Final Paper

Due: Friday, April 25th at 8:00AM

A statistical project as to the stock prices of companies within the...

Oil, Gas & Consumable Fuels Industry

Brianna L. Palmisano (X03625986)

St. John's University, The Peter J. Tobin College of Business

BUA 633: Predictive Analytics & Business Forecasting

Professor Manual G. Russon, PhD, CFA

Fall 2024

Table of Contents

I. Introduction

St. John's University is a school located in Queens, New York. This paper uses the stock prices of the companies within the Oil, Gas and Consumable Fuels (OGCF) industry of the Energy sector of the S&P1500 and the indicators within that industry, to examine the linear relationship between these variables. This includes the earnings per share (EPS), book value per share (BVPS), debt to total assets (DTA) and current ratio (CR) for each company within the OGCF industry. Information regarding the stock prices of companies within this industry could be valuable for not only our research within the climate change or sustainability literature, but for investors in energy technology or innovation, or stock traders following the Energy sector of investing as well.

II. Previous Research

There has been previous research on this industry and sector, though none directly in-line with the objective of this paper.

III. Methodology

The input data here has been collected from the Oil, Gas and Consumable Fuels (OGCF) industry, within the Energy sector of the S&P 1500. The S&P 1500 is a stock market index of the unique companies included in the S&P LargeCap 500, S&P MidCap 400, and S&P SmallCap 600 indices ("Cap" refers to market capitalization or equity). The S&P 1500 is inclusive of 11 sectors and 73 industries, with about 1507 publicly traded United States companies.

This is secondary data, as these observations have been adjusted for the integrity of the data. Here we have cross-section data, with a total of 51 observations. The dependent variable here is the price per share (Price), and the independent variables are the EPS, BVPS, DTA, and CR. Graphical techniques including both histograms and scatterplots are used in this analysis. This data has been analyzed using descriptive statistics (for scalable variables), as well as correlation and regression statistical analysis, via both Microsoft Excel and R-Script.

Listed below are the equations defining the functional specification (Eqn. 1), population regression equation (Eqn. 2) and sample regression equation (Eqn. 3).

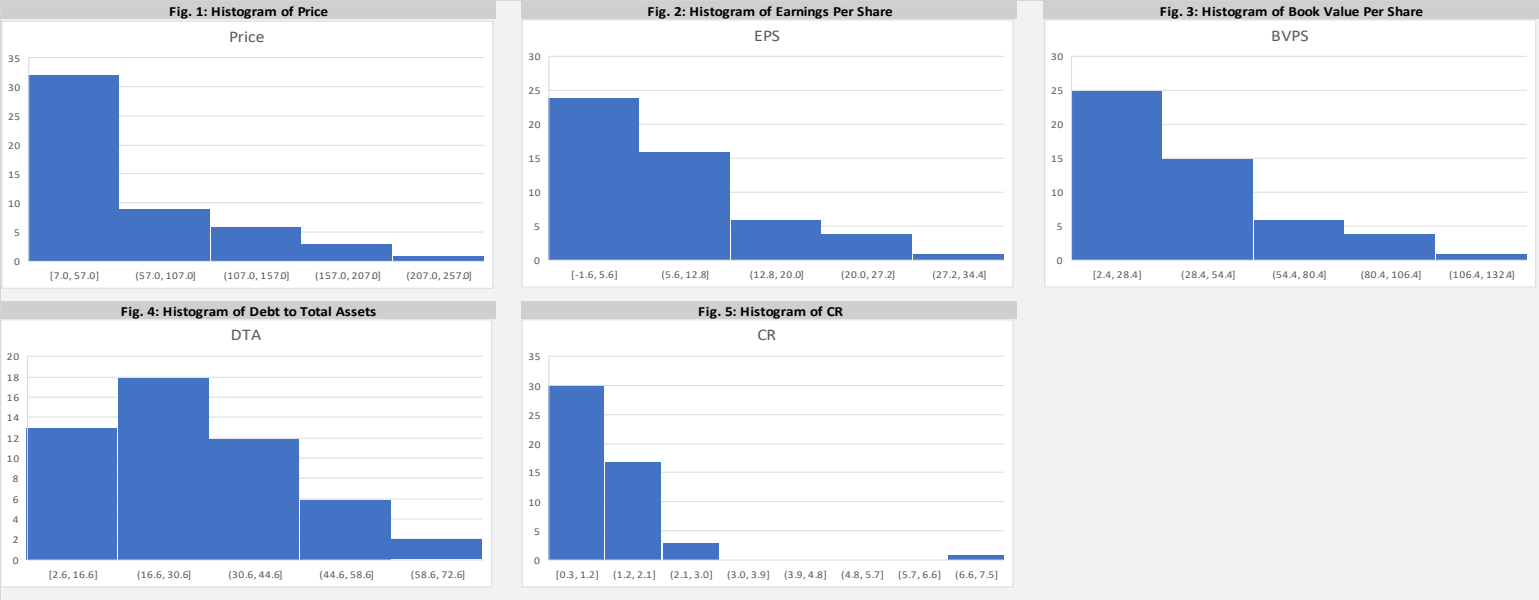
$$\begin{array}{llllll} & & + & + & + & - \\ \text{Eqn. 1} & \text{Price} = f(\text{EPS}, & \text{BVPS}, & \text{DTA}, & \text{CR}) & \\ \text{Eqn. 2} & \text{Price} = \alpha + \beta_{eps}\,\text{EPS} + \beta_{bvps}\,\text{BVPS} + \beta_{dta}\,\text{DTA} + \beta_{cr}\,\text{CR} & & & & \\ \text{Eqn. 3} & \text{Price} = a + b_{eps}\,\text{EPS} + b_{bvps}\,\text{BVPS} + b_{dta}\,\text{DTA} + b_{cr}\,\text{CR} & & & & \end{array}$$

The objective of this model is to evaluate the association or lack thereof that the OGCF industry stock prices have with EPS, BVPS, DTA, and CR.

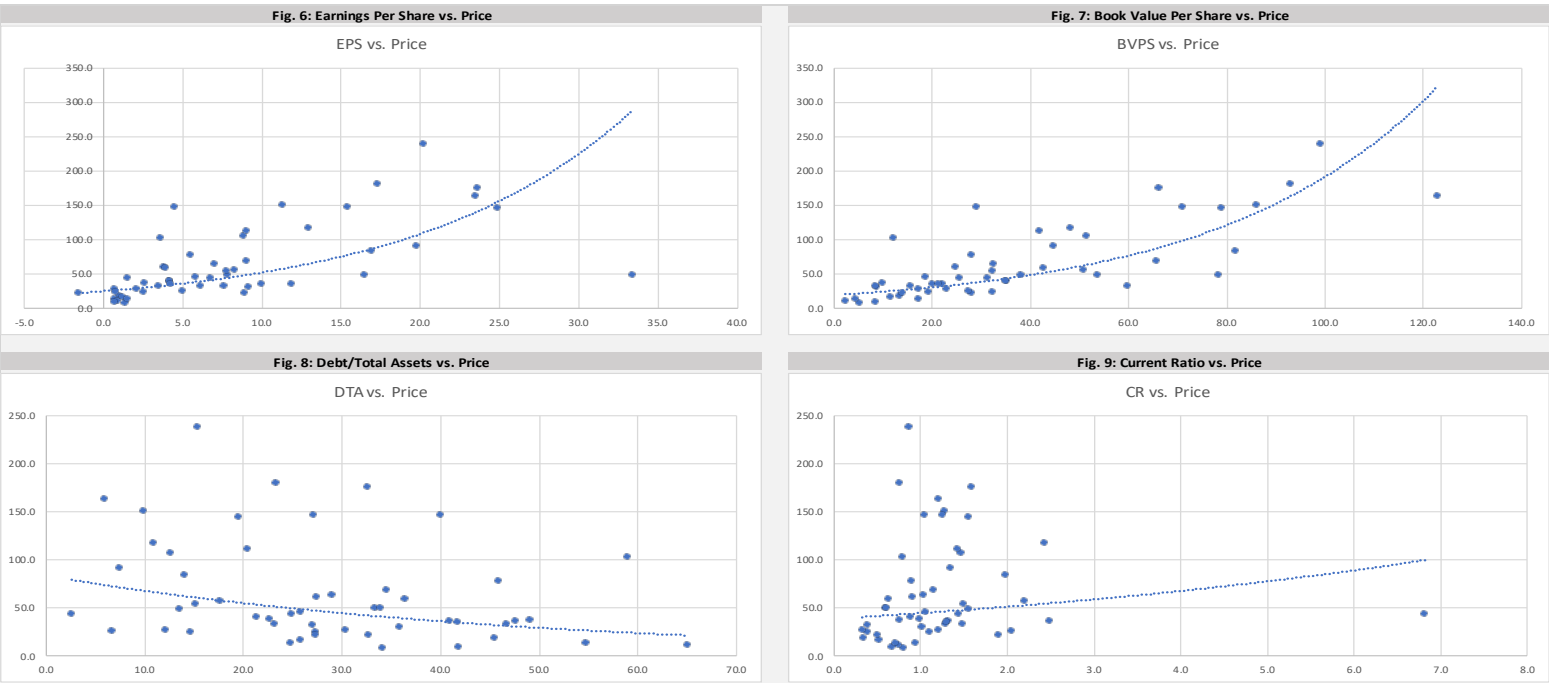
IV. Results

Graphical Analysis

Fig. 1 through *Fig. 5* show the histograms of each variable being analyzed. All histograms are negative and skewed-left. Though there is less of recognizable pattern in *Fig. 4* and *Fig. 5*, the graphs of DTA and C, having more peaks (possibly due to outliers) and irregular symmetry.



The relationships visible via the scatterplots in *Fig. 6* through *Fig. 9* are the independent variables each relative to the dependent variable.



The relationships seen in *Fig. 6* and *Fig. 7* are stronger or more positively correlated, heteroscedastic, linear and with few outliers. Whereas in *Fig. 8,* the pattern is more homoscedastic, linear and has some outliers, deviating from the general trend. Similarly, the data in *Fig. 9* is congested and clustered, though this could be due to the outlier company, REX American Resources Corporation.

Descriptive Statistics

In *Table 1*, are the descriptive statistics for all variables that are non-categorical, here, that is all variables. The standard deviation for each variable is relatively close to the mean, indicating that they all have moderate variability, as well as that they are all moderately volatile. The data is dispersed relatively close to the mean. The standard error around Price (at 7.49, with a mean of 64.12) can be considered moderate or higher, but the standard error around the independent variables is lower and reasonable, in this context.

Table 1: Descriptive Statistics

	price	*eps*	*bvps*	*dta*	*cr*
Mean	64.12	8.24	37.57	28.25	1.25
Standard Error	7.49	1.06	3.90	2.01	0.13
Median	43.62	6.19	29.26	27.17	1.07
Standard Deviation	53.47	7.59	27.88	14.35	0.95
Sample Variance	2858.77	57.55	777.17	205.84	0.90
Kurtosis	1.25	1.53	0.72	(0.19)	23.76
Skewness	1.36	1.33	1.11	0.43	4.21
Range	230.25	35.03	120.66	62.57	6.50
Minimum	7.00	(1.59)	2.41	2.59	0.33
Maximum	237.25	33.44	123.07	65.16	6.83
Sum	3270.13	420.22	1915.92	1440.75	63.89
Count	51	51	51	51	51

The measure of kurtosis is positive and slightly higher or more peaked than normal distribution in both Price and EPS. The measured kurtosis is negative around DTA, indicating that this data is more dispersed than the other variables. The higher kurtosis seen in the CR indicates that the data is also more dispersed but has a sharper peak and higher chance of outliers (because of its extreme value). The measure of skewness is asymmetric, right-skewed, and right-tailed across all variables in this model.

Correlation Matrix

Via the correlation matrix in *Table 2*, it can be observed that the Price is most correlated with the EPS (at 0.651) and BVPS (at 0.778), though BVPS and EPS were more correlated with one another (at 0.794). It should also be noted that the BVPS is correlated with DTA (at -0.518), indicating that although they are moving in opposing directions, there is still a relationship between these variables greater than others in this model.

Table 2: Correlation Matrix					
	price	***eps***	***bvps***	***dta***	***cr***
price	1				
eps	0.651261796	1			
bvps	0.778616365	0.794827434	1		
dta	-0.302772686	-0.317666527	-0.518158528	1	
cr	0.061486663	0.014943043	0.045591318	-0.397660152	1

The matrix gives a correlation that agrees with our original hypothesis, while the Price is more associated with BVPS, or has a higher correlation between this independent variable than the other variables. Due to high correlation coefficients, we can assume that there are signs of multicollinearity in this model.

Regression Results

Equation for Sample Regression Line.

			+	+	+	-	
Eqn. 4		Price =	f(EPS,	BVPS,	DTA,	CR)	
	t-stat	(-1.08)**	(0.31)***	(4.90)*	(1.51)	(0.94)	
	p-value	(0.29)	(0.75)	(0.00)	(0.14)	(0.35)	
	r (corr)		(0.33)	(1.59)	(0.67)	(5.31)	
	n = 51	r-sq. = 0.628	F = 19.442***		F-Prob = 0.000		SE = 33.984

Confidence Intervals.

*	Significant at the	0%	level of significance (90% Sure, or "are below")
**	Significant at the	5%	level of significance (95% Sure, or "are below")
***	Significant at the	1%	level of significance (99% Sure, or "are below")

Results of an F-test for the entire model.

	Ho: beps = bbvps = 0	(Null Hypothesis)
* 5%	Ha: at least 1 b_i not equal to 0 (19.442 > 4.99)	(Alternate Hypothesis)

To further interpret the regression statistics in *Table 3*, the above F-test includes the null (Ho) and alternative hypothesis (Ha) for the entire model. The results for the F-Statistic (19.442) indicates that the regression model is statistically significant, as it is above the cutoff (>4.99), signifying that the null hypothesis should be rejected. Rather, the higher value insists that the model explains a significant amount of the variation in the dependent variable with the predictor variables, than without.

Table 3: Regression Statistics

Regression Statistics	
Multiple R	0.793
R Square	0.628
Adjusted R Square	0.596
Standard Error	33.984
Observations	51

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	89,813.556	22,453.389	19.442	0.000
Residual	46	53,124.850	1,154.888		
Total	50	142,938.406			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	(24.01)	22.30	(1.08)	0.29	(68.89)	20.87	(68.89)	20.87
eps	0.33	1.06	0.31	0.75	(1.81)	2.47	(1.81)	2.47
bvps	1.59	0.32	4.90	0.00	0.94	2.25	0.94	2.25
dta	0.67	0.44	1.51	0.14	(0.22)	1.56	(0.22)	1.56
cr	5.31	5.64	0.94	0.35	(6.03)	16.66	(6.03)	16.66

Items of Import

1. Equation	*Intercept*	*EPS*	*BVPS*	*DTA*	*CR*
	(24.01)	0.33	1.59	0.67	5.31
2. F-Statistic	19.442				
3. R-Squard	0.628				
4. T-Statistics	(1.08)	0.31	4.90	1.51	0.94
5. Standard Error	33.984				

The results for our p-value show that the BVPS variable (at 0.00) is significant at the 5% level, having the most effect on the dependent variable, Price. This inclines us to reject the null hypothesis at the 95% confidence interval. Similarly, the p-value result for DTA (at 1.51) inclines us to reject the null hypothesis at the 90% confidence interval, indicating that Price is more effected by BVPS or DTA than without. All other independent variables (EPA and CR) were not statistically significant (ailing to reject the null hypothesis).

When looking at the results for the coefficient of determination, it can be concluded that that statistical model predicts an association between the dependent and independent variables that is strong. There is 62.8% of variation in the dependent variable can be explained by variations in the independent variables.

The standard error is considerably high for the variables at hand, indicating that there is a lot of room for uncertainty in this analysis. The high standard error value around the independent variable CR (5.64) tells us that the CR falls much further from the regression line than the other variables, especially EPE (at 1.06)). The greater Standard Error value seen in the intercept determines that the dependent variable is being affected by other independent variables, ones not included in this model.

The results for the t-Statistic indicate that the variables Price (intercept), EPS, DTA and CR are not significant, or have little effect on the dependent variable. Although, the high t-Statistic value

seen in BVPS (at 4.90) expresses that is variable is highly significant and impactful on the dependent variable, Price.

To discuss the statistically significant regression coefficients, the dependent variable is most impacted by the CR (at 5.31) and the BVPS (at 1.59). That is to say that for every unit increase in the independent variable CR, there is a $5.31 increase in the dependent variable (Price), and that there is a $1.59 increase in Price for every unit increase in BVPS, holding all other variables constant.

The fitted residuals are graphed in *Fig. 10* and *Fig. 11.* The residuals are not normally distributed, asymmetric, and skewed left. The scatterplot of the residuals shows that the data is weak, not heteroscedastic, nonlinear and may have some outliers. These results insist that the model is nonlinear and homoscedastic. This model should be re-run using nonlinear regression to become to improve its reliability (or become normally distributed), even though the homoscedasticity of the data plotted in *Fig. 11* insists that the model is reliable.

V. Conclusions

For this model, the null hypothesis is rejected, and the alternate hypothesis is accepted. The research presented here can be considered successful, as 62.8% of variation in the dependent variable can be explained by the independent variables. It can be determined true that the dependent variable is better supported by the independent variables, than if it were not. This model was not entirely predictive, as the variables here do not influence one-another enough but was somewhat predictive when looking at EPS and BVPS, objectively.

This model can be improved on by increasing the number of observations, as there were just enough observations to create a valuable model. More observations would increase the validity of any model, but it can be assumed that this would not change our overall results in this analysis. Should there be a need to break down the moments of closer associations, breaking this model up by time periods may add to the findings that this model is capable of.

Public policy does not entirely pertain to driving a relationship between these variables but could be influential in the context of regulating the terms of trade, changes in regulations that compromise operations within the Energy industry, as well as climate change efforts (including “green” incentives relative to carbon emissions). It could be beneficial to investors to have more insight on the variables that influence the stock price of the company at hand. This model or an expanded version of this model could also be beneficial to parties when looking at the impact of pertaining policies and the risks of those policies, as well as looking at the impact of the Energy industry and its environmental implications.

VI. Bibliography

Data Sources

"S&P Composite 1500®." S&P Dow Jones Indices, 2024. https://www.spglobal.com/spdji/en/indices/equity/sp-composite-1500/#data.

VII. Appendix I: Input Data

Dataset used in this model.

Input Data: Oil, Gas & Consumable Fuels Industry of the Energy Sector (S&P1500)

tkr	name	price	eps	bvps	dta	cr
XOM	Exxon Mobil Corporation	105.6	8.9	51.6	12.7	1.5
CVX	Chevron Corporation	149.6	11.4	86.3	10.0	1.3
COP	ConocoPhillips	111.2	9.1	41.8	20.5	1.4
EOG	EOG Resources, Inc.	116.2	13.0	48.4	10.9	2.4
MPC	Marathon Petroleum Corporation	175.2	23.6	66.3	32.6	1.6
PSX	Phillips 66	146.1	15.5	71.0	27.2	1.3
PXD	Pioneer Natural Resources Company	237.3	20.2	99.2	15.5	0.9
OXY	Occidental Petroleum Corporation	60.6	3.9	25.0	27.5	0.9
VLO	Valero Energy Corporation	144.3	25.0	79.0	19.6	1.6
HES	Hess Corporation	146.4	4.5	29.3	40.0	1.0
OKE	ONEOK, Inc.	77.5	5.5	28.3	45.9	0.9
WMB	Williams Companies, Inc.	36.7	2.6	10.2	49.1	0.8
KMI	Kinder Morgan Inc Class P	17.7	1.1	13.7	45.6	0.4
FANG	Diamondback Energy, Inc.	180.1	17.3	93.0	23.5	0.8
DVN	Devon Energy Corporation	45.1	5.8	19.0	25.9	1.1
TRGP	Targa Resources Corp.	102.5	3.7	12.3	59.1	0.8
CTRA	Coterra Energy Inc.	26.7	2.1	17.4	12.2	1.2
EQT	EQT Corporation	38.0	4.2	35.2	22.7	1.0
MRO	Marathon Oil Corporation	24.2	2.6	19.4	27.3	0.4
OVV	Ovintiv Inc	48.8	7.9	38.2	33.4	0.6
PR	Permian Resources Corporation Class A	15.6	1.2	11.7	25.9	0.5
DINO	HF Sinclair Corporation	55.8	8.3	50.8	17.7	2.2
CHK	Chesapeake Energy Corporation	82.9	16.9	82.0	14.2	2.0
APA	APA Corporation	30.5	9.2	8.7	35.9	1.0
AR	Antero Resources Corporation	27.0	0.8	23.0	30.4	0.3
RRC	Range Resources Corporation	32.8	3.6	15.6	23.3	1.5
SWN	Southwestern Energy Company	7.0	1.4	5.3	34.2	0.8
MTDR	Matador Resources Company	62.9	7.1	32.7	29.0	1.0
CIVI	Civitas Resources, Inc.	68.1	9.0	65.9	34.6	1.2
CHRD	Chord Energy Corporation	163.2	23.5	123.1	5.9	1.2
AM	Antero Midstream Corp.	13.6	0.8	4.5	54.9	0.9
MUR	Murphy Oil Corporation	39.2	4.2	35.1	21.4	0.9
PBF	PBF Energy, Inc. Class A	47.9	16.5	53.9	13.6	1.6
DTM	DT Midstream, Inc.	58.2	3.9	42.7	36.4	0.6
SM	SM Energy Company	43.6	6.9	31.2	24.9	1.4
ETRN	Equitrans Midstream Corporation	10.9	0.9	2.4	65.2	0.8
CRC	California Resources Corp	53.3	7.8	32.3	15.3	1.5
NOG	Northern Oil and Gas, Inc.	35.6	10.0	20.3	40.9	1.3
CNX	CNX Resources Corporation	21.5	9.0	28.2	27.4	0.5
BTU	Peabody Energy Corporation	25.6	5.0	27.6	6.7	2.1
CVI	CVR Energy, Inc.	32.2	7.7	8.4	46.8	1.3
CEIX	CONSOL Energy Inc	90.5	19.8	44.9	7.5	1.4
CRK	Comstock Resources, Inc.	8.7	0.8	8.5	41.9	0.7
PARR	Par Pacific Holdings Inc	35.0	11.9	22.3	41.8	1.3
TALO	Talos Energy, Inc.	13.0	1.6	17.4	24.8	0.7
CPE	Callon Petroleum Company	31.7	6.2	60.0	27.1	0.4
VTLE	Vital Energy, Inc.	49.2	33.4	78.6	34.0	0.6
WKC	World Kinect Corporation	24.3	0.9	32.5	14.6	1.1
LPG	Dorian LPG Ltd.	36.1	4.3	21.6	47.7	2.5
GPRE	Green Plains Inc.	21.5	-1.6	14.2	32.8	1.9
REX	REX American Resources Corporation	43.1	1.6	25.8	2.6	6.8

VIII. Appendix II: R-Script

R-Script:

```
# PREDICTIVE ANALYTICS - Project 1, Regression Analysis
# SECTOR: Energy
# INDUSTRY: Oil Gas & Consumable Fuels Industry of the Energy
library(YRmisc)
library(readxl)
sp1500 <- read_excel("Downloads/sp1500.xlsx", sheet = "Tab1")
View(sp1500)
dim(sp1500)
names(sp1500)
unique(sp1500$sector)
data.frame(unique(sp1500$sector))
unique(sp1500$industry)
data.frame(unique(sp1500$industry))
idf<-sp1500[sp1500$industry=="Oil Gas & Consumable Fuels",]
dim(idf)
# FIGURES 1-5: Histograms
par(mfrow=c(3,3))
hist(idf$price, col="blue", xlab="Price", ylab="Frequency", main="Fig. 1 Hist of Price")
hist(idf$eps, col="blue", xlab="EPS", ylab="Frequency", main="Fig. 2 Hist of EPS")
hist(idf$bvps, col="blue", xlab="BVPS", ylab="Frequency", main="Fig. 3 Hist of BVPS")
hist(idf$dta, col="blue", xlab="Total Assets", ylab="Frequency", main="Fig. 4 Hist of Debt/TotAssets")
hist(idf$cr, col="blue", xlab="Current Ratio", ylab="Frequency", main="Fig. 5 Hist of Current Ratio")
# FIGURES 6-9: SCATTERPLOTS
par(mfrow=c(2,2))
plot(idf$eps, idf$price, xlab="EPS", ylab="Price", main="Fig. 6 EPS vs. Price", type="n")
text(idf$eps, idf$price, as.character(idf$tkr), cex=0.5)
plot(idf$bvps, idf$price, xlab="BVPS", ylab="Price", main="Fig. 7 BVPS vs. Price", type="n")
text(idf$bvps, idf$price, as.character(idf$tkr), cex=0.5)
plot(idf$dta, idf$price, xlab="Total Assets", ylab="Price", main="Fig. 8 Total Assets vs. Price", type="n")
text(idf$dta, idf$price, as.character(idf$tkr), cex=0.5)
plot(idf$cr, idf$price, xlab="Current Ratio", ylab="Price", main="Fig. 9 Current Ratio vs. Price", type="n")
text(idf$cr, idf$price, as.character(idf$tkr), cex=0.5)
# DESCRIPTIVE STATISTICS
ds.summ(idf[,c("price","eps","bvps","dta","cr")],3)[,-c(7,8)]
# CORRELATION MATRIX
round(cor(idf[,c("price","eps","bvps","dta","cr")]),3)
# REGRESSION ANALYSIS
idf1<-na.omit(idf)
fit<-lm(price~eps+bvps+dta+cr,na.action=na.omit,data=idf1)
summary(fit)
# FIGURES 10 & 11: FITTED RESIDUALS
idf$price
fit$fitted.values
fit$residuals
par(mfrow=c(2,2))
hist(fit$residuals,col="blue",xlab="Fitted Residuals",ylab="Frequency",main="Fig. 10 Histogram of Fitted
Residuals",type="n")
plot(fit$residuals,xlab="Index",ylab="Fitted Residuals",main="Fig. 11 Scatterplot of Fitted
Residuals",type="n")
text(fit$residuals,as.character(idf$tkr),cex=.5)
```

Graphical Outputs:

Fig. 1 Hist of Price

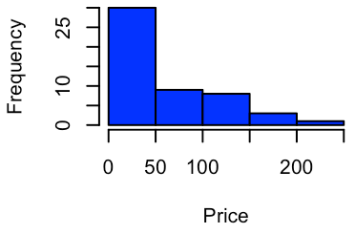


Fig. 2 Hist of EPS

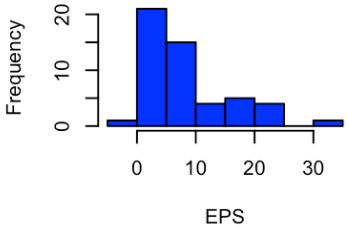


Fig. 3 Hist of BVPS

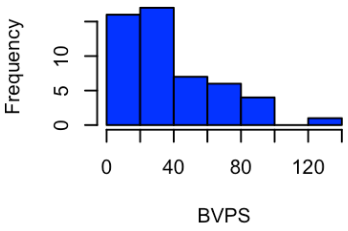


Fig. 4 Hist of Debt/TotAssets

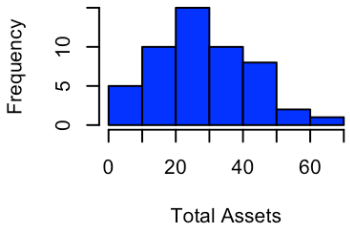


Fig. 5 Hist of Current Ratio

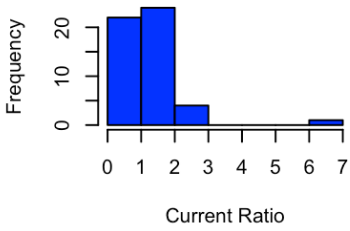


Fig. 6 EPS vs. Price

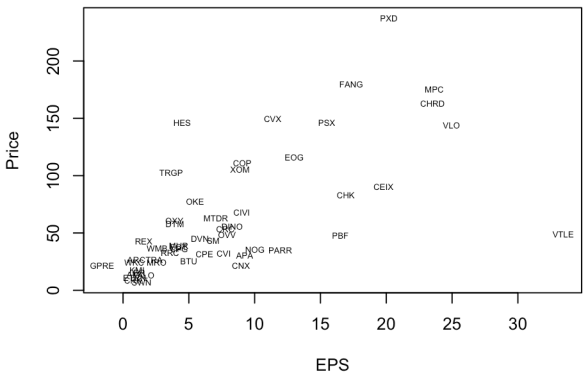


Fig. 7 BVPS vs. Price

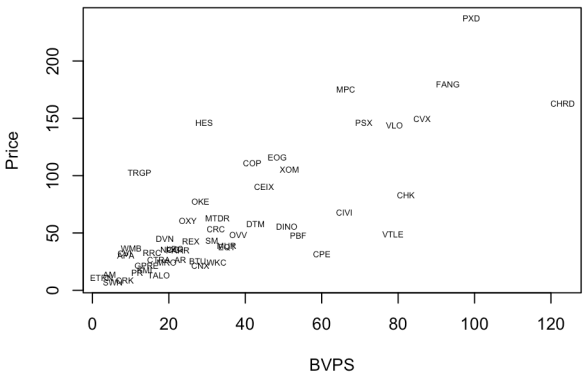


Fig. 8 Total Assets vs. Price

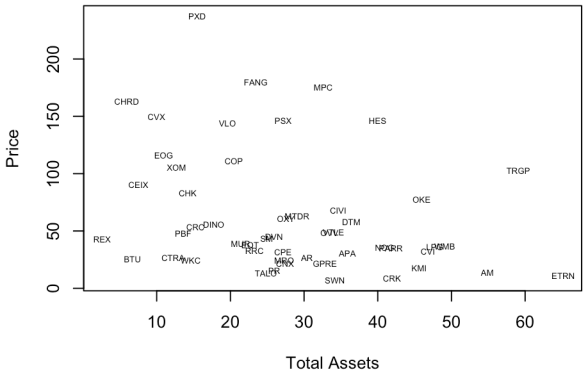


Fig. 9 Current Ratio vs. Price

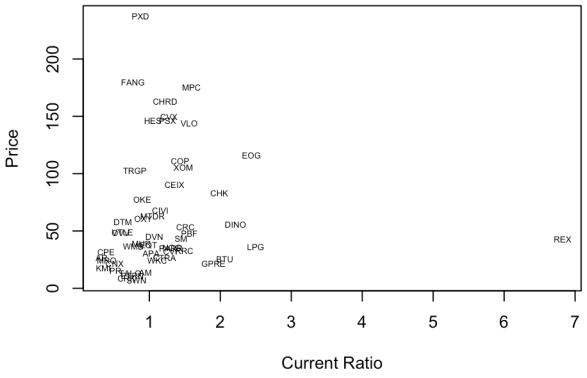


Fig. 10 Histogram of Fitted Residuals

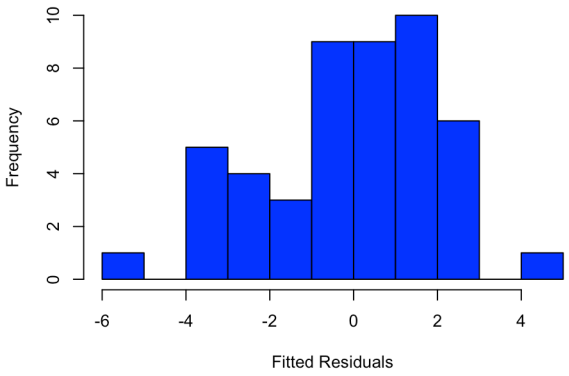


Fig. 11 Scatterplot of Fitted Residuals