ST300 PROJECT

Candidate number: 17915

Introduction

The objective of this project was to use R to analyse the dataset provided and develop a model to explain how the Price-earnings ratio (PE) was related to other financial variables. The PE ratio is used to evaluate whether the price of a stock is high/low compared to similar stocks and measures the current price of a stock relative to its earnings per-share (EPS), where EPS is defined as the company's profit divided by the outstanding shares of its common stock. [1]

The dataset used to build the model comprised of 12 financial variables corresponding to 282 datapoints across 94 industries. These financial variables included:

Categorical variables: Region, Industry

Continuous variables: Number of firms, ROI (Return on Equity as a %), EPS Growth (Earnings per share growth as a %), PBV (Price-Book Value), PS (Price to sales ratio), Beta (a value estimated from the CAPM model), Cost of Equity, CEO holding, Institutional holding

Dependent variable: PE (Price to earnings ratio)

The relationship between PE and the other financial variables was determined by carrying out an exploratory data analysis, transforming the required variables (all log transforms used were with respect to base e), creating the first model, removing the influential points and developing the final model from the initial model.

From the final model we can infer that there was a positive relationship between the PE ratio and the following variables: Price-Book Value (PBV), Beta and the log (Number of firms). Further, a negative relationship was observed between Return on Equity (ROE) and the PE ratio.

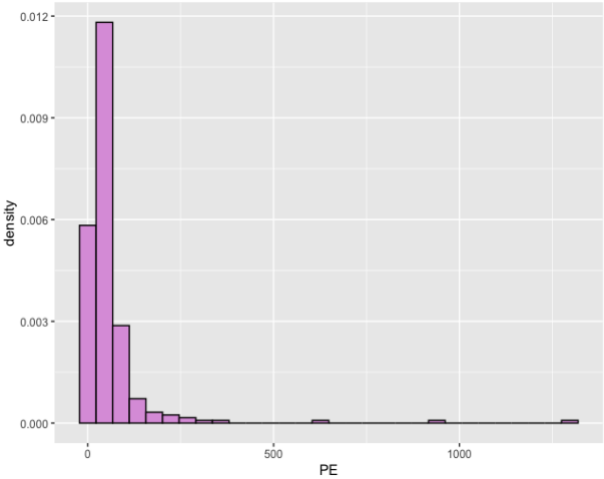
The final model obtained was

$\log(PE) = 2.1308 - 0.0184\ ROE + 0.1527\ PBV + 0.3812\ Beta + 0.2269 \log(\text{Number of firms}) + \varepsilon$

Exploratory data analysis

The dataset was pre-processed prior to conducting an exploratory data analysis. The continuous predictors wrongly coded as factors were converted to numerical variables. Upon observing the summary of the data, a missing data value (N/A entry) for PBV was identified and the row containing this value was removed.

Figure 1 – Histogram of PE



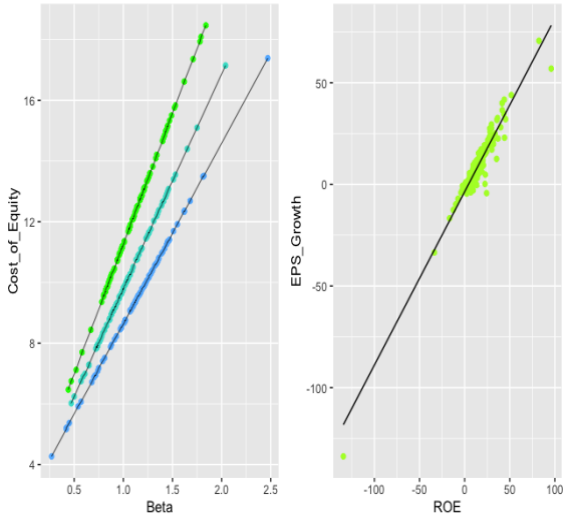
The first step of the exploratory analysis was to conduct a univariate analysis of the PE ratio. The histogram in Figure 1 reveals that PE was heavily positively skewed with a mean value of (59.78) exceeding its median (35.22). Further, the values of PE range from 7.05 to 1304.34. Figure 1 also shows the presence of three datapoints that corresponded to a value of PE exceeding 600. Datapoint 81 (80 after removing 69) corresponded to the largest value of PE. The presence of such extreme values was unsurprising due to the stochastic nature of financial variables. A log transform was applied to PE for multiple reasons. Firstly, the correlation matrix and the scatterplot matrix revealed that the correlation among the independent predictors (all except EPS Growth and PBV) and log (PE) was more prominent than the relationship between the independent predictors and PE. Secondly, applying a log transform made the data more gaussian (evenly spread out) and the mean of the data was approximately equal to the median.

Afterwards, univariate analyses were conducted to observe the nature of the independent variables. Bar charts were used to plot categorical variables and histograms were used to observe continuous predictors. Some predictors such as the Number of firms (left - Figure 4), PBV and PS were heavily positively skewed. Datapoint

263 (262 after removing 69) was less than the bulk of the data for ROE and EPS Growth. Datapoints 83, 71 and 89 (82, 70 and 88 after removing 69) were noticeably larger than the rest of the data for PBV.

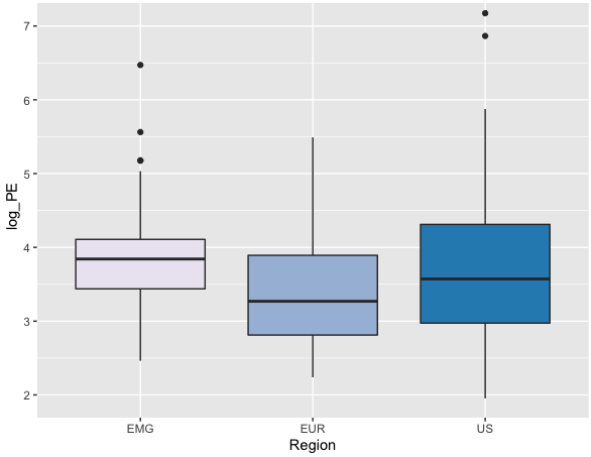
Next, the relationships among the independent variables were explored. A scatterplot matrix and a correlation matrix were plotted to identify such relationships. From the correlation matrix, Beta and Cost of Equity had a very strong positive correlation of 0.8538. Beta measures the sensitivity of a stock relative to its market risk. Therefore, when the risk associated with the stock (Beta) increases, the cost of equity will also increase. Upon closer inspection, all points of the scatterplot lay on 3 distinct lines (left - Figure 2). This indicates that there was a lurking predictor which affected this relationship. ROE and EPS Growth also had a very strong positive correlation of 0.9588. (right - Figure 2). This maybe because both variables are profitability ratios and an increase/decrease in profitability will have the same impact on both variables. Interestingly, ROE and EPS Growth also had similar ranges.

Figure 2 – Relationship between independent predictors



Then, the relationship between the independent variables and the log (PE) was examined. The correlation among the independent variables and log (PE) was first observed using a correlation matrix. Afterwards, boxplots were plotted for the categorical predictors against log (PE), scatterplots for the continuous predictors against log (PE) and simple linear regressions were conducted for all variables with log (PE) as the dependent variable. As per the correlation matrix, variables such as PS, EPS Growth and Number of firms had the strongest relationship with log (PE). However, these were rather weak correlations (0.2909, -0.2598 and 0.2310 respectively).

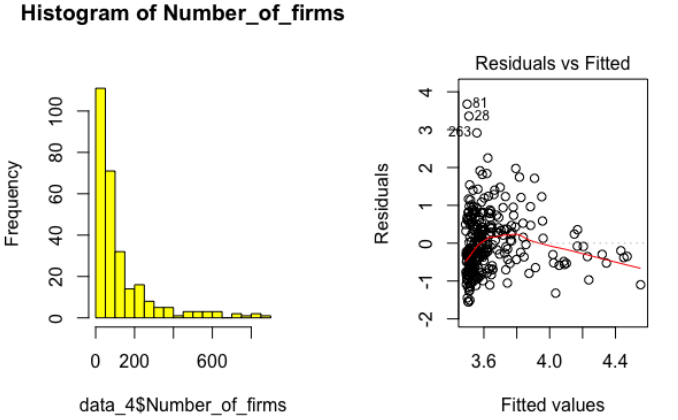
Figure 3 – Boxplot of Region vs log (PE)



The boxplot for Region vs log (PE) reveals that EMG and US had similar medians whilst EMG and EUR did not (Figure 3). Further, the US had the largest inter quartile range. The summary of Region revealed that this variable was bi-modal as both EMG and EUR had the highest number of data points (94). When a simple linear regression (SLR) was used to model Region against PE, the default base level for Region was EMG (as it was alphabetically first). As this was one of the modes, the base level was not changed.

Due to factors such as risk and growth rate, the PE ratio greatly varies across Industries. [2] Therefore, it was extremely difficult to merge levels of this predictor. The difference in log (PE) across Industries was clearly observed from the boxplot of log (PE) against Industry.

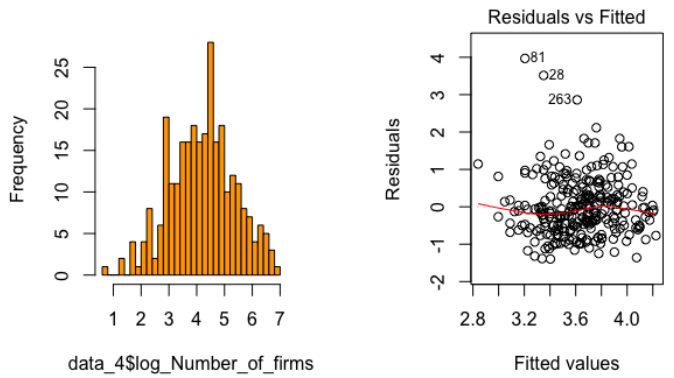
Figure 4 – Number of firms



Regressing log (PE) on the Number of firms returned a significant t-statistic. However, a funnel shape was observed in the residuals vs fitted plot (right - Figure 4). This signalled that the residual assumption of homoscedasticity (A4) was violated. A log transform was used on the Number of firms as the residual assumption of homoscedasticity was violated, the recorded Number of firms exceeded 1 and the variable was heavily skewed

(left - Figure 4). After the transformation, the significance of the predictor increased, and a constant variance of residuals was observed (right - Figure 5). Further, the log transformation made the data evenly spread out (left - Figure 5). Similarly, a log transformation was applied to the variable PS for the same reasons (Refer appendix – 3).

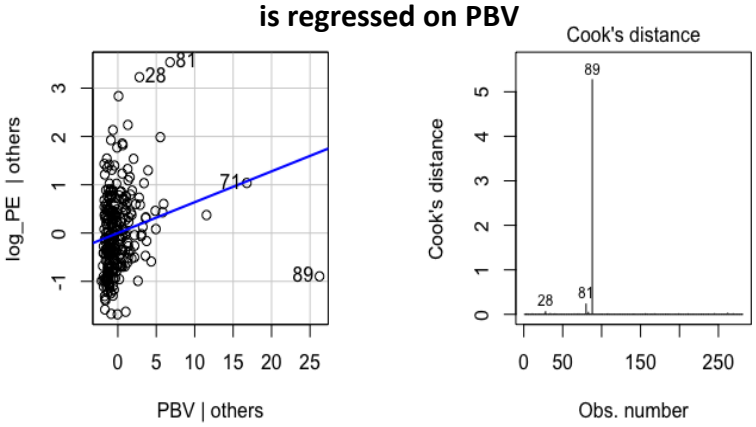
Figure 5 – log (Number of firms)



The scatterplots of ROE (Return on Equity) vs log (PE) and EPS Growth vs log (PE) were extremely similar (Refer appendix – 4). Both plots showed a negative relationship with log (PE). The negative correlation between EPS Growth and log (PE) was unsurprising as the PE ratio is inversely proportional to EPS Growth (from the formula used to calculate PE).

When a simple linear regression was created with PBV as the independent variable and log (PE) as the dependent variable, the slope coefficient was highly significant as per the t-test. However, datapoint 89 (88 after removing 69) was identified as a potential outlier after observing the added - variable plot (left - Figure 6) and the Cook's distance plot (right - Figure 6). This unusual datapoint corresponded to a high PBV, a low log (PE) and an abnormally large Cook's distance greater than 5. A simple linear regression of PBV vs log (PE) was run after removing this point. As a result, the t-statistic significantly improved, and the coefficient estimate increased from 0.0639 to 0.1256. Therefore, this data point was removed from the model.

Figure 6 – AVP and Cook's distance plot when log (PE) is regressed on PBV

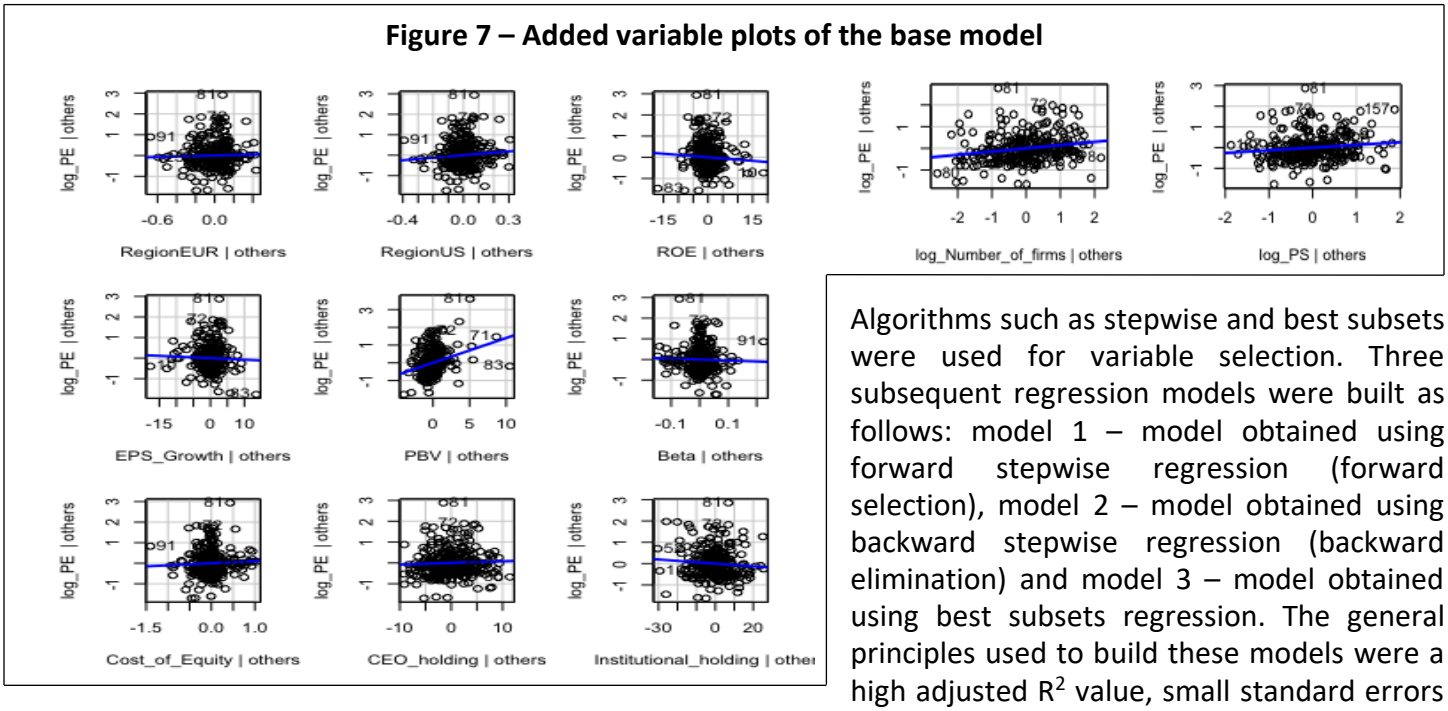


From the exploratory data analysis, it was decided that a multiple linear regression would best capture the relationship between log (PE) and the other variables for two reasons. Firstly, the dependent variable (log (PE)) was continuous. Secondly, an approximately linear relationship was observed between log (PE) and the independent financial variables.

Initial to final model

The base model was thus created with all independent variables bar industry. Including Industry would have increased the precision of the base model but reduced the interpretability of the model as Industry had 94 levels. As a result, Industry was excluded from the base model. The variables with significant t-statistics in the base model were PBV, log (Number of firms) and log (PS). The model assumptions A1 to A6 were satisfied (these model assumptions will be discussed in greater length on page 5). However, 81 (80 after removing 69 and 89) was identified as a candidate for an influential point after looking at the added variable plot (Figure 7) and Cook's distance plot. The base regression was then rerun without this point. The new regression had a marginally better adjusted R^2, lower AIC (Akaike information criterion) and an additional significant predictor compared to the base model. Therefore, datapoint 81 was classified as an outlier and removed. Variable selection was then done based off this dataset.

Figure 7 – Added variable plots of the base model



Algorithms such as stepwise and best subsets were used for variable selection. Three subsequent regression models were built as follows: model 1 – model obtained using forward stepwise regression (forward selection), model 2 – model obtained using backward stepwise regression (backward elimination) and model 3 – model obtained using best subsets regression. The general principles used to build these models were a high adjusted R^2 value, small standard errors for estimated coefficients and a low AIC (Akaike information criterion).

A forward stepwise regression starts with no predictors in the model and iteratively added the most contributive predictor at each step to create a model which explained the data. This process stopped when the AIC increases. Region, Institutional holding and log (PS) were significant predictors in the models obtained using forward selection and backward elimination (Refer Table 1). The predictor CEO Holding was only included in model 1 and this was the only predictor with a non-significant t-statistic that was included in model 1.

A backward stepwise regression is similar to forward selection but starts with all the predictors and then iteratively eliminates the least contributive predictor until the AIC increases. All predictors (including Industry) were included in the initial backward elimination model. As Industry had the highest AIC, it was the first variable to be removed. All predictors were statistically significant in the final backward elimination model (Refer Table 1). Surprisingly, this was the only model that included log (PS), EPS Growth and log (Number of firms) as significant predictors (As per the exploratory analysis, these variables had the strongest relationship with log (PE).

A best subset regression compared all possible models using a specified set of predictors and displayed the best-fitting model. Best subset regressions used both forward selection and backward elimination. As Industry had 94 levels, running a subset regression with this variable would take significant time and computing power. Further, excluding Industry as a predictor from the best subset regression was justified as it was not included in the final models obtained using forward selection and backward elimination. All predictors included in model 3 were statistically significant (Refer Table 1). Interestingly, EPS Growth was not a significant predictor whereas ROE was. The log (Number of firms) and PBV were the only two predictors that were statistically significant in all 3 models.

I chose the model obtained using best subsets (model 3) as the model that explained the data best for a number of reasons. Firstly, the best subsets algorithm fits all possible models based on the specified predictors and selects the best model with the lowest AIC or adjusted R^2. Secondly, all predictors included in model 3 had significant t-statistics. Thirdly, this model contained only 4 predictors and this reduced the danger of overfitting.

Table 1: Comparing the three models

Name of predictor	Coefficient of the predictor included in the model obtained using stepwise regression based on AIC		Coefficient of the predictor included in the model obtained using best subsets (model 3)
	Forward selection (model 1)	Backwards elimination (model 2)	
(Intercept)	2.5698***	2.3629***	2.1308***
Region EUR	0.0016	0.0551	-
Region US	0.3685.	0.5109**	
Industry	-	-	-
ROE	- 0.0170***	-	- 0.0184***
EPS Growth	-	- 0.0186***	-
PBV	0.1236***	0.1077***	0.1527***
Beta	0.3002*	-	0.3812**
Cost of Equity	-	0.0473*	-
CEO holding	0.0097	-	-
Institutional holding	- 0.0077.	- 0.0086*	-
Log (Number of firms)	0.1606***	0.1679***	0.2269***
Log (PS)	0.1270*	0.1298*	-
Significance codes: '***' 0.001 | '**' 0.01 | '*' 0.05 | '.' 0.1			
Diagnostics			
Residual standard error	0.6463	0.6457	0.6581
Multiple R^2 / Adjusted R^2	0.3618/ 0.3405	0.3605/ 0.3416	0.3260/ 0.3162
Significance of the F-statistic	< 2.2e-16	< 2.2e-16	< 2.2e-16

Diagnostics of model 3: All predictors were significant at the 1% level. The F- test measures the overall significance of a specified model by comparing it to a model with 0 predictors (the null hypothesis). [3] Model 3 had a very small p - value for the F - test of overall significance. Therefore, there was strong evidence against the null hypothesis. An adjusted R^2 value of 0.3162 (model 3) signified that there was a lot of variability in the data. However, a model cannot be rejected simply because it had a low adjusted R^2 value. The residual standard error was 0.6581 and this was relatively small compared to the range of log (PE) (6.865 – 1.953 = 4.912). The credibility of the model is high as the residual standard error was only 13% of the range of the data.

Model assumptions of model 3 : As the residuals vs fitted plot shows no fitted pattern (The red line in Figure 8 is approximately horizontal at 0), the model assumption of linearity (**A1**) was not violated. From the residuals vs fitted plot (Figure 8), the data points were approximately symmetrically concentrated around zero. Therefore, the zero mean assumption (**A3**) holds. As a random scatter of points was observed in the residuals vs fitted plot (Figure 8), the residuals had a constant variance (**A4**) and were not correlated (**A5**).

Figure 8 – Residuals vs Fitted Plot

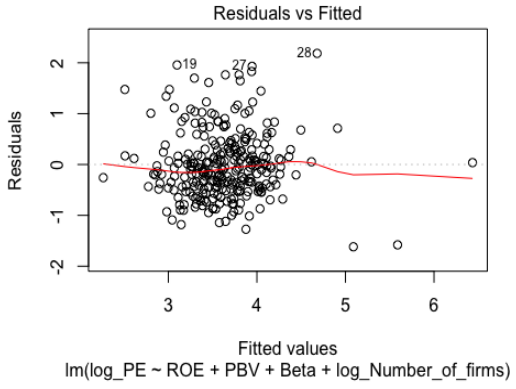
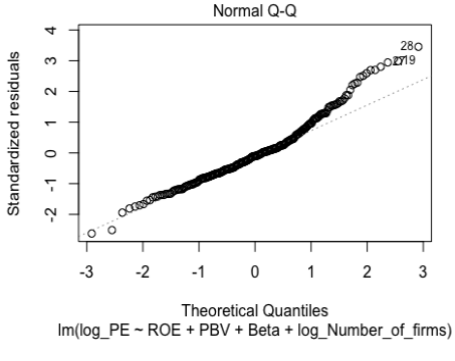


Figure 9 – Normal Q-Q Plot



Some kurtosis was observed in the model as the normal q-q plot (Figure 9) had heavy tails. However, the residuals approximately followed a normal distribution (**A6**).

As per the added variable plots, datapoints 28 and 83 (28 and 81 after removing datapoints 69,80 and 89) exhibited a different trend to the rest of the data. Therefore, they were identified as candidates for influential point. However, the Cook's distance of these points was less than 0.3. Further, the regression run without these points did not have a significant improvement. As a result, 28 and 83 were included in the model.

Next, I checked for multicollinearity in the model using the variance inflation factor (VIF). As model 3 consisted only of continuous predictors, the VIF could be used to detect multicollinearity in the model. As the VIF of all predictors was less than 1.2 (Refer appendix - 6), there was no multicollinearity in the model. Thus, model assumption **A2** was satisfied.

Interpretation of the final model

The final model obtained was:

log (PE) = 2.1308 - 0.0184 ROE + 0.1527 PBV + 0.3812 Beta + 0.2269 log (Number of firms) + ε

The coefficient estimate of the intercept was interpreted as the expected mean value of Log (PE) when ROE, PBV, Beta and log (Number of firms) equal 0. Therefore, the expected mean value of PE was approximately 8.4216 ($e^{2.1308}$) when ROE, PBV and Beta were 0 and Number of firms was 1.

The coefficient estimate of ROE was - 0.0184 and this means that there was a 1.82% ((1-$e^{-0.0184}$) *100) expected decrease in PE for a unit increase of ROE, given that all other factors remained a constant. Since ROE is a measure of profitability, one would expect companies with a higher ROE to obtain a higher PE. However, the relationship between these two ratios may not be too meaningful at a point in time. It's the trend that makes sense. This could be the reason for a negative correlation at the given point in time. The coefficient estimate for PBV was interpreted as a unit increase in PBV led to a 16.50% (($e^{0.1527}$-1) *100) expected increase in PE ceteris paribus. The coefficient estimate of Beta was interpreted as a unit increase in Beta multiplied the expected value of PE by 1.46 ($e^{0.3812}$). Whilst there is no direct link between the PE ratio and Beta, this positive correlation is not surprising. Since Beta measures risk, a higher risk indicates that shareholders will expect a higher return. Higher PE also indicates that there is an expectation of a higher future growth in the market price. The coefficient estimate of the log (Number of firms) reveals that a 1% increase in the number of firms leads on average to an expected change (increase) in PE by 0.23% (0.2269/100). Beta had the largest impact on PE.

The largest recorded value of log (PE) corresponded to datapoint 28. Its values for ROE, PBV, Beta and the log (Number of firms) were -33.49, 5.31, 1.19 and 2.944 respectively. Using the final model, the expected log (PE) of datapoint 28 was 4.679 and the expected PE was 107.71($e^{4.679}$). The actual value of log (PE) corresponding to datapoint 28 was 6.865.

Calculation of log (PE) for datapoint 28 using the model:

log (PE) = 4.679 = 2.1308 - 0.0184(-33.49) + 0.1527(5.31) + 0.3812(1.19) + 0.2269(2.9444).

Conclusion and understanding the limitations of the model

The final model developed was satisfactory to explain the relationship between the independent variables and the PE ratio. However, there were some limitations of the model. Firstly, it was difficult to infer how good the model was without carrying out cross-validation. Secondly, this model was created based on a cross sectional dataset. However, most financial variables (PE, ROE, PB etc.) are monitored over a period of time. Therefore, the model could have been improved if a time series dataset was used to build the same. Accordingly, this sample may not be representative of the market behaviour. Further, there is a large variation among financial data. As a result of all the above, extrapolation based on this model maybe difficult. Lack of expertise on the subject matter was a limiting factor when conducting the analysis and this may have impacted the model.

Appendix

1. References

[1] Khandelwal, Disa Vivek. (22/05/2020). Important financial parameters you should know before investing. Internet Source. Retrieved from https://www.caclubindia.com/articles/important-financial-parameters-know-before-investing-41626.asp

[2] Frankel, Matthew. (18/10/2017). Why is there so much variation in the P/E ratio of different companies? Internet Source. Retrieved from https://www.fool.com/investing/2017/10/18/why-is-there-so-much-variation-in-the-pe-ratio-of.aspx

[3] Unknown Author. (11/06/2015). What is the F-test of overall significance in regression analysis? Internet Source. Retrieved from https://blog.minitab.com/blog/adventures-in-statistics-2/what-is-the-f-test-of-overall-significance-in-regression-analysis

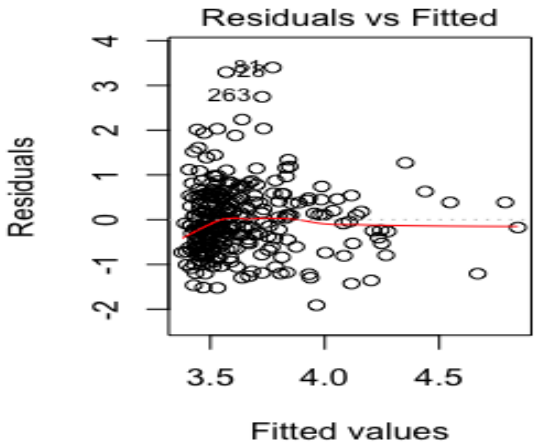
[4] Damodaran, Aswath. (25/08/2014). Session 15: PE Ratios. YouTube video. Retrieved from https://www.youtube.com/watch?v=42iyR6Geqiw&feature=youtu.be

2. Variable Definitions

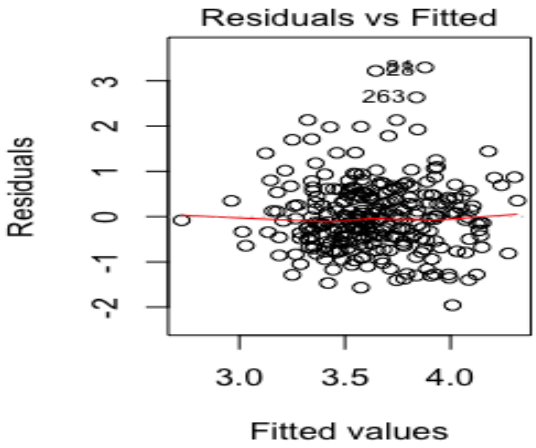
Variable	Data type	Description
Region	Factor	Levels are: US (USA), EUR (Europe) and EMG (Emerging markets)
Industry	Factor	List of sectors
Institutional holding	Numeric	Percentage of shares held by institutions
Number of firms	Numeric/ Integer	Number of firms in each sector
EPS Growth	Numeric	Earnings-per-share growth in %
PBV	Numeric	Price-Book value
Beta	Numeric	Estimated from CAPM
CEO holding	Numeric	Percentage of shares held by CEO
Cost of holding	Numeric	In percentages
ROE	Numeric	Return-on-equity (%)
PE	Numeric	Price-to-earnings ratio
PS	Numeric	Price-sales ratio

3. Residuals vs Fitted Plots for before and after a log transformation was applied to PS

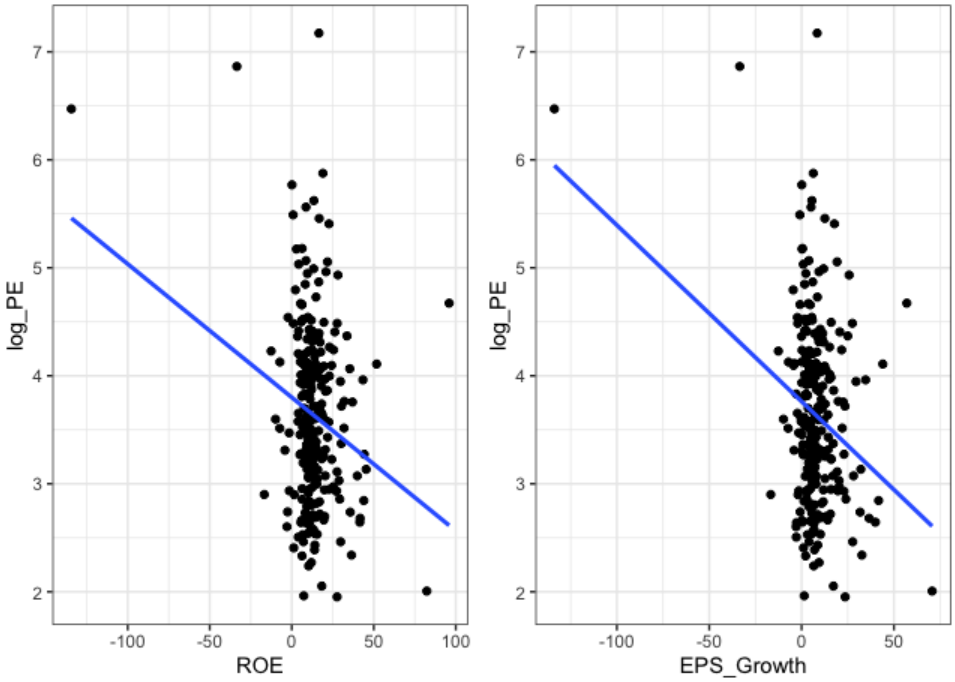
Before the transformation



After the transformation



4. Scatterplots for ROE vs log (PE) (left) and EPS Growth vs log (PE) (right)



5. Model 3 – R Output

```
Call:
lm(formula = log_PE ~ ROE + PBV + Beta + log_Number_of_firms,
    data = data_5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61803 -0.43671 -0.06649  0.29345  2.18526

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.130818   0.209692  10.162  < 2e-16 ***
ROE               -0.018401   0.002881  -6.387 7.21e-10 ***
PBV                0.152744   0.022055   6.925 3.09e-11 ***
Beta               0.381171   0.134020   2.844  0.00479 **
log_Number_of_firms  0.226897   0.034344   6.607 2.04e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6581 on 274 degrees of freedom
Multiple R-squared:  0.326,     Adjusted R-squared:  0.3162
F-statistic: 33.13 on 4 and 274 DF,  p-value: < 2.2e-16
```

6. Detecting multicollinearity in the model using vif

Variable	VIF Value
ROE	1.155452
PBV	1.179398
Beta	1.037709
Log (Number of Firms)	1.015100

7. R code

```
# Data Pre-processing
data_1 <- read.csv("Market_data2020.csv", header=T)
head(data_1)
summary(data_1)


## ROE, EPS_Growth, Cost_of_Equity, CEO_holding and Institutional_holding are wrongly recognised as factors
```

```
## Convert these factors to numbers(i.e. as.numeric)

## Strip out % and convert factor to numeric overwriting existing variable

data_1$ROE <-  as.numeric(gsub("[\\%,]", "", data_1$ROE))

data_1$EPS_Growth <-  as.numeric(gsub("[\\%,]", "", data_1$EPS_Growth))

data_1$Cost_of_Equity <-  as.numeric(gsub("[\\%,]", "", data_1$Cost_of_Equity))

data_1$CEO_holding <-  as.numeric(gsub("[\\%,]", "", data_1$CEO_holding))

data_1$Institutional_holding <-  as.numeric(gsub("[\\%,]", "", data_1$Institutional_holding))

str(data_1)


write.csv(data_1,'Market_data_clean.csv') ## After the changes, save the  changes to a new data file


# Install packages and formatting the dataset

library(ggplot2)

library(dplyr)

library(gridExtra)

library(car)

library(leaps)


data_2 <- read.csv("Market_data_clean.csv", header=T)

summary(data_2) ## There is a N/A entry in PBV

data_2[69,] ## The N/A entry is in row 69


## Remove the row with the N/A entry and create data_3

data_3 <- data_2[-69,] # Removing this datapoint makes point 70 correspond to datapoint 69

summary(data_3) ## The column X is unnecessary
```

```
## Create data_4 by removing the variable X

data_4 <- data_3[,-1]

summary(data_4) ## There are no missing values in the data


# Exploratory Data Analysis (EDA)


# 1. PE vs log_PE

summary(data_4$PE)

Histogram_PE <-ggplot(data=data_4)+

  geom_histogram(aes(x=PE,y=..density..),fillcolour= "plum")

Histogram_PE

round(cor(data_4[,-c(1,2)]), digits=4) ## Check the correlation among the independant variables vs PE/log_PE.

data_4$log_PE <- log(data_4$PE) ## Create a log version of PE

summary(data_4$log_PE)

Histogram_log_PE <- ggplot(data=data_4)+

  geom_histogram(aes(x=log_PE,y=..density..))+ labs(title="Histogram of log(PE)")

Histogram_log_PE

options(max.print=1000000)

data_4[order(data_4$log_PE),]

# Datapoints 263 (262 after removing point 69), 28 and 81 (80 after removing point 69) lie away from the others.


# Histograms of PE and a log version of PE

grid.arrange(Histogram_PE, Histogram_log_PE, ncol=2)
```

```
# 2. Observe the nature of the independent predictors
Bar_chart_1 <- ggplot(data=data_4)+
  geom_bar(aes(x=Region))+
  labs(title = "Region", x = "Region", y = "Frequency")+
  theme(plot.title = element_text(hjust = 0.5))
Bar_chart_1
RegionTab<-table(data_4$Region)
RegionTab
nlevels(data_4$Industry) ## 94
Bar_chart_2 <- ggplot(data=data_4)+
  geom_bar(aes(x=Industry))+
  labs(title = "Industry", x = "Industry", y = "Frequency")+
  theme(plot.title = element_text(hjust = 0.5))
Bar_chart_2
IndustryTab<-table(data_4$Industry)
IndustryTab
Histogram_1 <- ggplot(data=data_4)+
  geom_histogram(aes(x=Number_of_firms,y=..density..))
data_4[order(data_4$Number_of_firms),]
Histogram_2 <- ggplot(data=data_4)+
  geom_histogram(aes(x=ROE,y=..density..))
data_4[order(data_4$ROE),] ## 263 (262 after removing 69) was less than the bulk of the data
Histogram_3 <- ggplot(data=data_4)+
  geom_histogram(aes(x=EPS_Growth,y=..density..))
data_4[order(data_4$EPS_Growth),] ## 263 (262 after removing 69) was less than the bulk of the data
```

```
Histogram_4 <- ggplot(data=data_4)+

  geom_histogram(aes(x=PBV,y=..density..)) ## 71 and 89

data_4[order(data_4$PBV),] ## 83(82 after removing 69), 71(70 after removing 69) and 89(88 after removing 69) were further than the rest of the data

Histogram_5 <- ggplot(data=data_4)+

  geom_histogram(aes(x=PS,y=..density..))

data_4[order(data_4$PS),] ## No datapoint was significantly away from the rest of the data

Histogram_6 <- ggplot(data=data_4)+

  geom_histogram(aes(x=Beta,y=..density..))

data_4[order(data_4$Beta),] ## 91 (90 after removing 69) was further than the rest of the data

Histogram_7 <- ggplot(data=data_4)+

  geom_histogram(aes(x=Cost_of_Equity,y=..density..))

data_4[order(data_4$Cost_of_Equity),] ## No datapoint was significantly away from the rest of the data

Histogram_8 <- ggplot(data=data_4)+

  geom_histogram(aes(x=CEO_holding,y=..density..)) ## No datapoint was significantly away from the rest of the data

Histogram_9 <- ggplot(data=data_4)+

  geom_histogram(aes(x=Institutional_holding,y=..density..))

data_4[order(data_4$Institutional_holding),] ## No datapoint was significantly away from the rest of the data


grid.arrange(Histogram_1, Histogram_2, Histogram_3, Histogram_4, Histogram_5, Histogram_6 ,Histogram_7,

          Histogram_8, Histogram_9,nrow=3, ncol=3)

## Number_of_firms, PBV, PS, CEO_holding and Institutional_holding are positively skewed

## ROE, EPS_Growth, Beta, Cost_of_Equity are not skewed
```

```
# 3. Looking for relationships between the independant variables


round(cor(data_4[,-c(1,2)]), digits=4) ## Produce a correlation matrix

pairs(data_4[,-c(1,2)]) ## Produce a scatterplot matrix


## From the correlation matrix and the scatterplot matrix, strong correlationss are observed between Beta vs Cost_of_Equity and ROE vs EPS_Growth,


# Beta vs Cost_of_Equity

cor(data_4$Cost_of_Equity, data_4$Beta) ## 0.8538 so clear positive correlation

## The points lie on 3 clear lines

dat_1 <- data_4[1:93,]

dat_2 <- data_4[94:187,]

dat_3 <- data_4[188:281,]

ggplot(mapping = aes(x = Beta, y = Cost_of_Equity)) +

  geom_point(data = dat_1, color = "steelblue1") + geom_line(data = dat_1, color = "black",lwd = 0.25) +

  geom_point(data = dat_2, color = "turquoise") + geom_line(data = dat_2, color = "black",lwd = 0.25) +

  geom_point(data = dat_3, color = "green") + geom_line(data = dat_3, color = "black",lwd = 0.25)

cor(data_4[1:93,c(9,10)]) ## 0.9999651

cor(data_4[94:187,c(9,10)]) ## 0.9999446

cor(data_4[188:281,c(9,10)]) ## 0.9999534


## ROE vs EPS_Growth

## They have a similar range as well

ggplot(data_4,aes(x=ROE, y=EPS_Growth))+
```

```
  geom_point(color = "greenyellow")+
  geom_smooth(method=lm,se=F, color = "black", lwd = 0.5)
cor(data_4$ROE, data_4$EPS_Growth) ## 0.9588 so almost perfect positive correlation

# 4. Relationship between the independent and log_PE
Boxplot_Region <- ggplot(data=data_4,aes(x=Region, y=log_PE), colour = Region) +
  geom_boxplot()
Boxplot_Region ## EMG, EUR and US have similar medians
tapply(data_4$log_PE,data_4$Region,median, na.rm=T)
Boxplot_Industry <- ggplot(data=data_4)+
  geom_boxplot(aes(x=Industry, y=log_PE))

Boxplot_Industry ## Medians greatly differ across industries
tapply(data_4$log_PE,data_4$Industry,median, na.rm=T)
summary(data_4$Industry)

Scatterplot_Number_of_firms <- ggplot(data_4,aes(x=Number_of_firms, y=log_PE))+
  geom_point()+
  theme_bw()+
  geom_smooth(method=lm,se=F)
Scatterplot_Number_of_firms
m.Number_of_firms <- lm(log_PE~Number_of_firms, data_4)
summary(m.Number_of_firms) ## The slope coefficient is highly significant (t=3.966, p<0.01).
par(mfrow=c(2,2))
plot(m.Number_of_firms) ## Residuals vs fitted has a funnel shape
```

```
summary(data_4$Number_of_firms)

# Use a log transform the assumption of constant variance is violated and the values are greater than 0

data_4$log_Number_of_firms <- log(data_4$Number_of_firms)

m.log_Number_of_firms <- lm(log_PE~log_Number_of_firms, data_4)

summary(m.log_Number_of_firms) ## The slope coefficient is highly significant (t=5.647, p<0.01).

par(mfrow=c(2,2))

plot(m.log_Number_of_firms) ## Residuals vs fitted did not have a funnel shape

plot(m.log_Number_of_firms, which = 4) ## No points with a Cook's distance greater than 1

car::avPlots(m.log_Number_of_firms) ## 81 and 28 are candidates for influential points


Scatterplot_ROE <- ggplot(data_4,aes(x=ROE, y=log_PE))+

  geom_point()+

  theme_bw()+

  geom_smooth(method=lm,se=F) # Most of the points are -25 to 50 but there are a few points outside this range

Scatterplot_ROE

m.ROE <- lm(log_PE~ROE, data_4)

summary(m.ROE) ## The slope coefficient is highly significant (t=-3.782, p<0.01).

par(mfrow=c(2,2))

plot(m.ROE)

plot(m.ROE, which = 4) ## No points with a Cook's distance greater than 1

car::avPlots(m.ROE) ## 28,71,81 and 263 are candidates for influential plots


Scatterplot_EPS_Growth <- ggplot(data_4,aes(x=EPS_Growth, y=log_PE))+

  geom_point()+
```

```
  theme_bw()+
  geom_smooth(method=lm,se=F) # Most of the points are -25 to 50 but there are a few points outside this range
Scatterplot_EPS_Growth
m.EPS_Growth <- lm(log_PE~EPS_Growth, data_4)
summary(m.EPS_Growth) # The slope coefficient is highly significant (t=-4.494, p<0.01).
par(mfrow=c(2,2))
plot(m.EPS_Growth)
plot(m.EPS_Growth, which = 4) ## No points with a Cook's distance greater than 1
car::avPlots(m.EPS_Growth) ## 28,81 and 263 are candidates for influential plots


Scatterplot_PBV <- ggplot(data_4,aes(x=PBV, y=log_PE))+
  geom_point()+
  theme_bw()+
  geom_smooth(method=lm,se=F) # Most of the points are 0 to 10 but there are a few points outside this range
Scatterplot_PBV
m.PBV <- lm(log_PE~PBV, data_4)
summary(m.PBV) # The slope coefficient is highly significant (t=3.346, p<0.01).
par(mfrow=c(2,2))
plot(m.PBV)
plot(m.PBV, which = 4) ## 89 has a Cook's distance greater than 1
car::avPlots(m.PBV)
m.PBV.without.88 <- lm(log_PE~PBV, data_4[-88,])
summary(m.PBV.without.88)
data_4 <- data_4[-88,]
```

```
Scatterplot_PS <- ggplot(data_4,aes(x=PS, y=log_PE))+
  geom_point()+
  theme_bw()+
  geom_smooth(method=lm,se=F)
Scatterplot_PS
m.PS <- lm(log_PE~PS, data_4)
summary(m.PS) ## The slope coefficient is highly significant (t=5.079, p<0.01).
par(mfrow=c(2,2))
plot(m.PS) ## Residuals vs fitted has a funnel shape
summary(data_4$PS)
# Use a log transform the assumption of constant variance is violated and the values are greater than 0
data_4$log_PS <- log(data_4$PS)
m.log_PS <- lm(log_PE~log_PS, data_4)
summary(m.log_PS) ## The slope coefficient is highly significant (t=5.636, p<0.01).
par(mfrow=c(2,2))
plot(m.log_PS) ## Residuals vs fitted did not have a funnel shape
plot(m.log_PS, which = 4) ## No points with a Cook's distance greater than 1
car::avPlots(m.log_PS) ## 81 and 28 are candidates for influential points

Scatterplot_Beta <- ggplot(data_4,aes(x=Beta, y=log_PE))+
  geom_point()+
  theme_bw()+
  geom_smooth(method=lm,se=F)
Scatterplot_Beta
```

```
m.Beta <- lm(log_PE~Beta, data_4)
summary(m.Beta) # The slope coefficient is highly significant (t=2.957, p<0.01).
par(mfrow=c(2,2))
plot(m.Beta)
plot(m.Beta, which = 4) # No points with a Cook's distance greater than 1
car::avPlots(m.Beta) ## 28 and 81 are candidates for influential points
Scatterplot_Cost_of_Equity <- ggplot(data_4,aes(x=Cost_of_Equity, y=log_PE))+
  geom_point()+
  theme_bw()+
  geom_smooth(method=lm,se=F)

Scatterplot_Cost_of_Equity
m.Cost_of_Equity <- lm(log_PE~Cost_of_Equity, data_4)
summary(m.Cost_of_Equity) # The slope coefficient is highly significant (t=3.514, p<0.01).
plot(m.Cost_of_Equity)
plot(m.Cost_of_Equity, which = 4) # No points with a Cook's distance greater than 1
car::avPlots(m.Cost_of_Equity) ## 28 and 81 are candidates for influential points
Scatterplot_CEO_holding <- ggplot(data_4,aes(x=CEO_holding, y=log_PE))+
  geom_point()+
  theme_bw()+
  geom_smooth(method=lm,se=F)

Scatterplot_CEO_holding
m.CEO_holding <- lm(log_PE~CEO_holding, data_4)
summary(m.CEO_holding) # The slope coefficient is highly significant (t=3.764, p<0.01).
```

```
par(mfrow=c(2,2))

plot(m.CEO_holding)

plot(m.CEO_holding, which = 4) # No points with a Cook's distance greater than 1

car::avPlots(m.CEO_holding)


Scatterplot_Institutional_holding <- ggplot(data_4,aes(x=Institutional_holding, y=log_PE))+

  geom_point()+

  theme_bw()+

  geom_smooth(method=lm,se=F)

Scatterplot_Institutional_holding

m.Institutional_holding <- lm(log_PE~Institutional_holding, data_4)

summary(m.Institutional_holding) # The slope coefficient is highly significant (t=-2.831, p<0.01).

plot(m.Institutional_holding)

plot(m.Institutional_holding, which = 4) # No points with a Cook's distance greater than 1

car::avPlots(m.Institutional_holding) ## 28 and 81 are candidates for influential points


# Create Mlrs to predict log_PE


data_5 <- data_4[,-c(6,3,8)]


## Base plot

log_PE.lm <- lm(log_PE~.-Industry,data=data_5)

summary(log_PE.lm) ## Adjusted R-squared:  0.3313

par(mfrow=c(2,2))

plot(log_PE.lm) ## Seems fine
```

```
plot(log_PE.lm, which=4)

car::avPlots(log_PE.lm)# # 81 seems to be an influential points


## Rerun the regression without 80

log_PE.lm.without.80 <- lm(log_PE~.-Industry,data=data_5[-80,])

summary(log_PE.lm.without.80) ## Adjusted R-squared:  0.3379


AIC(log_PE.lm)

AIC(log_PE.lm.without.80)


## Remove datapoint 80

data_5 <- data_5[-80,]


## Model 1 - Using stepwise regression based on AIC ( Forward Selection )

## Initial model

null<-lm(log_PE~1, data=data_5)

full<-lm(log_PE~.,data=data_5)

log_PE.lm.1<-step(null, scope=list(lower=null, upper=full),direction="forward", trace=0)

summary(log_PE.lm.1)


## Model 2 - Using stepwise regression based on AIC (Backward elimination)

## Initial model

log_PE.lm<-lm(log_PE~.,data = data_5)

log_PE.lm.2 <-step(log_PE.lm)

summary(log_PE.lm.2) ## Step selects a formula-based model by AIC
```

```
## Model 3 - Using Best subsets

## Initial model

log_PE.lm.3<-leaps::regsubsets(log_PE~.-Industry, nvmax=10,really.big = TRUE,data=data_5)

plot(log_PE.lm.3, scale="bic") ## by default uses BIC; may use "adjr2"

## Final model

log_PE.lm.3.Final <-lm(log_PE~ ROE + PBV + Beta + log_Number_of_firms, data = data_5)

summary(log_PE.lm.3.Final)


## Model 3 is the final model


# Check the model assumptions for model 3

## Residual vs fitted plot and Normal Q-Q Plot

par(mfrow=c(1,2))

plot(log_PE.lm.3.Final, which = c(1,2))


#  Outlier analysis

## Cook's Distance

par(mfrow=c(1,1))

plot(log_PE.lm.3.Final, which=4) ## 28 and 83 (28 and 81 after removing 69,88 and 80) are candidates for influential points

## AVP Plot

car::avPlots(log_PE.lm.3.Final)


log_PE.lm.3.Final.without <-lm(log_PE~ ROE + PBV + Beta + log_Number_of_firms, data = data_5[-c(28,81),])
```

```
summary(log_PE.lm.3.Final.without) ## No significant difference. Therefore, stick to log_PE.lm.3.Final


#  Checking for multicollinearity

vif(log_PE.lm.3.Final)

# No multi-collinearity in the model
```
