

1st Question

$$\lim_{\lambda \rightarrow 0} \frac{y^\lambda - 1}{\lambda} = \log y$$

$\nearrow f(\lambda)$
 $\searrow g(\lambda)$

since

$$\lim_{\lambda \rightarrow 0} f(\lambda) = \lim_{\lambda \rightarrow 0} (y^\lambda - 1) = 1 - 1 = 0$$

and

$$\lim_{\lambda \rightarrow 0} g(\lambda) = \lim_{\lambda \rightarrow 0} (\lambda) = 0.$$

so we can use L'Hopital rule

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \frac{f(\lambda)}{g(\lambda)} &= \lim_{\lambda \rightarrow 0} \frac{f'(\lambda)}{g'(\lambda)} = \lim_{\lambda \rightarrow 0} \frac{y^\lambda \log(y)}{1} \\ &= \frac{1 \cdot \log(y)}{1} = \log(y) \end{aligned}$$

Assignment_2

Mustafa Mohammadi | a1838795

06/04/2021

```
pacman::p_load(tidyverse, ggglm)
```

Section A

Reading data into R.

```
companies <- read.table("companies.txt", header = TRUE)  
companies
```

##	Assets	Sales	MarketValue	Profits	CashFlow	Employees
## 1	2687	1870	1890	145.7	352.2	18.2
## 2	13271	9115	8190	-279.0	83.0	143.8
## 3	13621	4848	4572	485.0	898.9	23.4
## 4	3614	367	90	14.1	24.6	1.1
## 5	6425	6131	2448	345.8	682.5	49.5
## 6	1022	1754	1370	72.0	119.5	4.8
## 7	1093	1679	1070	100.9	164.5	20.8
## 8	1529	1295	444	25.6	137.0	19.4
## 9	2788	271	304	23.5	28.9	2.1
## 10	19788	9084	10636	1092.9	2576.8	79.4
## 11	327	542	959	54.1	72.5	2.8
## 12	1117	1038	478	59.7	91.7	3.8
## 13	5401	550	376	25.6	37.5	4.1
## 14	1128	1516	430	-47.0	26.7	13.2
## 15	1633	701	679	74.3	135.9	2.8
## 16	44736	16197	4653	-732.5	-651.9	48.5
## 17	5651	1254	2002	310.7	407.9	6.2
## 18	5835	4053	1601	-93.8	173.8	10.8
## 19	278	205	853	44.8	50.5	3.8
## 20	5074	2557	1892	239.9	578.3	21.9
## 21	866	1487	944	71.7	115.4	12.6
## 22	4418	8793	4459	283.6	456.5	128.0
## 23	6914	7029	7957	400.6	754.7	87.3
## 24	862	1601	1093	66.9	106.8	16.0
## 25	401	176	1084	55.6	57.0	0.7
## 26	430	1155	1045	55.7	70.8	22.5
## 27	799	1140	683	57.6	89.2	15.4
## 28	4789	453	367	40.2	51.4	3.0
## 29	2548	264	181	22.2	26.2	2.1
## 30	5249	527	346	37.8	56.2	4.1
## 31	3494	1653	1442	160.9	320.3	6.4
## 32	1804	2564	483	70.5	164.9	26.6
## 33	26432	28285	33172	2336.0	3562.0	304.0
## 34	623	2247	797	57.0	93.8	18.6
## 35	1608	6615	829	56.1	134.0	65.0
## 36	4662	4781	2988	28.7	371.5	66.2
## 37	5769	6571	9462	482.0	792.0	83.0
## 38	6259	4152	3090	283.7	524.5	62.0
## 39	1654	451	779	84.8	130.4	1.6
## 40	52634	50056	95697	6555.0	9874.0	400.2
## 41	999	1878	393	-173.5	-108.1	23.3
## 42	1679	1354	687	93.8	154.6	4.6
## 43	4178	17124	2091	180.8	390.4	164.6
## 44	223	557	1040	60.6	63.7	1.9
## 45	6307	8199	598	-771.5	-524.3	57.5
## 46	3720	356	211	26.6	34.8	2.4
## 47	3442	5080	2673	235.4	361.5	77.3
## 48	33406	3222	1413	201.7	246.7	15.8
## 49	1257	355	181	167.5	304.0	0.6
## 50	1743	597	717	121.6	172.4	3.5
## 51	12505	1302	702	108.4	131.4	9.0
## 52	3940	4317	3940	315.2	566.3	62.0
## 53	8998	882	988	93.0	119.0	7.4
## 54	21419	2516	930	107.6	164.7	15.6
## 55	2366	3305	1117	131.2	256.5	25.2
## 56	2448	3484	1036	48.8	257.1	25.4
## 57	1440	1617	639	81.7	126.4	3.5
## 58	14045	15636	2754	418.0	1462.0	27.3
## 59	4084	4346	3023	302.7	521.7	37.5
## 60	3010	749	1120	146.3	209.2	3.4
## 61	1286	1734	361	69.2	145.7	14.3
## 62	707	706	275	61.4	77.8	6.1

```
## 63 3086 1739 1507 202.7 335.2 4.9
## 64 252 312 883 41.7 60.6 3.3
## 65 11052 1097 606 64.9 97.6 7.0
## 66 9672 1037 829 92.6 118.2 8.2
## 67 1112 3689 542 30.3 96.9 43.5
## 68 1104 5123 910 63.7 133.3 48.5
## 69 478 672 866 67.1 101.6 5.4
## 70 10348 5721 1915 223.6 322.5 49.5
## 71 2769 3725 663 -208.4 12.4 29.1
## 72 752 2149 101 11.1 15.2 2.6
## 73 4989 518 53 -3.1 -0.3 0.8
## 74 10528 14992 5377 312.7 710.7 184.8
## 75 1995 2662 341 34.7 100.7 2.3
## 76 2286 2235 2306 195.3 219.0 8.0
## 77 952 1307 309 35.4 92.8 10.3
## 78 2957 2806 457 40.6 93.5 50.0
## 79 2535 5958 1921 177.0 288.0 118.1
```

Section B







EDA: Perform an EDA of the data - skim() and histogram of response is fine.

```
skimr::skim(companies)
```

Data summary

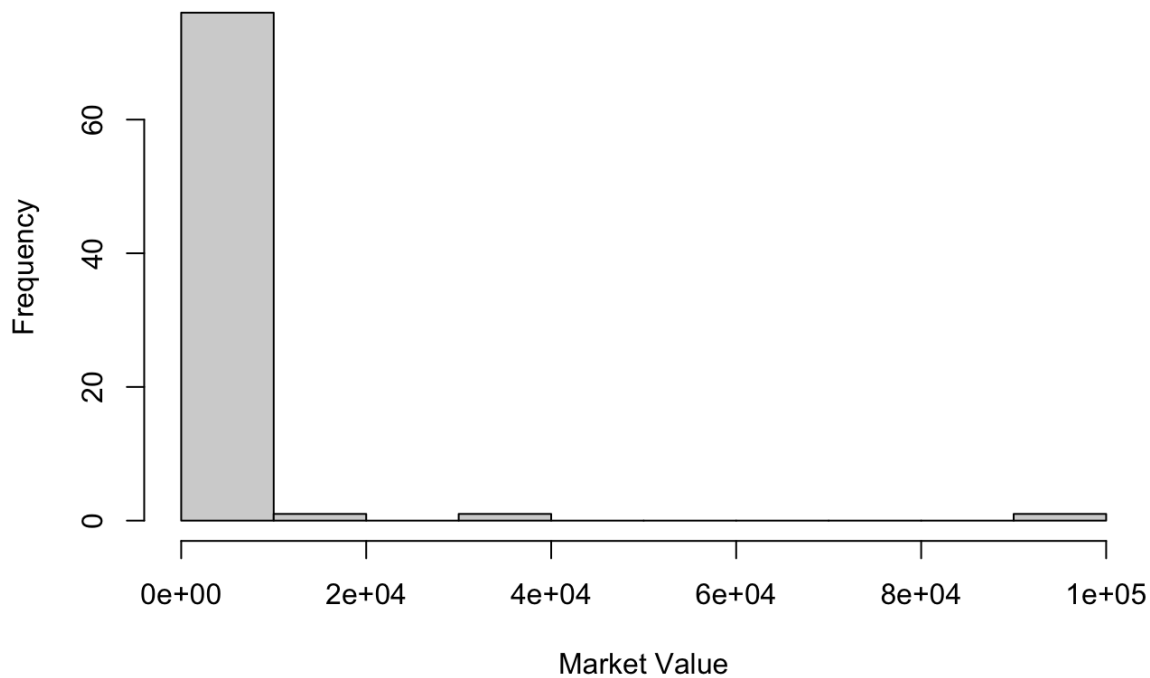
Name	companies
Number of rows	79
Number of columns	6
Column type frequency:	
numeric	6
Group variables	
None	

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

```
hist(companies$MarketValue, main = "Histogram of Market Values", xlab = "Market Value")
```

Histogram of Market Values

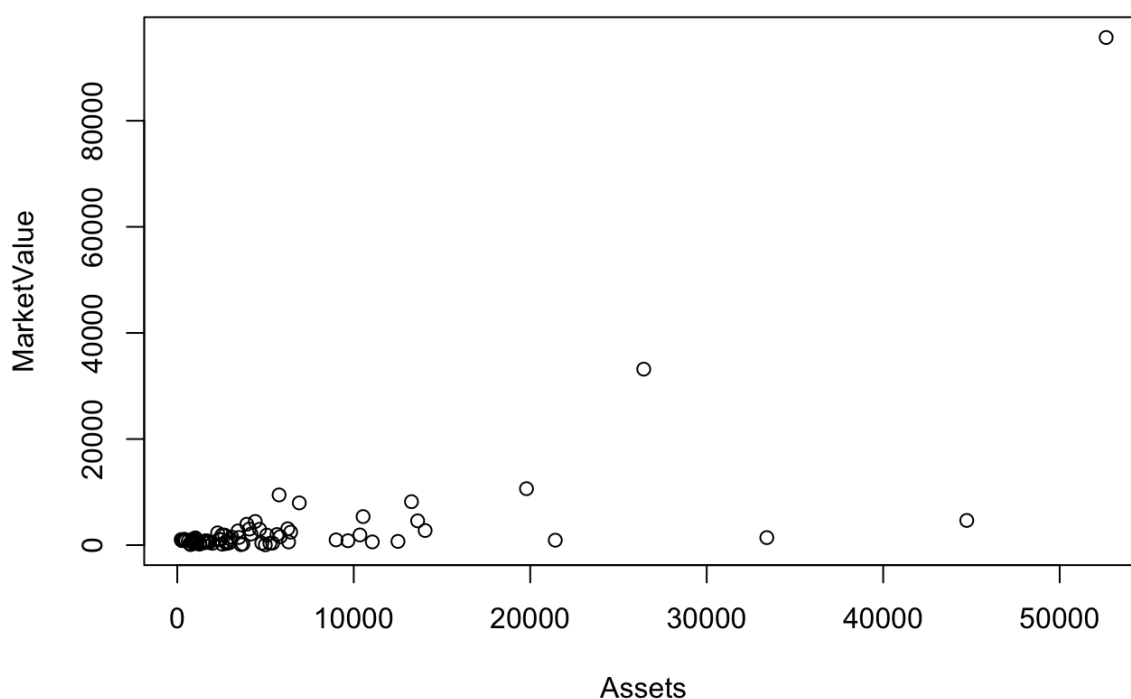


Section C

Produce scatterplots of MarketValue against each of the other predictors.

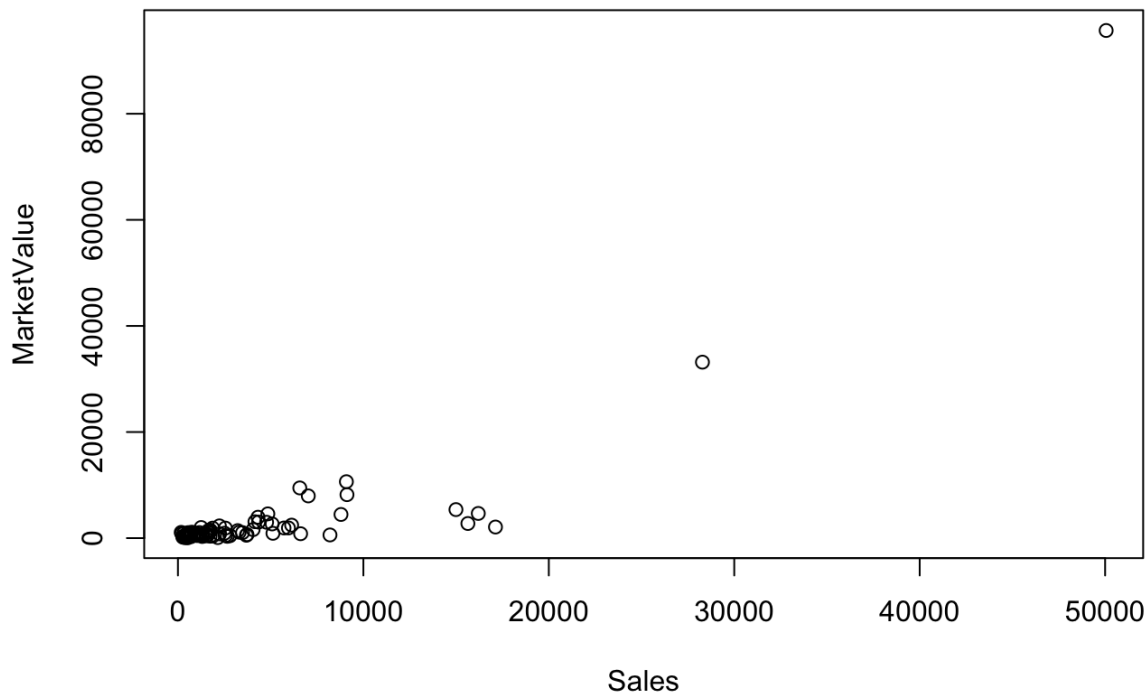
```
plot(MarketValue ~ Assets, data = companies, main = "Market Value vs Assets")
```

Market Value vs Assets



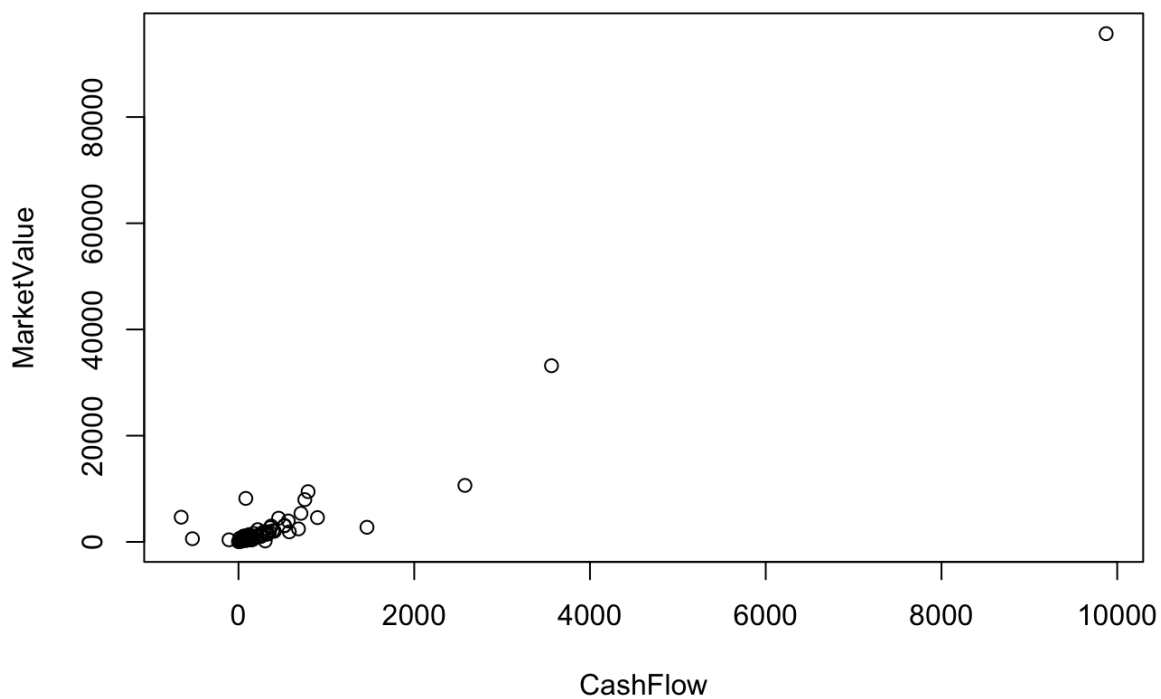
```
plot(MarketValue ~ Sales, data = companies, main = "Market Value vs Sales")
```

Market Value vs Sales



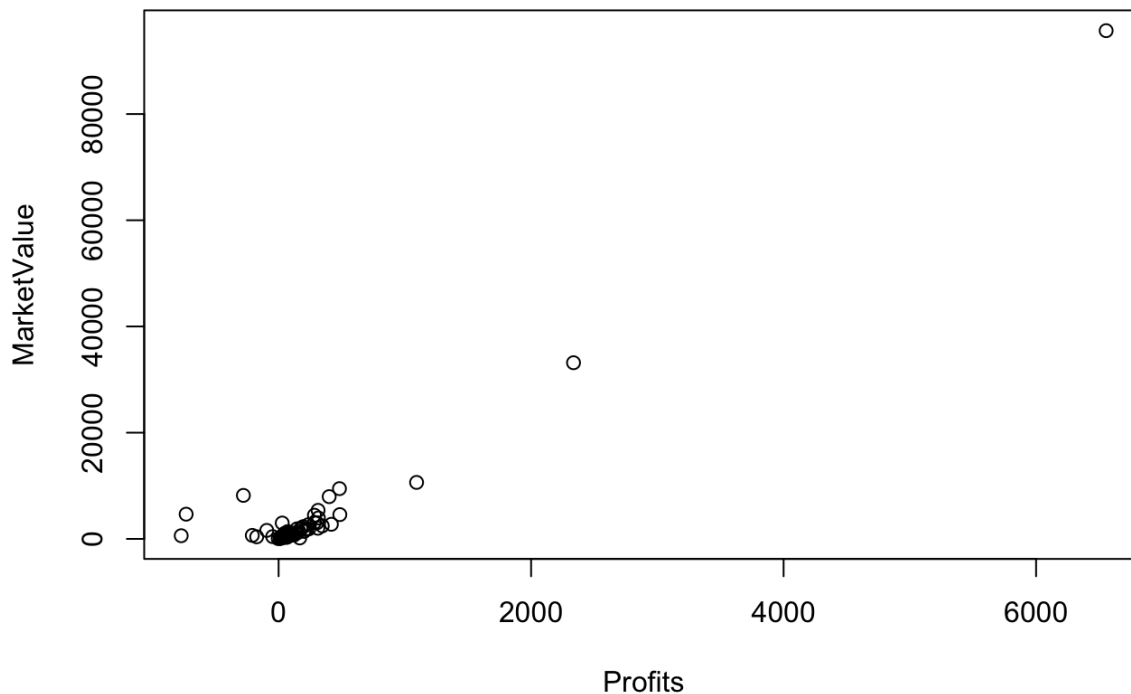
```
plot(MarketValue ~ CashFlow, data = companies, main = "Market Value vs CashFlow")
```

Market Value vs CashFlow



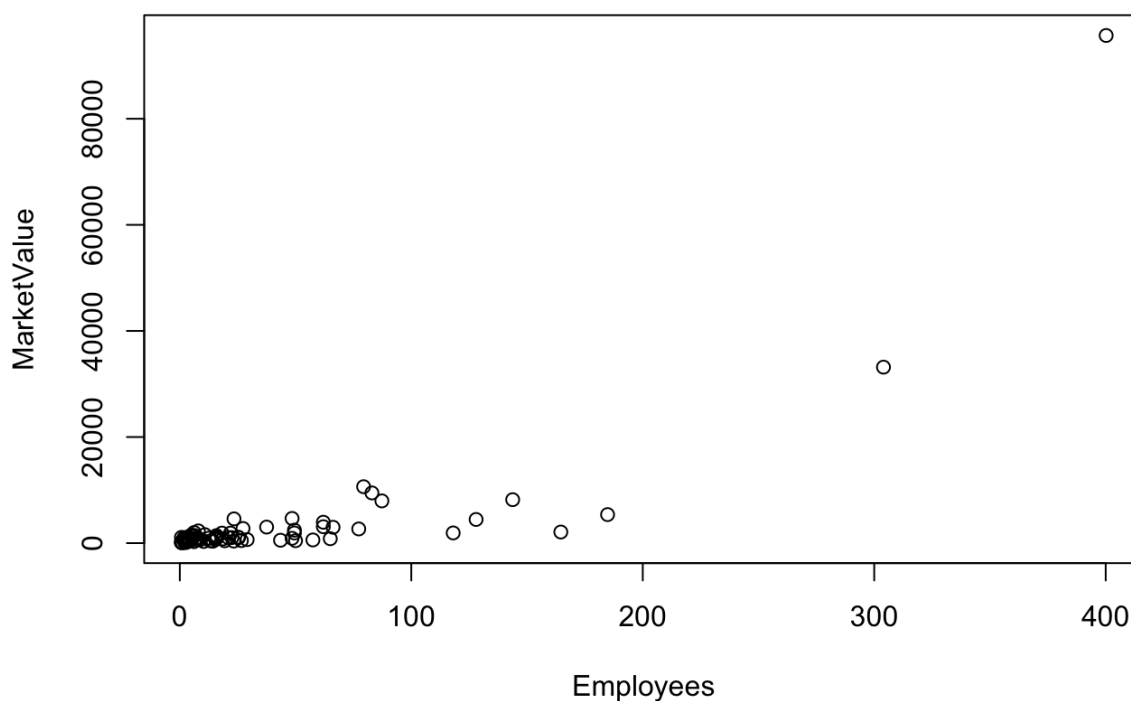
```
plot(MarketValue ~ Profits, data = companies, main = "Market Value vs Profits")
```

Market Value vs Profits



```
plot(MarketValue ~ Employees, data = companies, main = "Market Value vs Employees")
```

Market Value vs Employees

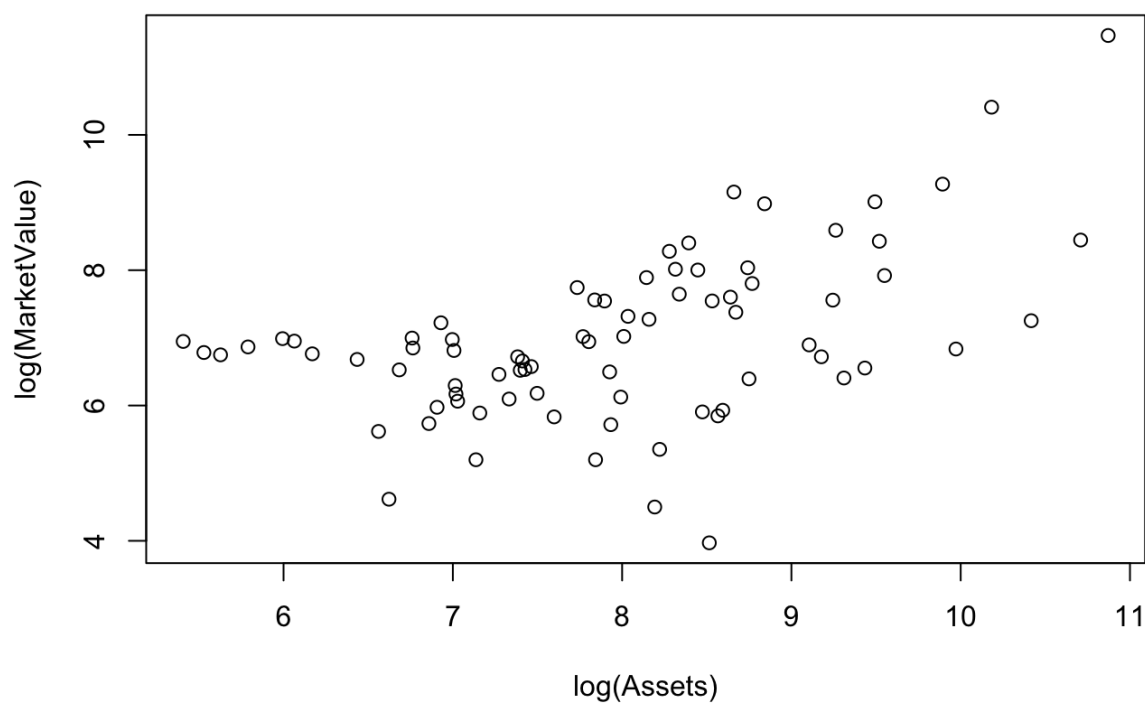


Section D

Produce scatter plots of Market Value against each of the other predictors with both the y-axis and x-axis on a log scale.

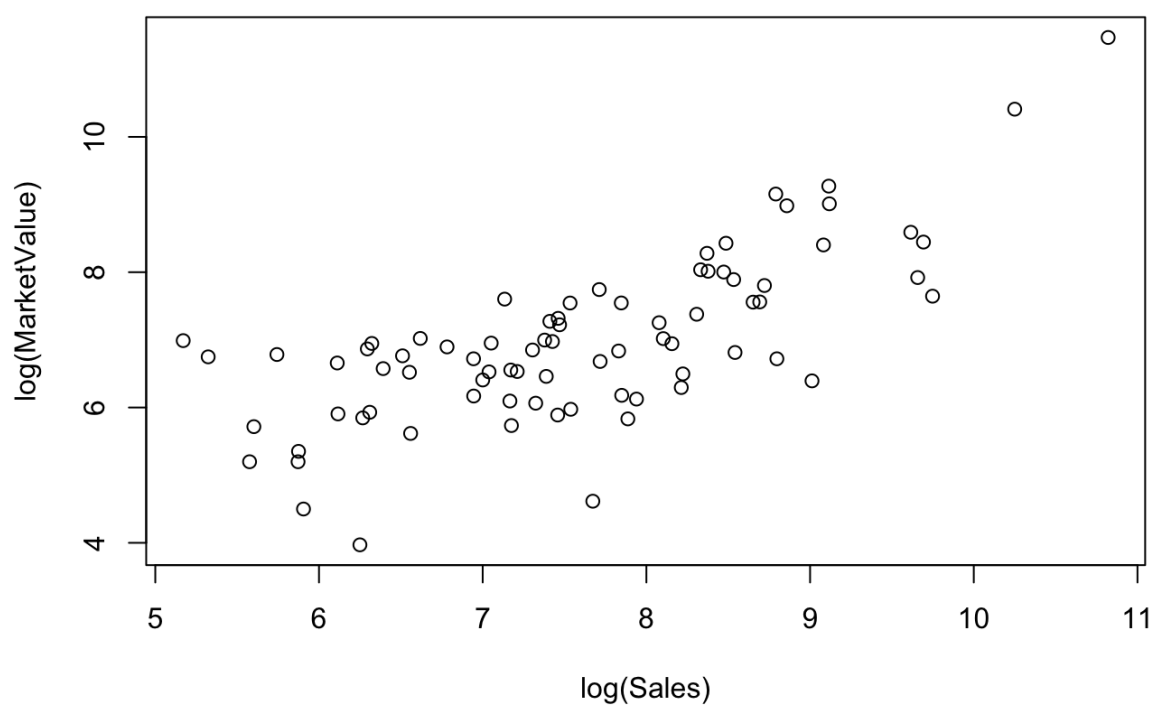
```
plot(log(MarketValue) ~ log(Assets), data = companies, main = "Log(Market Value) vs Log(Assets)")
```

Log(Market Value) vs Log(Assets)



```
plot(log(MarketValue) ~ log(Sales), data = companies, main = "Log(Market Value) vs Log(Sales)")
```

Log(Market Value) vs Log(Sales)

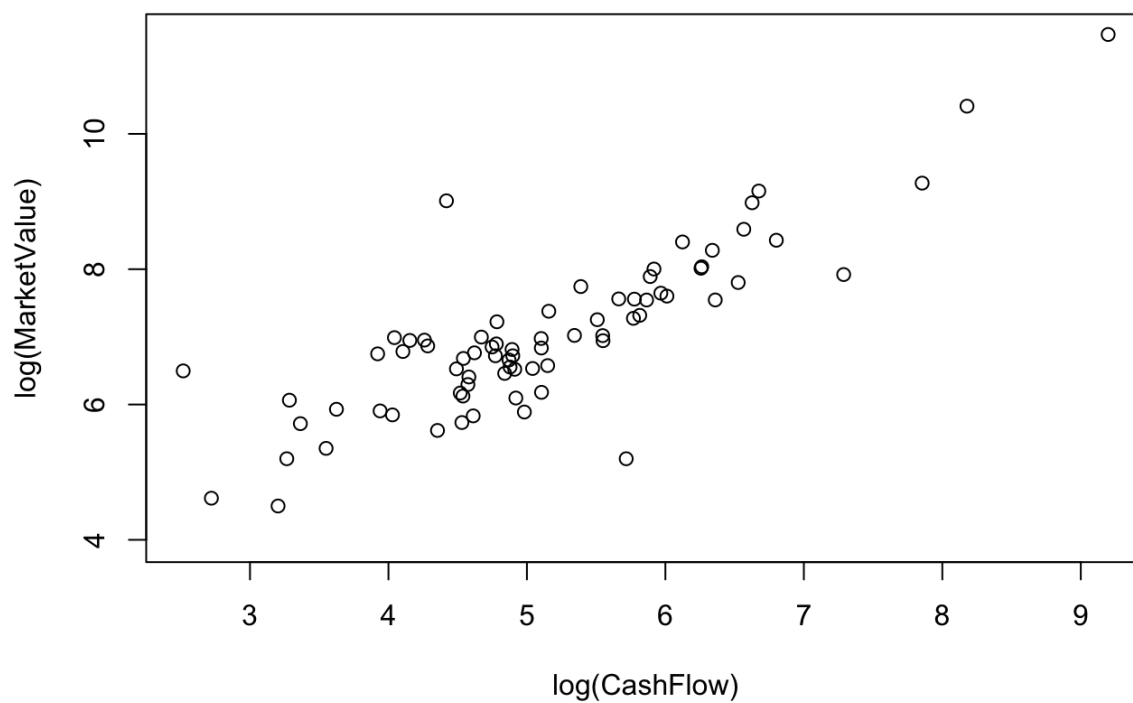


```
plot(log(MarketValue) ~ log(CashFlow), data = companies, main = "Log(Market Value) vs Log(CashFlow)")
```



```
## Warning in log(CashFlow): NaNs produced
```

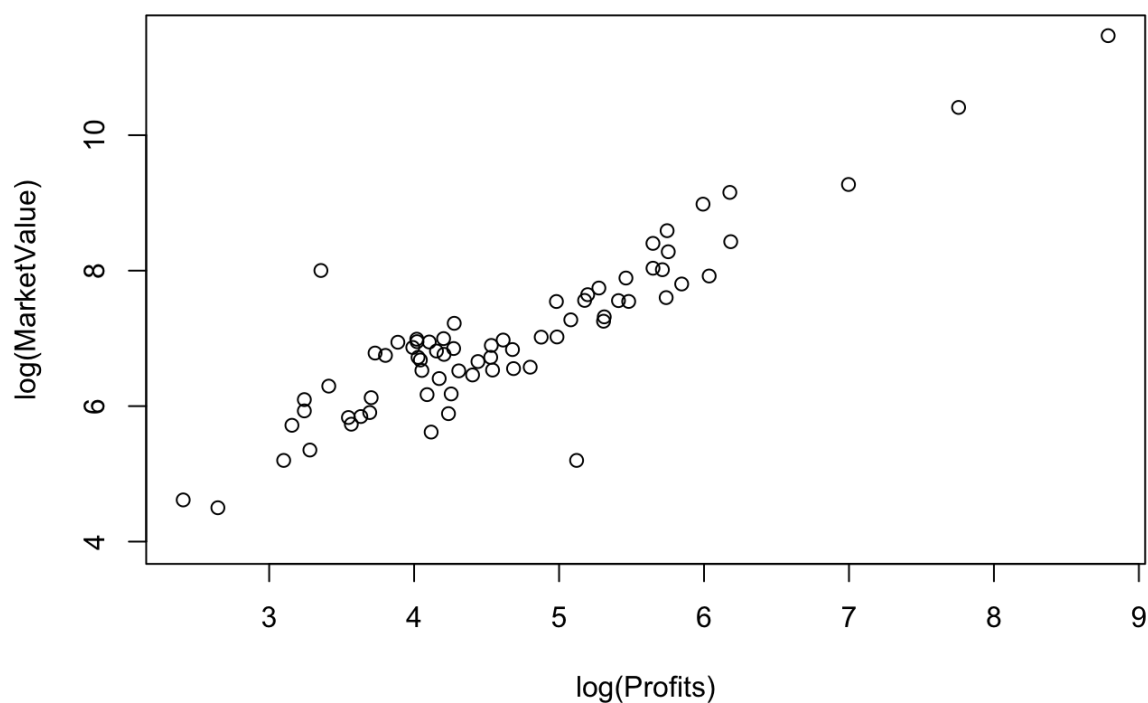
Log(Market Value) vs Log(CashFlow)



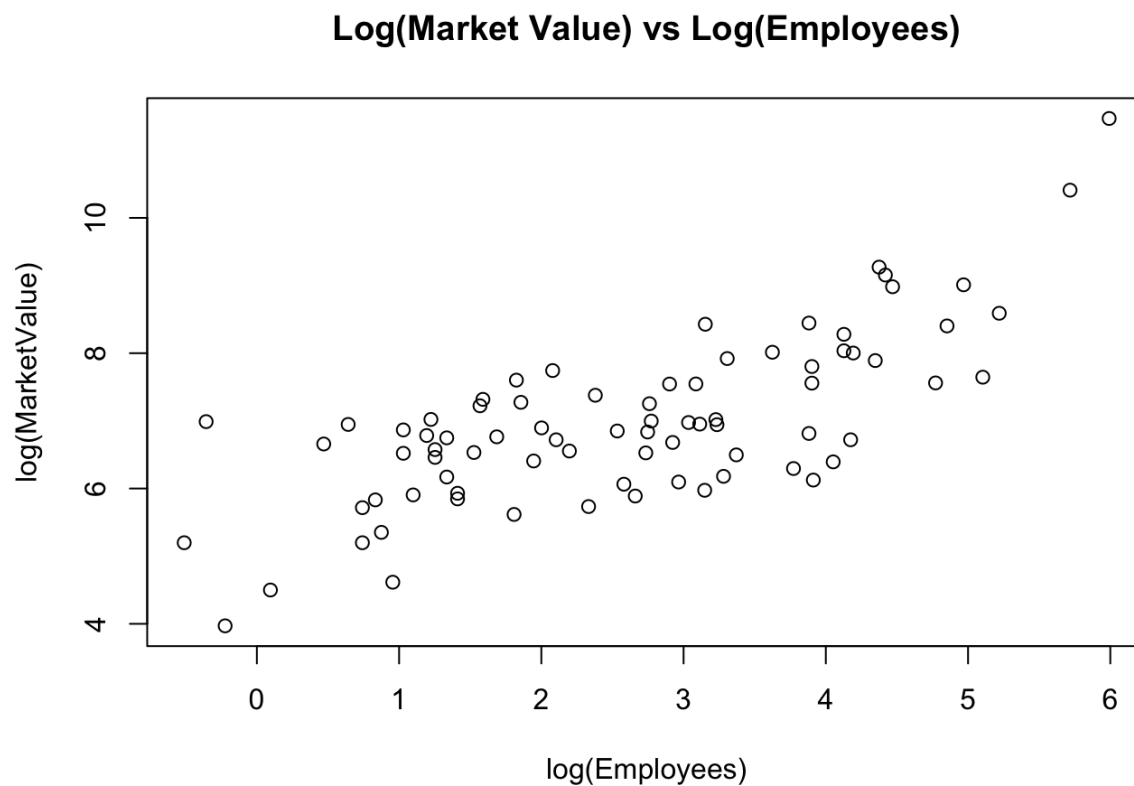
```
plot(log(MarketValue) ~ log(Profits), data = companies, main = "Log(Market Value) vs Log(Profits)")
```

```
## Warning in log(Profits): NaNs produced
```

Log(Market Value) vs Log(Profits)



```
plot(log(MarketValue)~ log(Employees), data = companies,main = "Log(Market Value) vs Log(Employees)")
```



Section E

Fit the model

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

```
##
## Call:
## lm(formula = MarketValue ~ log(Assets) + log(Sales) + Profits +
##     CashFlow + log(Employees), data = companies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8331.8  -974.1  -164.2   643.5 11501.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3650.342   3952.267  -0.924   0.3587
## log(Assets)    157.792    320.628   0.492   0.6241
## log(Sales)     257.400    676.877   0.380   0.7048
## Profits         8.436      3.045   2.770   0.0071 **
## CashFlow        3.275      2.103   1.557   0.1237
## log(Employees) 240.260    471.913   0.509   0.6122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2674 on 73 degrees of freedom
## Multiple R-squared:  0.9476, Adjusted R-squared:  0.944
## F-statistic: 264.1 on 5 and 73 DF,  p-value: < 2.2e-16
```

Section F

Why did we log some of the variables and not the others.

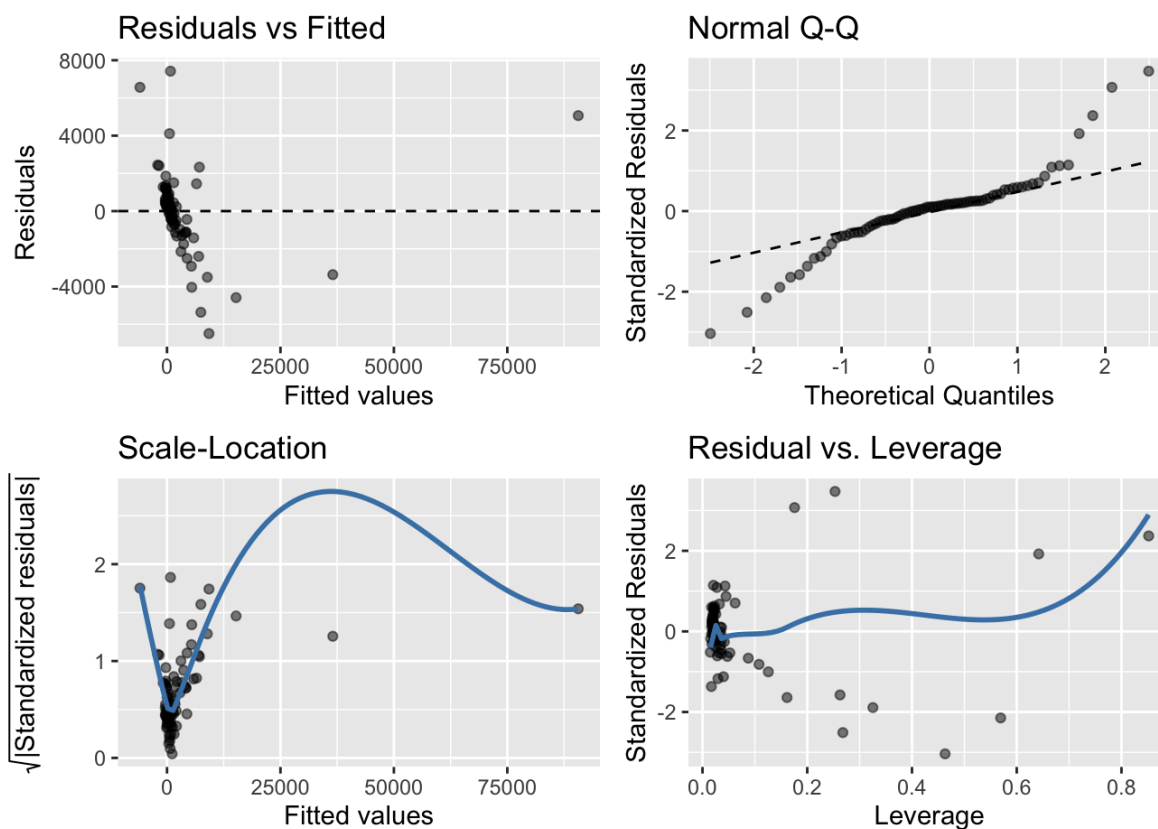
The variables `Asset`, `Sales` and `Employees` are log transformed to scale down their values; and fit the values to the lines relatively better. For example, Residual vs Fitted seems more clustered when logged. Normal Q-Q seems more fitted around the line when logged. Residual Vs leverage, when logged, the data seems to be more clustered, reducing the influence of higher leverage points. Furthermore, we did not apply the log transform for `Profits` and `CashFlow` since they have negative values.

Below is the comparison between fitting the model without performing log transformation for any of the predictor variables and fitting M1.

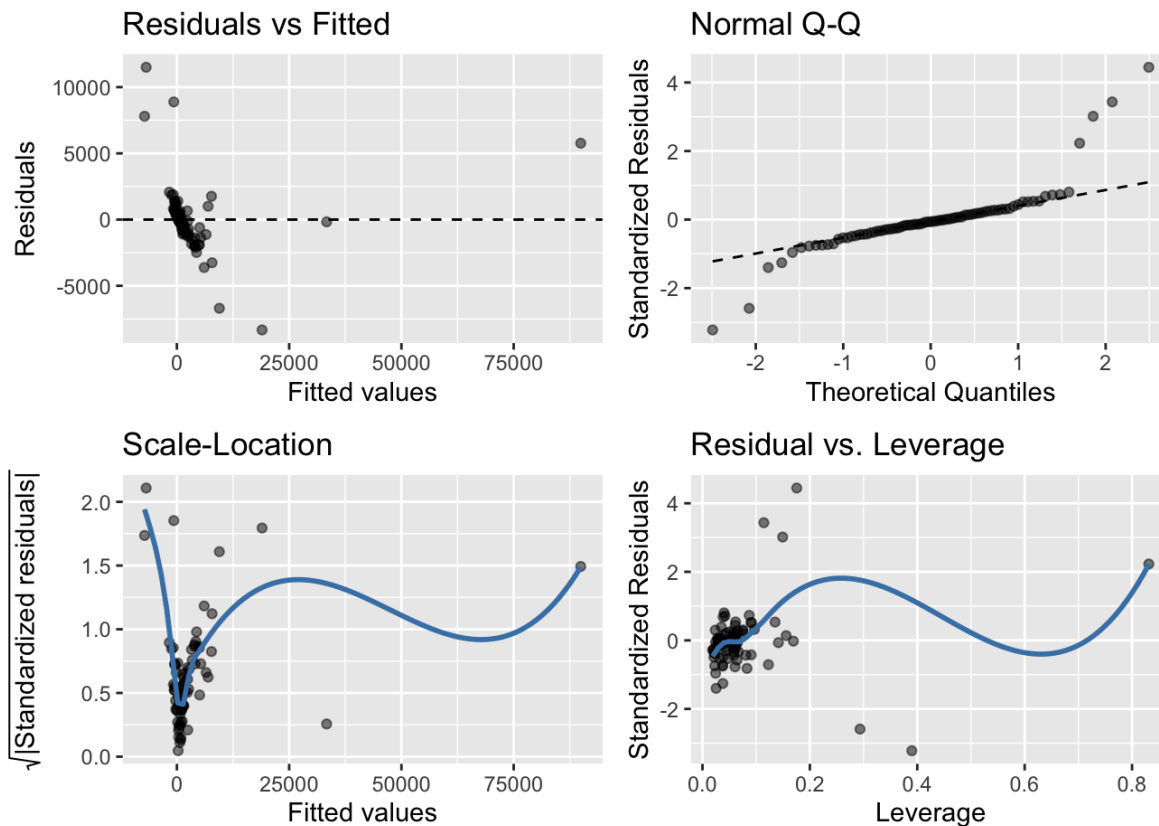
```
m1_without_log_transformation <- lm(MarketValue ~ Assets + Sales + Profits + CashFlow + Employe
es, data = companies)
summary(m1_without_log_transformation)
```

```
##
## Call:
## lm(formula = MarketValue ~ Assets + Sales + Profits + CashFlow +
##     Employees, data = companies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6493.0  -782.6   218.2   666.6  7421.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -956.04163   322.01647  -2.969  0.00404 **
## Assets         0.09558    0.04428    2.159  0.03418 *
## Sales         0.29949    0.14680    2.040  0.04496 *
## Profits       11.25226    2.38524    4.717 1.12e-05 ***
## CashFlow      -0.47715    1.76501   -0.270  0.78766
## Employees      6.29761    11.29525    0.558  0.57886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2209 on 73 degrees of freedom
## Multiple R-squared:  0.9643, Adjusted R-squared:  0.9618
## F-statistic: 393.9 on 5 and 73 DF,  p-value: < 2.2e-16
```

```
ggglm(m1_without_log_transformation)
```



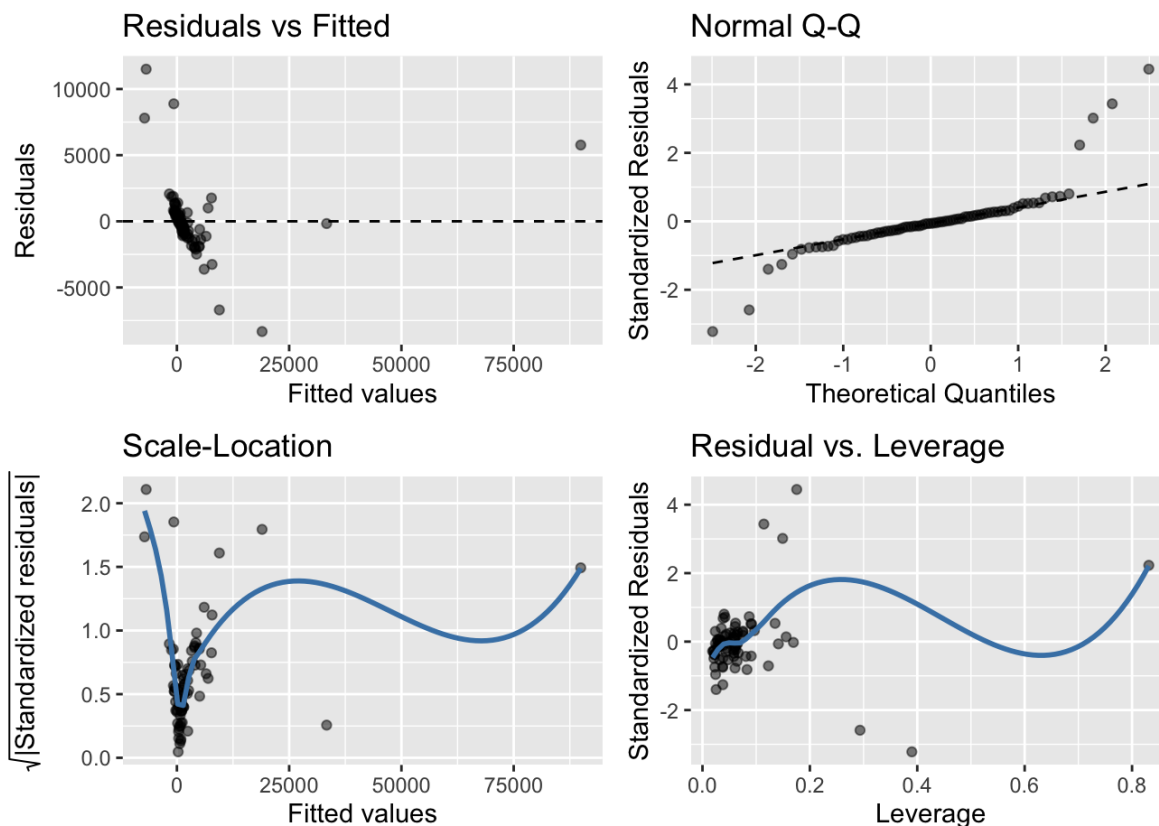
```
ggglm(M1)
```



Section G

Check the assumptions of the linear model M1.

```
gglm(M1)
```



Independence:

According to the dataset `companies.txt`, there is no dependencies can be seen. The dataset is not ordered in ordinal manner. Assuming that the data was randomly sampled according to the correct standard of data collection, we can safely assume there is no chronological or spacial relationship present.

Linearity (Residual Vs the Fitted plot):

As can be seen in the Residuals vs Fitted plot, there is a visible trend of a line with a downward slope. Residual points are not randomly distributed around the zero line showing signs of non-linearity.

Homoscedascity:

A curvature in the line is clearly visible in the scale-location plot with residuals not randomly distributed around $y = 1$. In fact the residuals are clustered in somewhat of a V-shape, suggesting that the constant error assumption is not met.

Normality (QQ plot):

While the majority of residuals fall on the line in QQ plot, we see many points depart at the two ends of the line suggesting some degrees of skewness or the present of unusual observations.

Leverage:

It can be in the Leverage point that there is one data point with substantially higher leverage comparing to the rest of the data.

It can quickly be found that the mentioned data point is an outlier in the dataset where each of its attribute is the maximum value for each of the corresponding variable. The difference can be seen more clearly when we investigate the distance between the minimum value to the 3rd quantile and from the 3rd quantile to the maximum value. For instance, the distance between the minimum value to the 3rd quantile in `Assets` is 5579 while the distance between the 3rd quantile to the maximum value is 46,832, which is 8.3 times more. This large distance can be seen across all variables.

In addition, the Cook's Distance plot confirm this information where the influence of data point 40 is evidently high. Elimination of this data point should be carefully considered.

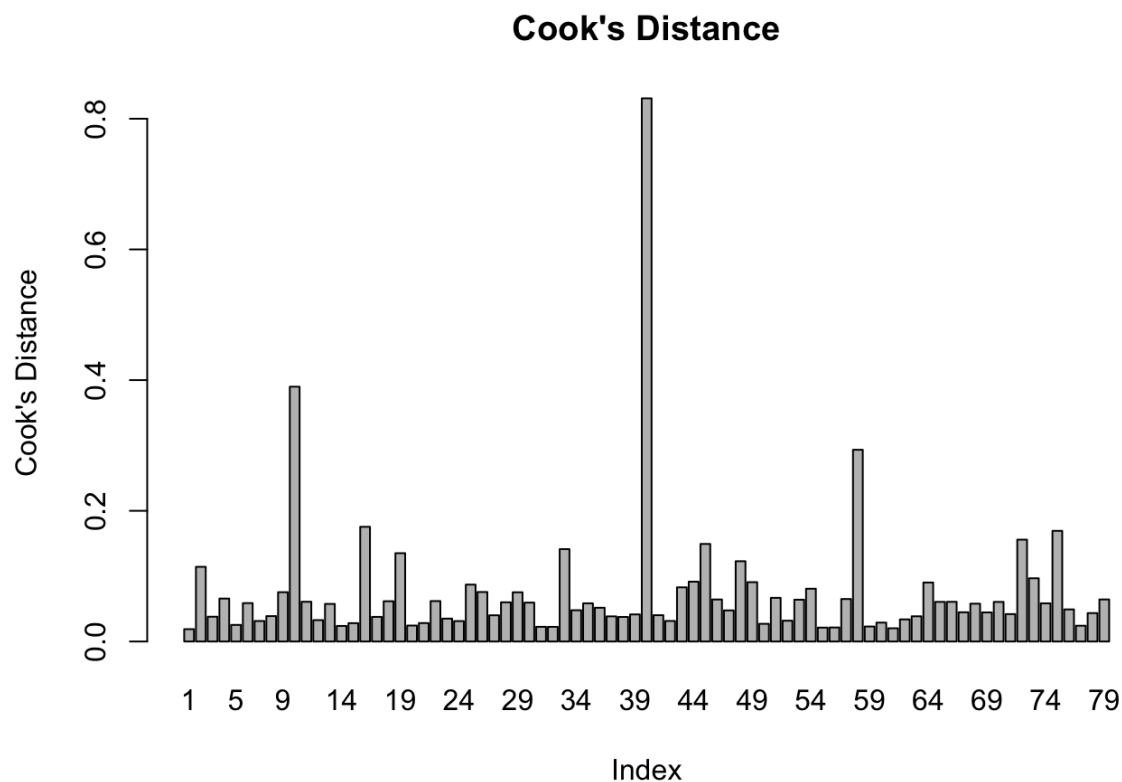
```
companies[40,]
```

```
##      Assets Sales MarketValue Profits CashFlow Employees
## 40  52634 50056      95697    6555      9874      400.2
```

```
summary(companies)
```

```
##      Assets      Sales      MarketValue      Profits
## Min.   : 223   Min.   : 176.0   Min.   : 53.0   Min.   : -771.5
## 1st Qu.: 1122  1st Qu.: 815.5   1st Qu.: 512.5  1st Qu.: 39.0
## Median : 2788  Median : 1754.0   Median : 944.0  Median : 70.5
## Mean   : 5941  Mean   : 4178.3   Mean   : 3269.8  Mean   : 209.8
## 3rd Qu.: 5802  3rd Qu.: 4563.5   3rd Qu.: 1961.5  3rd Qu.: 188.1
## Max.   :52634  Max.   :50056.0   Max.   :95697.0  Max.   :6555.0
##      CashFlow      Employees
## Min.   : -651.90   Min.   : 0.60
## 1st Qu.: 75.15    1st Qu.: 3.95
## Median : 133.30    Median : 15.40
## Mean   : 400.93    Mean   : 37.60
## 3rd Qu.: 328.85    3rd Qu.: 48.50
## Max.   :9874.00    Max.   :400.20
```

```
barplot(hatvalues(M1), main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Index")
```



Section H

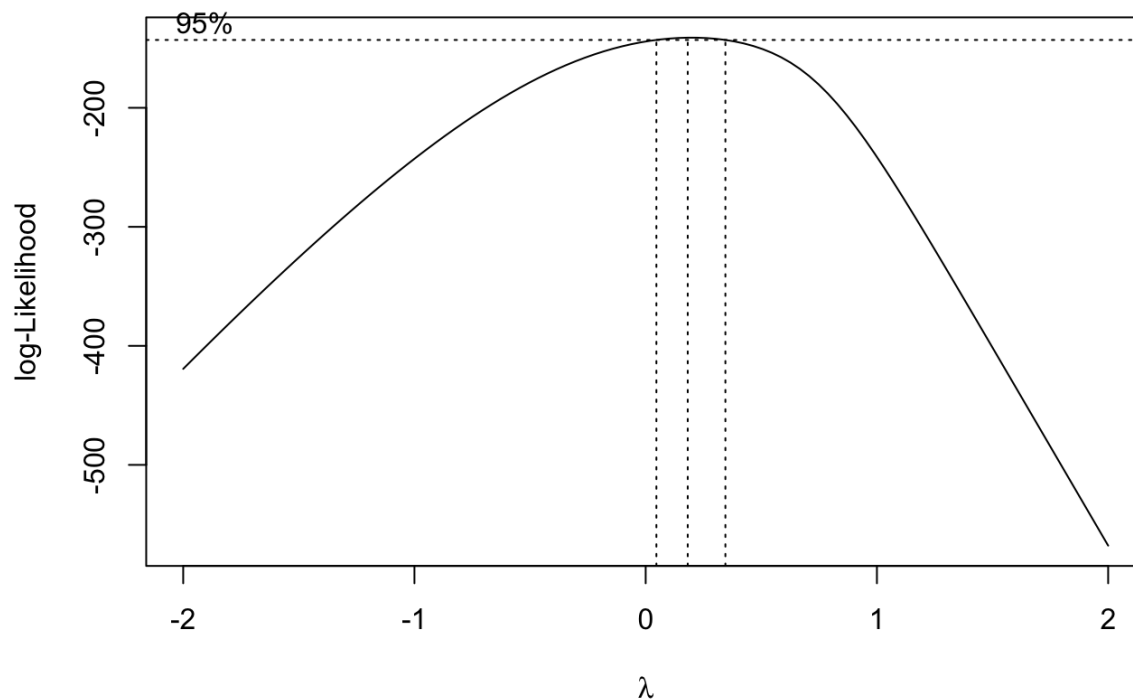
Use the Box-Cox method to find a suitable transformation of the data in the context of the model. State, with justification, your chosen value of λ .

```
library(MASS)
```

```
##  
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':  
##  
##      select
```

```
M1_box_cox <- boxcox(lm(MarketValue ~ log(Assets) + log(Sales) + Profits + CashFlow + log(Employees), data = companies))
```



```
lambda <- M1_box_cox$x[which.max(M1_box_cox$y)]
lambda
```

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

Given that the Box-Cox function plots the log-likelihood function with the two outer dashed lines represent the 95% confidence interval for the estimate of λ and that $\lambda = 0.182$ is the value that maximises this likelihood function, we can ensure that the chosen value of $\lambda = 0.182$ is an appropriate value.

Section I

Refit the model M1 but with the transformed response variable: denote this M2.

```
MarketValue_transformed = (companies$MarketValue^(0.182)-1)/(0.182)
M2 = lm(MarketValue_transformed ~ log(Assets) + log(Sales) + Profits + CashFlow + log(Employees), data = companies)
summary(M2)
```

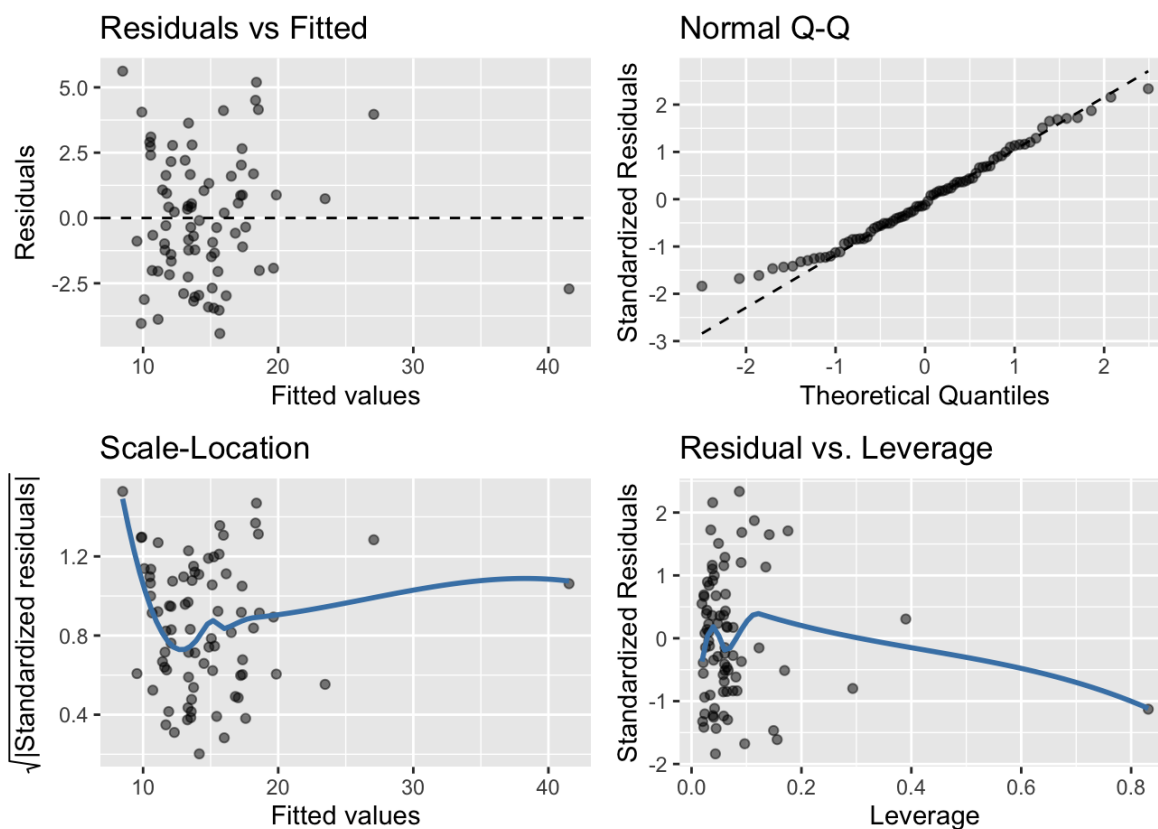


```
##
## Call:
## lm(formula = MarketValue_transformed ~ log(Assets) + log(Sales) +
##     Profits + CashFlow + log(Employees), data = companies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.4237 -1.9656 -0.2927  1.6464  5.6205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.927816   3.676033   1.341  0.18423
## log(Assets)    0.382807   0.298219   1.284  0.20333
## log(Sales)     0.312683   0.629569   0.497  0.62092
## Profits       -0.001208   0.002833  -0.427  0.67097
## CashFlow       0.002972   0.001956   1.519  0.13308
## log(Employees) 1.273342   0.438929   2.901  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 73 degrees of freedom
## Multiple R-squared:  0.7716, Adjusted R-squared:  0.756
## F-statistic: 49.32 on 5 and 73 DF,  p-value: < 2.2e-16
```

Section J

Check the assumptions for M2.

```
ggglm(M2)
```



independence:

Independence assumption is met, as discussed above.

Linearity:

The residuals are now randomly distributed around the zero line, with the exception of a few outliers.

Homoscedascity:

The V-shape cluster as seen in M1 is no longer visible as the residuals are spread out more evenly. Within the fitted values of 3 to 5, the values are distributed in a fairly random manner. However, we see some inconsistencies in the two ends of the plot which causes the line to be not flat and suggests the assumption of constant variance is not met.

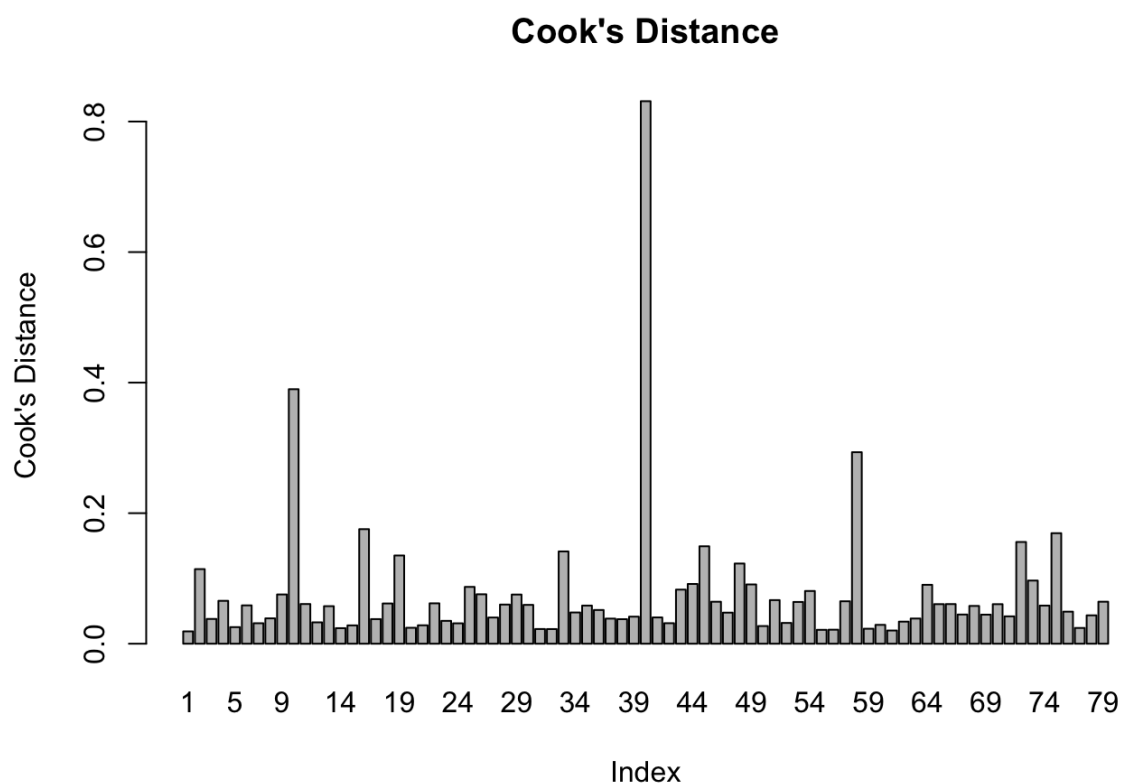
Normality:

Similar to M1, the majority of data points lie on the QQ line. We continue to see some departures at the two ends, however the departures are now of smaller values which indicates the normality assumption is met.

Leverage:

The outlier as mentioned above is still present and largely influential to the fit of M2.

```
barplot(hatvalues(M2), main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Index")
```



Section K

What is your preferred model M1 or M2? Justify your answer.

My preferred model is M2 due to the following reasons:

- The new model M2 has now met the normality assumption, which is the very first reason why we decided to perform Box-Cox transformation.
- The violation on the linearity and homoscedascity is largely due to the outliers.

Section L

Using your preferred model, obtain a 95% prediction interval for MarketValue for a company with

```
companies_pred <- predict(M2, newdata = data.frame(Assets = 1065, Sales = 642, Profits = 30, CashFlow = 59, Employees = 3.5), interval = "prediction")
```

```
companies_pred_res = exp(log(companies_pred*0.182 + 1) / 0.182)
companies_pred_res
```

```
##          fit      lwr      upr
## 1 471.5302 67.36555 1978.401
```

Thus, M2 predicts that the Market Value of the company with the given attributes (in the original scale) to be 471.530 million dollars with the 95% prediction interval (67.366, 1978.401) million dollars.