Statistical Modelling III
Assignment 2

Gia Bao Hoang - a1814824

Semester 1 2023

Q1

For $y > 0$, we have:

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

Since we have the limit of the form 0/0, we can apply the L'Hospital's rule by taking the derivative of the both the numerator and the denominator with respect to λ:

$$\begin{aligned}
\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{\frac{d}{d\lambda}\left(y^\lambda - 1\right)}{\frac{d}{d\lambda}\lambda} \\
&= \lim_{\lambda \to 0} \frac{y^\lambda \ln(y)}{1} \\
&= \frac{y^0 \ln(y)}{1} \\
&= \ln(y) \\
&= \log(y) \\
\therefore \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} &= \log(y)
\end{aligned}$$

Q2

Load the package

```
pacman::p_load(tidyverse, gglm, skimr, MASS)
```

(a)

```
data <- read_delim(("companies.txt"), delim = "\t")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 79 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Employees
## dbl (5): Assets, Sales, MarketValue, Profits, CashFlow
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data
```

```
## # A tibble: 79 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
##  7   1093  1679        1070   101.     164.  20.8
##  8   1529  1295         444    25.6    137   19.4
##  9   2788   271         304    23.5     28.9 2.1
## 10  19788  9084       10636  1093.    2577.  79.4
## # i 69 more rows
```

Convert Employees to numeric

```
data <- data %>%
  mutate(Employees = as.numeric(Employees))
```

(b)

EDA

```
skim(data)
```

Table 1: Data summary

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

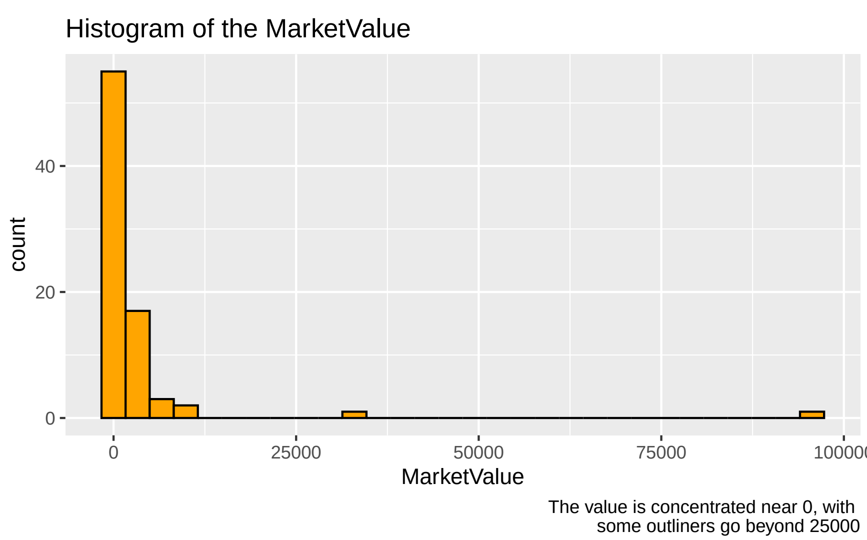
Variable type: numeric

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	

Histogram of the response variable

```
data %>%
  ggplot(aes(MarketValue)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(
    title = "Histogram of the MarketValue",
    caption = "The value is concentrated near 0, with
    some outliners go beyond 25000"
  )
```

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



Based on the histogram, the MarketValue is concentrated near 0, with some outliners go beyond 25000.

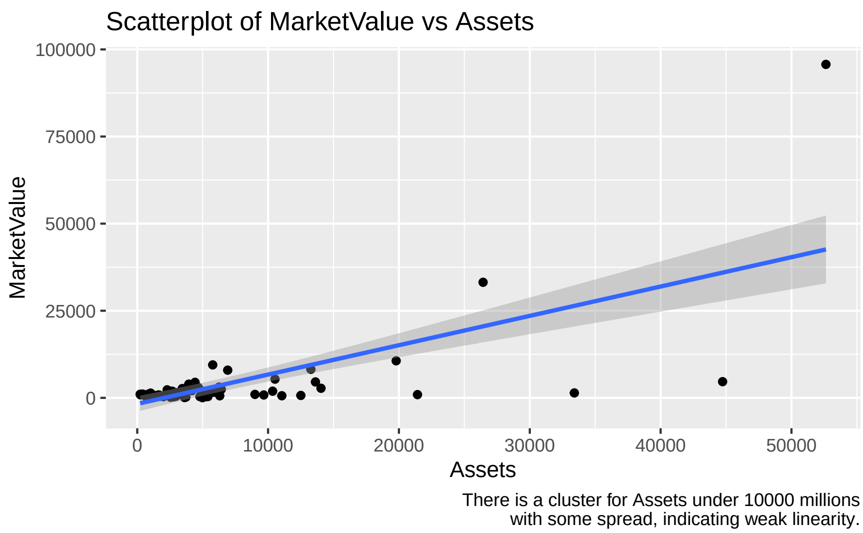
(c)

Scatterplots of MarketValue against each of the other predictors:

```
data %>%
  ggplot(aes(Assets, MarketValue)) +
  geom_point() +
```

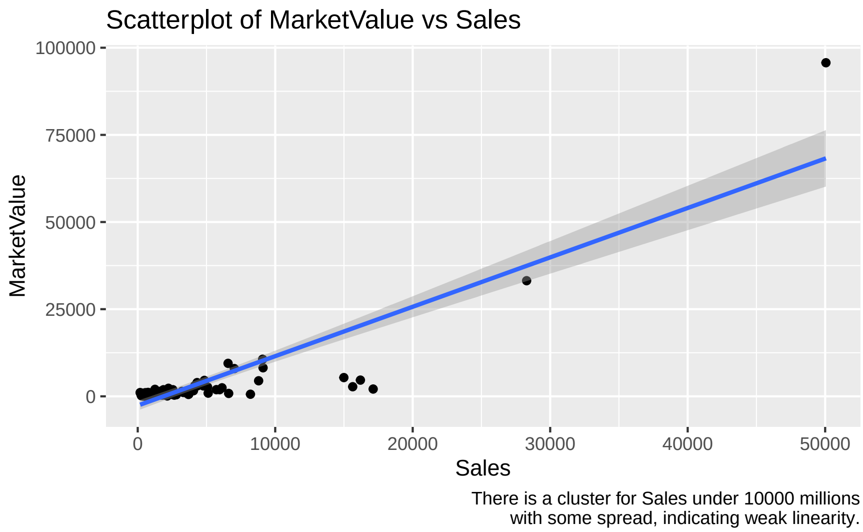
```
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of MarketValue vs Assets",
    caption = "There is a cluster for Assets under 10000 millions
    with some spread, indicating weak linearity."
  )
```

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



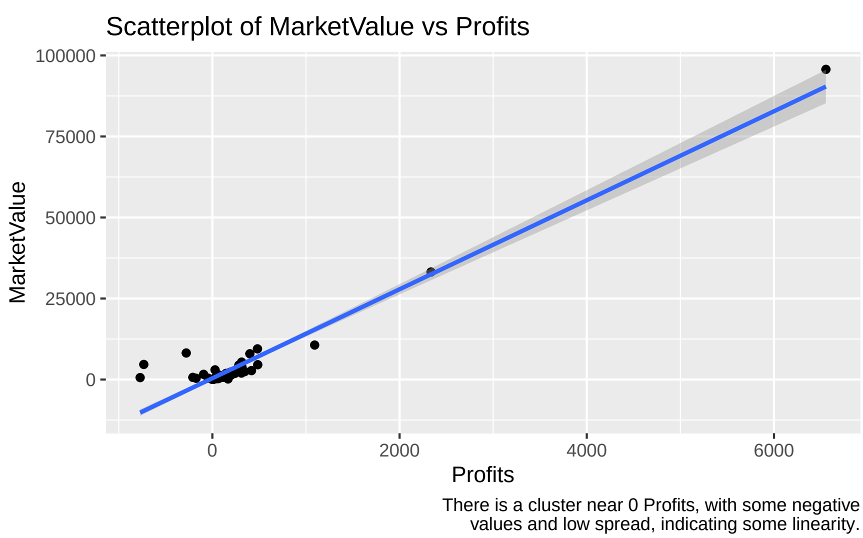

```
data %>%
  ggplot(aes(Sales, MarketValue)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of MarketValue vs Sales",
    caption = "There is a cluster for Sales under 10000 millions
    with some spread, indicating weak linearity."
  )
```

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




```
data %>%
  ggplot(aes(Profits, MarketValue)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of MarketValue vs Profits",
    caption = "There is a cluster near 0 Profits, with some negative
    values and low spread, indicating some linearity."
  )
```

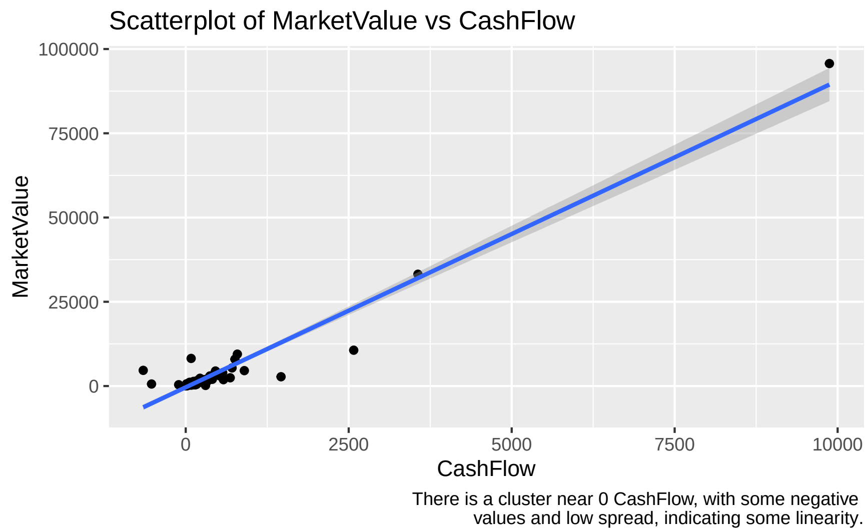
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
data %>%
  ggplot(aes(CashFlow, MarketValue)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
```

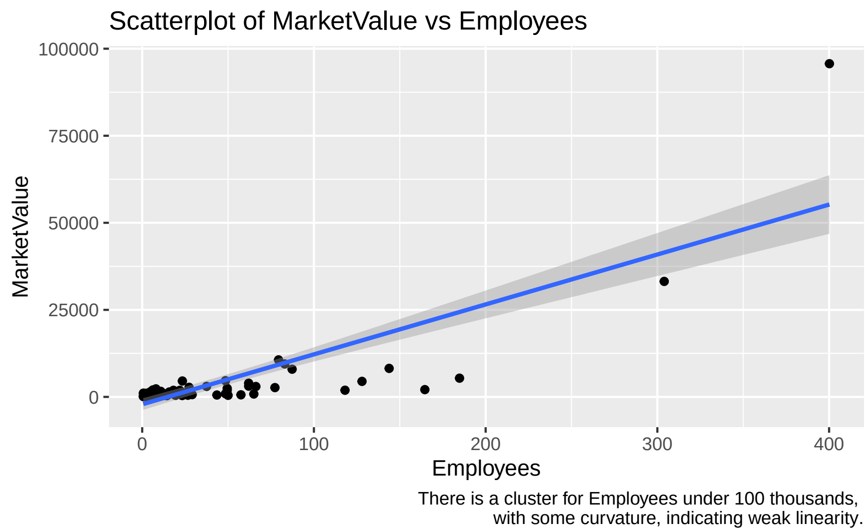
```
    title = "Scatterplot of MarketValue vs CashFlow",
    caption = "There is a cluster near 0 CashFlow, with some negative
    values and low spread, indicating some linearity."
  )
```

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




```
data %>%
  ggplot(aes(Employees, MarketValue)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of MarketValue vs Employees",
    caption = "There is a cluster for Employees under 100 thousands,
    with some curvature, indicating weak linearity."
  )
```

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

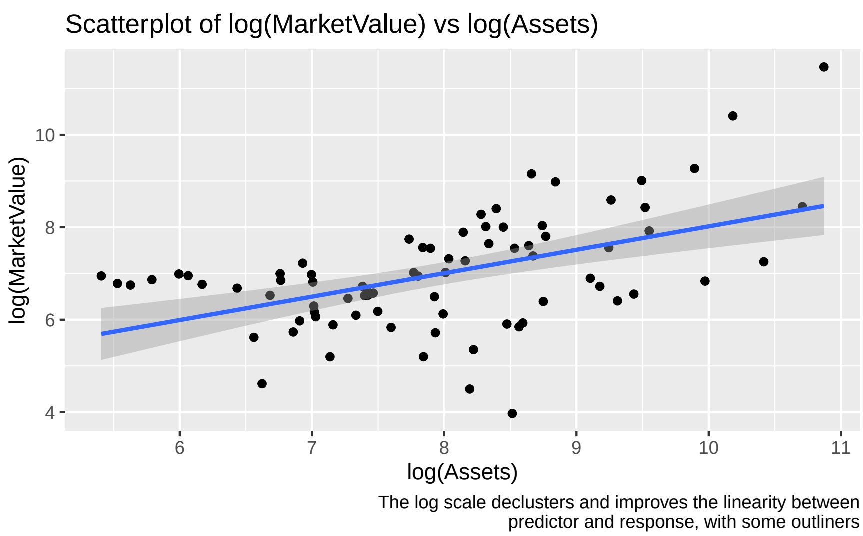


(d)

Scatterplots of MarketValue against each of other predictors on a log scale

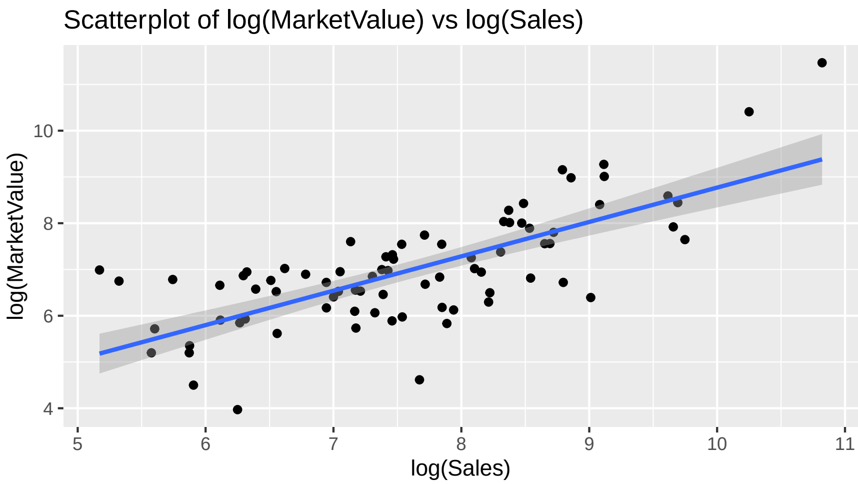
```
data %>%
  ggplot(aes(log(Assets), log(MarketValue))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of log(MarketValue) vs log(Assets)",
    caption = "The log scale declusters and improves the linearity between
    predictor and response, with some outliners"
  )
```

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




```
data %>%
  ggplot(aes(log(Sales), log(MarketValue))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of log(MarketValue) vs log(Sales)",
    caption = "The log scale declusters and improves the linearity between
    predictor and response, with some outliners"
  )
```

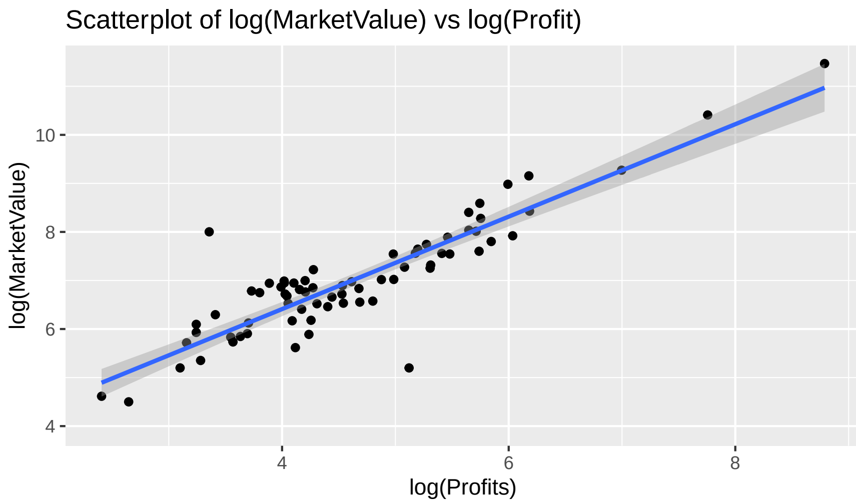
```
## `geom_smooth()` using formula = 'y ~ x'
```



The log scale declusters and improves the linearity between predictor and response, with some outliners

```
data %>%
  ggplot(aes(log(Profits), log(MarketValue))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of log(MarketValue) vs log(Profit)",
    caption = "There are NaNs produced due to some negative values in Profits."
  )
```

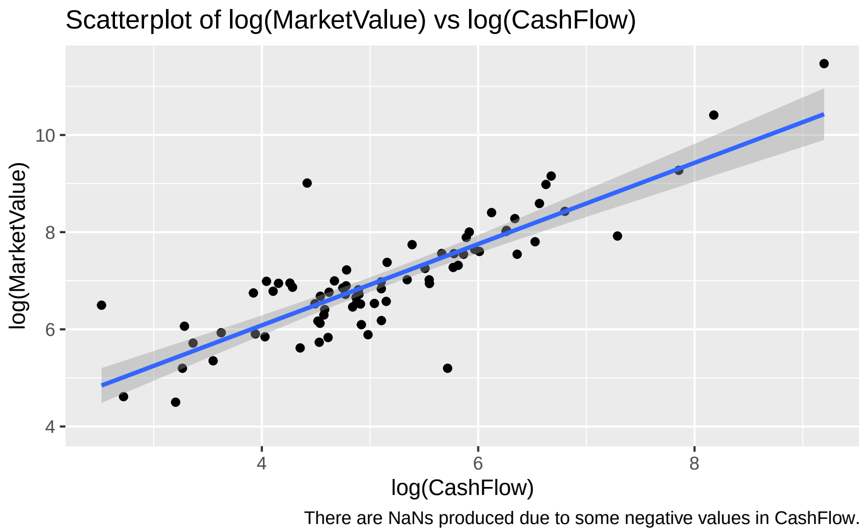
```
## Warning in log(Profits): NaNs produced

## Warning in log(Profits): NaNs produced

## Warning in log(Profits): NaNs produced

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

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 8 rows containing missing values (`geom_point()`).
```



There are NaNs produced due to some negative values in Profits.

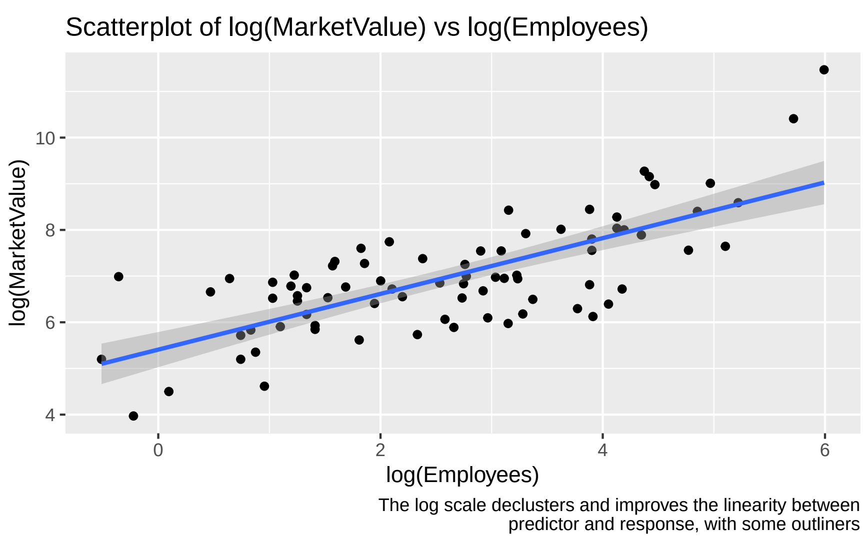
```
data %>%
  ggplot(aes(log(CashFlow), log(MarketValue))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of log(MarketValue) vs log(CashFlow)",
    caption = "There are NaNs produced due to some negative values in CashFlow."
  )
```

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

## Warning in log(CashFlow): NaNs produced

## Warning in log(CashFlow): NaNs produced

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

## Warning: Removed 4 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 4 rows containing missing values (`geom_point()`).
```




```r
data %>%
  ggplot(aes(log(Employees), log(MarketValue))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(
    title = "Scatterplot of log(MarketValue) vs log(Employees)",
    caption = "The log scale declusters and improves the linearity between
    predictor and response, with some outliners"
  )
```

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



(e)

Fit the model

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

```
##
## Call:
## lm(formula = MarketValue ~ log(Assets) + log(Sales) + Profits +
##     CashFlow + log(Employees), data = data)
##
## 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
```

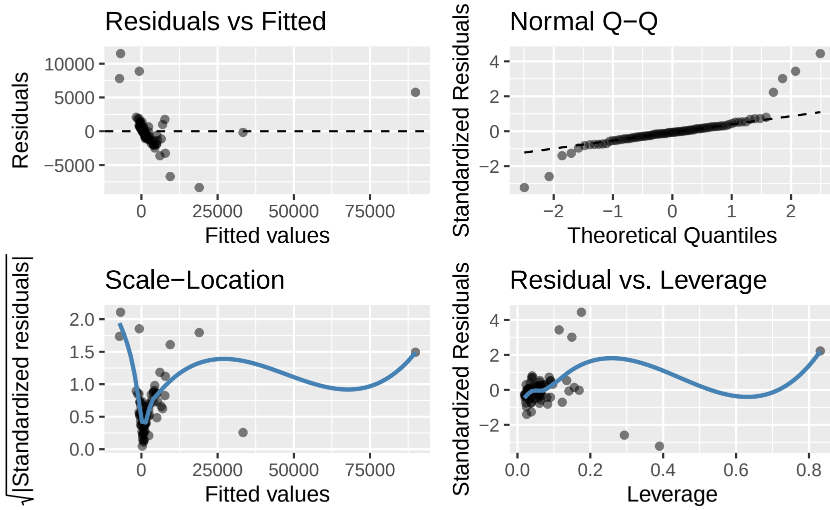
(f)

We apply log transformation to variables like Assets, Sales, and Employees to decluster and improve the linearity between them and the response variable, MarketValue. Scatterplots on a log scale demonstrate this enhanced linearity. However, attempting to log-transform Profits and CashFlow would generate warnings about NaN values due to the presence of negative values. Thus, it is not appropriate to apply logarithm transformations to these predictors in the model.

(e)

Check the assumptions of the model M1

```r
gglm(M1)
```



Regression assumptions

Linearity: The residuals are clustered around the (0,0) point with some outliners, especially there is one beyond the 75000 range. Hence, the linearity assumption is violated.

Homoscedasticity: There is a cluster at the 0 value with some outlines, especially there is one beyond the 75000 range. There is no constant spread about the zero line in the scale-location plot. Hence, the assumption of constant variance is violated.

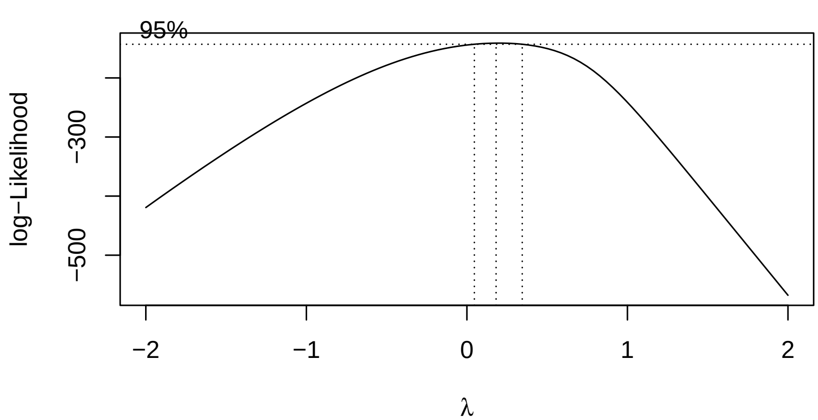
Normality: There is some departure from normality in both tails of the distribution of residuals. However, the majority of the data is close to normally distributed. Hence, normality assumption is reasonable.

Independence: The plots can not verify this assumption.

(h)

Use Box-Cox method

```
bc <- boxcox(M1)
```



We will choose the λ value that maximizes the likelihood function within the 95% confidence interval.

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

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

(i)

Refit the model with the transformed response variable

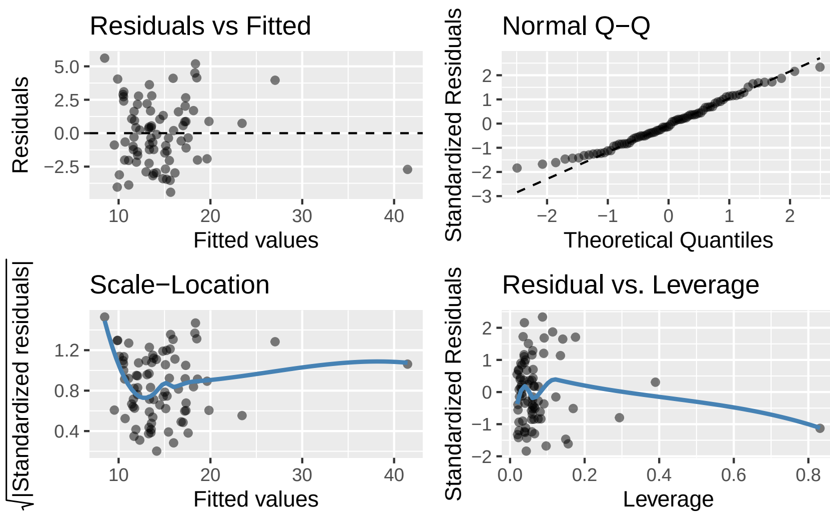
```
M2 <- lm(((MarketValue^lambda - 1)/lambda) ~ log(Assets) + log(Sales) + Profits
         + CashFlow + log(Employees), data=data)
summary(M2)
```

```
##
## Call:
## lm(formula = ((MarketValue^lambda - 1)/lambda) ~ log(Assets) +
##     log(Sales) + Profits + CashFlow + log(Employees), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4181 -1.9622 -0.2923  1.6446  5.6135
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.929556   3.671394   1.343  0.18353
## log(Assets)     0.382213   0.297842   1.283  0.20346
## log(Sales)      0.312235   0.628774   0.497  0.62098
## Profits        -0.001207   0.002829  -0.427  0.67082
## CashFlow        0.002967   0.001954   1.519  0.13317
## log(Employees)  1.271800   0.438375   2.901  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.484 on 73 degrees of freedom
## Multiple R-squared:  0.7715, Adjusted R-squared:  0.7559
## F-statistic:  49.3 on 5 and 73 DF,  p-value: < 2.2e-16
```

(j)

Check the assumption of the model M2

```
gglm(M2)
```



Regression assumptions

Linearity: The residuals are roughly randomly scattered about the zero line with the exception of some outliners in the residual versus fitted values plot. Hence, the linearity assumption is close to reasonable.

Homoscedasticity: The spread about the zero line appears roughly constant with the exception of some outliners in the scale-location plot. Hence, the assumption of constant variance is close to reasonable.

Normality: There is some departure from normality in both tails of the distribution of residuals. However, the majority of the data is close to normally distributed. Hence, normality assumption is reasonable.

Independence: The plots can not verify this assumption.

(k)

I would prefer model M2 over M1, since the regression assumptions (Linearity, Homoscedasticity and Normality) of M2 is more reasonable then M1.

(l)

Create a tibble with the data for the new company

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

Obtain the 95% prediction interval for transformed MarketValue of the new company, using the M2 model. Then find the 95% prediction interval for the original MarketValue of the new company

```
transform_pred <- predict(M2, newdata=new_company, interval="prediction", level=0.95)
market_pred <- exp(log(transform_pred*lambda + 1)/lambda)
market_pred
```

```
##        fit      lwr      upr
## 1 471.5377 67.41586 1978.406
```

The 95% prediction interval for the MarketValue of the new company: (67.42, 1978.41) in millions