Stat 3340 Final Project

Zonglin Wu (B00764717) Ziwei Wang (B00776666) Yuchan Zhong (B00791155)

2020-12-07

Abstract

The goal of this project is to find which company is undervalued by using 2013 and 2014 data. By using stepwise feature selection, we choose Depreciation, Net.Income, Retained,Earnings and Estimated.Shares.Outstanding as predictors. We build Market.Capital=6.366Net.Income+0.347Retained.Earnings+13.152Estimated.Shares.OutstandingMarket.Capital As a result, the top 20 undervalued stocks which investors can buy into are found by using 2013 and 2014 data, which include PBCT, EW, IDXX, HBAN, NFLX, SPLS, HPQ, PBI, AIZ, CHD, MAS, DNB, HRL, PDCO, MPC, CNC, WU, XRX, FLIR and ARNC.

Introduction

The goal of this project is to find which company is undervalued by using 2013 and 2014 data. To predict undervalued stocks, we use multiple linear regression model's predicted market capitalization from the model, dividing the difference by shares outstanding , then selecting the top 20 stocks as the undervalued stocks.

```
#packages input
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.3
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.3
```

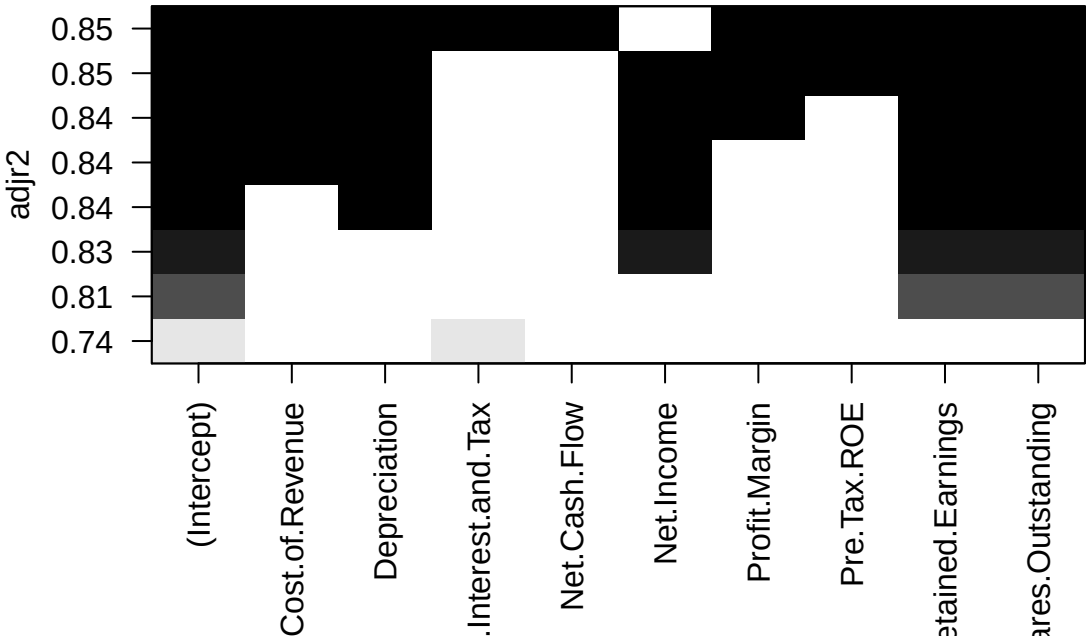
```
#data input

fundamental<-read.csv("D:/WUDOU/course/STAT_3340_Project/fundamentals.csv",header=TRUE)
price_s<-read.csv("D:/WUDOU/course/STAT_3340_Project/prices-split-adjusted.csv",header=TRUE)
security<-read.csv("D:/WUDOU/course/STAT_3340_Project/securities.csv",header=TRUE)
```

Data Pre-processing

```
#Creating yearly stock price mean
price_s$date<-ymd(price_s$date)
price_s$year<-year(price_s$date)
sprice_year<-aggregate(close~symbol+year,data=price_s,FUN=mean) #the mean of annual stock price

#Add Industry from "security" to "fundamental"
names(security)[1]<-"Ticker.Symbol"
information<-security[,c(1,4,5)] #the basic information
data<-merge(fundamental,information,by="Ticker.Symbol",all.x=TRUE) #grouped by Ticker.Symbol

#Date procession
data$Period.Ending<-ymd(data$Period.Ending)
data$year.Ending<-year(data$Period.Ending)

#Add yearly split-adj stock price to "fundamental"
sprice_year$num<-paste(as.character(sprice_year$symbol),sprice_year$year,sep="")
data$num<-paste(as.character(data$Ticker.Symbol),data$year.Ending,sep="")
data<-merge(data,sprice_year[,c(3,4)],by="num",all.x=TRUE) #grouped by the Ticker.Symbol and year.Endin

#Choose data and compute the market capital
data<-subset(data,year.Ending>2012&year.Ending< 2015) #choose 2013 and 2014
data<-subset(data,select=c(Ticker.Symbol,year.Ending,GICS.Sector,Cost.of.Revenue,Depreciation,Earnings.
data$Market.Capital<-data$close*data$Estimated.Shares.Outstanding
data<-subset(data,data$Estimated.Shares.Outstanding>0)

#Missing value processing
data<-na.omit(data)
data$GICS.Sector<-as.factor(as.character(data$GICS.Sector))
t<-table(data$Ticker.Symbol)
data<-subset(data,Ticker.Symbol %in% names(t[t==2]))
data$Ticker.Symbol<-as.factor(as.character(data$Ticker.Symbol))
data<-data[,-c(12,14)]
```

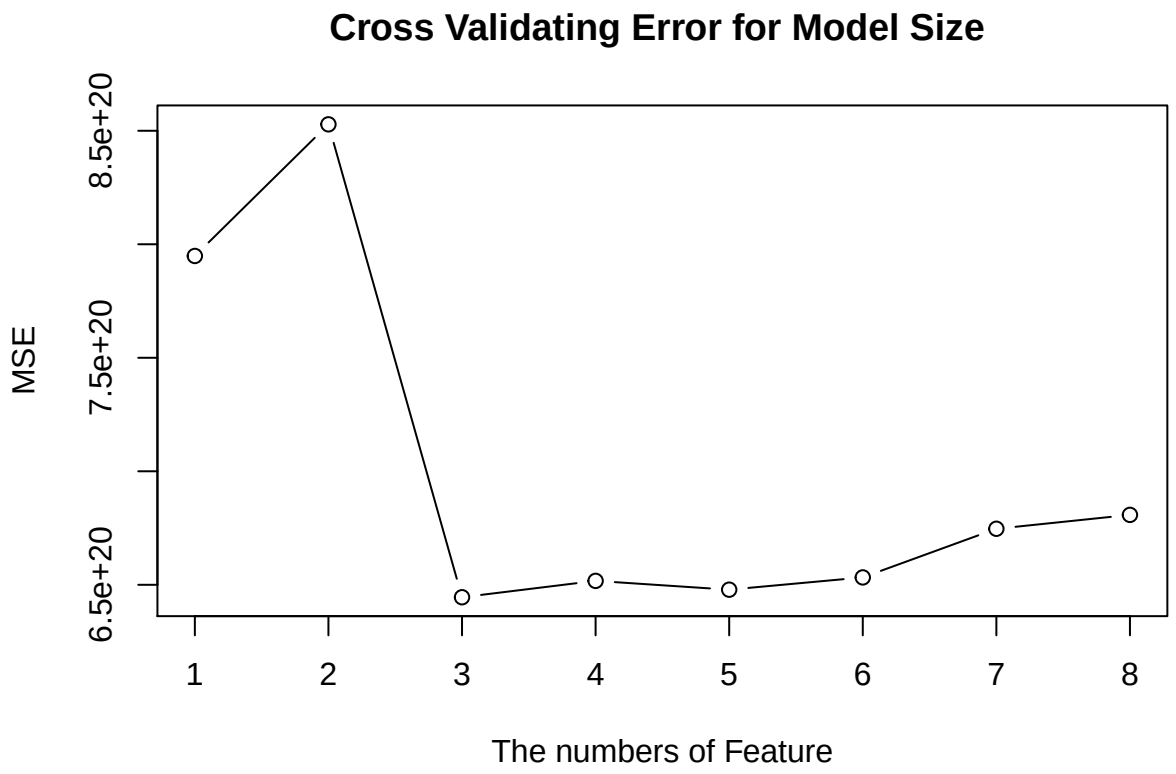
Stepwise Feature Selection

```
#train and test
train<-data[data$year.Ending==2013,-c(1,2,3)]
test<-data[data$year.Ending==2014,-c(1,2,3)]

#Exhaustive feature search
set.seed(1234)
p<-8
best<-regsubsets(Market.Capital~.,data=train,nvmax=p,method="exhaustive")
plot(best,scale="adjr2")
```



This plot shows the model strength according to features. From that, we may choose Depreciation, Net.Income, Retained.Earnings and Estimated.Shares.Outstanding.

```
#Cross-validation
train.matrix<-model.matrix(Market.Capital~.,data=train)
test.matrix<-model.matrix(Market.Capital~.,data=test)
train.error<-c()
MSE<-c()
for(i in 1:p){
  coefficient<-coef(best,id=i)
  pred<-test.matrix[,names(coefficient)]%*%coefficient
  MSE[i]<-mean((test$Market.Capital-pred)^2)
}

plot(MSE, type="b", xlab="The numbers of Feature", ylab="MSE", main="Cross Validating Error for Model S
```




```
which.min(MSE)
```

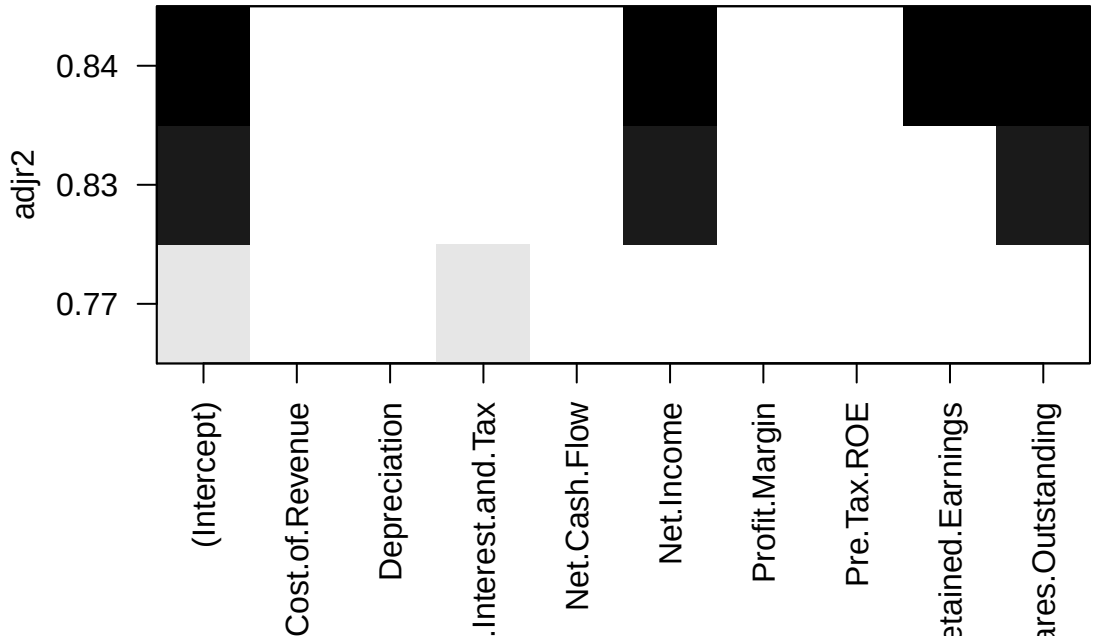
```
## [1] 3
```

```
MSE[which.min(MSE)]
```

```
## [1] 6.445156e+20
```

From this plot, we can choose 3 features to build the model and the smallest MSE equals 6.445555e+20.

```
#For full data
best_all<-regsubsets(Market.Capital~.,data=data[,-c(1,2,3)],nvmax=3)
plot(best_all,scale="adjr2")
```




```
names(coef(best_all,id=3))[-1]
```

```
## [1] "Net.Income"                    "Retained.Earnings"
## [3] "Estimated.Shares.Outstanding"
```

From this plot, we can build the model which has the best variables with 3 predictors by using 2013 and 2014 data.

Multiple regression model

```
best_l<-lm(Market.Capital~Net.Income+Retained.Earnings+Estimated.Shares.Outstanding,data=data)
```

1.Regression Coefficients

```
#Regression coefficient
best_l$coefficient[-1]
```

```
##                   Net.Income             Retained.Earnings
##                     6.367643                      0.346726
## Estimated.Shares.Outstanding
##                    15.152613
```

```
#95% CI for regression coefficient
confint(best_l)[4,]
```

```
##    2.5 %   97.5 %
## 13.47724 16.82798
```

Since $\hat{\beta} = (X'X)^{-1}X'y$, the estimated coefficients of Net.Income, Retained.Earnings and Esti-

mated.Shares.Outstanding are respectively 6.366, 0.347 and 13.152.The 95% CIs for the coefficients of Net.Income,Retained.Earnings and Estimated.Shares.Outstanding are respectively [5.690,7.041], [0.264,0.430] and [13.477,16.827].

The model can be built as

$$Market.Capital = 6.366 Net.Income + 0.347 Retained.Earnings + 13.152 Estimated.Shares.Outstanding$$

2.Hypothesis Testing on the Slope

We formulate this hypothesis test as follows:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0; H_1 : \beta_j \neq 0 \ at \ least \ one \ j$$

Test statistic:

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

```
summary(best_1)
```

```
##
## Call:
## lm(formula = Market.Capital ~ Net.Income + Retained.Earnings +
##     Estimated.Shares.Outstanding, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.303e+11 -6.165e+09 -3.009e+09  2.124e+09  1.809e+11
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.752e+09  8.207e+08   8.227 7.92e-16 ***
## Net.Income                    6.368e+00  3.441e-01  18.504  < 2e-16 ***
## Retained.Earnings             3.467e-01  4.214e-02   8.227 7.93e-16 ***
## Estimated.Shares.Outstanding 1.515e+01  8.535e-01  17.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01e+10 on 786 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8392
## F-statistic:  1374 on 3 and 786 DF,  p-value: < 2.2e-16
```

The t-statistics of regression coefficients are respectively 18.500, 8.236 and 17.755 and the p-values are all smaller than 0.05. Thus, we reject the null hypothesis, which means the three regressor contributes significantly to the model.

3.Multicollinearity

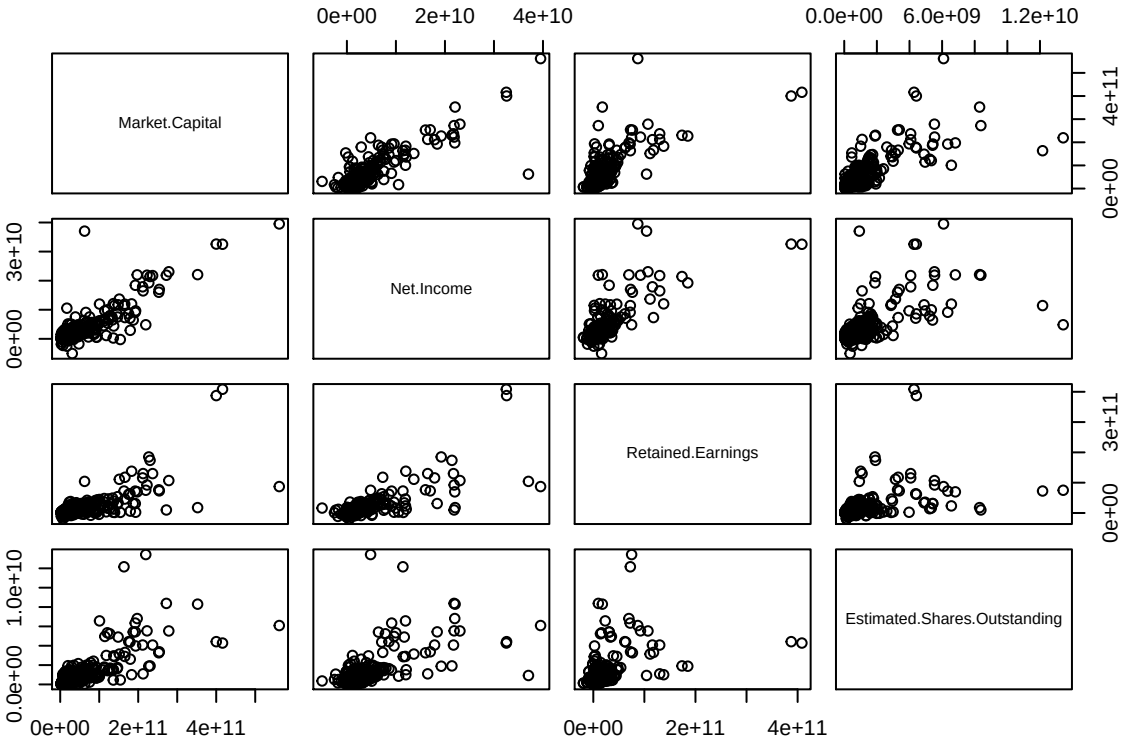
```
print(paste("AIC:",AIC(best_1)))
```

```
## [1] "AIC: 39731.7747312099"
```

```
print(paste("VIF:",vif(best_1)))
```

```
## [1] "VIF: 3.54722431736565" "VIF: 2.52341935619683" "VIF: 1.8536482496265"
```

```
pairs(Market.Capital~Net.Income+Retained.Earnings+Estimated.Shares.Outstanding,data=data)
```



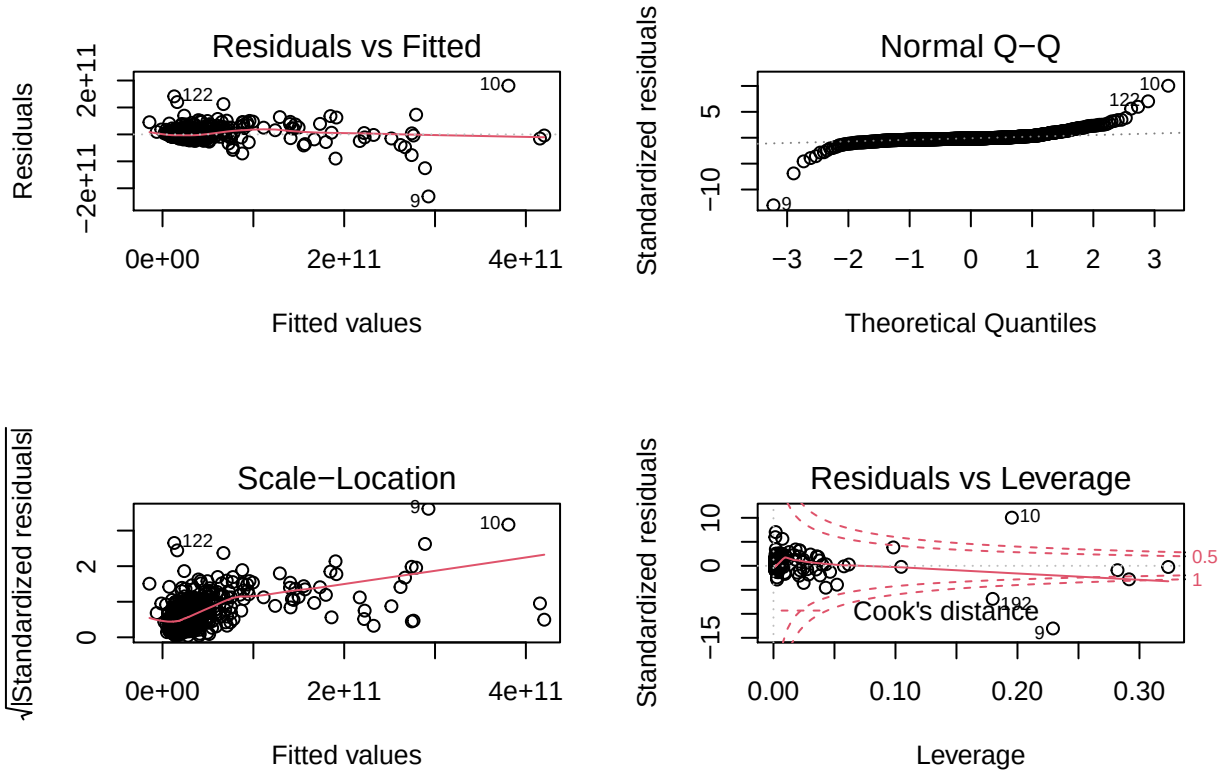
The VIFs mean that Net.Income and Retained.Earnings are nearly linearly dependent on some of the other regressors.

4.Residual analysis

```
#MSE
n<-length(data$Market.Capital)
b<-length(best_l$coefficients)
(MSE<-sum(best_l$residuals*best_l$residuals)/(n-b))
```

```
## [1] 4.039919e+20
```

```
# Model Adequacy Checking
di<-rstandard(best_l) #standardized residuals
ri<-rstudent(best_l) #studentized residuals
par(mfrow=c(2,2))
plot(best_l)
```



From the plot,

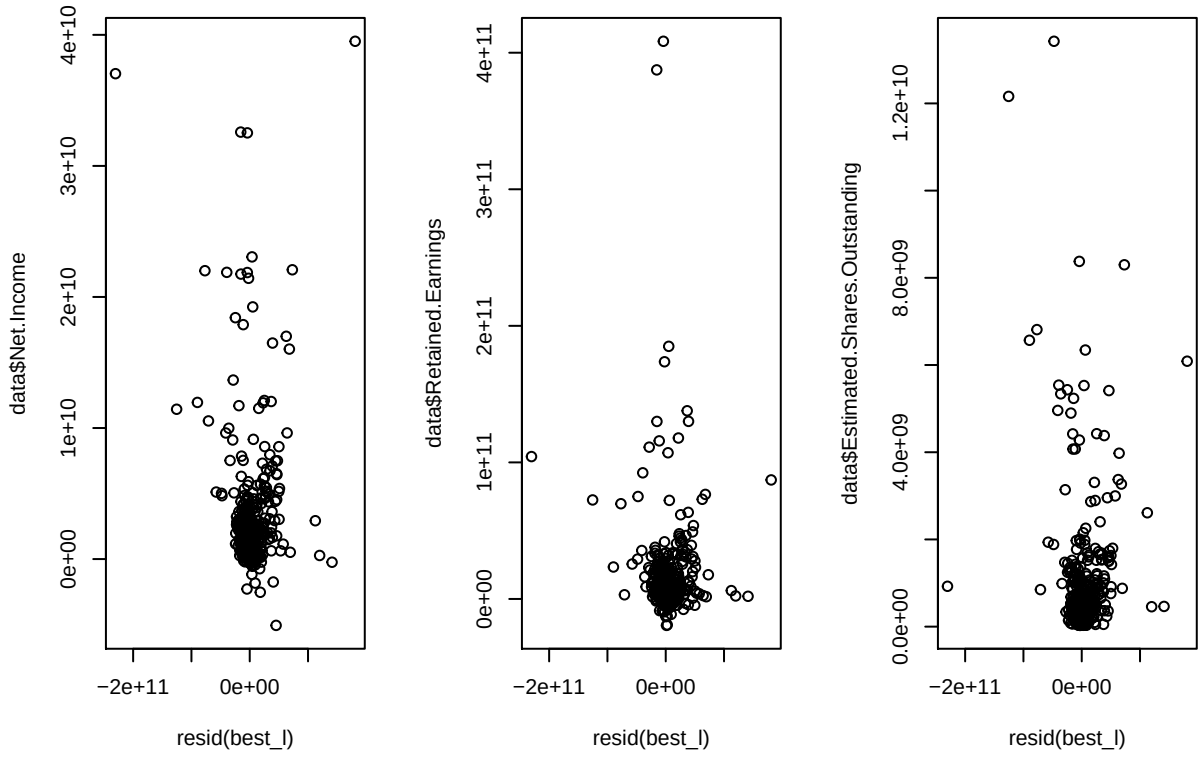
1) Residual vs. Fitted: Line remains fairly straight, which means that the model does not break the assumption of linearity.

2) Normal Q-Q: The runoff tails on either end of the plot suggest that the residuals are not normally distributed. This is another limitation of the model.

3) Scale-Location: There appears to be high residual outliers, breaking the assumption of constant variance. x-axis imbalance.

4) Residuals vs Leverage: It appears that residual 10 has large leverage and effect on the regression line.

```
par(mfrow=c(1,3))
plot(resid(best_1),data$Net.Income)
plot(resid(best_1),data$Retained.Earnings)
plot(resid(best_1),data$Estimated.Shares.Outstanding)
```




```
residtable<-cbind(resid(best_l),rstandard(best_l),rstudent(best_l))
```

Find undervalued companies

```
#Highest Residuals
resi_h<-best_l$residuals[order(best_l$residuals,decreasing = TRUE)[1:20]]
com<-c()
for (i in 1:20){
  com[i]<-as.character(data[as.numeric(attributes(resi_h)$names)[i],1])
}
com
```

```
##  [1] "ABT"  "CAT"  "CAT"  "ROP"  NA     NA     NA     NA     NA     "MCHP"
## [11] "AIV"  "WYN"  NA     NA     "BSX"  "MAT"  NA     "WYN"  "ADM"  NA
```

```
#Undervalued Companies
ud<-predict(best_l)-data$Market.Capital
frame.ud<-data.frame(data$Ticker.Symbol,data$Market.Capital,ud,data$year.Ending)
undervalue<-subset(frame.ud,data.year.Ending==2014)
names(undervalue)<-c("Ticker", "Market.Capital","Total.Undervalue", "Year")

#Top Companies
undervalue$Percent.Undervalue <-undervalue$Total.Undervalue/undervalue$Market.Capital
top20<-undervalue[order(undervalue$Percent.Undervalue, decreasing=TRUE),]
top20$Company<-unlist(lapply(top20[,1],function(x) security$Security[match(x,security$Ticker.Symbol)]))
top20<-subset(top20,Percent.Undervalue<2)
top20$Year<-NULL
```

```
top20<-top20[,c(5,1,2,3,4)]
row.names(top20)<-NULL
kable(top20[1:20,])
```

Company	Ticker	Market.Capital	Total.Undervalue	Percent.Undervalue
People's United Financial	PBCT	4393989016	8787701058	1.999937
Edwards Lifesciences	EW	4883521157	9631476442	1.972240
IDEXX Laboratories	IDXX	3216266747	6034184581	1.876146
Huntington Bancshares	HBAN	8376856391	15163628871	1.810181
Netflix Inc.	NFLX	3454873933	6190536060	1.791827
Staples Inc.	SPLS	8257369259	14759277405	1.787407
HP Inc.	HPQ	28999530390	48341733182	1.666983
Pitney-Bowes	PBI	5182714277	8457631170	1.631892
Assurant Inc	AIZ	4773459149	7738954548	1.621247
Church & Dwight	CHD	4687206020	7587177928	1.618699
Masco Corp.	MAS	7043475226	10802843549	1.533738
Dun & Bradstreet	DNB	4072840610	6088818362	1.494981
Hormel Foods Corp.	HRL	6479129131	9435286448	1.456258
Patterson Companies	PDCO	4191389860	5896421299	1.406794
Marathon Petroleum	MPC	12454318892	17301580487	1.389203
Centene Corporation	CNC	4445344480	6165369731	1.386927
Western Union Co	WU	8926268549	11661881873	1.306468
Xerox Corp.	XRX	15404031746	19823387420	1.286896
FLIR Systems	FLIR	4701527674	6042214223	1.285160
Arconic Inc	ARNC	13824762177	17223273691	1.245828

Result

We can build the linear regression model $Market.Capital = 6.366 Net.Income + 0.347 Retained.Earnings + 13.152 Estimated.Shares.Outstanding$ using 2013 and 2014 year, although there are some limitations in this model like the large MSE. The adjusted R-squared is 0.84 so that the model can fit the data well.

We can make some actionable recommendations to for potential investors. Abovetable shows the top 20 undervalued stocks which investors can buy into, which include PBCT, EW, IDXX, HBAN, NFLX, SPLS, HPQ, PBI, AIZ, CHD, MAS, DNB, HRL, PDCO, MPC, CNC, WU, XRX, FLIR and ARNC.