Southern Connecticut State University
MAT 326, Spring 2021

Firm Profits

James Weaver and Amy Rodriguez

May 13^{th}, 2021

Instructor .... Andrew Bartlett

Contents

1 Introduction

1.1 Introduction

A firm's profit is important to employees, customers and to its respective shareholders. When a firm decreases it's profits over time, or fails to increase its profits, it can negatively impact the same group of people. The negative impact can result in budget cuts, decreasing the quantity of products/selling existing products at a discounted rate, and having shareholders sell their shares and bringing the average cost per share down. On the other hand, an increase in firm's profits is beneficial to its employees, customers and prospective shareholders, as it results in raises for employees, customer's ability to consume products, and demonstrates firm's growth to shareholders. For these reasons a firm will be most interested in finding which factors/variables are related to its profits. Return on total assets ratio, also known as the average profit per total assets(PT), is considered to be an indicator of how effectively a company is at using its assets to generate earnings. Typically, interest expense and taxes are added back to income to calculate the profitability ratio without the effects of interest and taxes. PT is focused on operating earnings without the influence of tax or financing differences when compared to similar companies. The higher the ratio, the more effective the company is at using their assets to generate income. Hence, this research seeks to find a 'best' multiple linear regression model to predict PT based on a set of economic factors.

2 Initial Observations

The data used in this analysis are profit rates and market structure of 57 advertising intensive firms from 2013 to 2018. We are interested in finding what factors from the data are contributing to predict PT. At first glance, one can assume the average profit per shareholder would be a good predictor of PT because when its average increases, it demonstrates how effectively a firm is using their net assets. Since our objective is to find the best model to predict PT, we will perform a multiple linear regression analysis. The formula for the multiple linear regression line is as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} + \hat{\epsilon}_i$$

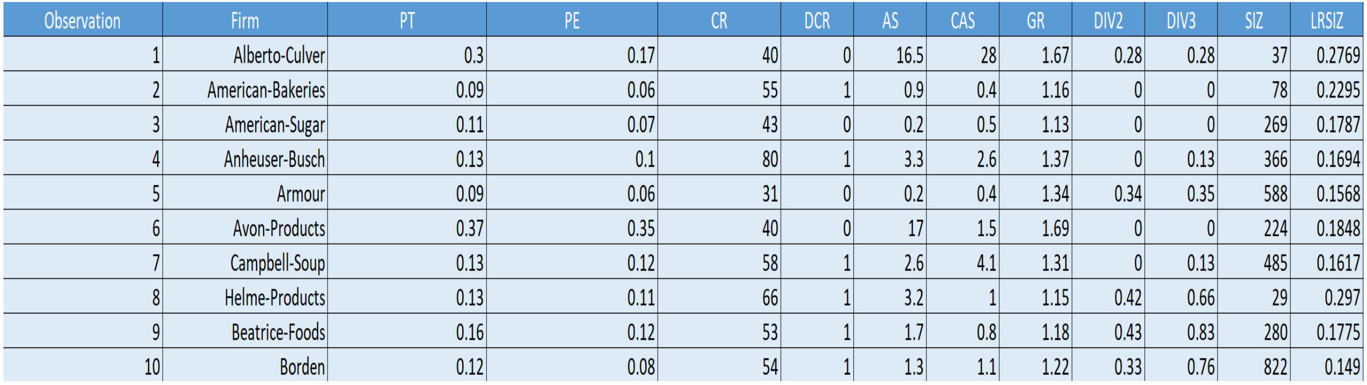
where

$$\begin{aligned}
\hat{y}_i &= \textit{The response(dependent) variable we want to predict.} \\
x_{ik} &= \textit{The explanatory(independent) variables to make the predictions.} \\
\hat{\beta}_0 &= \textit{The prediction(y-intercept) when } x_{ik} = 0. \\
\hat{\beta}_k &= \textit{The slope coefficients for each independent variable.} \\
\hat{\epsilon}_i &= \textit{The error term.}
\end{aligned}$$

$$\textit{where for } i = n \textit{ observations.}$$

3 Exploratory Data Analysis

Figure 1: *First 10 Observations*



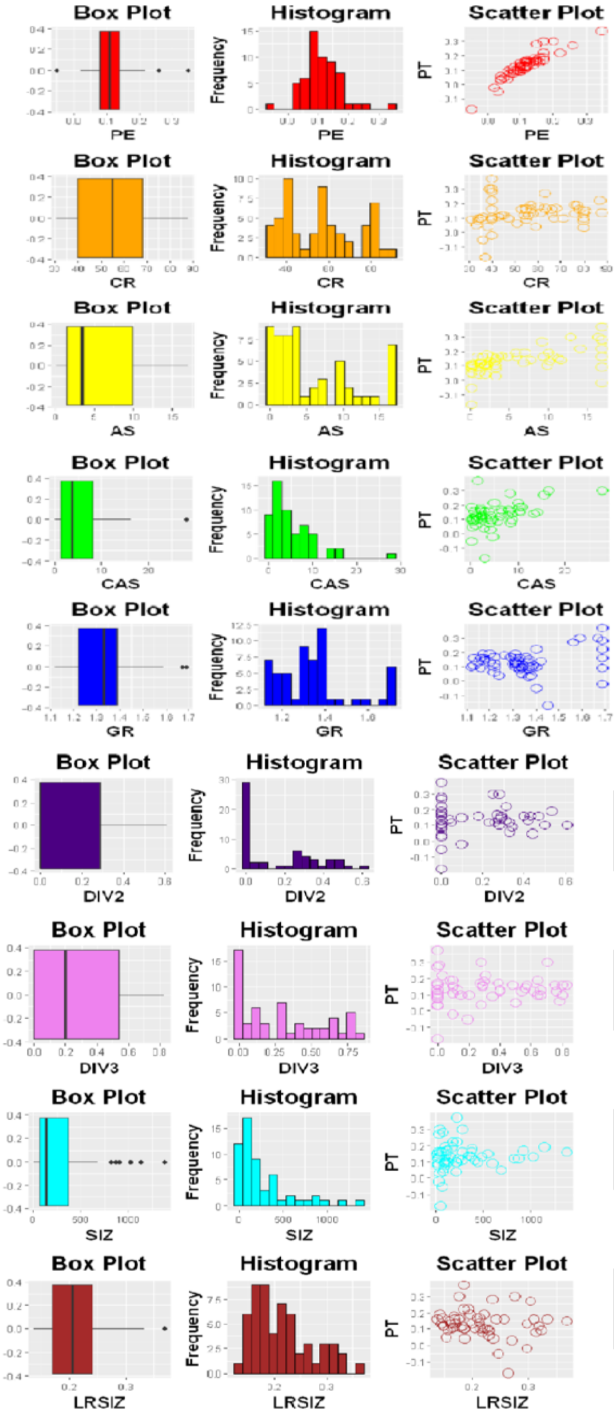
Observation	Firm	PT	PE	CR	DCR	AS	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
1	Alberto-Culver	0.3	0.17	40	0	16.5	28	1.67	0.28	0.28	37	0.2769
2	American-Bakeries	0.09	0.06	55	1	0.9	0.4	1.16	0	0	78	0.2295
3	American-Sugar	0.11	0.07	43	0	0.2	0.5	1.13	0	0	269	0.1787
4	Anheuser-Busch	0.13	0.1	80	1	3.3	2.6	1.37	0	0.13	366	0.1694
5	Armour	0.09	0.06	31	0	0.2	0.4	1.34	0.34	0.35	588	0.1568
6	Avon-Products	0.37	0.35	40	0	17	1.5	1.69	0	0	224	0.1848
7	Campbell-Soup	0.13	0.12	58	1	2.6	4.1	1.31	0	0.13	485	0.1617
8	Helme-Products	0.13	0.11	66	1	3.2	1	1.15	0.42	0.66	29	0.297
9	Beatrice-Foods	0.16	0.12	53	1	1.7	0.8	1.18	0.43	0.83	280	0.1775
10	Borden	0.12	0.08	54	1	1.3	1.1	1.22	0.33	0.76	822	0.149

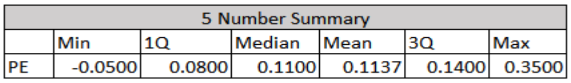
Our data contains 57 observations; however, only the first 10 are shown in figure 1. Since we are interested in predicting PT, this will be our dependent variable for our model. Below are the descriptions of the potentially independent variables we will use to predict PT:

- PT $\hat{y}$ The average profit per total assets.(quantitative & dependent variable),
- Firm x_1 Name of firm(categorical),
- PE x_2 Average profit per shareholder equity(quantitative),
- CR x_3 Weighted concentration ratio of firm's product markets(quantitative),
- DRC x_4 Weighted concentration ratio of firm's product markets greater than 50(categorical),
- AS x_5 Weighted average Industry advertising-to-sales ratio of firm's product markets(quantitative),
- CAS x_6 Overall advertising-to-sales ratio of the firm(quantitative),
- GR x_7 Weighted average percent changes in industry sales in the firm's market(quantitative),
- DIV2 x_8 Firm's diversification in the more broad industries(quantitative),
- DIV3 x_9 Firm's diversification in the less broad industries(quantitative),
- SIZ x_{10} Firms 2018 total assets (millions)(quantitative),
- LRSIZ x_{11} Inversely proportional to market shares(quantitative).

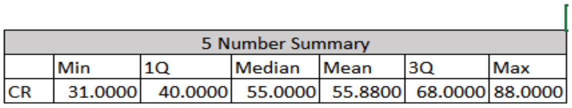
The data illustrates that there are 11 independent variables to choose from to make the 'best' multiple linear regression model.This implies that there are $2^{11} = 2048$ possible models as the null model is used when all other models are not significant.However, it can also be seen that *Firm* and *DCR* are labeled as categorical variables. Since we do not have the knowledge on how to build models with both categorical and quantitative variables,we will exclude x_1 and x_4 for the analysis. This leaves us with $2^9 = 512$ possible models.

Figure 2: *Graphical Representation Of The Data*

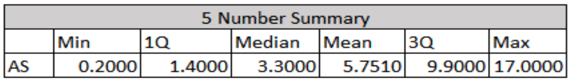




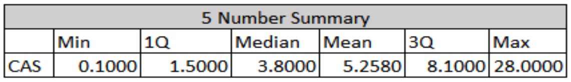
	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
PE	-0.0500	0.0800	0.1100	0.1137	0.1400	0.3500



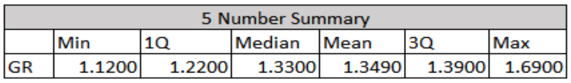
	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
CR	31.0000	40.0000	55.0000	55.8800	68.0000	88.0000



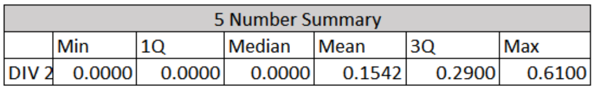
	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
AS	0.2000	1.4000	3.3000	5.7510	9.9000	17.0000



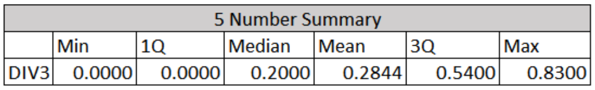
	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
CAS	0.1000	1.5000	3.8000	5.2580	8.1000	28.0000



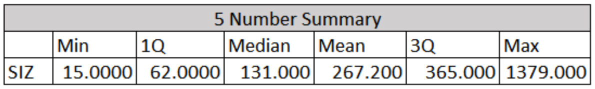
	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
GR	1.1200	1.2200	1.3300	1.3490	1.3900	1.6900



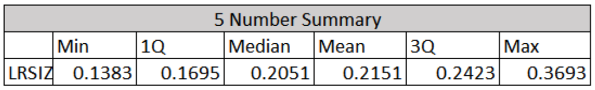
	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
DIV 2	0.0000	0.0000	0.0000	0.1542	0.2900	0.6100



	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
DIV3	0.0000	0.0000	0.2000	0.2844	0.5400	0.8300



	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
SIZ	15.0000	62.0000	131.000	267.200	365.000	1379.000



	5 Number Summary					
	Min	1Q	Median	Mean	3Q	Max
LRSIZ	0.1383	0.1695	0.2051	0.2151	0.2423	0.3693

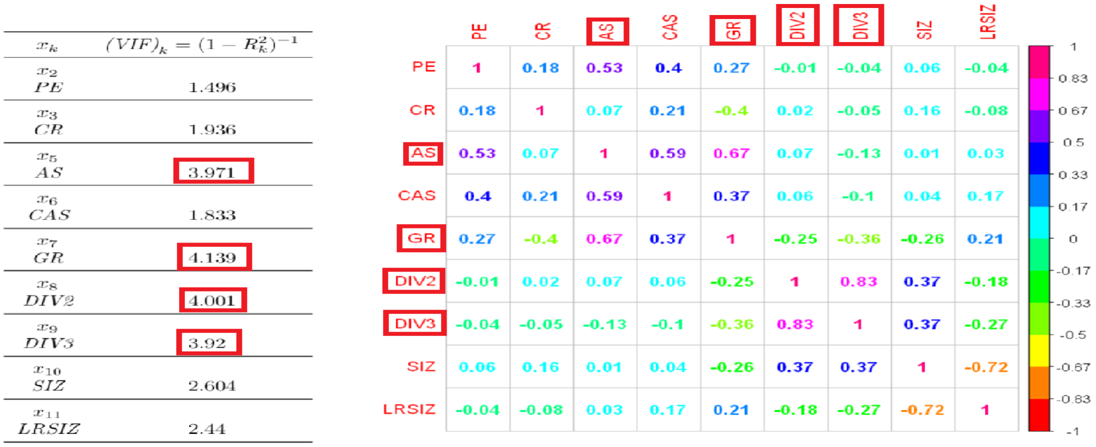
Just by observation from the graphs in figure 2, it appears that PT has the strongest linear relationship with PE. For building the 'best' model, it might be best to include this variable. Since we are only making predictions based on the graphs in figure 2, we still need to do the multiple linear regression analysis for confirmation.

4 Analysis

4.1 Checking For Multicollinearity

For the multiple linear regression analysis, first we must make sure that the independent variables are independent. In other words, the independent variables cannot be correlated or it will cause a multicollinearity problem. If we ignore multicollinearity, some of the consequences are that the standard error of $\hat{\beta}_i$ will be inflated , there will be an effect on the signs(±)for the estimated coefficients,and an inaccurate t-tests/F-test results. To detect multicollinearity, we need to calculate the variance inflation factors(VIF) and do a correlation analysis on the explanatory variables. Our criteria is that if an explanatory variable has a VIF greater than 4, or is correlated with any of the other independent variables, we will consider removing it from the analysis to avoid multicollinearity.

Figure 3: *Correlation Analysis 1*

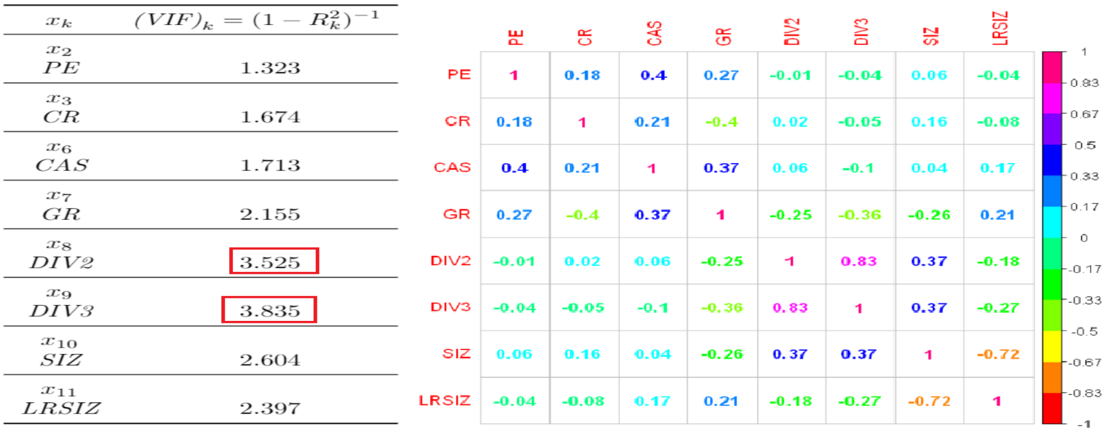


x_k	$(VIF)_k = (1 - R_k^2)^{-1}$
x_2 PE	1.496
x_3 CR	1.936
x_5 AS	3.971
x_6 CAS	1.833
x_7 GR	4.139
x_8 $DIV2$	4.001
x_9 $DIV3$	3.92
x_{10} SIZ	2.604
x_{11} $LRSIZ$	2.44

	PE	CR	AS	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
PE	1	0.18	0.53	0.4	0.27	-0.01	-0.04	0.06	-0.04
CR	0.18	1	0.07	0.21	-0.4	0.02	-0.05	0.16	-0.08
AS	0.53	0.07	1	0.59	0.67	0.07	-0.13	0.01	0.03
CAS	0.4	0.21	0.59	1	0.37	0.06	-0.1	0.04	0.17
GR	0.27	-0.4	0.67	0.37	1	-0.25	-0.36	-0.26	0.21
DIV2	-0.01	0.02	0.07	0.06	-0.25	1	0.83	0.37	-0.18
DIV3	-0.04	-0.05	-0.13	-0.1	-0.36	0.83	1	0.37	-0.27
SIZ	0.06	0.16	0.01	0.04	-0.26	0.37	0.37	1	-0.72
LRSIZ	-0.04	-0.08	0.03	0.17	0.21	-0.18	-0.27	-0.72	1

According to the results in figure 3, GR had the highest VIF of 4.139. If we look at the correlation matrix, we can also see that GR has a moderate correlation with the variable AS. However, AS has a VIF of 3.971 and has a moderate correlation with the variables PE, CAS, and GR. Therefore, we will remove x_5 from the analysis to avoid multicollinearity. We will now reset the process with the remaining explanatory variables. Since an explanatory variable was dropped, the VIFs will either stay the same or decrease.

Figure 4: *Correlation Analysis 2*

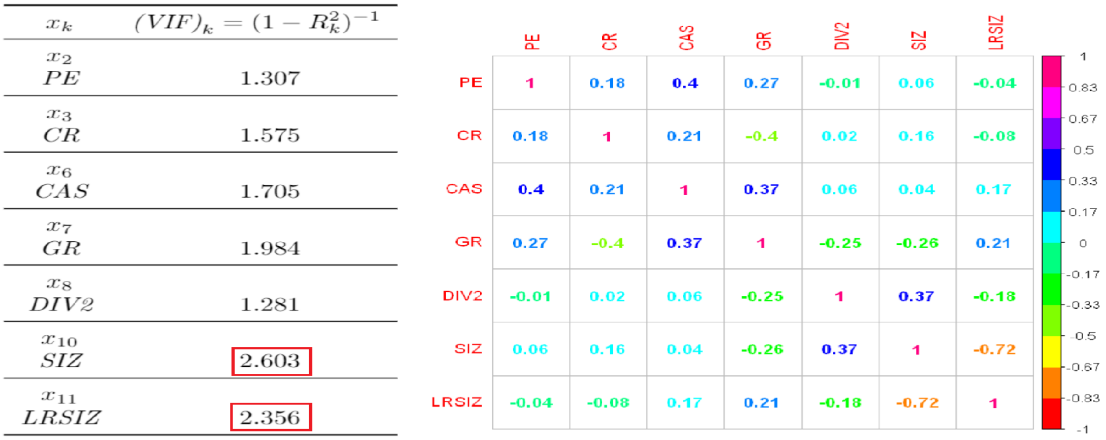


x_k	$(VIF)_k = (1 - R_k^2)^{-1}$
x_2 *PE*	1.323
x_3 *CR*	1.674
x_6 *CAS*	1.713
x_7 *GR*	2.155
x_8 *DIV2*	3.525
x_9 *DIV3*	3.835
x_{10} *SIZ*	2.604
x_{11} *LRSIZ*	2.397

	PE	CR	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
PE	1	0.18	0.4	0.27	-0.01	-0.04	0.06	-0.04
CR	0.18	1	0.21	-0.4	0.02	-0.05	0.16	-0.08
CAS	0.4	0.21	1	0.37	0.06	-0.1	0.04	0.17
GR	0.27	-0.4	0.37	1	-0.25	-0.36	-0.26	0.21
DIV2	-0.01	0.02	0.06	-0.25	1	0.83	0.37	-0.18
DIV3	-0.04	-0.05	-0.1	-0.36	0.83	1	0.37	-0.27
SIZ	0.06	0.16	0.04	-0.26	0.37	0.37	1	-0.72
LRSIZ	-0.04	-0.08	0.17	0.21	-0.18	-0.27	-0.72	1

You can see in figure 4 that none of the independent variables exceeded the threshold. However, *DIV2*'s and *DIV3*'s VIFs are close to 4 and need further investigation. Based on the correlation matrix, *DIV2* and *DIV3* are highly correlated. Since *DIV2* is less correlated with other explanatory variables than *DIV3*, x_9 will be removed from the analysis.

Figure 5: *Correlation Analysis 3*

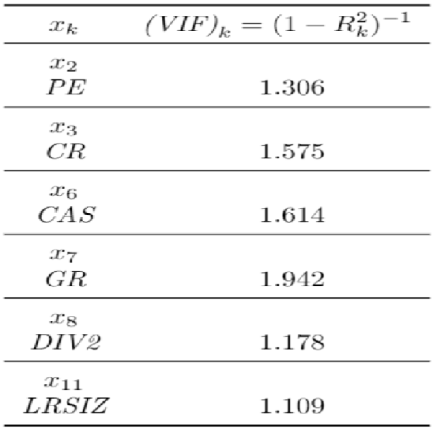


x_k	$(VIF)_k = (1 - R_k^2)^{-1}$
x_2 *PE*	1.307
x_3 *CR*	1.575
x_6 *CAS*	1.705
x_7 *GR*	1.984
x_8 *DIV2*	1.281
x_{10} *SIZ*	2.603
x_{11} *LRSIZ*	2.356

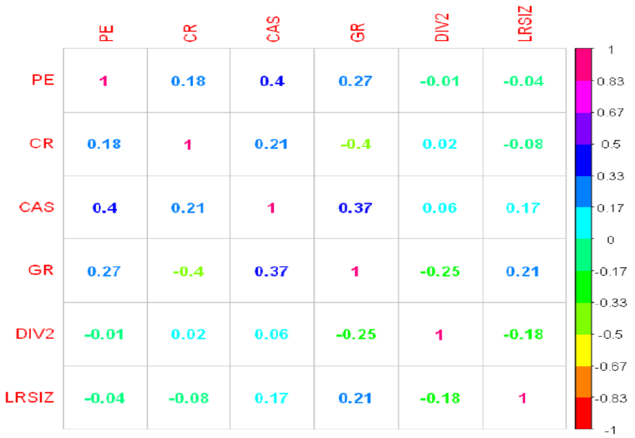
	PE	CR	CAS	GR	DIV2	SIZ	LRSIZ
PE	1	0.18	0.4	0.27	-0.01	0.06	-0.04
CR	0.18	1	0.21	-0.4	0.02	0.16	-0.08
CAS	0.4	0.21	1	0.37	0.06	0.04	0.17
GR	0.27	-0.4	0.37	1	-0.25	-0.26	0.21
DIV2	-0.01	0.02	0.06	-0.25	1	0.37	-0.18
SIZ	0.06	0.16	0.04	-0.26	0.37	1	-0.72
LRSIZ	-0.04	-0.08	0.17	0.21	-0.18	-0.72	1

You can see in figure 5 that the independent variables' VIFs decreased. However, the correlation matrix shows that *SIZ* and *LRSIZ* are highly correlated. Since *LRSIZ* is less correlated with other independent variables than *LIZ*, x_{10} will be removed from the analysis.

Figure 6: *Correlation Analysis 4*



x_k	$(VIF)_k = (1 - R_k^2)^{-1}$
x_2 PE	1.306
x_3 CR	1.575
x_6 CAS	1.614
x_7 GR	1.942
x_8 $DIV2$	1.178
x_{11} $LRSIZ$	1.109



	PE	CR	CAS	GR	DIV2	LRSIZ
PE	1	0.18	0.4	0.27	-0.01	-0.04
CR	0.18	1	0.21	-0.4	0.02	-0.08
CAS	0.4	0.21	1	0.37	0.06	0.17
GR	0.27	-0.4	0.37	1	-0.25	0.21
DIV2	-0.01	0.02	0.06	-0.25	1	-0.18
LRSIZ	-0.04	-0.08	0.17	0.21	-0.18	1

You can see in figure 6 that the correlation matrix shows that all independent variables are weakly correlated with each other. This implies that multicollinearity will no longer be a concern for the analysis.

4.2 Model Selection

There are 6 explanatory variables remaining which leaves us with $2^6 = 64$ possible models including the null model when all other models are not significant. For each model, we will use the R-studio software to conduct t-tests for their independent variables to determine if they are significant at a significance level of 5%. If the model has an independent variable that is not significant, we will remove it from the analysis because we want the 'best' model to make predictions. However,we should always try to limit the number of t-tests conducted to avoid the potential problem of making too many Type 1 errors.

Two-Tailed t-test of an Individual Parameter Coefficient

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k + \hat{\epsilon} \quad i : 1, 2, ..., k$$

The null hypothesis: $H_0 : \beta_i = 0$

The alternative hypothesis: $H_a : \beta_i \neq 0$

Test statistic: $t = \dfrac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$

P-value: $p = 2P(T > |t| \quad | \quad \beta_i = 0)$

where β_i *is one of the predictors from the model*

For the t-tests, if the p-value is less than the significance level, then we reject the null hypothesis, meaning that there is sufficient evidence to conclude that the predictor is useful for the model to predict PT after controlling the other predictors.

Model's variables	R_a^2	*Residual Standard Error*
Model 1: x_2,x_6	0.8502786	0.03266
Model 2: x_2,x_8	0.8491832	0.03278
Model 3: x_6	0.2081542633	0.07512
Model 4: x_2,x_6,x_{11}	0.8604218	0.03154
Model 5: x_2	0.8401529	0.03375
Model 6: x_3,x_7	0.0887896490	0.08058

You can see in the table above that out of the 64 possible models, only 6 were significant. To find the 'best' model, we will choose the set of independent variables with the largest adjusted multiple coefficient of determination, which is denoted as R_a^2. This number is a measure of the model's adequacy and takes into 'account' the sample size n, the number of explanatory variables k, and the coefficient of determination that is used in simple linear regression R^2.

$$R_a^2 = 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1-R^2)$$

This number ranges from 1 to 0 but can sometimes be negative if the model is poor-fitting. The larger the R_a^2, the better the model will be at predicting the dependent variable. Still looking at the table, the set x_2, x_6, and x_{11} have the largest R_a^2 of 0.8604. If the model is significant, this means that 86.04% of the sample variation in average profit per total assets can be explained by average profit per shareholder equity, the overall advertising-to-sales ratio of the firm, and the inversely proportional to market shares. These set of variables also have the lowest residual standard error. This number measures the standard deviation of the residuals; the smaller the number, the better the model fits the dataset. Given that we now have an appropriate set of independent variables to make the 'best' model,we will use the method of least squares to write our multiple linear regression equation. Since the calculations are complicated to do by hand, we used the R-studio software to get the equation.

The Multiple Linear Regression Equation

$$\hat{y} = 0.020178 + 1.184759x_2 + 0.002391x_6 - 0.167515x_{11} + \hat{\epsilon}$$

$$\hat{y} = \textit{the average profit per total assets}$$

$$x_2 = \textit{average profit per shareholder equity}$$

$$x_6 = \textit{overall advertising-to-sales ratio of the firm}$$

$$x_{11} = \textit{inversely proportional to market shares}$$

$$\hat{\epsilon} = \textit{the error term}$$

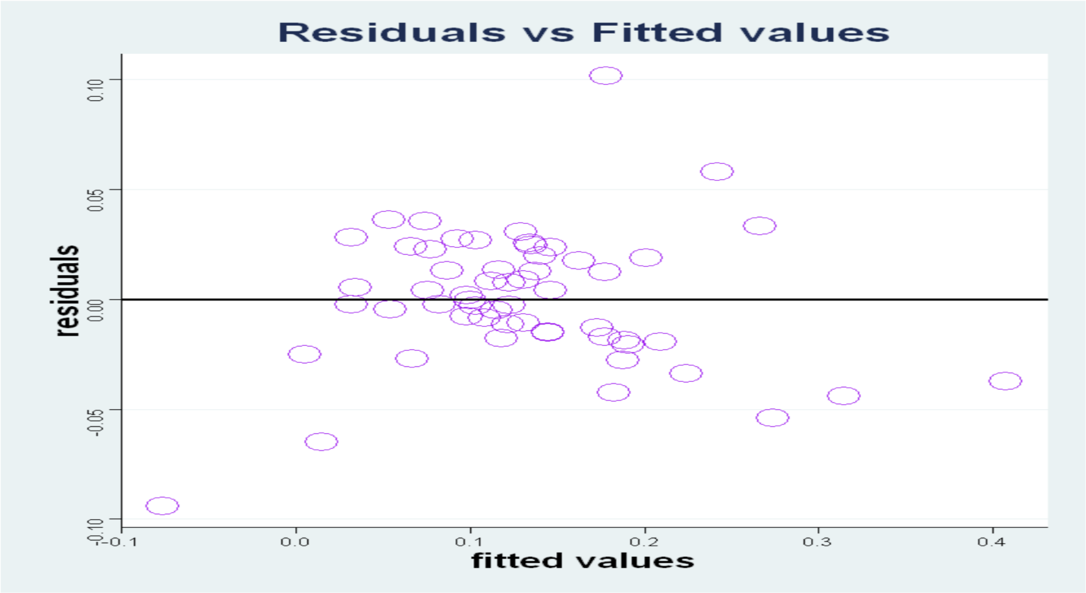
4.3 Checking Model Assumptions

For the multiple linear regression analysis, certain assumptions need to be met for the model to be valid and to make inferences about model's adequacy. All estimates, intervals, and hypothesis tests arising in a regression analysis have been developed assuming that the model is correct. If the model is incorrect, the methods we use are at high risk of being incorrect. These assumptions are as follows: (1) there is a linear relationship between the dependent variable and the independent variables, (2) the residuals are independent, (3) the residuals are normally distributed, and (4) the residuals have equal variance. A residual can be defines as the difference between an observed value and a predicted value of the dependent variable.

4.3.1 Linearity Assumption

For the linearity assumption, the mean of the residual needs to be 0. We detect nonlinearity from the residual vs fitted plot.

Figure 7: *Residual vs Fitted Plot*

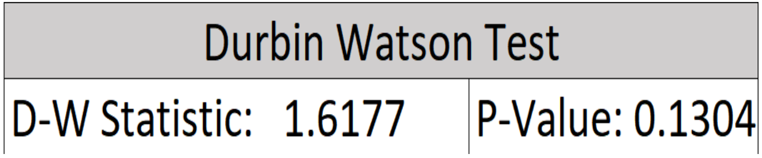


The residuals vs fitted plot = (figure 7) shows that for the majority of time, the points "bounces randomly" around the horizontal zero line. This is an indication that the mean of the residuals is approximately zero. Therefore, the linearity assumption is met.

4.3.2 Independence Assumption

For the independence assumption, the residuals need to be independent or have no correlation with each other. If the data for this analysis was a time series data, we would check the independence assumption by looking at the residuals vs order plot. Since the data is not a time series data, we will use the R-Studio software to perform the Durbin-Watson Test to determine if there is correlation between the residuals.

Figure 8: *Durbin-Watson Test*



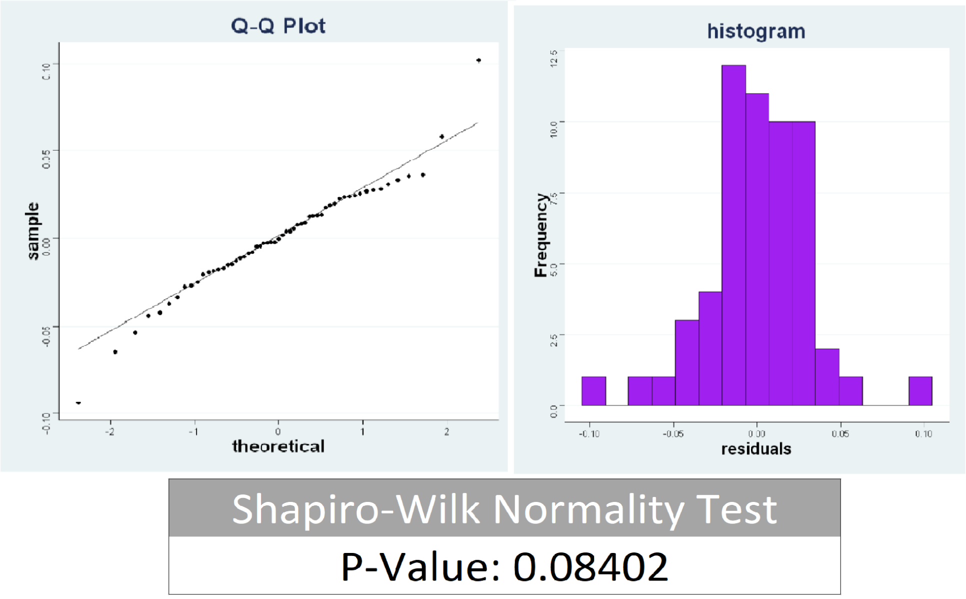
Durbin Watson Test	
D-W Statistic: 1.6177	P-Value: 0.1304

For the Durbin-Watson Test, the null hypothesis is as follows: H_0 : there is no correlation between the residuals vs the alternate hypothesis. The alternative hypothesis is as follows: H_a : the residuals have a positive or negative correlation.We choose our significant level to be $\alpha = 0.05$. Since the p-value= $0.1304 > 0.05$, we fail to reject the null hypothesis, which means there is sufficient evidence that there is no auto- correlation between the residuals. Therefore, the independence assumption is met.

4.3.3 Normality Assumption

For the normality assumption, the residuals need to follow a normal distribution. We can check this assumption using a q-q plot, histogram graph, or performing the Shapiro-Wilk Test.

Figure 9: *Q-Q Plot,Histogram,Shapiro-Wilk Test*



Shapiro-Wilk Normality Test
P-Value: 0.08402

In figure 9, you can see that the majority of the points for the q-q plot are on the straight diagonal line indicating that the residuals follow an approximately normal distribution. The histogram also looks approximately normal because it looks bell-shaped and symmetric around the mean of 0. Lastly, we used R-Studio to conduct the Shapiro-Wilk normality test. The null hypothesis is: H_0 : the residuals follow a normal distribution vs the alternative hypothesis is: H_a : the residuals does not follow a normal distribution. We choose our significant level to be $\alpha = 0.05$. Since the p-value=0.08402 > 0.05, we fail to reject the null hypothesis, which means that there is sufficient evidence that the residuals are normally distributed. Therefore, the normality assumption is met.

4.3.4 Equal Variance Assumption

For the equal variance assumption, the variances of the residuals need to be constant and can be checked by the residual vs fitted plot. As it can be seen in figure 7, the plots are randomly scattered and show no obvious patterns. This is an indication that the variances of the residuals are constant. Therefore, the equal variance assumption is met.

Since all of the assumptions are met, the model for the multiple linear regression is valid.

4.4 Transformations

Based on the fact that all of the linear regression assumptions were met, no transformations are needed for the analysis.

4.5 Detecting Outliers(Influential observations)

An outlier is an unusual observation from the data and should be treated with caution.Four methods we will use to detect them are:

- Standardized Residuals

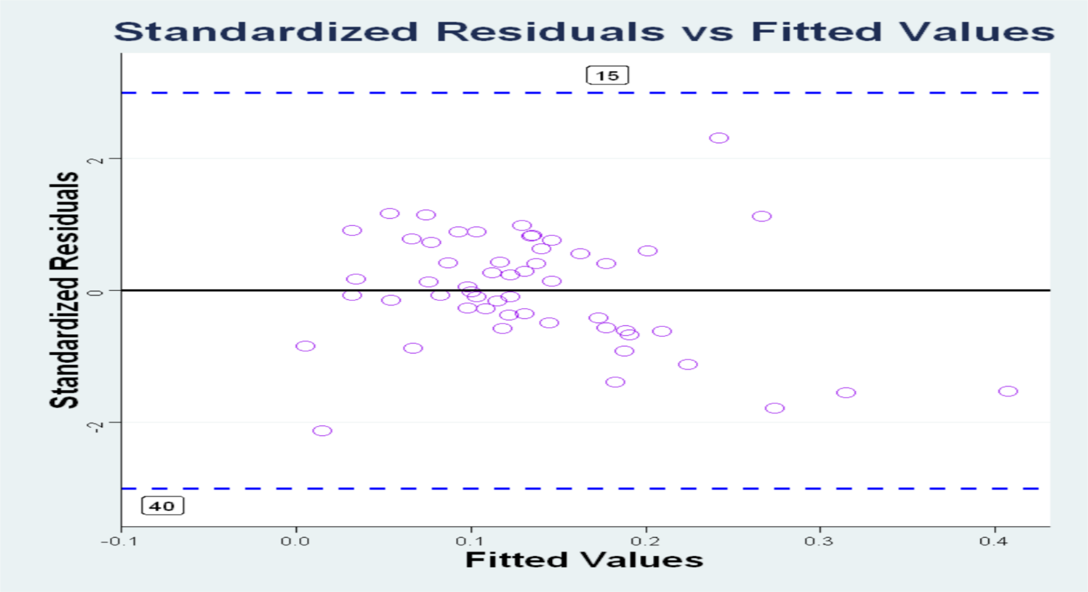
$$r_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}$$

$$\hat{\epsilon}_i : \quad \textit{the residual of the } i^{th}\textit{observation}$$

$$\hat{\sigma} : \quad \textit{the leverage of the } i^{th} \textit{ observation}$$

$$\textit{A residual } r_i^* > 3 \textit{ standard deviations(in absolute value) will be considered an outlier}$$

Figure 10: *Standardized Residuals*



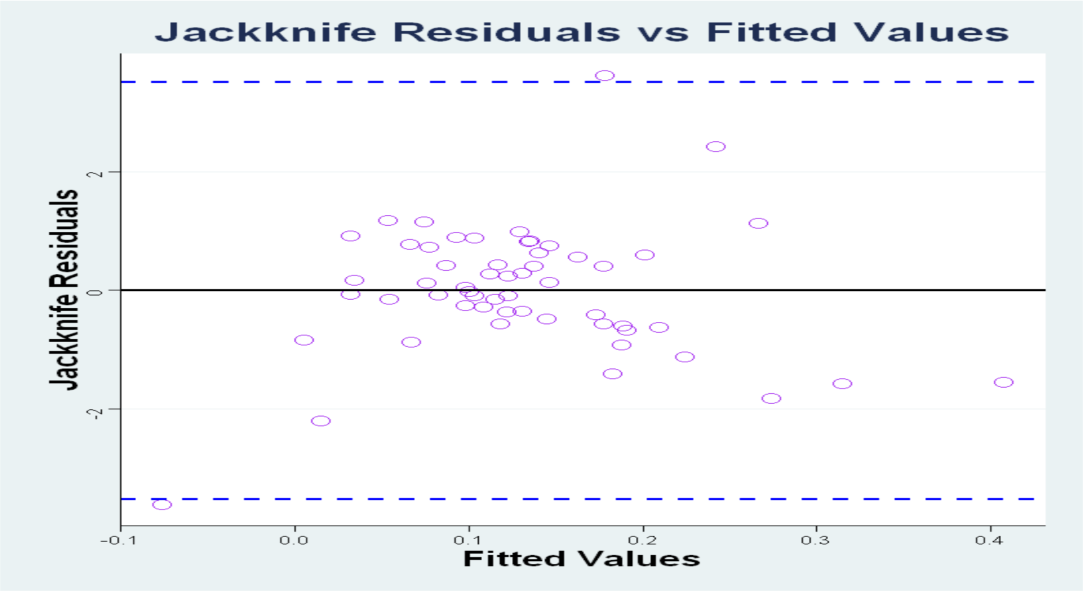
- <span>JACKKNIVES RESIDUALS</span>

$$r^*_{(i)} = \frac{\hat{\epsilon}_{(i)}}{\hat{\sigma_{(i)}}\sqrt{1-h_i}}$$

Bonferroni critical value:qt(.025/n,df=n-2,lower.tail = FALSE)

A residual $r^*_{(i)} >$ *Bonferroni critical value(in absolute value) will be considered an influential observation*

Figure 11: *Jackknives Residuals*



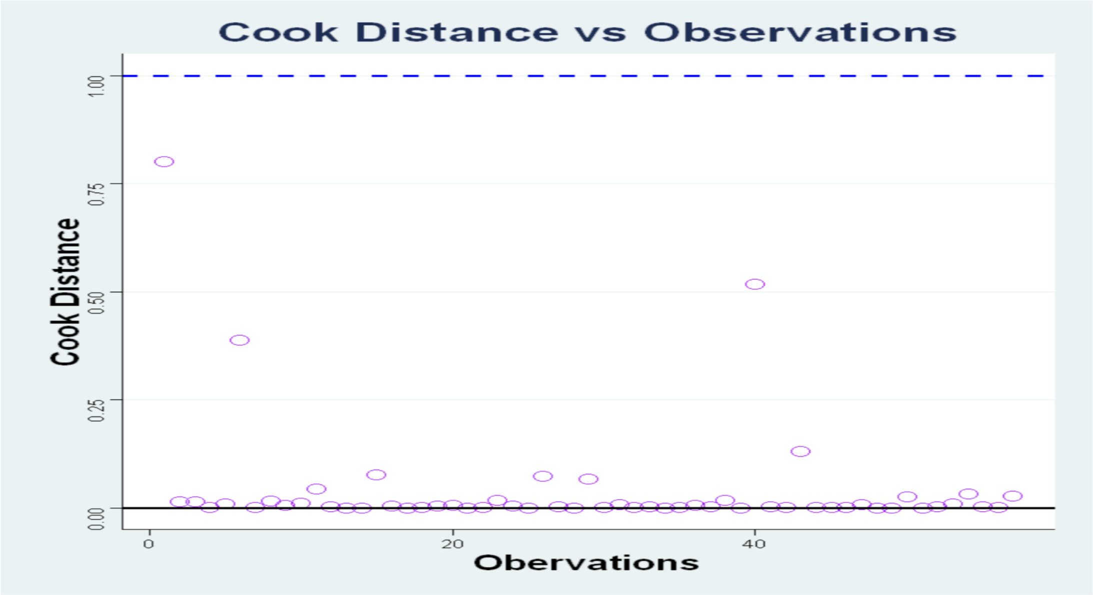
- Cook's Distance Statistic

$$D_i = \frac{\Sigma_j(\hat{y}_j - \hat{y}_{(j)(i)})^2}{p\hat{\sigma}^2}$$

$$D_i > 1 \textit{ will be considered an influential observation(small samples).}$$

$$D_i > \frac{4}{n} \textit{ will be considered an influential observation(large samples).}$$

Figure 12: *Cook's Distance*



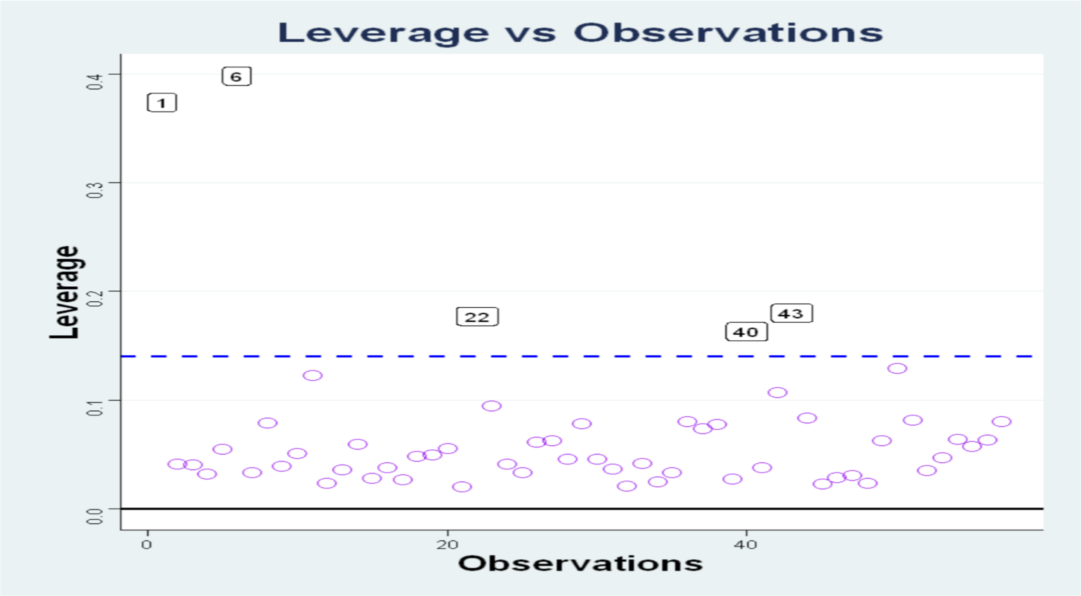
- Leverage

$$h_i \text{ is the leverage for the ith term}$$

$$k\text{= the number of } \beta\text{'s in the model excluding } \beta_0$$

$$h_i > \frac{2(k+1)}{n} \text{ will be considered an influential observation}$$

Figure 13: *Leverage*



For the standardized residuals in figure 10, the analysis shows that observations 15 and 40 are outliers. In figure 13, an observation is considered to have a high value of leverage if its greater than 0.14. The analysis detects that observations 1, 6, 22, 40,and 43 are influential observations. Since observation 40 is an outlier but also an influential observation, we considered removing it because it was the only observation that had a negative PE value. However,it is possible for a firm's PE to be negative. Given that the sample size of the data is small, we felt this observation represents the population and decided to keep it. Also, we removed some outliers to see if there were changes that would make the model perform better but the changes were not significant enough to be deleted from the analysis.

4.6 Testing the Utility of a Model

Due to the fact that all assumptions about the model are met, in multiple linear regression we need to conduct an analysis of variance F-test to determine whether the model is adequate for predicting PT for the researcher's question.If the overall model is not significant, we will rerun the analysis by using the model with the second highest R_a^2 from section 4.2.

F-test for Overall Model Adequacy

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k + \hat{\epsilon}$$

The null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_i = 0$ *(All predictors are unimportant for predicting y)*

The alternative hypothesis: *At least one predictor* $\beta_i \neq 0$ *(for i=1,2,...,k) is useful for predicting y)*

Test statistic: $$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

P-value: $$p = P(F > f, df_{numerator}, df_{denominator})$$

Degree of freedom: df

Sample size: n

The number of terms in the model: k

Recall that our 'best' model is

$$\hat{y} = 0.020178 + 1.184759x_2 + 0.002391x_6 - 0.167515x_{11}$$

$H_0 : \beta_2 = \beta_6 = \beta_{11} = 0$ *(the predictors PE, CAS, and LRSIZ are unimportant for predicting PT)*

H_a : *At least one* $\beta_2, \beta_6, \beta_{11} \neq 0$ *(at least one predictor from PE, CAS, or LRSIZ is useful for predicting PT)*

Test statistic: $F = 116.1$

P-value $= 2.2 \times 10^{-16}$

For this test, we choose our significance level $\alpha = 0.05$.Since the p-value$< \alpha$, we reject the null hypothesis. This provides strong evidence that at least one of the model's predictors is nonzero. Therefore, there is significant evidence that the overall model is useful for predicting the average profit per total assets.

5 Conclusion

After checking for the multicollinearity, we were able to determine that PE, CR, CAS, GR,DIV2, and LRSIZ were the set of variables that may contain the 'best' multiple linear regression model. Since there are 6 variables, then it is noted that there are 64 different combinations to make a model. Of those 64 models, only 6 were found to be significant such that there is evidence that all of the predictors in each model are useful to predict PT. After selecting the top six models, we then looked at their R_a^2's. We observed that model 4 had the highest R_a^2 and the lowest residual standard error.This is important, as the R_a^2 tells us that 86.04% of the sample variation in PT can be explained by the variables average profit per shareholder equity (x_2), overall advertising-to-sales ratio of the firm (x_6), and inversely proportional to market shares (x_{11}). Finding the model with the lowest residual standard error was also important, as having a low residual standard error can indicate how well the model fits the dataset. After carefully considering and concluding which model met our requirements, we can write our best multiple linear regression model as:

$$\hat{y} = 0.020178 + 1.184759x_2 + 0.002391x_6 - 0.167515x_{11}$$

Although we determined this as our best multiple linear regression model, we checked the four line assumptions to make sure the model is valid for the regression. After verifying the model's assumptions, we used statistical methods to see if there were outliers. Lastly, we conducted an F-test to see if the overall model is significant.

Since the model is finalized, we can now interpret each variable and make predictions. We note that $\hat{\beta}_0$ is equal to .020178 , $\hat{\beta}_2$ is equal to 1.184759, $\hat{\beta}_6$ is equal to .002391 and $\hat{\beta}_{11}$ is equal to -0.167515. The interpretation is as follows:

$\hat{\beta}_0$**:** Is a junk term that helps capture the data, it does not provide any good interpretation of the model as we cannot have a PT of .020178 when PE, CAS and LRSIZ are 0.

$\hat{\beta}_2$: We estimate that the mean profit per total assets $E(y)$ for a company will increase by 1.184759 for every 1 unit increase in average profit per shareholder equity(x_2) when the overall advertising-to-sales ratio of the firm(x_6) and the inversely proportional to market shares(x_{11}) are held fixed.

$\hat{\beta}_6$: We estimate that the mean profit per total assets $E(y)$ for a company will increase by 0.002391 for every 1 unit increase in the overall advertising-to-sales ratio when the average profit per shareholder equity(x_2) and the inversely proportional to market shares(x_{11}) are held fixed.

$\hat{\beta}_{11}$: We estimate that the mean profit per total assets $E(y)$ for a company will decrease by 0.167515 for every 1 unit increase in the inversely proportional to market shares when the average profit per shareholder equity(x_2) and the overall advertising-to-sales ratio of the firm(x_6) are held fixed.

Using different values for each of our variables in our model, we can test our model.
Note:

The scope of our data is as follows:

$$\hat{y} = 0.020178 + 1.184759x_2 + 0.002391x_6 - 0.167515x_{11}$$

PE: $(-.05, 0.35)$, **CAS:** (0.1,28), **LRSIZ:** (0.1383, 0.3693)

PE (x_2)	*CAS* (x_6)	*LRSIZ* (x_{11})	*PT* $\hat{y} =$	*Confident Interval*	*Prediction Interval*
0.20	4	.20	0.2331902	$(0.2167559 \leq \mu \leq 0.2496245)$	$(0.1678309 \leq y \leq 0.2985495)$
0.50	30	.50	0.6005235	$(0.5292747 \leq \mu \leq 0.6717723)$	$(0.5052442 \leq y \leq 0.6958027)$
.75	20	.45	0.8811811	$(0.782076 \leq \mu \leq 0.9802862)$	$(0.7636075 \leq y \leq 0.9987548)$

Since our best fit model can only make prediction for the average profit per total assets when PE, CAS, and LRSIZ are in the scope of the data, it would be useful to interpret the CI or PI when they are in the scope. For example in the first row, we are 95% confident that the expected average profit per total assets is between 0.2167559 and 0.2496245 when the average profit per shareholder equity is 0.20, the overall advertising-to-sales ratio of the firm is 4, and the inversely proportional to market shares is .20. For the predicition interval, we are 95% confident that the average profit per total assets is between 0.1678309 and 0.2985495 when the average profit per shareholder equity is 0.20, the overall advertising-to-sales ratio of the firm is 4, and the inversely proportional to market shares is .20.

6 Weakness

One weakness is the small sample size. Due to the limited sample size we had observations that seemed to be outliers but they really represented the population. For future work, we may want to have a larger sample size. As this would give us more information and a better representation of the population. Additionally, having a larger sample size also limits the influence of outliers or extreme observations and removes biases that a small sample size may have. Another weakness in the analysis is that we had categorical variables that we did not use from the dataset, as we did not have the knowledge on how to create models with those types of variables. Including categorical variables could have potentially created a better model, especially if the categorical variables were important. For example, DRC is a variable that measures whether it's CR is greater than 50. This variable is important as it tells us which companies are more competitive than others. With more knowledge, we would be able to incorporate the categorical variables in our original set of independent variables rather than taking them out.

7 Appendix

In [1]:

```
.libPaths()
```

```
'C:/Users/james/OneDrive/Documents/R/win-library/4.0' · 'C:/R-4.0.5/library'
```

In [2]:

```
# create a variable for the library paths
# autocomplete press tab
lib_path = 'C:/Users/james/OneDrive/Documents/R/win-library/4.0'
```

In [3]:

```
# install libraries
install.packages('ggplot2', lib = lib_path)
install.packages('lubridate', lib = lib_path)
install.packages('GGally', lib = lib_path)
install.packages("ggpubr",lib = lib_path)
install.packages("qpcR",lib = lib_path)
install.packages("matlib",lib = lib_path)
install.packages("car",lib = lib_path)
install.packages("dplyr ",lib = lib_path)
install.packages("ggrepel ",lib = lib_path)
install.packages("latex2exp",lib = lib_path)
install.packages("graphics",lib = lib_path)
install.packages("knitr",lib = lib_path)
install.packages("cowplot",lib = lib_path)
install.packages("gridExtra",lib = lib_path)
install.packages("patchwork",lib = lib_path)
install.packages("corrplot",lib = lib_path)
install.packages("faraway",lib = lib_path)
install.packages("ggthemes",lib = lib_path)
install.packages("lmtest",lib = lib_path)
```

```
package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'lubridate' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'GGally' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'ggpubr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'qpcR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'matlib' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'car' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages

Warning message:
"package 'dplyr ' is not available for this version of R

A version of this package for your version of R might be available elsewhere,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages" (https://cran.r-project.
org/doc/manuals/r-patched/R-admin.html#Installing-packages")
Warning message:
"package 'ggrepel ' is not available for this version of R

A version of this package for your version of R might be available elsewhere,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages" (https://cran.r-project.
org/doc/manuals/r-patched/R-admin.html#Installing-packages")

package 'latex2exp' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages

Warning message:
"package 'graphics' is in use and will not be installed"

package 'knitr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'cowplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'gridExtra' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'patchwork' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages

  There is a binary version available but the source version is later:
         binary source needs_compilation
corrplot   0.84   0.88             FALSE


installing the source package 'corrplot'



package 'faraway' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'ggthemes' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
package 'lmtest' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\james\AppData\Local\Temp\RtmpY5fBBA\downloaded_packages
```

In [4]:

```
# load libraries
library('ggplot2')
library('lubridate')
library('GGally')
library('ggpubr')
library('qpcR')
library('matlib')
library('car')
library('dplyr')
library('ggrepel')
library('latex2exp')
library('graphics')
library('rgl')
library('cowplot')
library('gridExtra')
library('patchwork')
library('corrplot')
library('faraway')
library('ggthemes')
library('lmtest')
```

```
Loading required package: Matrix


Attaching package: 'matlib'


The following object is masked from 'package:rgl':

    GramSchmidt


Loading required package: carData


Attaching package: 'dplyr'


The following object is masked from 'package:car':
```

In [5]:

```
# function to find all adjusted r^squared #

### input are strings ###
combination_function = function(dependent_var,independent_vars_vector,data_frame){

    # empty list
    equations = list()

    for(i in 1:length(independent_vars_vector)){

        vector_com = combn(independent_vars_vector,i)
        for (j in 1:ncol(vector_com)){

            model_f = as.formula(paste0(dependent_var,"~",paste0(vector_com[,j],collapse = "+")))
            equations = c(equations, model_f)
    }
}
    equation_output = NULL

    for(k in 1:length(equations)){

        equation = lm( equations[[k]], data= data_frame)
        terms = length(equation$coefficients)
        independent_var = c()

        independent_var[1] = paste0(dependent_var," = (",round(as.numeric(equation$coefficients[1]),4),") +

        for(i in 2:terms){

            independent_var[i] = paste0("(",round(as.numeric(equation$coefficients[i]),4),") ",names(equatio

        }

        eq = paste(independent_var,collapse = "")
        adj_rsquare = summary(equation)$adj.r.squared
        equation_df = data.frame("Equation"=eq, "adjusted r^2"=adj_rsquare)
        equation_output = rbind(equation_output,equation_df)

    }

    equation_output = equation_output[order(-equation_output$adjusted.r.2),]
    return(equation_output)
    }
```

In [ ]:

```

```

Dataset #2F – Firm Profits (Source: accessed publicly and merged from a secret database)
##
##

□ Description: Profit rates and Market structure of Advertising Intensive firms from (2013 – 2018) are provided.
###
##

□ Variables: □ Firm - name of firm

□ PT – (Net Income + Interest Expense)/total assets

□ PE – (Net Income/shareholder equity)

□ CR – Weighted concentration ratio of firm’s product markets

□ DRC – Dummy variable for CR > 50

□ AS - Weighted average Industry advertising-to-sales ratio of firm's product markets

□ CAS – Overall advertising-to-sales ratio of the firm

□ GR – Weighted average percent changes in industry sales in the firm’s market

□ DIV2 - Firm’s diversification in the more broad industries

□ DIV3 - Firm’s diversification in the less broad industries

□ SIZ - Firms 2018 total assets (millions)

□ LRSIZ – inversely proportional to market shares
##
##

□ Research Question*: Find a ‘best’ multiple linear regression model to predict the average profit per total assets (PT)

In [28]:

```
# upload data

firm_profit_df = read.csv('2F - FirmProfit.csv', header=TRUE)
head(firm_profit_df,5)
```

A data.frame: 5 × 12

	FIRM	PT	PE	CR	DCR	AS	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
	<chr>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>
1	Alberto-Culver	0.30	0.17	40	0	16.5	28.0	1.67	0.28	0.28	37	0.2769
2	American-Bakeries	0.09	0.06	55	1	0.9	0.4	1.16	0.00	0.00	78	0.2295
3	American-Sugar	0.11	0.07	43	0	0.2	0.5	1.13	0.00	0.00	269	0.1787
4	Anheuser-Busch	0.13	0.10	80	1	3.3	2.6	1.37	0.00	0.13	366	0.1694
5	Armour	0.09	0.06	31	0	0.2	0.4	1.34	0.34	0.35	588	0.1568

In [116]:

```
#####################################################################
######################## PE boxplot #################################
pe.box = ggplot(data=firm_profit_df, aes(x=PE)) +

geom_boxplot(fill='red') +

#### lable names ####
labs(x='PE',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## PE histogram #################################
pe.his = ggplot(data=firm_profit_df, aes(x=PE))+

#histogram
geom_histogram(bins = 15, fill = 'red', color = 'black')+

# lable names
labs(x='PE', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs PE scatt ################################
pe.scat = ggplot(data=firm_profit_df, aes(x=PE,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="red") +
#### lable names ####
labs(x='PE', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [117]:

```
#####################################################################
######################## CR boxplot #################################
cr.box = ggplot(data=firm_profit_df, aes(x=CR)) +

geom_boxplot(fill='orange') +

#### lable names ####
labs(x='CR',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## CR histogram #################################
cr.his = ggplot(data=firm_profit_df, aes(x=CR))+

#histogram
geom_histogram(bins = 15, fill = 'orange', color = 'black')+

# lable names
labs(x='CR', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs CR scatt ################################
cr.scat = ggplot(data=firm_profit_df, aes(x=CR,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="orange") +
#### lable names ####
labs(x='CR', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [118]:

```
#####################################################################
######################## AS boxplot #################################
as.box = ggplot(data=firm_profit_df, aes(x=AS)) +

geom_boxplot(fill='yellow') +

#### lable names ####
labs(x='AS',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## AS histogram #################################
as.his = ggplot(data=firm_profit_df, aes(x=AS))+

#histogram
geom_histogram(bins = 15, fill = 'yellow', color = 'black')+

# lable names
labs(x='AS', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs AS scatt ################################
as.scat = ggplot(data=firm_profit_df, aes(x=AS,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="yellow") +
#### lable names ####
labs(x='AS', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [119]:

```
#####################################################################
######################## CAS boxplot #################################
cas.box = ggplot(data=firm_profit_df, aes(x=CAS)) +

geom_boxplot(fill='green') +

#### lable names ####
labs(x='CAS',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## CAS histogram #################################
cas.his = ggplot(data=firm_profit_df, aes(x=CAS))+

#histogram
geom_histogram(bins = 15, fill = 'green', color = 'black')+

# lable names
labs(x='CAS', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs CAS scatt ################################
cas.scat = ggplot(data=firm_profit_df, aes(x=CAS,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="green") +
#### lable names ####
labs(x='CAS', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [120]:

```
#####################################################################
######################## GR boxplot #################################
gr.box = ggplot(data=firm_profit_df, aes(x=GR)) +

geom_boxplot(fill='blue') +

#### lable names ####
labs(x='GR',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## GR histogram #################################
gr.his = ggplot(data=firm_profit_df, aes(x=GR))+

#histogram
geom_histogram(bins = 15, fill = 'blue', color = 'black')+

# lable names
labs(x='GR', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs GR scatt ################################
gr.scat = ggplot(data=firm_profit_df, aes(x=GR,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="blue") +
#### lable names ####
labs(x='GR', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [121]:

```
#####################################################################
######################## DIV2 boxplot #################################
div2.box = ggplot(data=firm_profit_df, aes(x=DIV2)) +

geom_boxplot(fill='#4b0082') +

#### lable names ####
labs(x='DIV2',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## DIV3 histogram #################################
div2.his = ggplot(data=firm_profit_df, aes(x=DIV2))+

#histogram
geom_histogram(bins = 15, fill = '#4b0082', color = 'black')+

# lable names
labs(x='DIV2', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs DIV2 scatt ################################
div2.scat = ggplot(data=firm_profit_df, aes(x=DIV2,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="#4b0082") +
#### lable names ####
labs(x='DIV2', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [122]:

```
#####################################################################
######################## DIV3 boxplot #################################
div3.box = ggplot(data=firm_profit_df, aes(x=DIV3)) +

geom_boxplot(fill='violet') +

#### lable names ####
labs(x='DIV3',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## DIV3 histogram #################################
div3.his = ggplot(data=firm_profit_df, aes(x=DIV3))+

#histogram
geom_histogram(bins = 15, fill = 'violet', color = 'black')+

# lable names
labs(x='DIV3', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs DIV3 scatt ################################
div3.scat = ggplot(data=firm_profit_df, aes(x=DIV3,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="violet") +
#### lable names ####
labs(x='DIV3', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [123]:

```
#####################################################################
######################## SIZ boxplot #################################
siz.box = ggplot(data=firm_profit_df, aes(x=SIZ)) +

geom_boxplot(fill='cyan') +

#### lable names ####
labs(x='SIZ',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#######################################################################
######################## SIZ histogram #################################
siz.his = ggplot(data=firm_profit_df, aes(x=SIZ))+

#histogram
geom_histogram(bins = 15, fill = 'cyan', color = 'black')+

# lable names
labs(x='SIZ', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## PT vs SIZ scatt ################################
siz.scat = ggplot(data=firm_profit_df, aes(x=SIZ,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="cyan") +
#### lable names ####
labs(x='SIZ', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

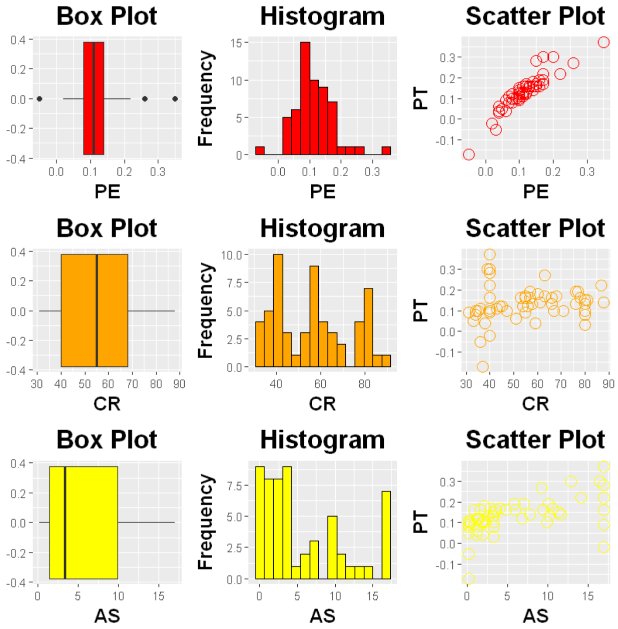
In [124]:

```
######################################################################
######################## LRSIZ boxplot #################################
lrsiz.box = ggplot(data=firm_profit_df, aes(x=LRSIZ)) +

geom_boxplot(fill='brown') +

#### lable names ####
labs(x='LRSIZ',y='', title='Box Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
########################################################################
######################## LRSIZ histogram #################################
lrsiz.his = ggplot(data=firm_profit_df, aes(x=LRSIZ))+

#histogram
geom_histogram(bins = 15, fill = 'brown', color = 'black')+

# lable names
labs(x='LRSIZ', y='Frequency',title='Histogram')+

# title-label
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold'))
#########################################################################
######################## PT vs LRSIZ scatt ################################
lrsiz.scat = ggplot(data=firm_profit_df, aes(x=LRSIZ,y=PT)) +

#### point settings ###
geom_point(shape=1,size=4,color="brown") +
#### lable names ####
labs(x='LRSIZ', y='PT', title='Scatter Plot') +

#### Title-label ####
theme(plot.title = element_text(hjust=.5, size=20, face='bold')) +

#### x-label ####
theme(axis.title.x = element_text(hjust=.5, size=15, face='bold')) +

#### y-label ####
theme(axis.title.y = element_text(hjust=.5, size=15, face='bold'))
```

In [ ]:

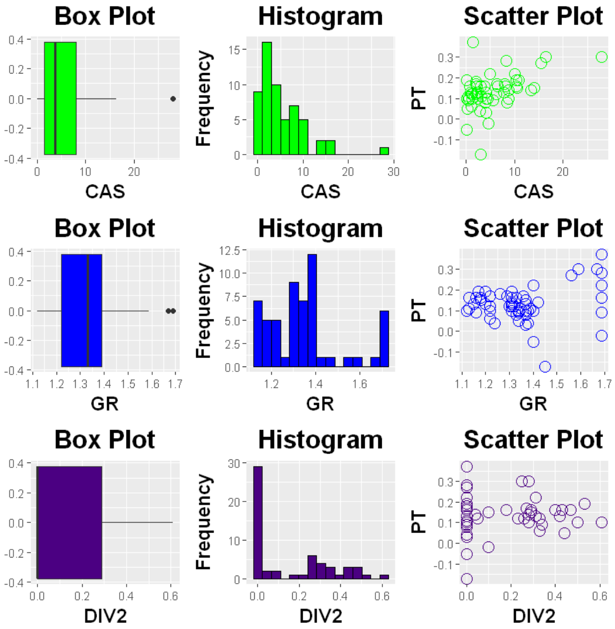
In [125]:

```
pe.box+pe.his+pe.scat+
cr.box+cr.his+cr.scat+
as.box+as.his+as.scat
```



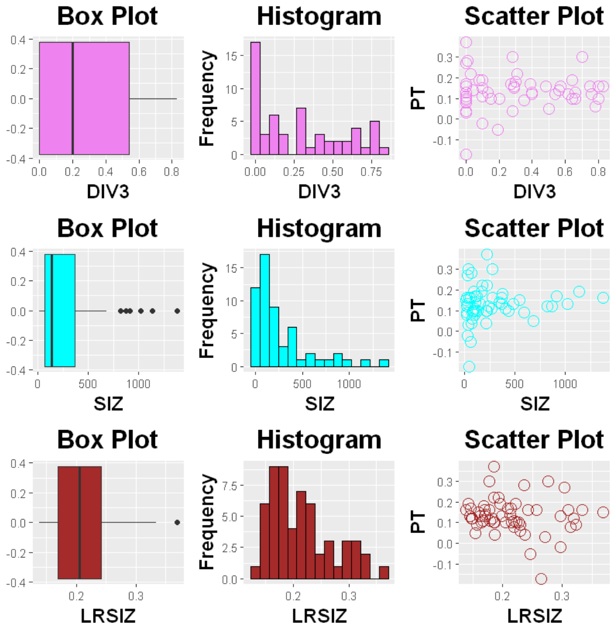
In [126]:

```
cas.box+cas.his+cas.scat+
gr.box+gr.his+gr.scat+
div2.box+div2.his+div2.scat
```



In [127]:

```
div3.box+div3.his+div3.scat+
siz.box+siz.his+siz.scat+
lrsiz.box+lrsiz.his+lrsiz.scat
```



In [ ]:

In [9]:

```
head(firm_profit_df,3)
```

A data.frame: 3 × 12

	FIRM	PT	PE	CR	DCR	AS	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
	<chr>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>
1	Alberto-Culver	0.30	0.17	40	0	16.5	28.0	1.67	0.28	0.28	37	0.2769
2	American-Bakeries	0.09	0.06	55	1	0.9	0.4	1.16	0.00	0.00	78	0.2295
3	American-Sugar	0.11	0.07	43	0	0.2	0.5	1.13	0.00	0.00	269	0.1787

In [ ]:

In [29]:

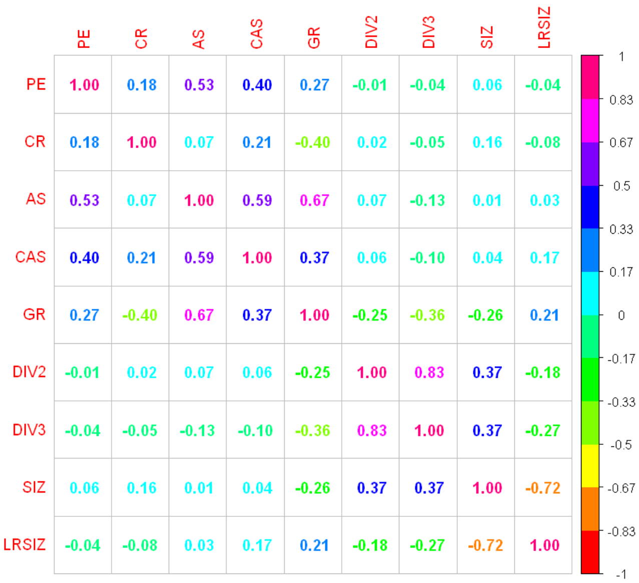
```
###### Full blown model #####
ind_vars = c('PE','CR','AS','CAS','GR','DIV2','DIV3','SIZ','LRSIZ')
model = lm(PT ~PE+CR+AS+CAS+GR+DIV2+DIV3+SIZ+LRSIZ,data=firm_profit_df)

######### eigenn values #########
X = model.matrix(model)[,-1]
eig = eigen(t(X)%*%X)
eig$val
K = sqrt(eig$val[1]/eig$val)
```

```
9573480.63155159 · 109519.064603014 · 2604.70192550467 · 657.56145197888 · 10.8370390898704 ·
4.99823468751705 · 0.413937853275911 · 0.137651174941192 · 0.0793350163191742
```

In [ ]:

In [30]:

```
### corr maxtrix trail 1###

X = model.matrix(model)[,-1] # remove the intercept part
M = cor(X)
corrplot(M, method="number",col = rainbow(12)) ### AS variable will be removed
```



	PE	CR	AS	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
PE	1.00	0.18	0.53	0.40	0.27	-0.01	-0.04	0.06	-0.04
CR	0.18	1.00	0.07	0.21	-0.40	0.02	-0.05	0.16	-0.08
AS	0.53	0.07	1.00	0.59	0.67	0.07	-0.13	0.01	0.03
CAS	0.40	0.21	0.59	1.00	0.37	0.06	-0.10	0.04	0.17
GR	0.27	-0.40	0.67	0.37	1.00	-0.25	-0.36	-0.26	0.21
DIV2	-0.01	0.02	0.07	0.06	-0.25	1.00	0.83	0.37	-0.18
DIV3	-0.04	-0.05	-0.13	-0.10	-0.36	0.83	1.00	0.37	-0.27
SIZ	0.06	0.16	0.01	0.04	-0.26	0.37	0.37	1.00	-0.72
LRSIZ	-0.04	-0.08	0.03	0.17	0.21	-0.18	-0.27	-0.72	1.00

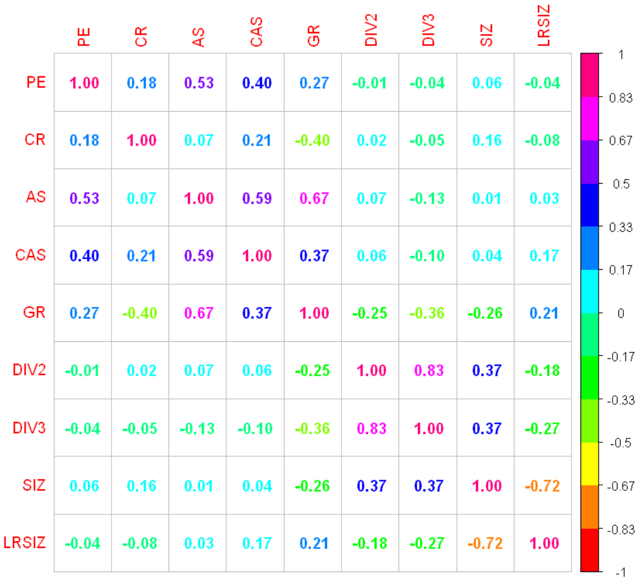
In [132]:

```
model = lm(PT ~PE+CR+CAS+GR+DIV2+DIV3+SIZ+LRSIZ,data=firm_profit_df)
round(vif(model),3)
```

PE: 1.323 **CR:** 1.674 **CAS:** 1.713 **GR:** 2.155 **DIV2:** 3.525 **DIV3:** 3.835 **SIZ:** 2.604 **LRSIZ:** 2.397

In [31]:

```
### corr maxtrix trail 2###

X = model.matrix(model)[,-1] # remove the intercept part
M = cor(X)
corrplot(M, method="number",col = rainbow(12)) ### DIV3 variable will be removed
```



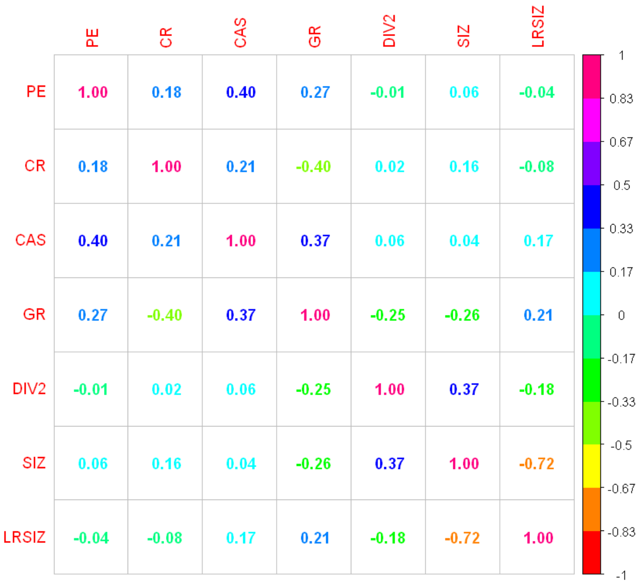
	PE	CR	AS	CAS	GR	DIV2	DIV3	SIZ	LRSIZ
PE	1.00	0.18	0.53	0.40	0.27	-0.01	-0.04	0.06	-0.04
CR	0.18	1.00	0.07	0.21	-0.40	0.02	-0.05	0.16	-0.08
AS	0.53	0.07	1.00	0.59	0.67	0.07	-0.13	0.01	0.03
CAS	0.40	0.21	0.59	1.00	0.37	0.06	-0.10	0.04	0.17
GR	0.27	-0.40	0.67	0.37	1.00	-0.25	-0.36	-0.26	0.21
DIV2	-0.01	0.02	0.07	0.06	-0.25	1.00	0.83	0.37	-0.18
DIV3	-0.04	-0.05	-0.13	-0.10	-0.36	0.83	1.00	0.37	-0.27
SIZ	0.06	0.16	0.01	0.04	-0.26	0.37	0.37	1.00	-0.72
LRSIZ	-0.04	-0.08	0.03	0.17	0.21	-0.18	-0.27	-0.72	1.00

In [32]:

```
### corr maxtrix trail 3###

model = lm(PT ~PE+CR+CAS+GR+DIV2+SIZ+LRSIZ,data=firm_profit_df)
round(vif(model),3)

### corr maxtrix trail 3###

X = model.matrix(model)[,-1] # remove the intercept part
M = cor(X)
corrplot(M, method="number",col = rainbow(12)) ### SIZ variable will be removed

X = model.matrix(model)[,-1]
eig = eigen(t(X)%*%X)
eig$val
```

PE: 1.307 **CR:** 1.575 **CAS:** 1.705 **GR:** 1.984 **DIV2:** 1.281 **SIZ:** 2.603 **LRSIZ:** 2.356

9572627.78965759 · 108528.06170664 · 1451.27764860324 · 12.1258002822566 · 1.71515761884628 · 0.170696562350851 · 0.092362595029722

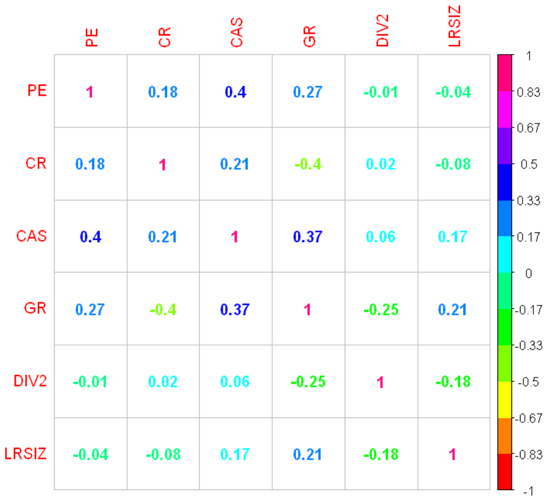


	PE	CR	CAS	GR	DIV2	SIZ	LRSIZ
PE	1.00	0.18	0.40	0.27	-0.01	0.06	-0.04
CR	0.18	1.00	0.21	-0.40	0.02	0.16	-0.08
CAS	0.40	0.21	1.00	0.37	0.06	0.04	0.17
GR	0.27	-0.40	0.37	1.00	-0.25	-0.26	0.21
DIV2	-0.01	0.02	0.06	-0.25	1.00	0.37	-0.18
SIZ	0.06	0.16	0.04	-0.26	0.37	1.00	-0.72
LRSIZ	-0.04	-0.08	0.17	0.21	-0.18	-0.72	1.00

In [135]:

```
model = lm(PT ~PE+CR+CAS+GR+DIV2+LRSIZ,data=firm_profit_df)
round(vif(model),3)

### corr maxtrix trail 4####

X = model.matrix(model)[,-1] # remove the intercept part
M = cor(X)
corrplot(M, method="number",col = rainbow(12)) ### Looks Great!
```

PE: 1.306 **CR:** 1.575 **CAS:** 1.614 **GR:** 1.942 **DIV2:** 1.178 **LRSIZ:** 1.109



	PE	CR	CAS	GR	DIV2	LRSIZ
PE	1	0.18	0.4	0.27	-0.01	-0.04
CR	0.18	1	0.21	-0.4	0.02	-0.08
CAS	0.4	0.21	1	0.37	0.06	0.17
GR	0.27	-0.4	0.37	1	-0.25	0.21
DIV2	-0.01	0.02	0.06	-0.25	1	-0.18
LRSIZ	-0.04	-0.08	0.17	0.21	-0.18	1

In [11]:

```
ind_vars = c('PE','CR','CAS','GR','DIV2','LRSIZ')
results = combination_function('PT',ind_vars,firm_profit_df)
```

In [34]:

```
##### top 16 from the list #####
head(results,16)
```

A data.frame: 16 × 2

	Equation	adjusted.r.2
	<chr>	<dbl>
63	PT = (0.0806) + (1.2197) PE(-4e-04) CR(0.0029) CAS(-0.0385) GR(0.0276) DIV2(-0.1448) LRSIZ	0.8645994
50	PT = (0.0093) + (1.1922) PE(0.0022) CAS(0.0366) DIV2(-0.1431) LRSIZ	0.8644472
59	PT = (0.0231) + (1.1988) PE(-2e-04) CR(0.0024) CAS(0.0363) DIV2(-0.1513) LRSIZ	0.8641758
58	PT = (0.1084) + (1.2222) PE(-5e-04) CR(0.0031) CAS(-0.0521) GR(-0.159) LRSIZ	0.8637812
61	PT = (0.022) + (1.1965) PE(0.0023) CAS(-0.0104) GR(0.0343) DIV2(-0.1396) LRSIZ	0.8621212
28	PT = (0.0202) + (1.1848) PE(0.0024) CAS(-0.1675) LRSIZ	0.8604218
44	PT = (0.0342) + (1.1915) PE(-2e-04) CR(0.0025) CAS(-0.1757) LRSIZ	0.8601874
49	PT = (0.0456) + (1.1948) PE(0.0026) CAS(-0.0221) GR(-0.1569) LRSIZ	0.8593117
27	PT = (-0.023) + (1.211) PE(0.0019) CAS(0.0452) DIV2	0.8579144
57	PT = (0.0535) + (1.2398) PE(-4e-04) CR(0.0025) CAS(-0.0439) GR(0.035) DIV2	0.8577760
43	PT = (-0.0139) + (1.2168) PE(-2e-04) CR(0.002) CAS(0.0453) DIV2	0.8565631
48	PT = (-5e-04) + (1.2174) PE(0.002) CAS(-0.0174) GR(0.041) DIV2	0.8561045
42	PT = (0.0862) + (1.2456) PE(-5e-04) CR(0.0029) CAS(-0.0623) GR	0.8548928
31	PT = (0.0022) + (1.2706) PE(0.0431) DIV2(-0.1018) LRSIZ	0.8511662
26	PT = (0.0249) + (1.2185) PE(0.0023) CAS(-0.0328) GR	0.8509660
8	PT = (-0.0161) + (1.2055) PE(0.002) CAS	0.8502786

In [35]:

```
### all models that are sig at 0.05 ###

model = lm(PT~PE+CAS+LRSIZ,firm_profit_df) # best model
#model = lm(PT~PE+CAS,firm_profit_df)
#model3 =lm(PT~PE+DIV2,firm_profit_df)
#model4 = lm(PT~PE,firm_profit_df)
#model5 = lm(PT~CAS,firm_profit_df)
#model6 = lm(PT~CR+GR,firm_profit_df)
summary(model)
```

```
Call:
lm(formula = PT ~ PE + CAS + LRSIZ, data = firm_profit_df)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.094078 -0.017058 -0.000543  0.019753  0.101697 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.020178   0.018659   1.081   0.2844    
PE           1.184759   0.076206  15.547   <2e-16 ***
CAS          0.002391   0.000902   2.651   0.0106 *  
LRSIZ       -0.167515   0.075489  -2.219   0.0308 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03154 on 53 degrees of freedom
Multiple R-squared:  0.8679,	Adjusted R-squared:  0.8604 
F-statistic: 116.1 on 3 and 53 DF,  p-value: < 2.2e-16
```

In [28]:

In [14]:

```
# getting the residuals
residuals = model$residuals

# standardize residuals
stand_residuals = rstandard(model)

# jackknife residuals
jack_residuals = rstudent(model)

#find Cook's distance for each observation
cooks_distance = cooks.distance(model)

#diangoal of the hat matrix
leverage = hatvalues(model)

#make a dataframe of interest

#empty dataframe
df = data.frame()


#add cols of interest
n = length(firm_profit_df$LRSIZ)
df = cbind( c(1:n),firm_profit_df$PE,firm_profit_df$CAS,firm_profit_df$LRSIZ,firm_profit_df$PT,fitted(model)


#cols name
col_labels=c('observations',"PE(x2)","CAS(x6)","LRSIZ(x11)","PT(y)",'PT(yhat)',"residuals","stand_residuals"
colnames(df) = col_labels

#convert matrix to dataframe
df = data.frame(df)
```

In [ ]:

In [37]:

```
head(df,5)
```

A data.frame: 5 × 11

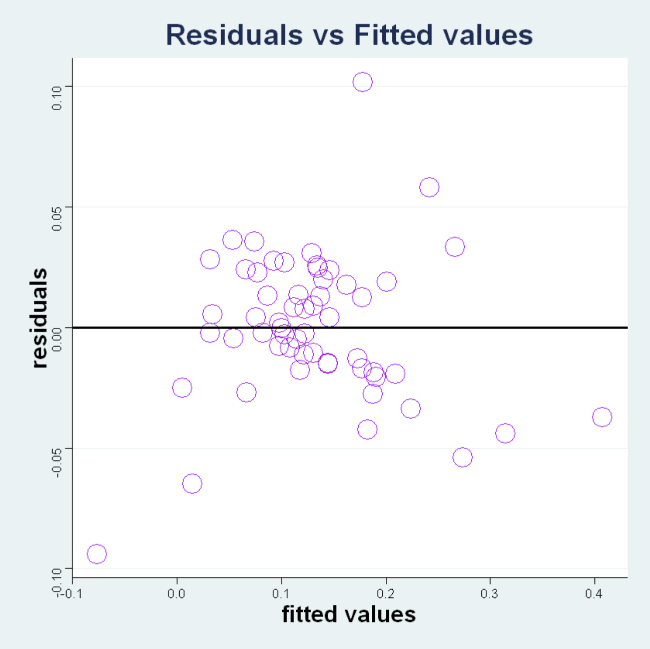
	observations	PE.x2.	CAS.x6.	LRSIZ.x11.	PT.y.	PT.yhat.	residuals	stand_residuals	jackknife_residuals	cooks_distance	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	1	0.17	28.0	0.2769	0.30	0.24214416	0.05785584	2.3179516	2.4220269	0.801446577	0.
2	2	0.06	0.4	0.2295	0.09	0.05377549	0.03622451	1.1726884	1.1769421	0.014596637	0.
3	3	0.07	0.5	0.1787	0.11	0.07437192	0.03562808	1.1531593	1.1568333	0.013981737	0.
4	4	0.10	2.6	0.1694	0.13	0.11649321	0.01350679	0.4352408	0.4318877	0.001557424	0.
5	5	0.06	0.4	0.1568	0.09	0.06595384	0.02404616	0.7842292	0.7813422	0.008920208	0.

In [ ]:

```

```

In [142]:

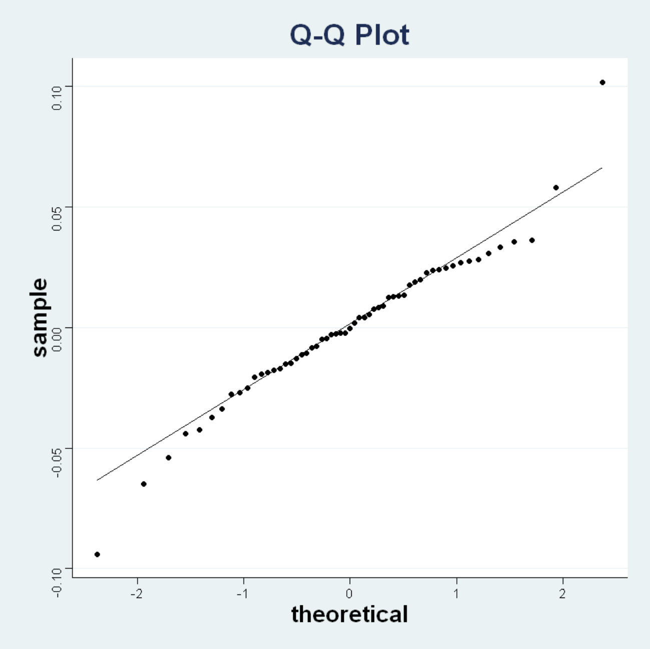
```
### Linearity ###

#upload data
ggplot(data=df, aes(x=PT.yhat., y=residuals)) +

#point settings
geom_point(shape=1,size=7,colour="purple") + # point settings

theme_stata() +


# lable names
labs(x='fitted values', y='residuals', title='Residuals vs Fitted values') +

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold')) +

# add hor line
geom_hline(yintercept=0, col = "black", size=1)
```


Residuals vs Fitted values
residuals
fitted values
0.10
0.05
0.00
-0.05
-0.10
-0.1
0.0
0.1
0.2
0.3
0.4

```
In [57]:
```

In [161]:

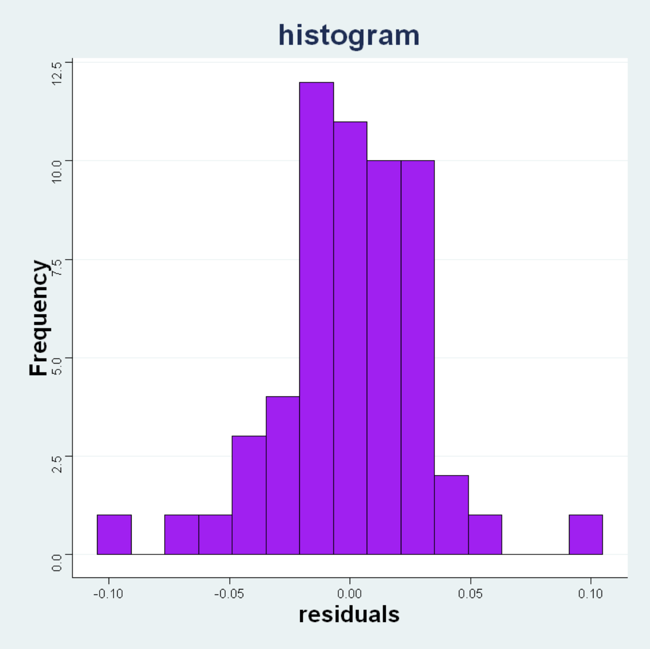
```
### normalirty1 ###

#upload data
ggplot(data=df, aes(sample=residuals)) +

# qq plot
stat_qq() +

# add line
geom_qq_line()+

scale_color_brewer(palette="Dark2")+

theme_stata() +

# lable names
labs( title='Q-Q Plot')+

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold'))  +
# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold'))
```



```
In [ ]:
```

In [160]:

```
### normalirty2 ###

#upload data
ggplot(data=df, aes(x=residuals))+

#histogram
geom_histogram(bins = 15, fill = 'purple', color = 'black')+

# lable names
labs(x='residuals', y='Frequency',title='histogram')+

theme_stata() +

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold'))
```



In [ ]:

In [38]:

```
### normalirty3 ###

shapiro.test(residuals)
```

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.9636, p-value = 0.08402
```

In [ ]:

```
```

In [39]:

```
#### independence assumption #####
dwtest(model, alternative=c("two.sided"))
```

```
	Durbin-Watson test

data:  model
DW = 1.6177, p-value = 0.1304
alternative hypothesis: true autocorrelation is not 0
```

In [ ]:

```
```

In [156]:

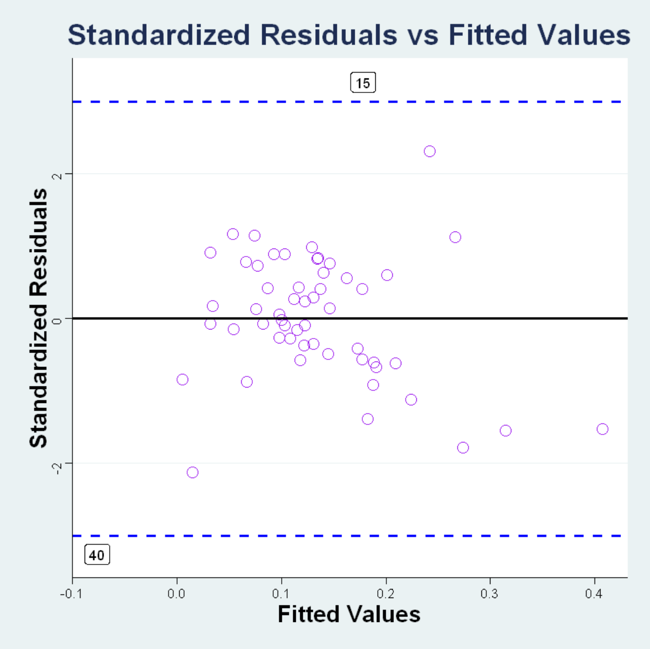
```
head(df,3)
```

A data.frame: 3 × 11

	observations	PE.x2.	CAS.x6.	LRSIZ.x11.	PT.y.	PT.yhat.	residuals	stand_residuals	jackknife_residuals	cooks_distance	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	1	0.17	28.0	0.2769	0.30	0.24214416	0.05785584	2.317952	2.422027	0.80144658	0.
2	2	0.06	0.4	0.2295	0.09	0.05377549	0.03622451	1.172688	1.176942	0.01459664	0.
3	3	0.07	0.5	0.1787	0.11	0.07437192	0.03562808	1.153159	1.156833	0.01398174	0.

In [29]:

```
##### standarize residuals #####

# need to be  > 3 std

#upload data
ggplot(data=df, aes(x=PT.yhat., y=stand_residuals)) +

theme_stata() +

#point settings
geom_point(shape=1,size=4,colour="purple") + # point settings

# lable names
labs(x='Fitted Values', y='Standardized Residuals', title='Standardized Residuals vs Fitted Values') +

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold')) +

# add hor line
geom_hline(yintercept=0, col = "black", size=1)+

# add hor line
geom_hline(yintercept=3, linetype='dashed',col = "blue", size=1)+

# add hor line
geom_hline(yintercept=-3,linetype='dashed' ,col = "blue", size=1)+

# add label to potential outliers under conditions
geom_label(data=df %>% filter(abs(stand_residuals)>3),aes(label=observations))


#new_df = df[order(stand_residuals),]
#tail(new_df,6)
```

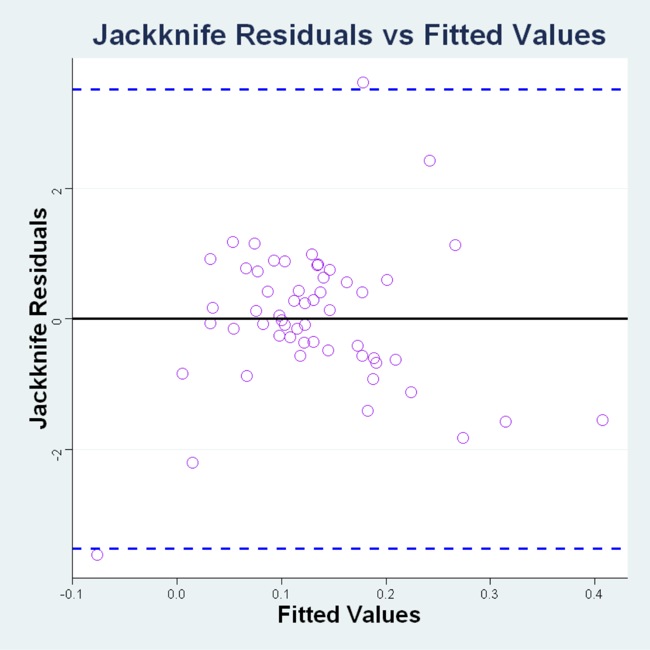

Standardized Residuals vs Fitted Values
Standardized Residuals
Fitted Values
15
40
2
0
-2
-0.1
0.0
0.1
0.2
0.3
0.4

```
In [ ]:
```

In [155]:

```
##### Jacknfive residual #####

#Bonferroni critical value
bonferroni=qt(.025/n,df=n-2,lower.tail = FALSE)


#upload data
ggplot(data=df, aes(x=PT.yhat., y=jackknife_residuals)) +

theme_stata() +

#point settings
geom_point(shape=1,size=4,colour="purple") + # point settings

# lable names
labs(x='Fitted Values', y='Jackknife Residuals', title='Jackknife Residuals vs Fitted Values') +

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold')) +

# add hor line
geom_hline(yintercept=0, col = "black", size=1)+

# add hor line bonferroni
geom_hline(yintercept=bonferroni, linetype='dashed',col = "blue", size=1)+

# add hor line bonferroni
geom_hline(yintercept=-bonferroni,linetype='dashed' ,col = "blue", size=1)+

# add label to potential outliers under conditions
geom_label( data=df %>% filter(abs(stand_residuals)>bonferroni),aes(label=observations))

bonferroni
```

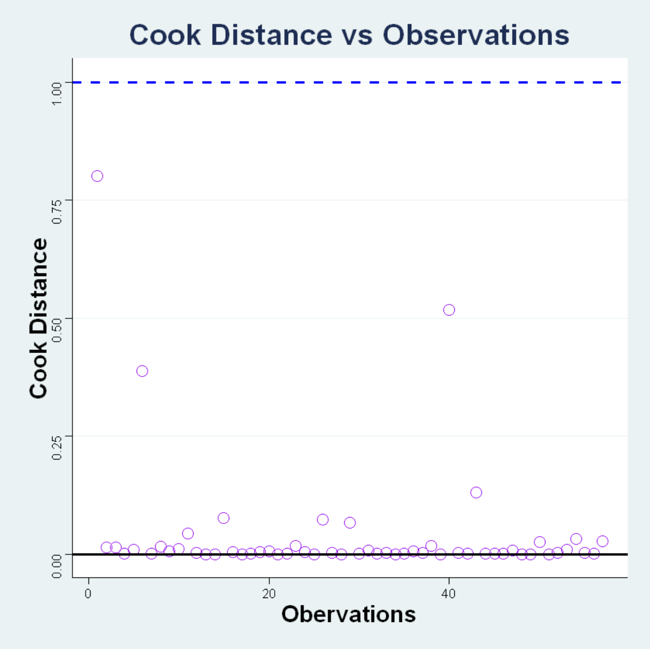
```
3.51919753815852
```


Jackknife Residuals vs Fitted Values
Jackknife Residuals
2
0
-2
-0.1
0.0
0.1
0.2
0.3
0.4
Fitted Values

```
In [ ]:
```

In [46]:

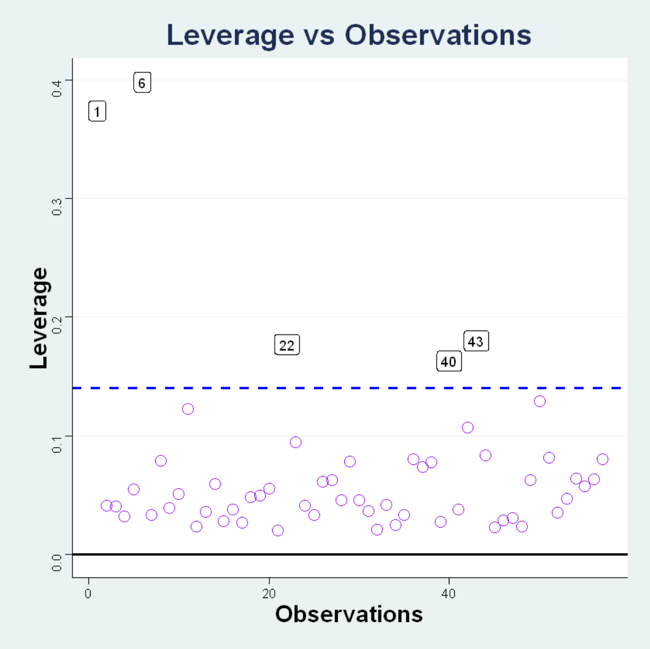
```
### cook distance ###

ggplot(data=df, aes(x=observations, y=cooks_distance)) +

#point settings
geom_point(shape=1,size=4,colour="purple") + # point settings

theme_stata() +

# lable names
labs(x='Obervations', y='Cook Distance', title='Cook Distance vs Observations') +

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold')) +

# add hor line
geom_hline(yintercept=0, col = "black", size=1)+

# add hor cook distance thresh hold
geom_hline(yintercept=1, linetype='dashed',col = "blue", size=1)+


# add label to potential outliers under conditions
geom_label( data=df %>% filter(cooks_distance>1),aes(label=observations))
```


Cook Distance vs Observations
Cook Distance
1.00
0.75
0.50
0.25
0.00
0
20
40
Obervations

```
In [ ]:
```

In [145]:

```
### leverage points graph ###

# Outlier = Leverage > 2*p/n, p = rank of X

X = model.matrix(model)
p = as.integer(rankMatrix(X))
c.thes = 2*p/n

#upload data
ggplot(data=df, aes(x=observations, y=leverage)) +

#point settings
geom_point(shape=1,size=4,colour="purple") + # point settings

theme_stata() +

# lable names
labs(x='Observations', y='Leverage', title='Leverage vs Observations') +

# title-label
theme(plot.title = element_text(hjust=.5, size=23, face='bold')) +

# x-label
theme(axis.title.x = element_text(hjust=.5, size=18, face='bold')) +

# y-label
theme(axis.title.y = element_text(hjust=.5, size=18, face='bold')) +

# add hor line
geom_hline(yintercept=0, col = "black", size=1)+

# add hor line for thres hold
geom_hline(yintercept=c.thes, linetype='dashed',col = "blue", size=1)+

# add label to potential outliers under conditions
geom_label( data=df %>% filter(leverage>c.thes),aes(label=observations))
```



```
In [146]: c.thes
```

```
0.140350877192982
```

In [ ]:

```
```

In [28]:

```
summary(model)
```

```

Call:
lm(formula = PT ~ PE + CAS + LRSIZ, data = firm_profit_df)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.094078 -0.017058 -0.000543  0.019753  0.101697 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.020178   0.018659   1.081   0.2844    
PE           1.184759   0.076206  15.547   <2e-16 ***
CAS          0.002391   0.000902   2.651   0.0106 *  
LRSIZ       -0.167515   0.075489  -2.219   0.0308 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03154 on 53 degrees of freedom
Multiple R-squared:  0.8679,	Adjusted R-squared:  0.8604 
F-statistic: 116.1 on 3 and 53 DF,  p-value: < 2.2e-16
```

In [7]:

```
```

```
Error in eval(expr, envir, enclos): object 'firm_profit_df' not found
Traceback:
```

In [ ]:

```
```

In [84]:

```
# before
summary(model)
```

```
Call:
lm(formula = PT ~ PE + CAS + LRSIZ, data = firm_profit_df)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.094078 -0.017058 -0.000543  0.019753  0.101697 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.020178   0.018659   1.081   0.2844    
PE           1.184759   0.076206  15.547   <2e-16 ***
CAS          0.002391   0.000902   2.651   0.0106 *  
LRSIZ       -0.167515   0.075489  -2.219   0.0308 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03154 on 53 degrees of freedom
Multiple R-squared:  0.8679,	Adjusted R-squared:  0.8604 
F-statistic: 116.1 on 3 and 53 DF,  p-value: < 2.2e-16
```

In [ ]:

```

```

In [24]:

```
### confindence int for population ###1

# Confidence interval
newdata=data.frame(PE=.20,CAS=4,LRSIZ=.20)

predict(model,newdata,inteval='confidence',interval="confidence",level=.95)

#we are 95% confindent that the expected _____ is between lb and hb when x2 is __, x6 is ___, and x11 is ___
```

```
A matrix: 1 × 3 of type dbl

        fit       lwr       upr
1 0.2331902 0.2167559 0.2496245
```

In [40]:

```
### confindence int for population ###2

# Confidence interval
newdata=data.frame(PE=.50,CAS=30,LRSIZ=.50)

predict(model,newdata,inteval='confidence',interval="confidence",level=.95)

#we are 95% confindent that the expected _____ is between lb and hb when x2 is __, x6 is ___, and x11 is ___
```

```
A matrix: 1 × 3 of type dbl

        fit       lwr       upr
1 0.6005235 0.5292747 0.6717723
```

In [41]:

```
### confindence int for population ###3

# Confidence interval
newdata=data.frame(PE=.75,CAS=20,LRSIZ=.45)

predict(model,newdata,inteval='confidence',interval="confidence",level=.95)

#we are 95% confindent that the expected _____ is between lb and hb when x2 is __, x6 is ___, and x11 is ___
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	0.8811811	0.782076	0.9802862

In [ ]:

In [27]:

```
### prediction int for an indivual y-value ###1

# prediction interval
newdata=data.frame(PE=.2,CAS=4,LRSIZ=.20)

predict(model,newdata,inteval='prediction',interval="prediction",level=.95)

#we are 95% confindent that the y _____ is between lb and ub when x2 is __, x6 is ___, and x11 is ____.
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	0.2331902	0.1678309	0.2985495

In [42]:

```
### prediction int for an indivual y-value ###2

# prediction interval
newdata=data.frame(PE=.5,CAS=30,LRSIZ=.50)

predict(model,newdata,inteval='prediction',interval="prediction",level=.95)

#we are 95% confindent that the y _____ is between lb and ub when x2 is __, x6 is ___, and x11 is ____.
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	0.6005235	0.5052442	0.6958027

In [43]:

```
### prediction int for an indivual y-value ###3

# prediction interval
newdata=data.frame(PE=.75,CAS=20,LRSIZ=.45)

predict(model,newdata,inteval='prediction',interval="prediction",level=.95)

#we are 95% confindent that the y _____ is between lb and ub when x2 is __, x6 is ___, and x11 is ____.
```

A matrix: 1 × 3 of type dbl

	fit	lwr	upr
1	0.8811811	0.7636075	0.9987548

In [ ]: