Assignment 1 for Statistical Modelling III

Abdul Malik Muhammad a1783655

Q1)

Using L'Hopital's Rule

$$\lim_{x\to\alpha}\frac{f(x)}{g(x)} = \lim_{x\to\alpha}\frac{f'(x)}{g'(x)}$$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = \lim_{\lambda\to 0}\frac{\frac{d}{d\lambda}(y^\lambda - 1)}{\frac{d}{d\lambda}(\lambda)}$$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = \lim_{\lambda\to 0}\frac{d}{d\lambda}(y^\lambda)$$

Using the exponent rule

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = \lim_{\lambda\to 0}\frac{d}{d\lambda}(e^{log(y^\lambda)})$$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = \lim_{\lambda\to 0}\frac{d}{d\lambda}(e^{\lambda log(y)})$$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = \lim_{\lambda\to 0} log(y)(e^{\lambda log(y)})$$

Let $\lambda = 0$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = log(y)(e^{(0)log(y)})$$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = log(y)(e^0)$$

$$\lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = log(y)(1)$$

$$\therefore \lim_{\lambda\to 0}\frac{y^\lambda - 1}{\lambda} = log(y)$$

Q2) First we want to load in the tidyverse, gglm & skimr libraries and set the working directory using the library() and setwd() functions.

Q2a) Read the data in as shown below

```
companies <- read.delim("companies.txt")
```

Q2b)

```
skimr::skim_without_charts(companies)
```

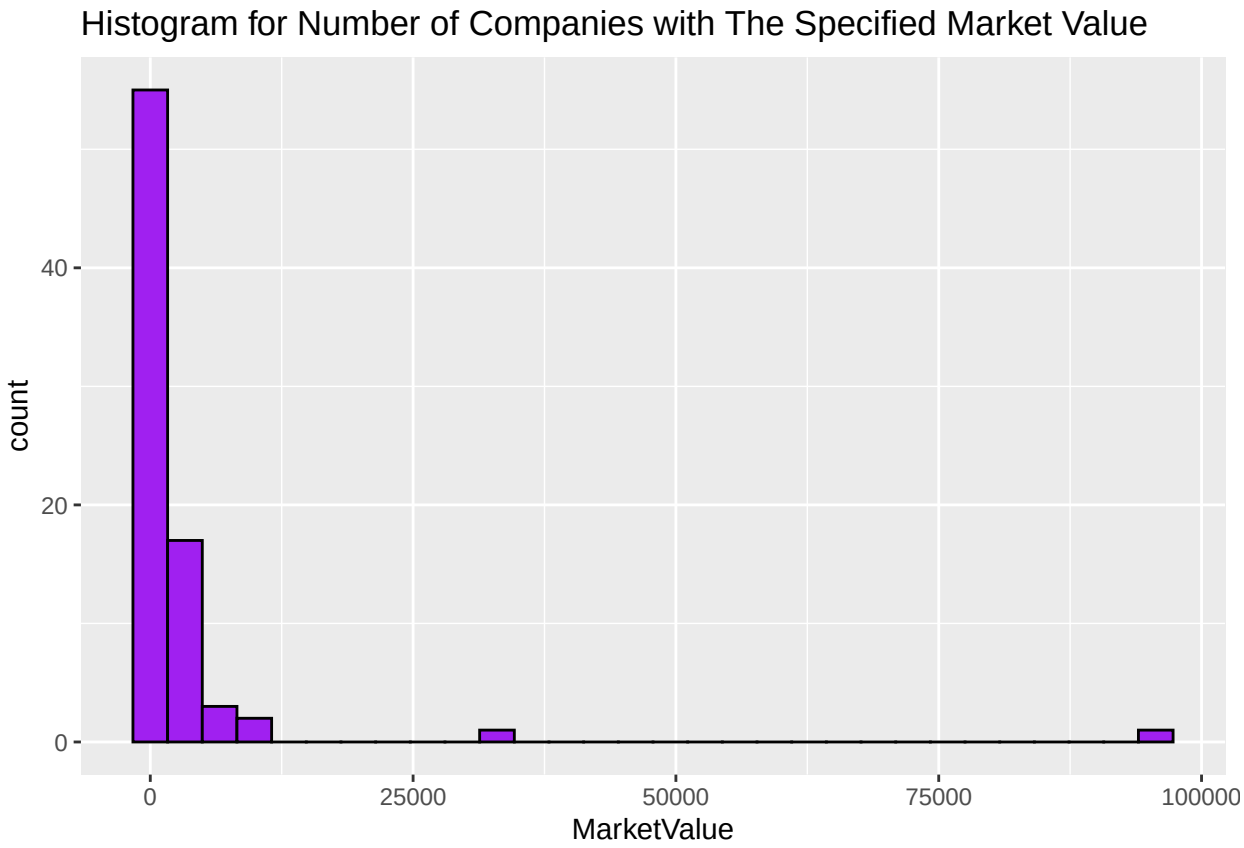
Table 1: Data summary

Name	companies
Number of rows	79
Number of columns	6
Column type frequency:	
numeric	6
Group variables	None

Variable type: numeric

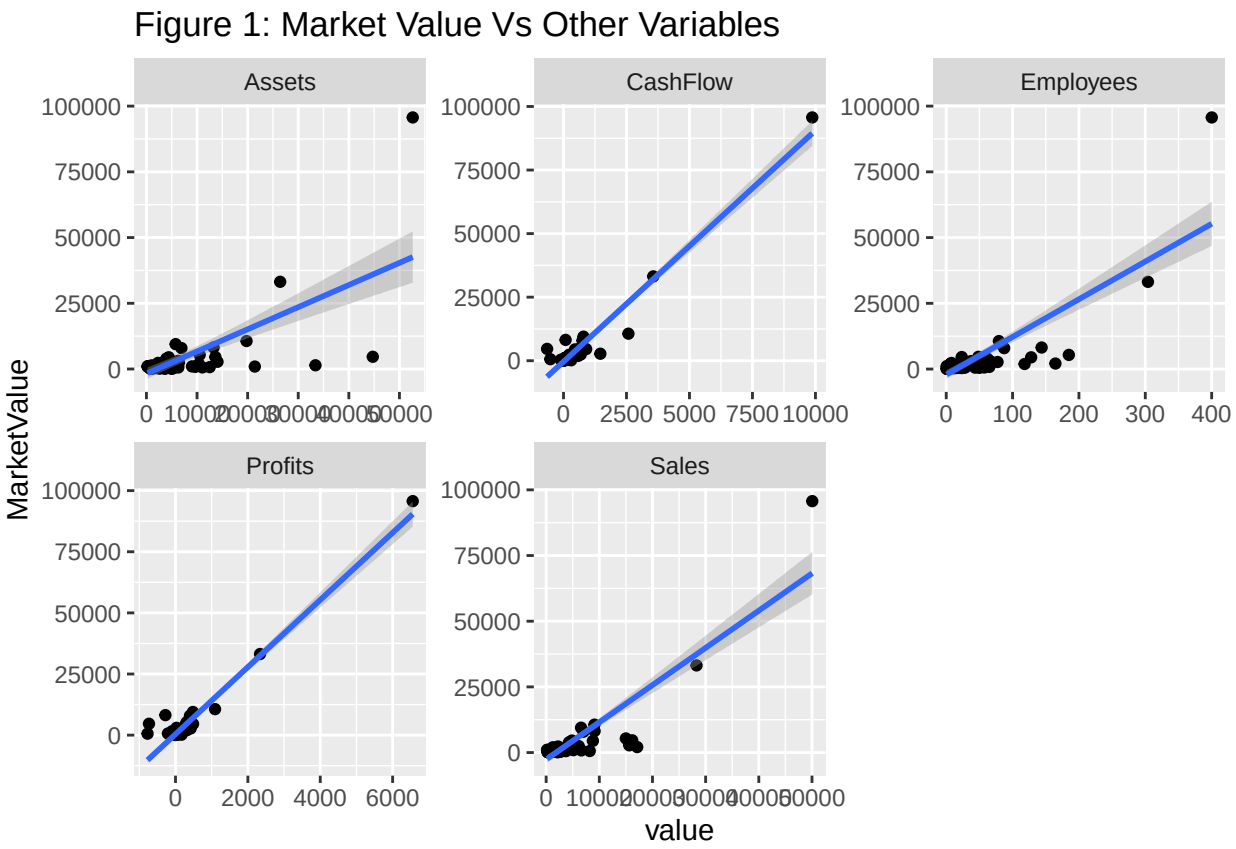
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Assets	0	1	5940.53	9156.78	223.0	1122.50	2788.0	5802.00	52634.0
Sales	0	1	4178.29	7011.63	176.0	815.50	1754.0	4563.50	50056.0
MarketValue	0	1	3269.75	11303.55	53.0	512.50	944.0	1961.50	95697.0
Profits	0	1	209.84	796.98	-771.5	39.00	70.5	188.05	6555.0
CashFlow	0	1	400.93	1205.53	-651.9	75.15	133.3	328.85	9874.0
Employees	0	1	37.60	64.50	0.6	3.95	15.4	48.50	400.2

```
companies%>%
  ggplot(aes(MarketValue))+
  geom_histogram(col = "black", fill = "purple")+
  ggtitle("Histogram for Number of Companies with The Specified Market Value")
```


Histogram for Number of Companies with The Specified Market Value
count
40
20
0
0
25000
50000
75000
100000
MarketValue

Q2c)

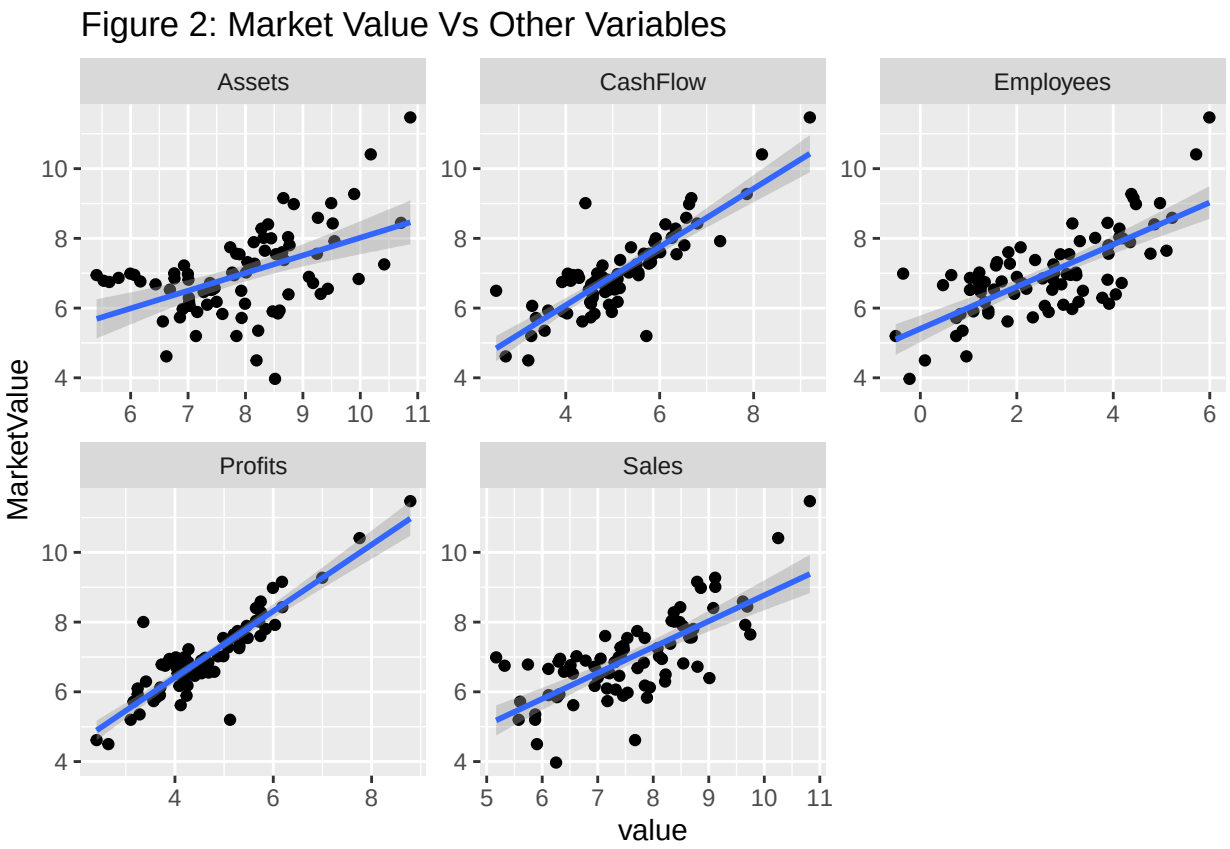
```
companies%>%
  select(Assets, Sales, Profits, CashFlow, Employees, MarketValue) %>%
  pivot_longer(-MarketValue) %>%
  ggplot(aes(value, MarketValue))+
  geom_point()+
   geom_smooth(method= lm) +
  facet_wrap(~name, scales = "free")+
  ggtitle("Figure 1: Market Value Vs Other Variables")
```



As shown above in Figure 1, all the variables do not look linear, however CashFlow and Profits may be linear.

Q2d)

```
log(companies)%>%
  select(Assets, Sales, Profits, CashFlow, Employees, MarketValue) %>%
  pivot_longer(-MarketValue) %>%
  ggplot(aes(value, MarketValue))+
  geom_point()+
   geom_smooth(method= lm) +
  facet_wrap(~name, scales = "free")+
  ggtitle("Figure 2: Market Value Vs Other Variables")
```



As shown above in Figure 2, all the variables look pretty linear.

Q2e)

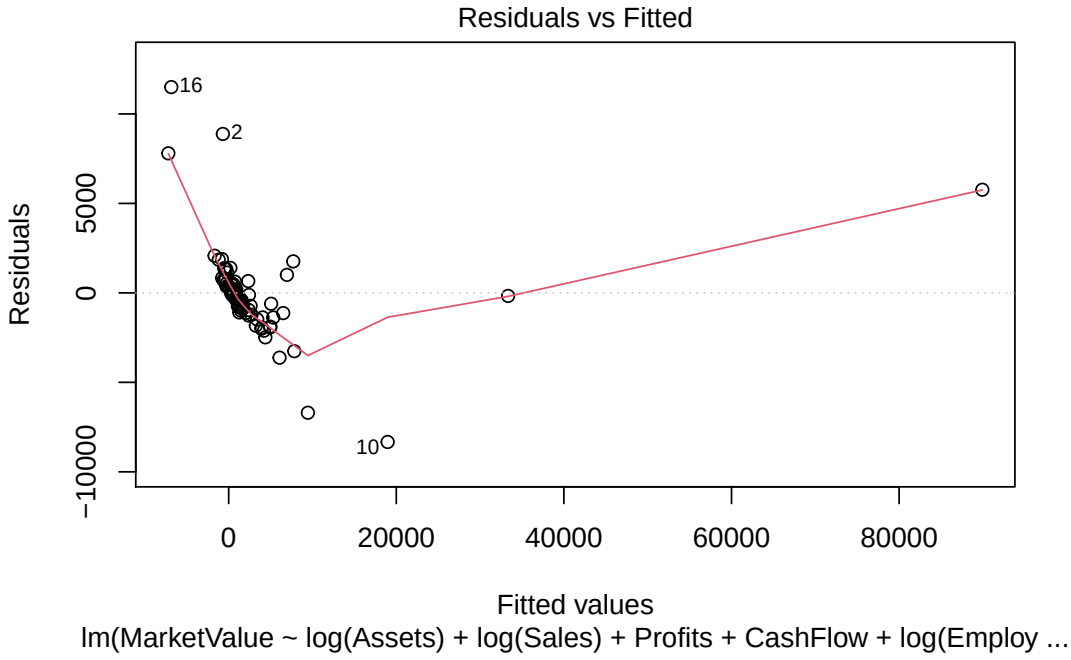
```
M1 <- lm(MarketValue~log(Assets)+log(Sales)+Profits+CashFlow+log(Employees),
         data = companies)
```

Q2f) We log-transformed some of the variables and not others due to the NaN's produced. These values occur due to some values being less than 0, and it is mathematically not possible to log a negative number. Therefore we did not log transorm all the variables. Also another reason is, as stated in Q2c) "in Figure 1, all the variables do not look linear, however CashFlow and Profits may be linear." So there is no need to transform a variable that already looks linear.

Q2g) Assumption checking

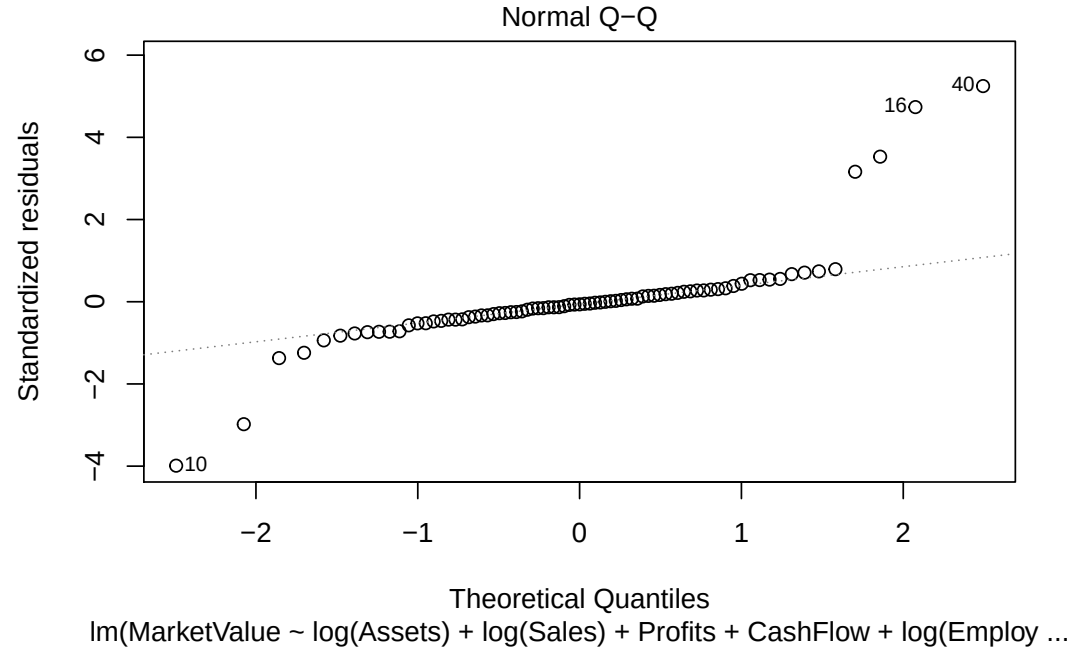
```
plot(M1, which = 1, main = "Figure 3: Residuals vs Fitted for Market Value")
```

Figure 3: Residuals vs Fitted for Market Value



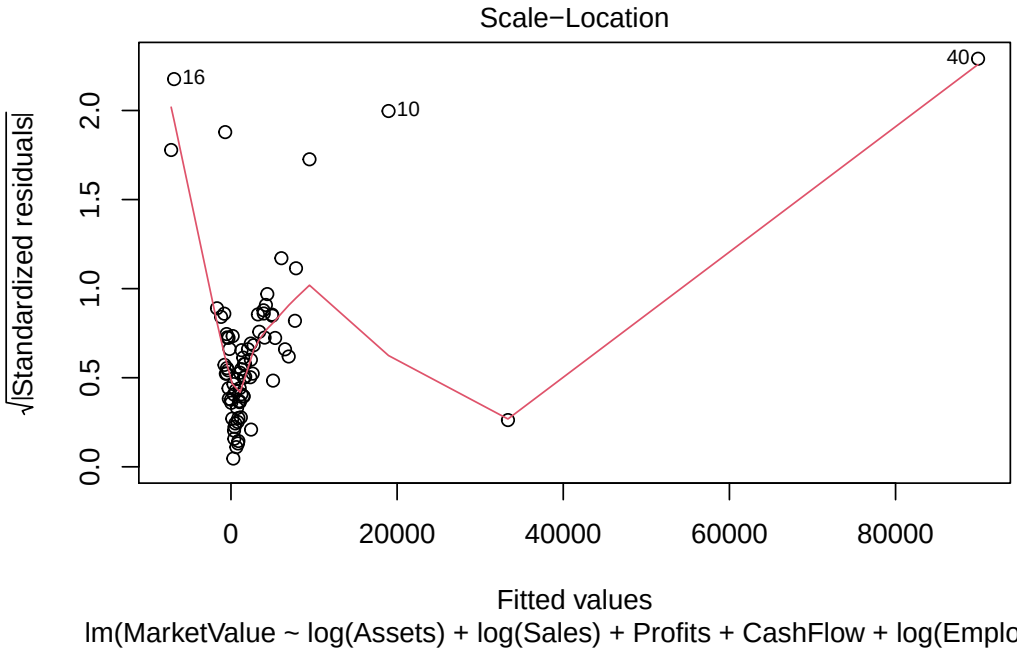

```
plot(M1, which = 2, main = "Figure 4: Normal-QQ Plot for Market Value")
```

Figure 4: Normal−QQ Plot for Market Value



```
plot(M1, which = 3, main = "Figure 5: Scale-Location Plot for Market Value")
```

Figure 5: Scale−Location Plot for Market Value



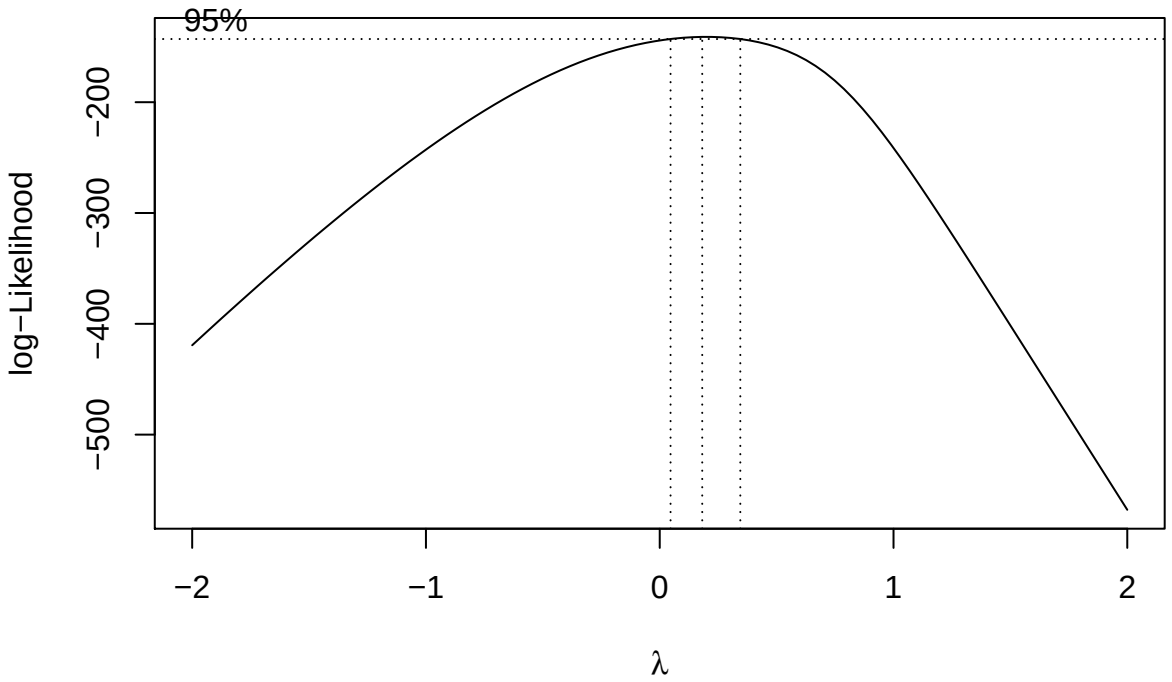
Linearity: Does NOT satisfy because the Residuals vs Fitted Plot (Figure 3) does not have a straight line of best fit, it looks like the line slopes down then up sharply. This may be due to the outlier at the top right of Residuals vs Fitted Plot.

Normality: Does NOT satisfy because, in the Normal QQ-Plot(Figure 4), the tails to the right and left of the plot are way off the dotted line.

Homoscedasticity: Does NOT satisfy because there is not random, vertically equal scatter about the middle of the plot on the Scale-Location Plot(Figure 5). Instead of random scatter, the data looks to be all bunched up towards the left of the Scale-Location Plot(Figure 5).

Q2h)

```
boxcox<-MASS::boxcox(M1)
```



λ cannot be 0, this is because the 95% confidence interval does does include zero, hence λ should be the mean/middle on the interval. For this we need to calculate the maximum log-likelihood of the plot shown above. To do this, we use the code below

```
boxcox$x[which.max(boxcox$y)]
```

```
## [1] 0.1818182
```

Therefore the my chosen value of λ is 0.18182

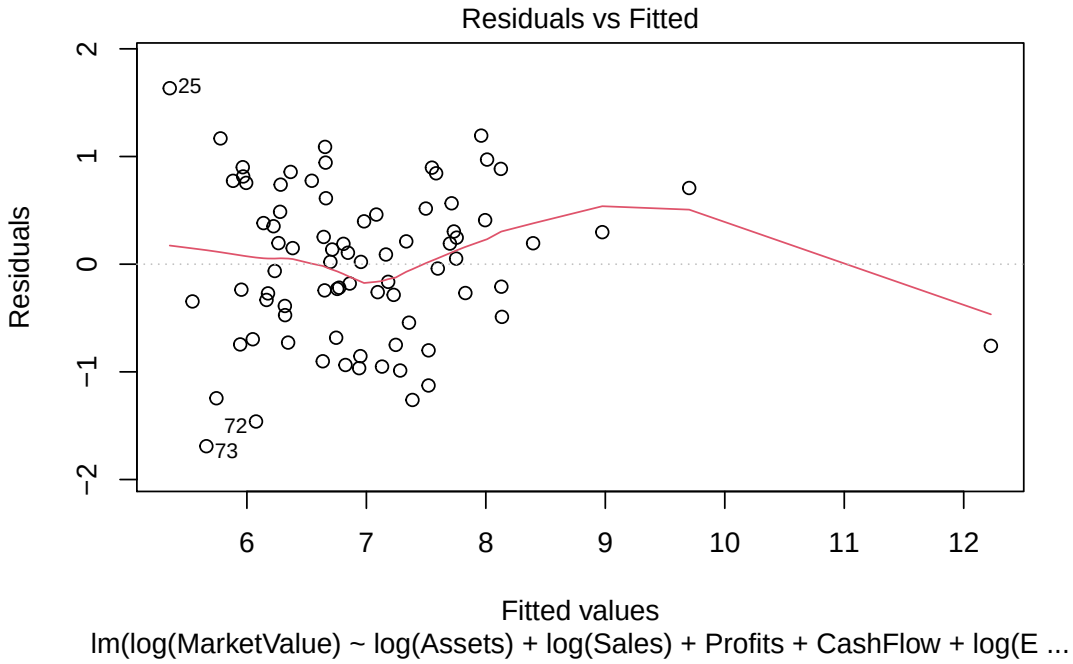
Q2i)

```
M2 <- lm(log(MarketValue)~log(Assets)+log(Sales)+Profits+CashFlow+log(Employees),
         data = companies)
```

Q2j)

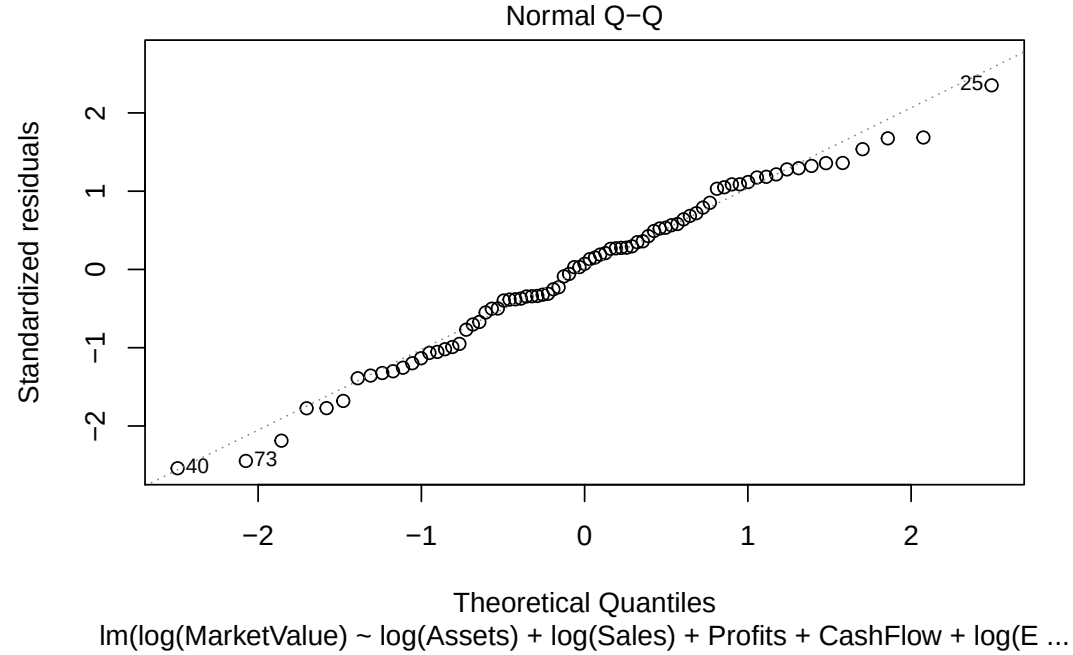
```
plot(M2, which = 1, main = "Figure 6: Residuals vs Fitted for Market Value")
```

Figure 6: Residuals vs Fitted for Market Value



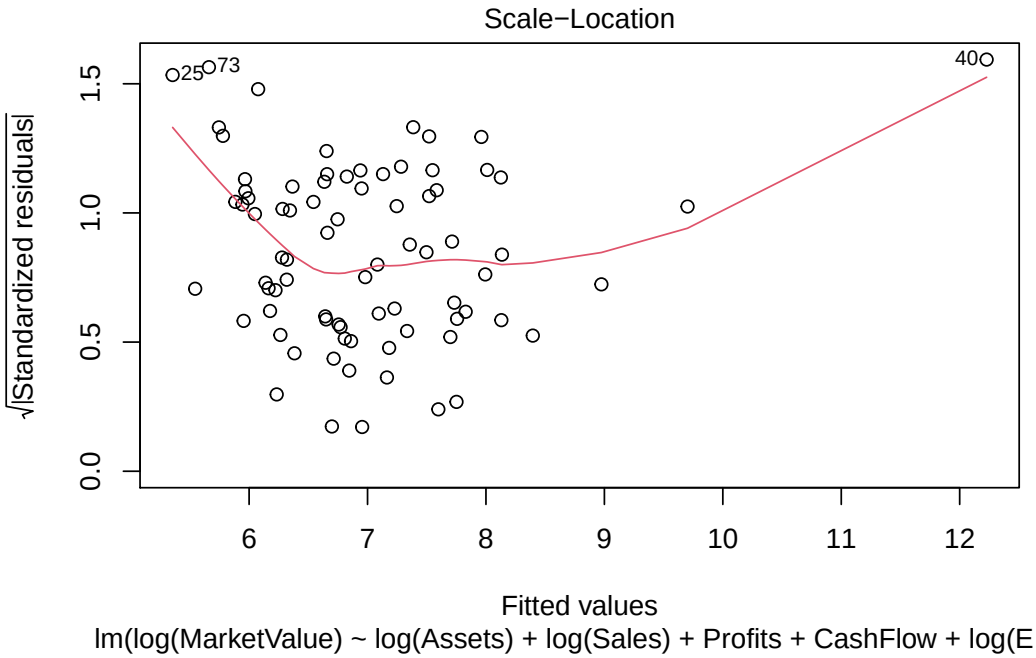

```
plot(M2, which = 2, main = "Figure 7: Normal-QQ Plot for Market Value")
```

Figure 7: Normal−QQ Plot for Market Value



```
plot(M2, which = 3, main = "Figure 8: Scale-Location Plot for Market Value")
```

Figure 8: Scale−Location Plot for Market Value



Linearity: Does satisfy linearity because the Residuals vs Fitted Plot(Figure 6) has a fairly straight line of best fit and all the data is scattered randomly about the zero line.

Normality: Does satisfy Normality because, in the Normal QQ-Plot(Figure 7), most of the data lies on the dotted line.

Homoscedasticity: Does satisfy Homoscedasticity because, there is random, vertically equal scatter about the middle of the plot on the Scale-Location Plot(Figure 8).

Q2k) M2 is the better choice because it satisfies the assumptions where as M1 does not.

Q2l) Using the code below we can caluculate the 95% prediction interval.

```
newdata = data.frame(Assets = 1065, Sales = 642, Profits = 30, CashFlow = 59,
                     Employees = 3.5)
predict(M2, newdata, interval='predict')
```

```
##        fit      lwr      upr
## 1 6.155464 4.687227 7.623701
```

Hence the 95% prediction interval is:

$$4.687 \leq log(MarketValue) \leq 7.624$$

$$e^{4.687} \leq MarketValue \leq e^{7.624}$$

$$108.5271 \leq MarketValue \leq 2046.733$$

Where the numbers are in millions of dollars.