*Candidate number: 16923*

*MLR for the PE ratio*

ST300

## Introduction

The price to earnings ratio (PE) is the current price of a stock divided by the company's earnings per share (EPS), where EPS is defined by the company's profit divided by the total number of stocks owned by shareholders. The PE indicates whether a stock is a good investment opportunity by predicting if a company is over or under-priced. A low PE indicates an under-priced stock that investors can buy cheaply, with low risk of loss and potentially high return on equity, ROE.

This paper creates a multiple linear regression model for PE ratio. If the given PE ratio by the model is higher than that of reality, it indicates that it is a good investment opportunity.

The categorical variables in the dataset are:

* Region
* Industry

The Continuous variables are:

* Institutional Holding
* Number of Firms
* Earnings per share growth (EPS Growth)
* **Price book Value (PBV)**
* **Beta**
* CEO Holding
* Cost of Equity
* **Returns on Equity (ROE)**
* **Price to earnings ratio (PE)**
* **Price to sales ratio (PS)**

*Variables in bold have reached the final model*

To reach the final model, an exploratory analysis was run. Transformations on heavily skewed variables were performed (all log transformations are to the base e) and outliers were spotted (but not removed). A model was run, and outliers were analysed and removed where appropriate. The final model was subject to diagnostics and checked against all underlying assumptions that it met. The model is as follows:

This means that while keeping all the other variables constant, an increase in ‘PBV’, ‘Beta’, ‘Number of Firms’, or ‘PS’, increases the expected value of ‘PE’. Or, as ‘ROE’ increases, the expected value of ‘PE’ decreases.

## The data set

Before any analysis could be ran, some of the variables had to be fixed as they were inappropriately coded as factors when they were numeric. This was the case for 5 variables, including ‘ROE’. Once the data types were rectified, the data was saved to a new file.

The data set contains 12 variables and 282 observations, taken from three different regions: ‘US’, ‘EUR’, and ‘EMG’; representing 94 different industries. For all variables, check the appendix\*1.

Chart, histogram

Description automatically generated

The histogram shows us the raw data plotted against the outcome variable in a histogram.

The data shows a strong positive (right) skew. The range of values is 7.05 to 1304.34. The mean, 59.78, is higher than the median, 35.22. There are three extremely large values. Due to the stochastic nature of financial variables, large values at the tails are to be expected. It is for this reason caution shall be taken before eliminating these points, ergo, the analysis shall carry on with them in, and they shall be investigated further. Within the data is a single missing value in row 69 for ‘PVB’; this causes issues for subset regression. As it's a large data set this observation shall be removed.

## Exploratory Analysis

To establish a relationship between each of the predictors and the ‘PE’ ratio, single variate analysis was conducted. Scatter plots were made for the nine continuous predictors and boxplots were made for the two factors; a simple linear regression was made for each predictor.

Many of the predictors' residual plots were funnel-shaped, indicating heteroskedasticity. This was confirmed by many of them having bow-shaped Q-Q plots. As there were issues with the majority of predictors, a log transformation of the data was carried out. This corrected the Q-Q plots from extremely bow-shaped to slightly bow-shaped, meaning that the change was not perfect, but still effective.

One variable it did not fix was ‘Number of Firms’. The residual vs fitted plot had a clear funnel shape and the distribution of the error terms showed a heavy positive skew. To correct this a log transformation was taken of that predictor variable.

Chart, scatter chart

Description automatically generatedBefore the log transformation:

Chart, histogram

Description automatically generated

After the log transformation:

Chart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

The new residual vs fitted plot presents its data in a band. While there is a lot of variation, it clearly shows the approximate constant variance of the error term, as required. Points 81,28,263 appear to be somewhat extreme values, but as previously mentioned, this is expected of financial data. Another predictor that benefitted from a transformation was ‘PS’; while the funnel shape in the data was less prominent, there was heavy positive skew (see appendix\*2).  
  
For the factors, box plots were made along with a Normal Q-Q of the simple linear regressions.

Chart, line chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

The box plot for ‘region’ shows that the median does not vary by a large amount between different regions and is approximately in the middle of the distribution for each, except in the case of EMG which is slightly higher. Those in ‘EMG’ are expected as the highest ‘PE’, and those in ‘EUR’ as the lowest ‘PE’. It also shows that there are a few potential outliers and the skew is not heavy (ignoring the outliers).

The normal Q-Q for ‘Region’ is very similar to the others where the upper end bends up slightly; this implies that there is positive skew. The normal Q-Q shows that the same three points are particularly extreme: 81, 28, 263.

Chart, scatter chart

Description automatically generated

After both transformations were performed, all of the graphs showed linearity. This suggests it satisfises the first condition [A1], linearity. However, some of the graphs visually needed outliers removed, this shall be addressed within in the model making and diagnostics. For the rest of the scatter plots, please see appendix\*3.

‘Industry’ is problematic due to its high number of factors. If these were merged in the most logical way, by industry type, there would still be too many variables to run best subset regression, thus not beneficial.

## Making the Model

After the transformations have been made and potential outliers have been identified, the first regression model is made. Three techniques are used to make the regression: backwards elimination, forwards selection, and best subsets. They are all based around minimizing AIC, but use different methods often resulting in different models.

Each time a model is created, forwards selection and backwards elimination run initially to check the variables included. If they do not include ‘Industry’, proceeding to make the model using best subsets is appropriate. This is because ‘Industry’ must be removed to run best subsets; best subsets calculates the AIC of every possible model, and as ‘Industry’ has 96 levels, the computing power needed to complete this is too high. Removing ‘Industry’ is justified as very few of the levels appear to be statistically significant in the single variate analysis. While it might be significant when combined with other variables, if it does not appear using either backwards elimination of forwards selection, it's highly unlikely that it helps minimize AIC.

For the base model, ‘Industry’ did not appear in either forward's selection or backwards elimination, so proceeding with best subsets is applicable. The variables in the initial model are as follows:

Residual Standard error: 0.7025 | Adjusted R-squared 0.2885 | P value of F-stat: p-value: < 2.2e-16

Then a model is made to remove the heavy outliers seen throughout, points: 28, 81, 263. While they are likely real values, this model is to predict ‘PE’ for average stocks, and these extreme points represent abnormally high values which are of no relevance to this model, and thus should be removed.

Neither backwards elimination nor forwards selection included ‘Industry’, so it may be left out of the best subsets.

The next model:

Residual Standard error: 0.66 | Adjusted R-squared 0.2367 | P value of F-stat: p-value: < 2.2e-16

This is a clear improvement; thus, this model should be kept over the first. Removing three values changed the model completely. Many variables have been eliminated in place of Beta. The residual standard error decreased from 0.7025 to 0.66. However, a decreased adjusted R-squared value is observed, from 0.2885 to 0.2367. This is explained by the first model having more parameters, possible being an overfitting model. For this reason, it is not of concern. However, this model only has three predictors, making it likely that this model is underfitting.

The next model produced removes the values with exceptionally high Cook’s distances. The only value that this applies to is point 89, with a value of around 2. The forwards selection and backwards elimination don't include ‘Industry’, and so best subsets regression may be used.

The next model:

Residual Standard error: 0.6376 | Adjusted R-squared 0.2869 | P value of F-stat: p-value: < 2.2e-16

In the next model, removing the 83rd value was attempted. This was due to its Cook’s distance being significantly higher than all the other points in the model, despite being under 1. ‘Industry’ does not appear in the first two selection processes, so best subsets may be used.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | P-value of t-stat |
| (Intercept) | 2.141 | 0.208 | 2e-16 |
| Beta | 0.364 | 0.130 | 0.005 |
| ROE | -0.014 | 0.004 | 0.0005 |
| PBV | 0.133 | 0.027 | 1.75e-06 |
| Log(Number of Firms) | 0.218 | 0.0344 | 1.06e-09 |
| Log(PS) | 0.135 | 0.0540 | 0.013 |

Residual Standard error: 0.6341 | Adjusted R-squared 0.2964 | P value of F-stat: p-value: < 2.2e-16

A single coefficient has changed; ‘EPS Growth’ has been changed for ‘ROE’. An interesting observation is that the correlation between these two variables is 0.963 which means there is a very strong correlation between the two predictors.

Most of the coefficient of each value remained fairly close to what they were previously. The largest change was ‘PBV’ going from 0.052 to 0.133, an increase of 156%. All the predictors except for Log(‘PS’) have increased from the previous model. All coefficients estimates are positive except for ‘ROE’. This means that as each of the variables except ‘ROE’ increases (while keeping the others constant), the expected value of ‘PE’ will also increase. Removing one point increases the adjusted r-squared value by 0.095 and decreases the residual standard error slightly. This is expected as r squared is a measure of the proportion of the data that is explained by the model, and in a moderately small data set removing a single non-explained value is likely to have this effect.

With a different set of predictors and a substantial change in ‘PBV’, it is clear that point 83 is an outlier. This model is an improvement upon the previous so shall be kept.

## Diagnostics and Model evaluation

The final model residual graph is as follows:

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

The first graph shows no problems as the residuals are approximately constant, forming a band on zero [A3] with no obvious pattern. This means that the data is free of heteroskedasticity [A4] and errors are pairwise uncorrelated [A5] thanks to the use of transformations. It also implies that observations are independent. Point 19 has a residual of just above two, but there are many values near it indicating that this is not an issue.

The normal Q-Q is adequate. While there is a long tail, skewing the data slightly positively, the data is approximately normally distributed so it satisfies the normality condition [A6].

Chart, histogram

Description automatically generated

The Cook's distance shows us that no points are highly influential, all being below 0.1. Point 24 is substantially higher than the others, so further analysis should be run with the absents of this point. As well as point 24, points 11, 69 and 169 need further analysis as they are apparent outliers of ‘ROE’/’Beta’, as shown by their extreme values in the plots.

Four additional models were created, each excluding a different one of these points and it's clear from the results that they do not have a substantial impact on the data. None of the models show a change in predictors or a large change in their estimates (For models, run the R code). For this reason, each of these models were rejected, and the prior model including these points is kept.

Checking the VIF within the model shows that all values are between 1 and 1.7 (see appendix\*4). The correlations between the variables are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ROE | PBV | Beta | Log(Number of Firms) | Log(PS) |
| ROE | 1 | 0.53926166 | 0.12493974 | -0.17083792 | 0.03035256 |
| PBV | 0.53926166 | 1 | 0.17092177 | -0.03524901 | 0.33775138 |
| Beta | 0.12493974 | 0.17092177 | 1 | 0.03606151 | 0.09619175 |
| Log(Number of Firms) | -0.17083792 | -0.0352490 | 0.03606151 | 1 | 0.22643204 |
| Log(PS) | 0.03035256 | 0.33775138 | 0.09619175 | 0.22643204 | 1 |

The correlations show that no variable is strongly correlated to another. However, there does appear to be some correlation between ‘ROE’ and ‘PBV’ with a value of 0.539. While that may appear high, it is not substantial due to the low VIF of the predictors. This moderate correlation may signify a lurking predictor.

With low VIFs and correlation scores between variables (except the one mentioned), the small changes in the estimates and std. errors when points are removed, and all the p values of t-stats in the model being significant: the requirement of no multicollinearity [A2] is satisfied.

Throughout this paper, it's been shown that the model satisfies the 6 underlying model assumptions (A1 – A6). We may conclude that the t-tests and F-tests are accurate, confirming the reliability of the results produced.

## Model Interpretation

The best model presented by this paper is:

This means that while keeping all the other predictors constant:

A one unit increase in the predictors ‘PBV’ or ‘Beta’ will cause the expected value of ‘PE’ to increase by 100%\*(e0.133-1) = 14.225; or 100%\*(e0.364-1) = 43.90742, respectively. ‘ROE’ is interpreted similarly but has negative relationship with ‘PE’. This means a one-unit increase in ‘ROE’ will lead to a decrease in ‘PE’ by 100%\*(e0.014 -1) = 1.410, each time. For the coefficients of Log[‘Number of Firms’] or Log[‘PS’], a one per cent change leads to a 0.217% or 0.135% change in ‘PE’ respectively.

The plausibility of the estimations can be examined. The ranges of values can be used to judge how much each predictor can change the estimate of Log(‘PE’).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Coefficient | PVB | ROE | Beta | Log(PS) | Log(Number of Firms) |
| Range | 18.9 | 112.69 | 2.2 | 4.6707 | 6.1071 |
| Possible impact on PE | 268.8525 | 158.8755 | 96.59632 | 21.7%  =75.78291 | 13.5%  =47.14605 |

The range of ‘PE’ is 349.23. All the variables seem realistic as they are all smaller than the value of ‘PE’, but cumulatively cover this range, and thus are neither too big nor small. Moreover, all variables are relevant, each large enough to have a considerable contribution on the model. The most impactful variable is ‘PVB’ followed by ‘ROE’. The least impactful included variable on ‘PE’ is the ‘Number of Firms’.

## Model Limitations

While the model appears effective, it has large 'random' variations, typical for financial data, as shown by the low R-squared value. This means the data is not tight to the regression line and that the true ‘PE’ may be notably higher or lower than predicted. There are questions as to how useful the ‘PE’ is as it evaluates past performance against past earnings, when this may not be relevant to future results. This said, ‘PE’ is used among other tools can be helpful in the right circumstance, and this is a simplistic model to generate the ‘PE’ from a few predictors.

Another limitation is not knowing where this data was collected from, as it might not be a representative sample of the population of stocks. This would cause the model to be wrong. Another issue with the data is the relatively small data set. There are hundreds of thousands of shares, so a sample of 282 is not large. To improve the reliability of the model, using a larger data set would help. This is also a pitfall in regard to extrapolation. The model may only be used for values that lie within the dataset’s range, and as it’s a small data set, the range of values this model may be used to interpolate is moderately small. This interpolation works for categorical variables too, data was only collected from the USA, EUR and EMG meaning that the model can only predict reliably for shares from these regions.

Lastly, the efficient market hypothesis states that asset prices reflect all the information available; if this is true, it would imply that no stock can be over or undervalued, meaning that the ‘PE’ would already be ‘priced in’ to the stock’s value.

# Appendix

## Variables:

Table

Description automatically generated

Credit for table: Phil Chan, D.1 Data description

<https://moodle.lse.ac.uk/pluginfile.php/1516457/mod_resource/content/1/ST300%20Project_data_description.pdf>

Last accessed 23 December 13:45

## Transformation of PS:

Before Transformation:

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, histogram

Description automatically generated

After Transformation:

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, histogram

Description automatically generated

1. Linearity plots

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

For the plots not in the model, please see the R code.

## VIF:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | PBV | Beta | Log(Number of Firms) | Log(PS) | ROE |
| VIF | 1.667 | 1.036 | 1.093 | 1.237 | 1.497 |