more than it did before it can observe the effect of normalising the data.

Data Analysis

Multivariate linear regression

Linear regression for more than one independent variable is called multiple linear regression. The mathematical modeling of multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon$$

where,
Y is the dependent variable
$X_1, X_2, ..., X_n$ are the n independent variables
β_0 is the unknown intercept
β_1 is the effect on Y of a change in X_1, holding, $X_2, ..., X_n$ constant and
ϵ is the the regression error (omitted factors)

The slope coefficient will be denoted as β and the predicted value will be $\hat{\beta}$. The error term, also called the residual.

The regression equation for this paper is as following:

$$Y = \beta_0 + \beta_1 MarketCap + \beta_2 P/Ettm + \beta_3 Beta + \beta_4 DivYield + \beta_5 EPSttm + \epsilon$$

Market Cap coefficient is the change in Y due to a change of one unit in the Market Capital (everything else held constant). P/E ttm coefficient is the change in Y due to a change of one unit in the P/E ttm rate (everything else held constant). Similarly all the other variables can be defined in terms of Y. Here standard Error reflects the level of accuracy of the coefficients.

Hypothesis testing

The null hypothesis is that there is no relationship between the predictor variables and the target variable (stock price). Alternative hypothesis H_1 is expressed as that there is sufficient evidence to suggest that independent variables are all correlated with stock price. The test will assess the probability of the null hypothesis being true or false. To analyse the hypothesis a test statistic is conducted and thereafter, H_0 shall be accepted or rejected.

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

P-value

Indicates the level of statistical significance of the variable. A p-value of less than 0.05 is considered to be statistically significant. H_0 is accepted if the p-value is larger than a defined significance level of $\alpha = 0.05$

R^2

R^2 measures the fraction of the variance of Y that is explained by X; it is unit-free and ranges between zero (no fit) and one (perfect fit) generally values that are at least larger than 0.5 are considered significant to consider that the data fits well. R^2 always increases when another regressor is added to the model.

Adjusted R^2

Adjusted R-square is a modified version of the R^2 that does not necessarily increase when a new regressor is added. It is an adjusted statistic based on the number of independent variables in the model.

Data Modeling

The model is evaluated in terms of significance of the co-variates. Firstly, the model is implemented with the original data set. Thereafter the data set is normalised but there is still a problem faced with uncorrelated variables. After adjusting the model for this through log-transformation, the final model is achieved. Lastly, an elimination process is used in which the initial model is improved by eliminating variables with little correlation. The model is tested for significance and statistical tests are used to analyze it further.

Low explanatory variables are removed using the covariance plot as seen in Figure 5

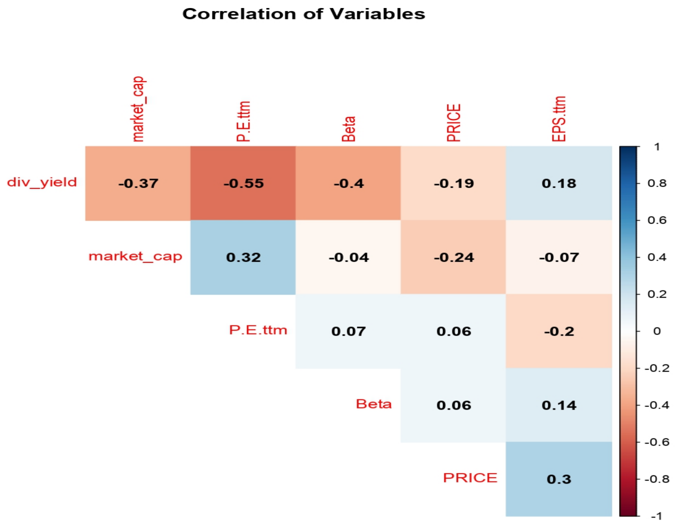


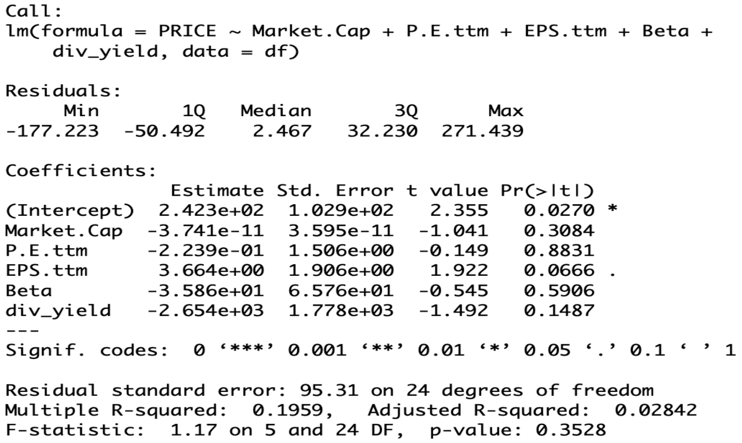
Figure 5. Correlation Plot

Noticeably, Beta and P.E.ttm show the lowest correlation with stock price. Hence they have been removed from the model and a final regression is run on them.

In order to obtain a better picture of the influence of several determinants on stock price, regression technique is executed. A regression approach is better suited in this context since it looks at several determinants of stock prices at the same time, even though it gives us the marginal effect of one variable at a time keeping all other variables in the model constant. Further, a regression approach gives us an idea of the contribution of variation in the dependent variable explained by the variations in the independent variables. This can also be used to generate predictions about the dependent variable and provides us with important summary statistics.

First implementation

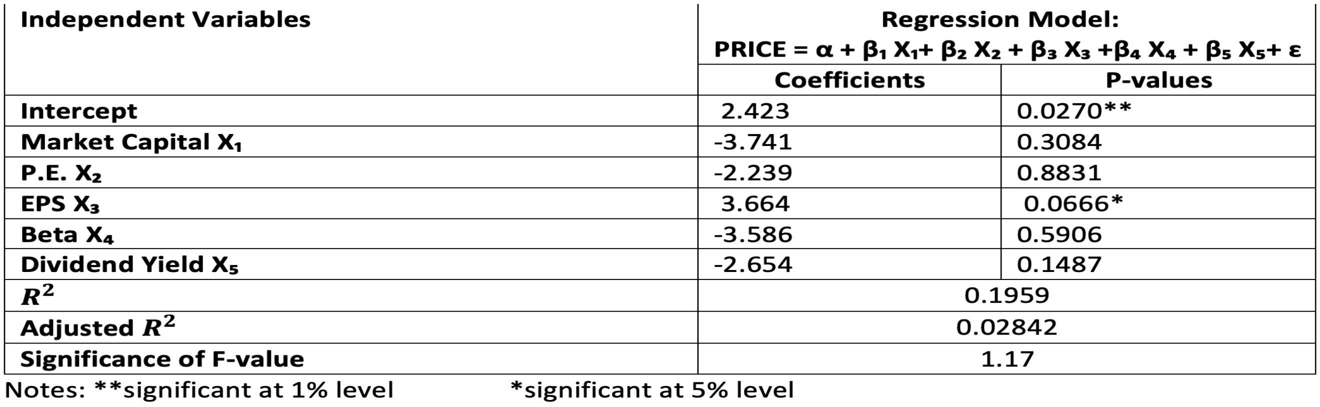
In the initial model all the covariates are used. The initial regression classification results are as shown below:



```
Call:
lm(formula = PRICE ~ Market.Cap + P.E.ttm + EPS.ttm + Beta +
    div_yield, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-177.223  -50.492    2.467   32.230  271.439

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.423e+02  1.029e+02   2.355   0.0270 *
Market.Cap  -3.741e-11  3.595e-11  -1.041   0.3084
P.E.ttm     -2.239e-01  1.506e+00  -0.149   0.8831
EPS.ttm      3.664e+00  1.906e+00   1.922   0.0666 .
Beta        -3.586e+01  6.576e+01  -0.545   0.5906
div_yield   -2.654e+03  1.778e+03  -1.492   0.1487
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.31 on 24 degrees of freedom
Multiple R-squared:  0.1959,   Adjusted R-squared:  0.02842
F-statistic:  1.17 on 5 and 24 DF,  p-value: 0.3528
```

Figure 6. Results of first regression from R programming

The logistic regression identifies five variables to be correlated with stock price. Since all had p-values greater than the level of significance, $\alpha = 0.05$. Hence there does not exist a correlation between these five variables and stock price. The model can be reduced further without losing much explanatory value to achieve better results.



Independent Variables	**Regression Model:** **PRICE = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$**	
	Coefficients	**P-values**
Intercept	2.423	0.0270**
Market Capital X_1	-3.741	0.3084
P.E. X_2	-2.239	0.8831
EPS X_3	3.664	0.0666*
Beta X_4	-3.586	0.5906
Dividend Yield X_5	-2.654	0.1487
R^2	0.1959	
Adjusted R^2	0.02842	
Significance of F-value	1.17	

Notes: **significant at 1% level *significant at 5% level

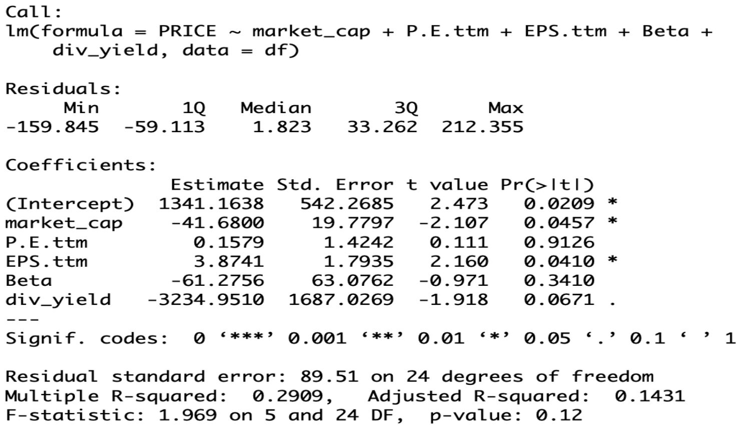
Figure 7. Table1: Summary statistics from first regression

The R^2 value for the model is 0.1959. R^2 is a measure of goodness of fit but R^2 always increases when another regressor is added to the model which is a bit of a problem for a measure of ”fit”. Thus we use Adjusted R^2 view the accuracy of the model. The Adjusted R^2 value for the model is 0.02842. The score is an overall accuracy measure for the model, which in our case is low, suggesting that this model does not perform well overall.

The null hypothesis for each of these tests is considered to be that the variables are not correlated. Since each test has a p-value greater than the level of significance, $\alpha = 0.05$, we accept the assumed null hypothesis that these variables are not correlated.

Second implementation

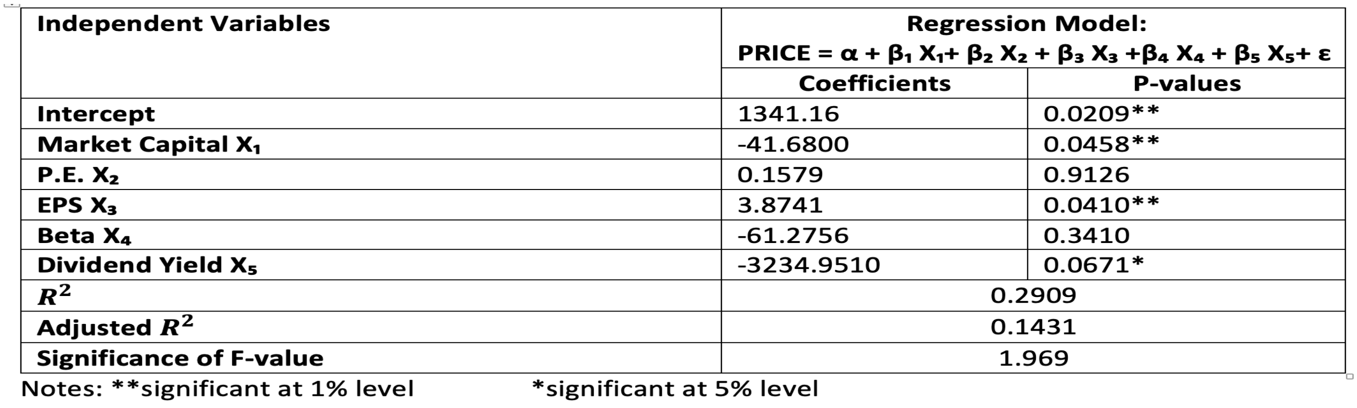
As a measure to improve the performance log-transformation of some of the variables is done. In this case, a log-transformation of Market capital is done. Results from this model are as pictured below:



```
Call:
lm(formula = PRICE ~ market_cap + P.E.ttm + EPS.ttm + Beta +
    div_yield, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-159.845  -59.113    1.823   33.262  212.355

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1341.1638   542.2685   2.473   0.0209 *
market_cap    -41.6800    19.7797  -2.107   0.0457 *
P.E.ttm         0.1579     1.4242   0.111   0.9126
EPS.ttm         3.8741     1.7935   2.160   0.0410 *
Beta          -61.2756    63.0762  -0.971   0.3410
div_yield   -3234.9510  1687.0269  -1.918   0.0671 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.51 on 24 degrees of freedom
Multiple R-squared:  0.2909,   Adjusted R-squared:  0.1431
F-statistic: 1.969 on 5 and 24 DF,  p-value: 0.12
```

Figure 8. Results of second regression from R programming

This time Market cap and EPS.ttm have p-values less than the level of significance, $\alpha = 0.05$. Hence there exists a correlation between these variables and stock price. The model can be further reduced by removing the uncorrelated variables.



Independent Variables	**Regression Model:** **PRICE = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$**	
	Coefficients	**P-values**
Intercept	1341.16	0.0209**
Market Capital X_1	-41.6800	0.0458**
P.E. X_2	0.1579	0.9126
EPS X_3	3.8741	0.0410**
Beta X_4	-61.2756	0.3410
Dividend Yield X_5	-3234.9510	0.0671*
R^2	0.2909	
Adjusted R^2	0.1431	
Significance of F-value	1.969	

Notes: **significant at 1% level *significant at 5% level

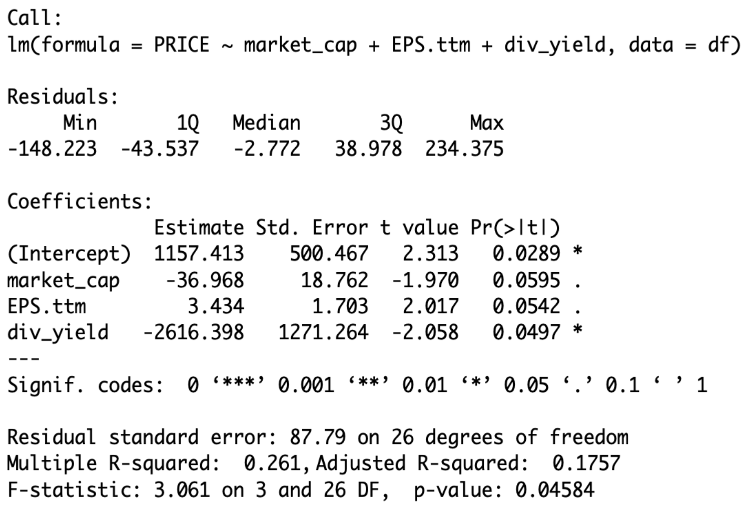
Figure 9. Table2: Summary statistics from second regression

The R^2 value for the model is 0.2909 and the Adjusted R^2 is given by 0.1431 which in our case has become better than the first attempt but is still low, suggesting that this model does not perform well overall.

Hence there is sufficient evidence to suggest that a correlation exists amongst Market cap, EPS.ttm and Stock Price.

Final implementation

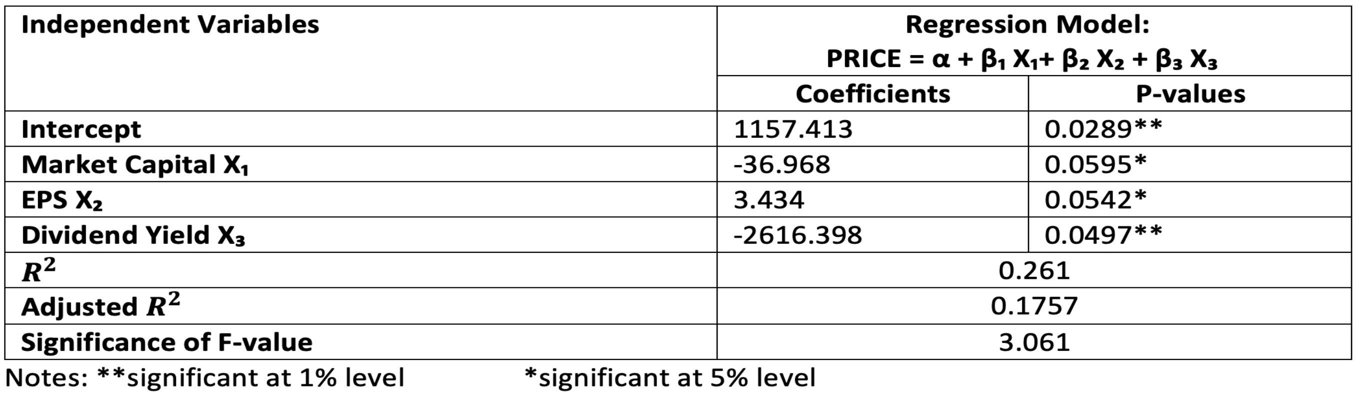
As a measure to improve the performance, variables that are not found to be significant are omitted.



```
Call:
lm(formula = PRICE ~ market_cap + EPS.ttm + div_yield, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-148.223  -43.537   -2.772   38.978  234.375 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1157.413    500.467   2.313   0.0289 *
market_cap    -36.968     18.762  -1.970   0.0595 .
EPS.ttm         3.434      1.703   2.017   0.0542 .
div_yield   -2616.398   1271.264  -2.058   0.0497 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.79 on 26 degrees of freedom
Multiple R-squared:  0.261,	Adjusted R-squared:  0.1757 
F-statistic: 3.061 on 3 and 26 DF,  p-value: 0.04584
```

Figure 10. Results of the final regression from R programming

After reducing the variables and log-transforming the final model yields the following results:



Independent Variables	**Regression Model:** **PRICE = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$**	
	Coefficients	**P-values**
Intercept	1157.413	0.0289**
Market Capital X_1	-36.968	0.0595*
EPS X_2	3.434	0.0542*
Dividend Yield X_3	-2616.398	0.0497**
R^2	0.261	
Adjusted R^2	0.1757	
Significance of F-value	3.061	

Notes: **significant at 1% level *significant at 5% level

Figure 11. Table3: Summary statistics from final regression

The R^2 value for the model is 0.261 and the Adjusted R^2 is given by 0.1757 which has become better than the second attempt but is still low, suggesting that this model does not perform well overall.

The null hypothesis for each of these tests is considered to be that the variables are not correlated. Since each test has a p-value greater than the level of significance, $\alpha = 0.05$ for Market cap, EPS.ttm and Stock Price, we accept the assumed null hypothesis that these variables are not correlated. But there is sufficient evidence to suggest that a correlation exists amongst dividend yield and Stock Price. Hence according to our result we conclude that dividend yield is the most important factor affecting the stock prices of the DOW 30 companies.

Analysis of final model

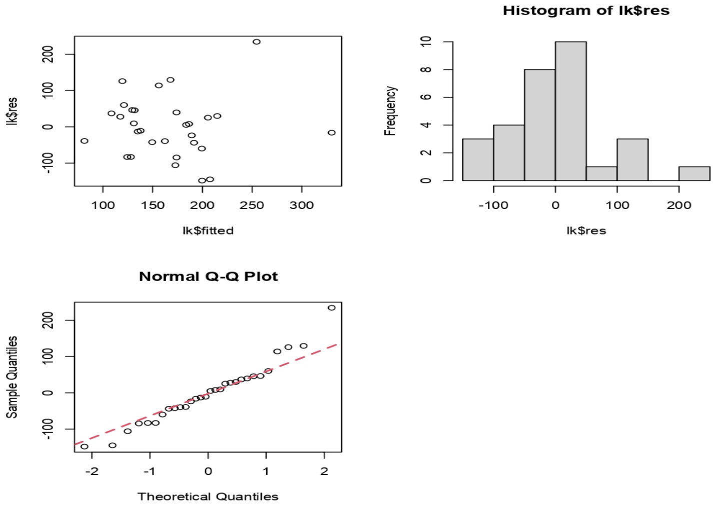


Figure 12. Normal Q-Q Plot

Q-Q plot can determine whether the dependent variable is a linear function of the independent, if the plotted values follow a straight line. We know that in this plot the residual is derived from the observed minus the fitted value. A good plot will have the scatters form a line.

The observations are much more close to the straight line as can be seen in Figure 12, which indicates that they are more close to a normal distribution. Upon examining our outliers, we concluded that removal of these outlier data points would be inappropriate and would compromise the model.

Shortcomings

Heteroscedasticity

Heteroscedasticity appears when the residual's variances are not constant in relation to the value of the variables. This leads to the approximations being unreliable and the F-test inappropriate. In order to avoid this, natural logarithm of the variables is taken.

Endogeneity

It is essential that the error terms are not correlated with the selected variables. Endogeneity appears when the expected value of the residual is not equal to zero, leading to inconsistent approximations. It usually occurs due to sample selection bias, missing relevant variables and most frequently due to omitted variables.

Since only three normalised variables are used in the final model endogeneity is more likely to occur.

Conclusion

The goal of this project was to identify variables that were associated with stock price and investigate if it is possible to build an accurate regression model to categorize the variable influencing it the most. This was achieved by implementing linear regression analysis in R. For estimating the linear regression model, the p-value and Adjusted R-squared score were seen. Including residuals vs. fitted value plot, histograms, and scatter plot for diagnostic analysis.

The explanatory value of the full regression model is low. The adjusted R^2 value of the regression model is 17.57% in the log-transformed state, and 2.84% without the log-transformation. Thus it can be concluded that qualitatively, there does not seem to be much use for the regression model when implementing on such a small , fluctuating data set. Or, it could be that the variables collected are not indicative of the overall goal and a new study could be conducted where the data collection is refocused on other variables along with a series of data points indexed in time order which could produce better results.

References

[1] Stocks:Dow Jones Indices-Barchart.com. (2022). Retrieved 27 June 2022, from https://www.barchart.com/stocks/indices/dowjones/industrials

[2] Team, T., Team, T. (2022). Exploring the Dow-Jones Industrial Average using Linear Regression. Retrieved 26 June 2022, from https://towardsai.net/p/l/exploring-the-dow-jones-industrial-average-using-linear-regression

[3] Linear regression on market data — Using Python and R. (2022). Retrieved 26 June 2022, from https://blog.quantinsti.com/linear-regression-market-data-python-r/

[4] Linear Regression Analysis in R. (2022). Retrieved 26 June 2022, from https://towardsdatascience.com/linear-regression-analysis-in-r-fdd59295d4a8

[5] Chen, S. (2022). Forecasting Daily Stock Market Return with Multiple Linear Regression. Retrieved 26 June 2022, from https://digitalcommons.latech.edu/mathematics-senior-capstone-papers/19/

[6] (2022). Retrieved 26 June 2022, from https://www.diva-portal.org/smash/get/diva2:942663/FULLTEXT01.pdf

[7] Example of Multiple Linear Regression in R - Data to Fish. (2022). Retrieved 26 June 2022, from https://datatofish.com/multiple-linear-regression-in-r/

[8] Modeling, M. (2022). Creating a multiple regression. Retrieved 26 June 2022, from https://community.rstudio.com/t/creating-a-multiple-regression/45295