Assignment_2 STATS 3001

Zhao Ming Soh

07/04/2022

Q1

If we were to do direct substitution of $\lambda = 0$ to

$$\begin{aligned} \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} &= \frac{y^0 - 1}{0} \\ &= \frac{0}{0} \qquad \text{(Indeterminant)} \end{aligned}$$

As you can see this would give us an indeterminant. Therefore the L'Hospital Rule needs to be employed. We start by finding the derivatives of the numerator and denominator of the $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda}$:

$$\begin{aligned} \lim_{\lambda \to 0} \frac{f'(y^\lambda - 1)}{f'(\lambda)} &= \lim_{\lambda \to 0} \frac{y^\lambda \times \ln y \times 1 - 0}{1} \\ &= \lim_{\lambda \to 0} (y^\lambda \times \ln y) \\ &= \lim_{\lambda \to 0} (y^\lambda \times \ln y) \qquad \text{direct substitution of } \lambda = 0 \\ &= y^0 \times \ln y \\ &= \ln y \end{aligned}$$

The result of $\ln y$ is essentially the same as $\log y$ assuming that the base of the log is e, therefore we have proven that $\lim_{\lambda \to 0} \frac{f'(y^\lambda - 1)}{f'(\lambda)} = \log y$.

Q2

a) Read data in

```
companies_dt <- read_delim("companies.txt")
head(companies_dt)
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
## # A tibble: 6 x 6
##   Assets Sales MarketValue Profits CashFlow Employees
##    <dbl> <dbl>       <dbl>   <dbl>    <dbl> <chr>
## 1   2687  1870        1890   146.      352.  18.2
## 2  13271  9115        8190  -279        83   143.8
## 3  13621  4848        4572   485      899.  23.4
## 4   3614   367          90    14.1     24.6 1.1
## 5   6425  6131        2448   346.     682.  49.5
## 6   1022  1754        1370    72      120.  4.8
```

b) EDA

```
skim_without_charts(companies_dt)
```

Table 1: Data summary

Name	companies_dt
Number of rows	79
Number of columns	6
Column type frequency:	
character	1
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Employees	0	1	3	5	0	72	0

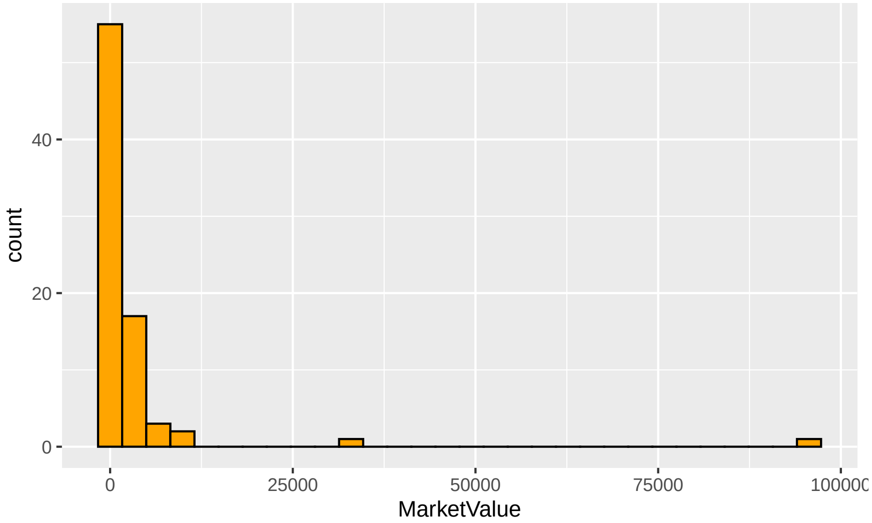
Variable type: numeric

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
Sales	0	1	4178.29	7011.63	176.0	815.50	1754.0	4563.50	50056
MarketValue	0	1	3269.75	11303.55	53.0	512.50	944.0	1961.50	95697
Profits	0	1	209.84	796.98	-771.5	39.00	70.5	188.05	6555
CashFlow	0	1	400.93	1205.53	-651.9	75.15	133.3	328.85	9874

(Histogram of the response variable - MarketValue)

```
companies_dt %>%
  ggplot(aes(MarketValue)) +
  geom_histogram(col = "black", fill = "orange")
```

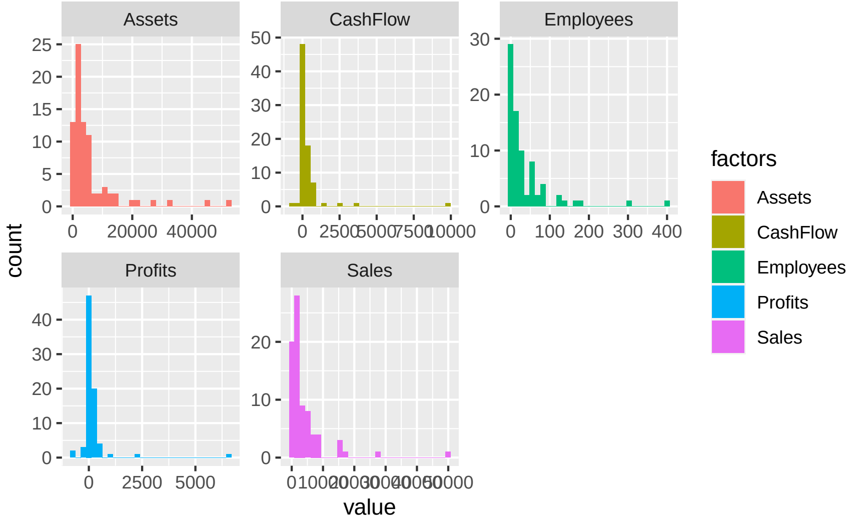
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(Histogram of the each predictor variables - Assets, Sales, Profits, CashFlow)

```
companies_dt$Employees <- as.double(companies_dt$Employees)
companies_dt %>%
  pivot_longer(cols = -MarketValue, names_to = "factors") %>%
  ggplot(aes(value)) +
  geom_histogram(aes(fill=factors)) +
  facet_wrap(~factors, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

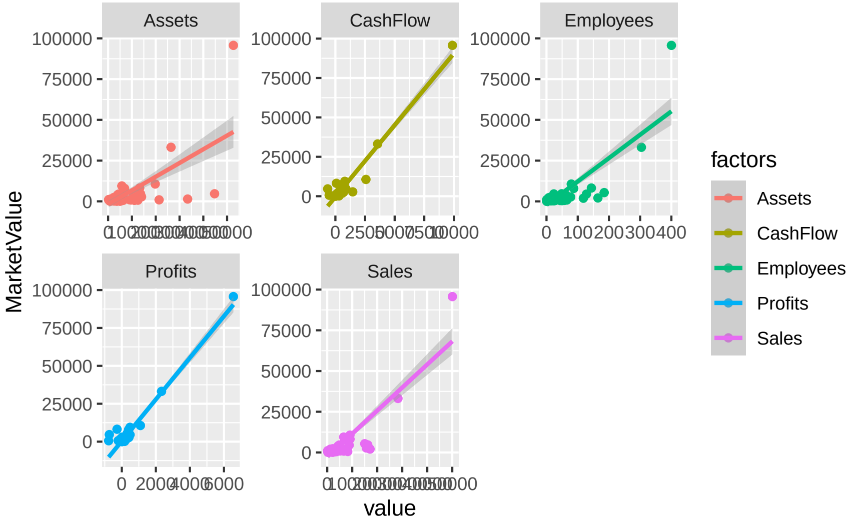


c) Scatterplots of MarketValue against each predictors

```
companies_dt %>%
  pivot_longer(cols = -MarketValue, names_to = "factors") %>%
```

```r
  ggplot(aes(value, MarketValue, col = factors)) +
  geom_point(aes(fill=factors)) +
  facet_wrap(~factors, scales = "free") +
  geom_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



d) Do a log transform on both the x and y and produce the similar plots to c)

```r
companies_dt_pivot <- companies_dt %>%
  pivot_longer(cols = -MarketValue, names_to = "factors")

# Log transfrom all the values of the predictors and response variable.
col_vars <- c("MarketValue", "value")
companies_dt_pivot[col_vars] <- log10(companies_dt_pivot[col_vars])
```

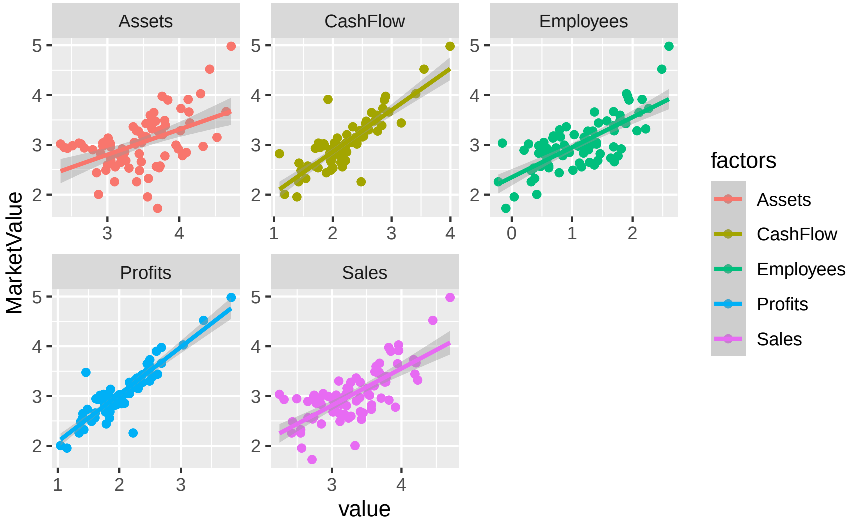
```
## Warning in lapply(X = x, FUN = .Generic, ...): NaNs produced
```

```r
companies_dt_pivot %>%
  ggplot(aes(value, MarketValue, col = factors)) +
  geom_point() +
  facet_wrap(~factors, scales = "free") +
  geom_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 12 rows containing non-finite values (stat_smooth).

## Warning: Removed 12 rows containing missing values (geom_point).
```



e) Fit the model

```
M1 <- lm(MarketValue ~ log10(Assets) + log10(Sales) + Profits + CashFlow + log10(Employees),
         data = companies_dt)
```

f) Log Transformation of Profits and Cashflow

```
log_transfrom <- function(x){
  result <- log10(x)
}

companies_dt_log <- companies_dt %>%
  select(where(is.double)) %>%
  mutate_all(~ log_transfrom(.))
```

```
## Warning in log_transfrom(.): NaNs produced

## Warning in log_transfrom(.): NaNs produced
```

As you can see predictor Profits and CashFlow produces NaNs after log transform, this is mainly due to both predictors having negative values. The negative value for both Profits and CashFlow is valid in the sense that it signifies the losses in a company which is also important for valuing the company. So, to log transform it would mean a company will never have any losses which made no sense and can lead to a very wrong model.

g) Checking the Assumptions of M1

- **Linearity:**
  - Linearity assumption is not reasonable because the zero line is curved and the residuals are not randomly scattered above and below the zero line as shown in Figure 1 below.

- **Homoscedasticity:**
  - Homoscedasticity assumption is not reasonable because the line is not horizontal and the residuals are not equally spread across the line as shown in Figure 3 below.
- **Normality**
  - Most of the residuals are on the line but there is a considerable amount of fanning from each ends of the line. Therefore, normality assumption is not met as shown in Figure 2 below.
- **Influence and Leverage**
  - Data Points that exceeded the cook's distance of 1 are 40 and 10 denoting high leverage points with large amount of residuals. This simply means data points 40 and 10 have high influence on the parameters of M1 and should be assessed further as shown in Figure 4 below.

```
plot(M1, which = 1)
```

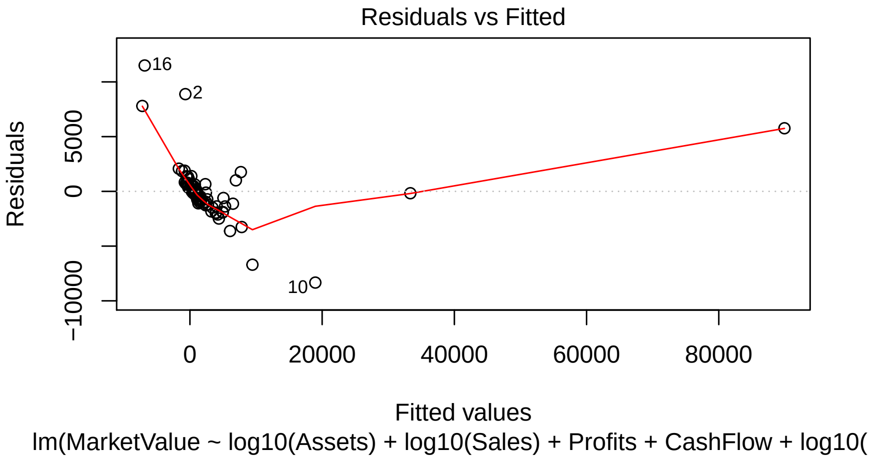


Figure 1: Shows the differences(residuals) in the Market Value of 79 companies between the one that is given in the dataset and the one that is predicted(fitted) with the multilinear regression model M1

```
plot(M1, which = 2)
```

```
plot(M1, which = 3)
```

```
plot(M1, which = 5)
```

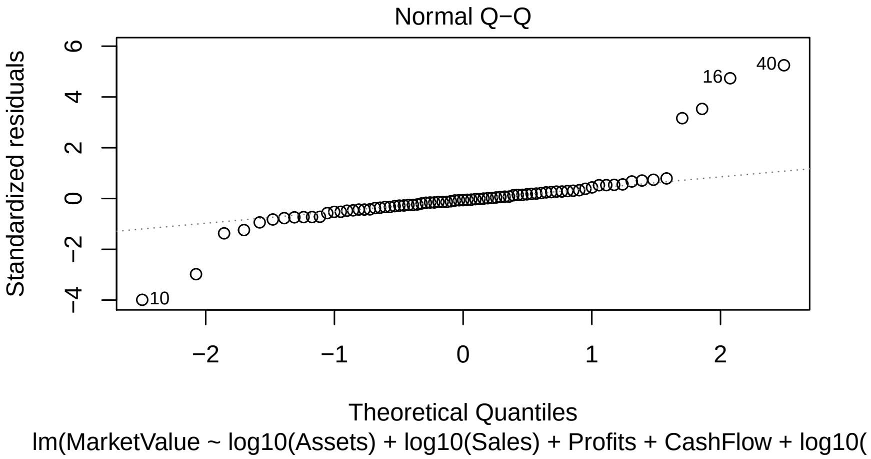


Figure 2: Shows how the differences(residuals) in the Market Value of 79 companies between the one that is given in the dataset and the one that is predicted with the multilinear regression model M1 is distributed

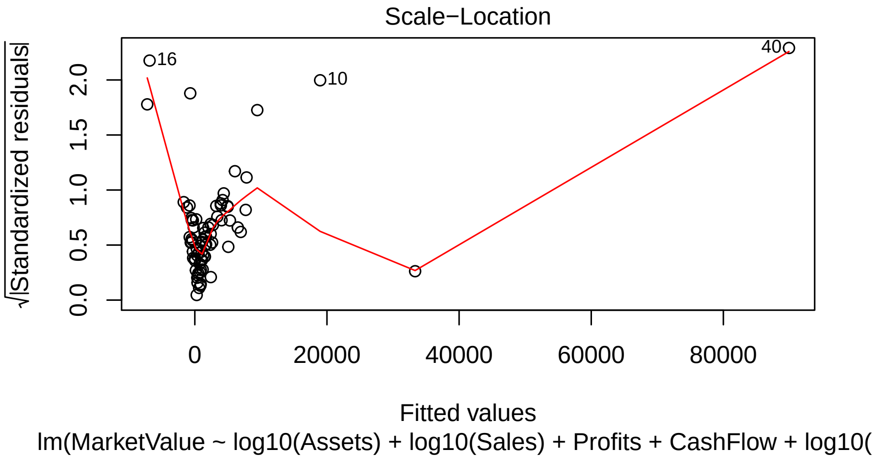


Figure 3: Shows the scaled standardised differences(residuals) in the Market Value of 79 companies between the one that is given in the dataset and the one that is predicted(fitted) with the multilinear regression model M1

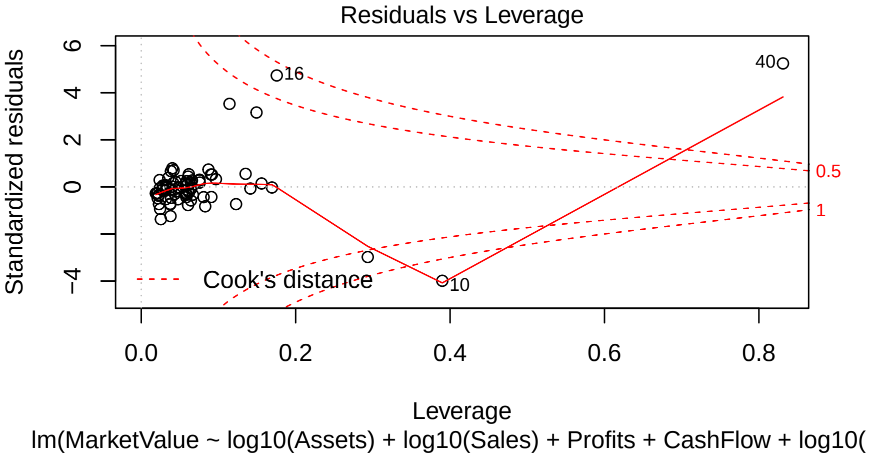
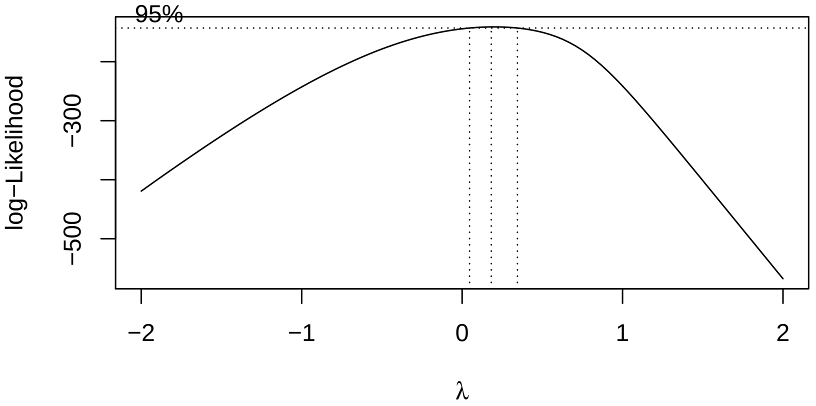


Figure 4: Shows how much influence each of the datapoint has on the multi linear regression model as indicated by the how big of a distance each datapoint has from the mean of the dataset given by the Cook's distance metric. Generally, if the Cook's distance of a datapoint is greater or equal to 1, it could be a problematic point.

h) Box-Cox Method of Estimation

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



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

The estimated lambda $\hat{\lambda}$ is 0.1818. This is show by the Figure above, the center of the dotted line represent the estimated lambda $\hat{\lambda}$ and the other 2 dotted lines represent the upper and lower boundary of the confident interval of the estimated lambda $\hat{\lambda}$. Since, 0 value is not in the 95% confidence interval of the estimated lambda $\hat{\lambda}$, we have to transform the response variable y with $\frac{y^{\lambda}-1}{\lambda}$.

i) Refit the model M1 with Box-Cox Transformation

```
cox_box <- function(x) {
  result <- ((x^(lambda)-1)/lambda)
}

companies_dt_cox_box <- companies_dt %>%
  mutate(MarketValue = cox_box(MarketValue))

M2 <- lm( MarketValue ~ log10(Assets) + log10(Sales) + Profits + CashFlow + log10(Employees),
          data = companies_dt_cox_box)
```

j) Check M2 model assumptions

- **Linearity:**
  - Although the zero line is not entirely horizontal but there isn't any obvious pattern in the dispersion of the data points apart from the 2 outliers on the right-hand side as shown in Figure 5 below. Therefore, linearity assumption is reasonable.
- **Homoscedasticity:**
  - From Figure 7 below, homoscedasticity assumption is reasonable because the points are disperse somewhat evenly along the zero line, although the zero line is not entirely horizontal due to the outliers in the dataset.
- **Normality**
  - From Figure 6 below, most of the data points are scattered around the dotted line and only minimal fanning of points at each ends of the dotted line, so normality assumption is reasonable.
- **Influence and Leverage**
  - From Figure 8 below, it appears that most of the influential data points lie on

```
plot(M2, which = 1)
```

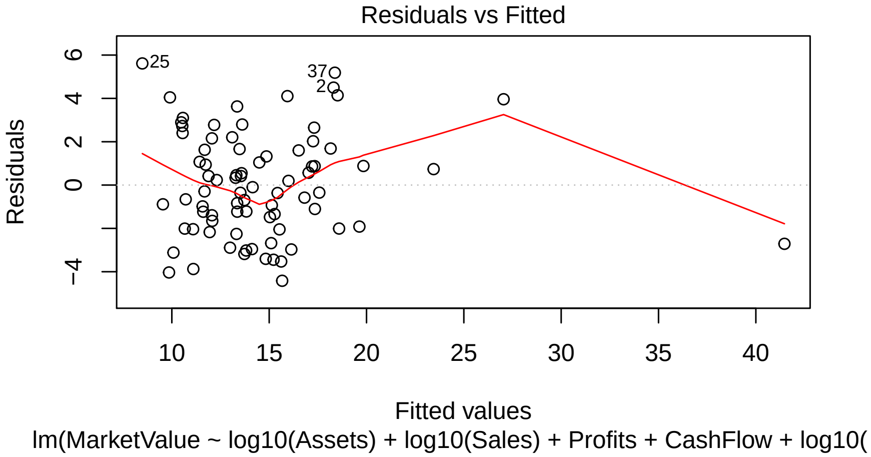


Figure 5: Shows the differences(residuals) in the Market Value of 79 companies between the one that is given in the dataset and the one that is predicted(fitted) with the multilinear regression model M2

```
plot(M2, which = 2)
```

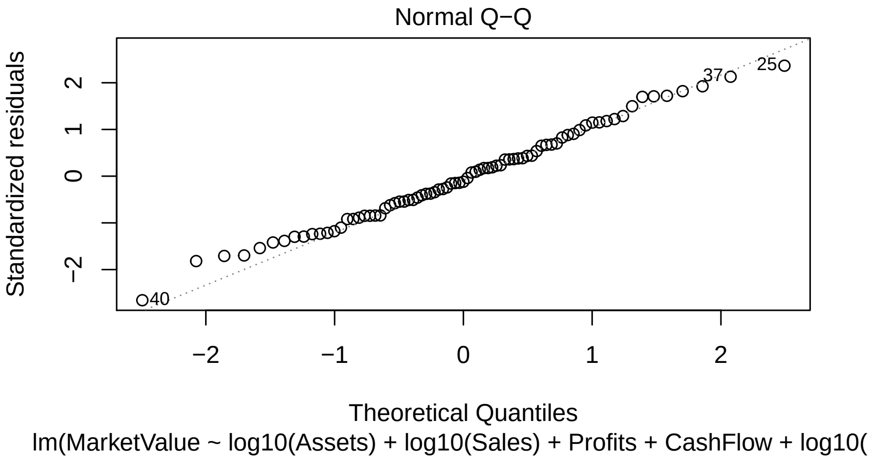


Figure 6: Shows how the differences(residuals) in the Market Value of 79 companies between the one that is given in the dataset and the one that is predicted with the multilinear regression model M2 is distributed

```
plot(M2, which = 3)
```

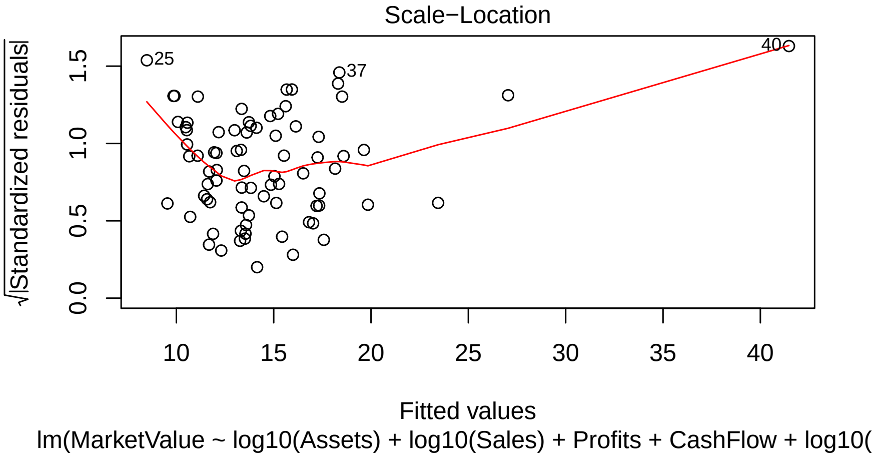


Figure 7: Shows the scaled standardised differences(residuals) in the Market Value of 79 companies between the one that is given in the dataset and the one that is predicted(fitted) with the multilinear regression model M2

```
plot(M2, which = 5)
```

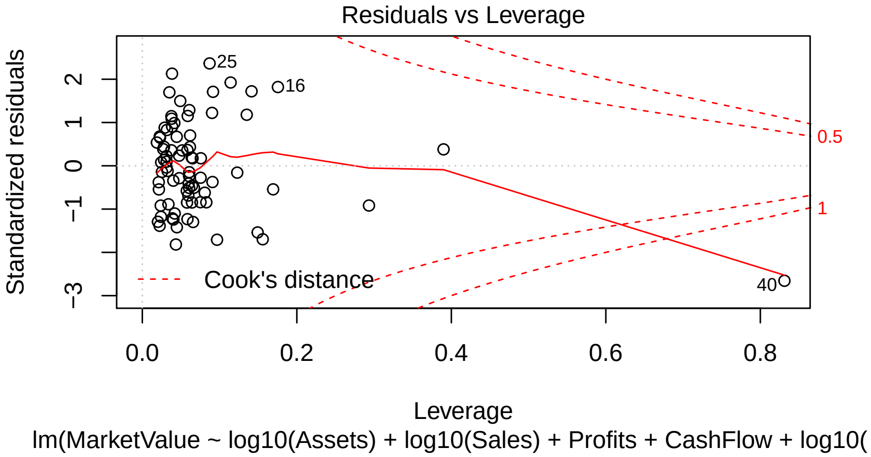


Figure 8: Shows how much influence each of the datapoint has on the multi linear regression model as indicated by the how big of a distance each datapoint has from the mean of the dataset given by the Cook's distance metric. Generally, if the Cook's distance of a datapoint is greater or equal to 1, it could be a problematic point.

k) Choosing Model M1 or M2 ?

Since, the model assumptions of M2 are met, M2 should be the preferred model as it would give us accurate result as opposed to M1 that has violated all model assumptions.

l) 95% prediction interval for MarketValue for a company with :

Assets	Sales	Profits	CashFlow	Employees
1065	642	30	59	3.5

```
newdata_1 <- tibble(
  Assets = 1065,
  Sales = 642,
  Profits = 30,
  CashFlow = 59,
  Employees = 3.5
)

prediction_val <- predict(M2, newdata_1, interval = "prediction" )
prediction_val
```

```
##        fit      lwr      upr
## 1 11.34444 6.326736 16.36214
```

The y here represents the MarketValue of the dataset : (Backtrans)

$$
\begin{aligned}
6.33 \leq \frac{y^{\lambda}-1}{\lambda} &\leq 16.36 \\
\text{Given that } \lambda &= 0.1818 \\
6.33 \leq \frac{y^{0.1818}-1}{0.1818} &\leq 16.36 \\
6.33 \times 0.1818 + 1 \leq y^{0.1818} &\leq 16.36 \times 0.1818 + 1 \\
\ln 2.1508 \leq \ln y^{0.1818} &\leq \ln 3.9742 && \text{multiply both sides by ln} \\
\frac{0.7658}{0.1818} \leq \ln y &\leq \frac{1.3798}{0.1818} && \text{divide both sides by 0.1818} \\
e^{4.2125} \leq y &\leq e^{7.5899} && \text{multiple both sides by e} \\
67.53 \leq y &\leq 1978.03
\end{aligned}
$$

We are 95% confident that the Market Value of a company with 1065 millions in assets, 642 millions in sales, 30 millions in profits, 59 millions in cash flow and 3.5 millions in employees will be between 67.53 millions and 1978.03 millions in USD.