Week 6 Assignment

2023-06-26

```
library(DescTools)
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(readr)
week_6_data <- read_csv("Downloads/week 6 data.csv")
```

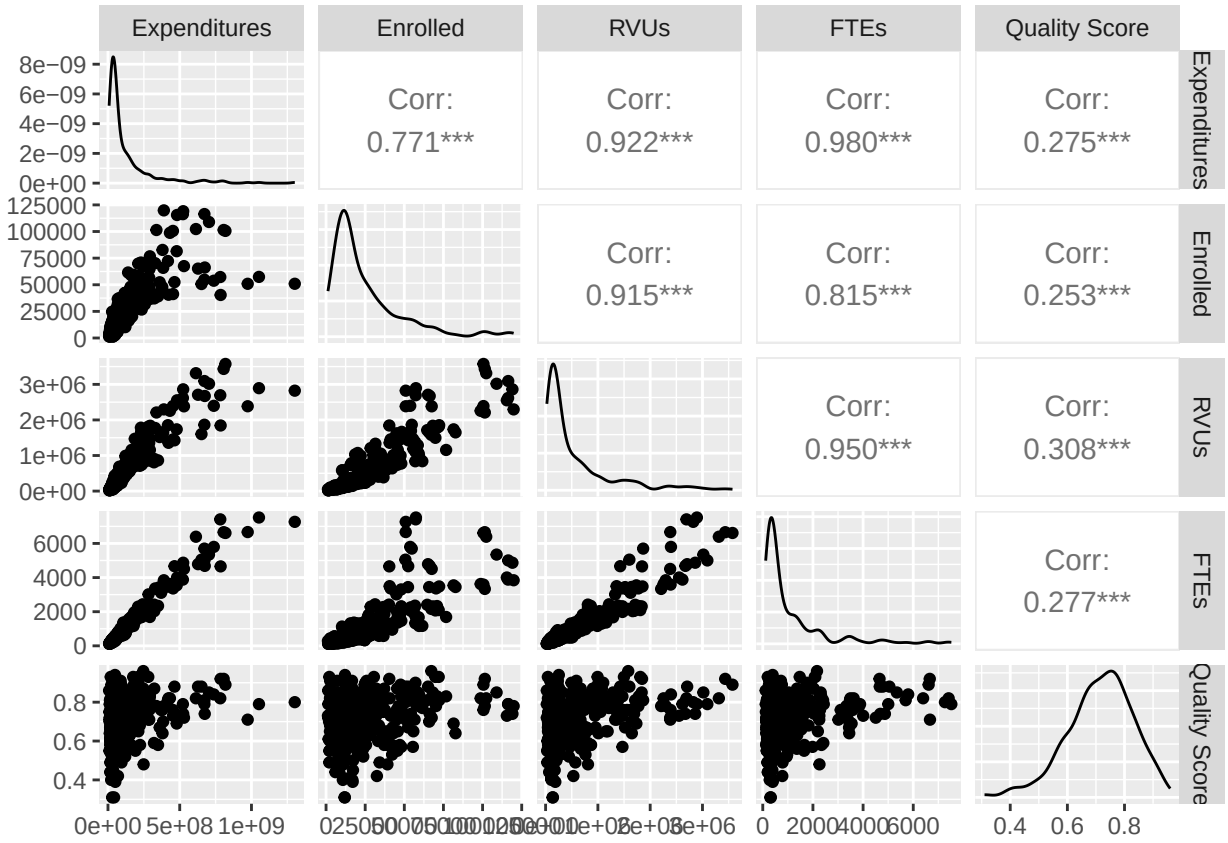
```
## Rows: 384 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (5): Expenditures, Enrolled, RVUs, FTEs, Quality Score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Correlation

```
mydata <- week_6_data
cor(mydata)
```

```
##               Expenditures  Enrolled      RVUs      FTEs Quality Score
## Expenditures     1.0000000 0.7707756 0.9217239 0.9796506     0.2749501
## Enrolled         0.7707756 1.0000000 0.9152024 0.8148491     0.2526991
## RVUs             0.9217239 0.9152024 1.0000000 0.9504093     0.3075742
## FTEs             0.9796506 0.8148491 0.9504093 1.0000000     0.2769058
## Quality Score    0.2749501 0.2526991 0.3075742 0.2769058     1.0000000
```

```
ggpairs(mydata)
```



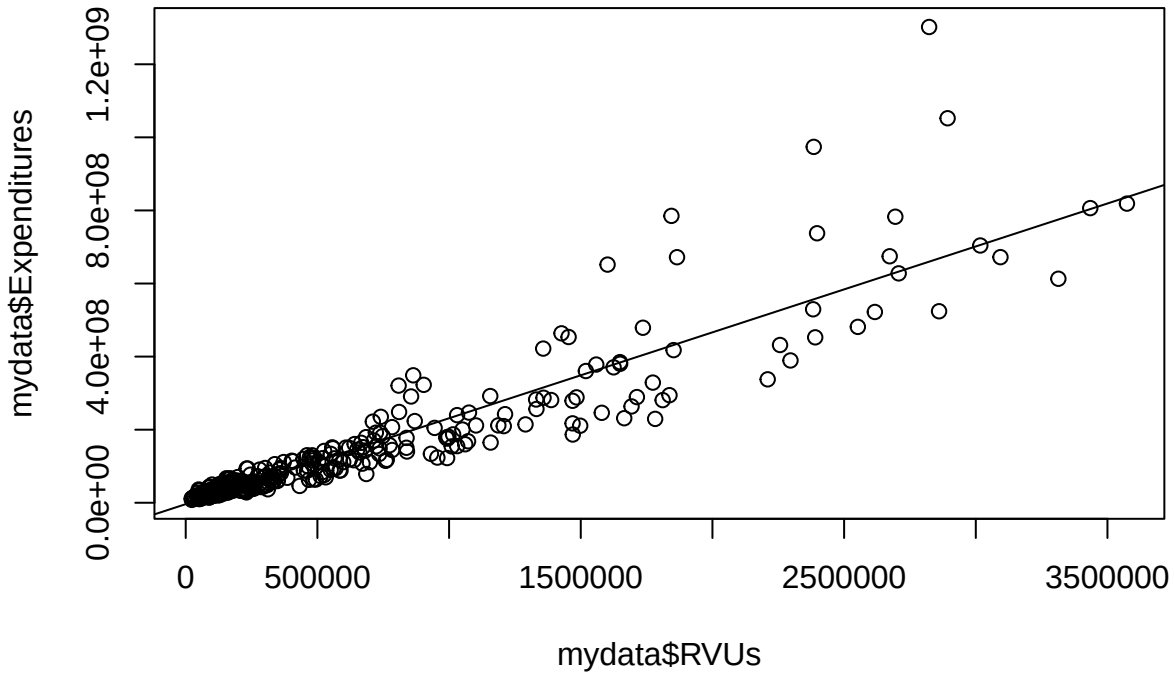
Interpretation: The relationship between RVUs and other variables has strong positive correlations except for Quality Score. The correlation between RVUs and Quality Score is 0.3076 which is considered weak and positively correlated.

Linear Model 1: Expenditures~RVUs

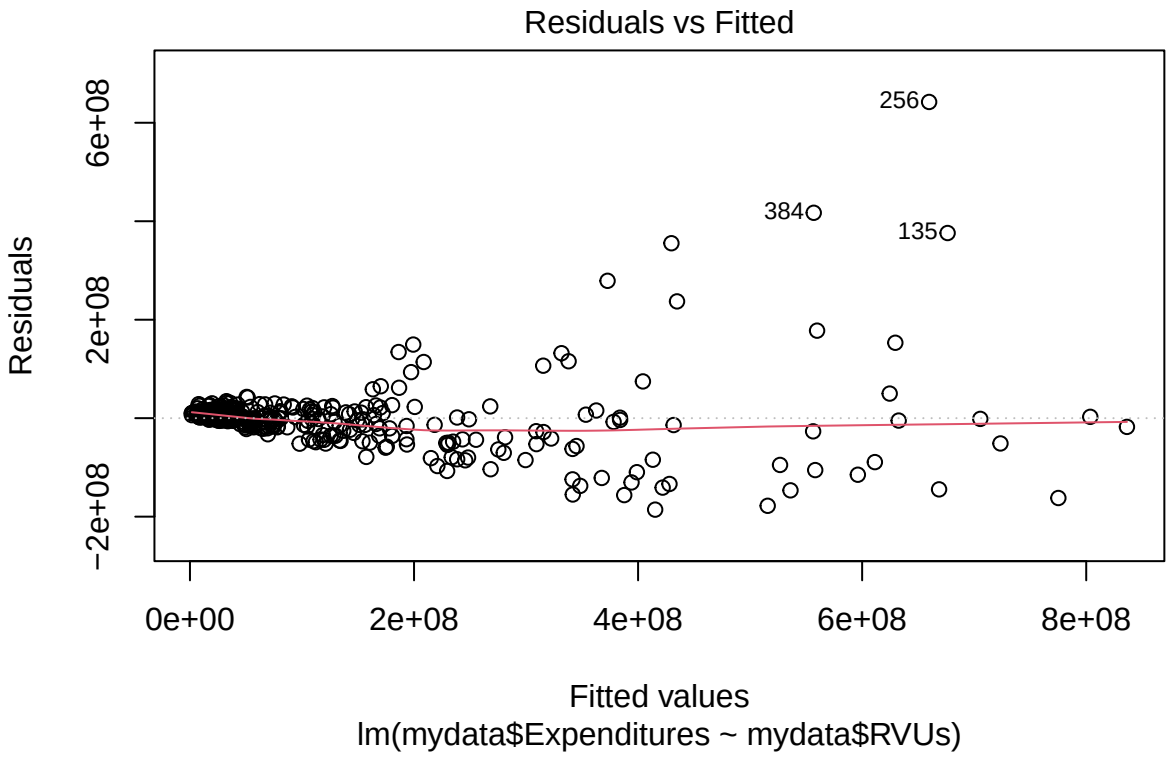
```
model1 <- lm(mydata$Expenditures~mydata$RVUs)
summary(model1)
```

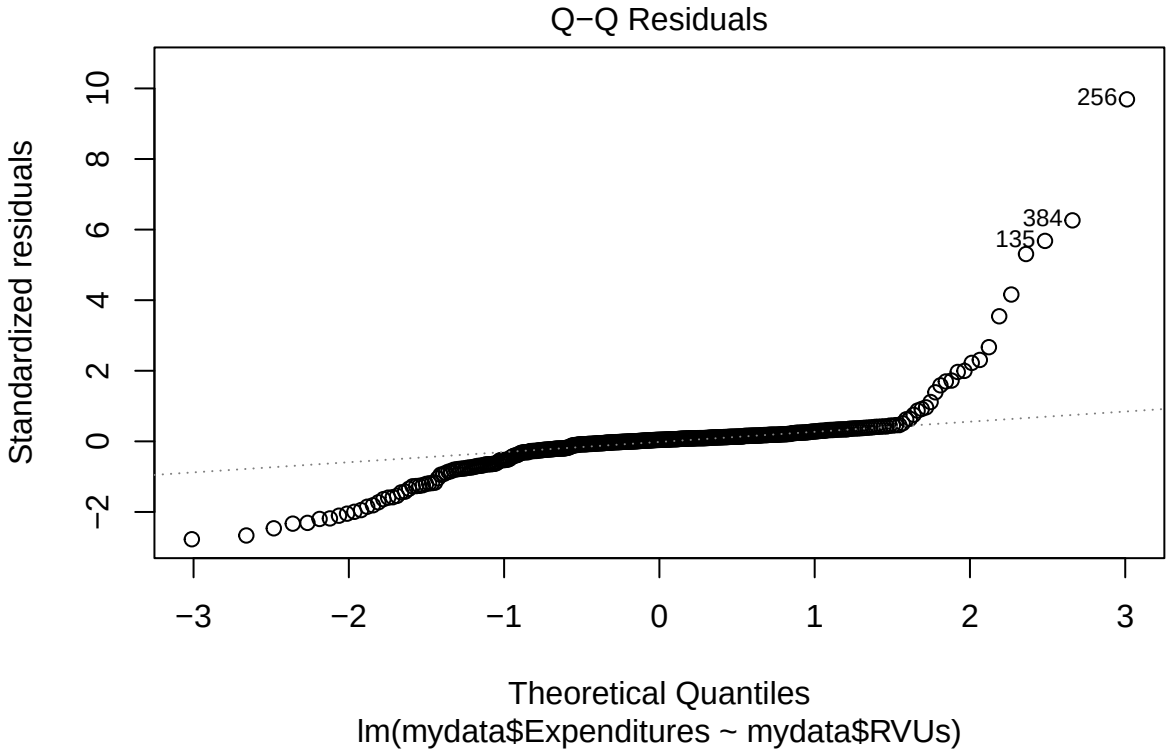
```
##
## Call:
## lm(formula = mydata$Expenditures ~ mydata$RVUs)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -185723026  -14097620    2813431   11919781  642218316
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.785e+06  4.413e+06  -0.858    0.392
## mydata$RVUs  2.351e+02  5.061e+00  46.449   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67350000 on 382 degrees of freedom
## Multiple R-squared:  0.8496, Adjusted R-squared:  0.8492
## F-statistic:  2157 on 1 and 382 DF,  p-value: < 2.2e-16
```

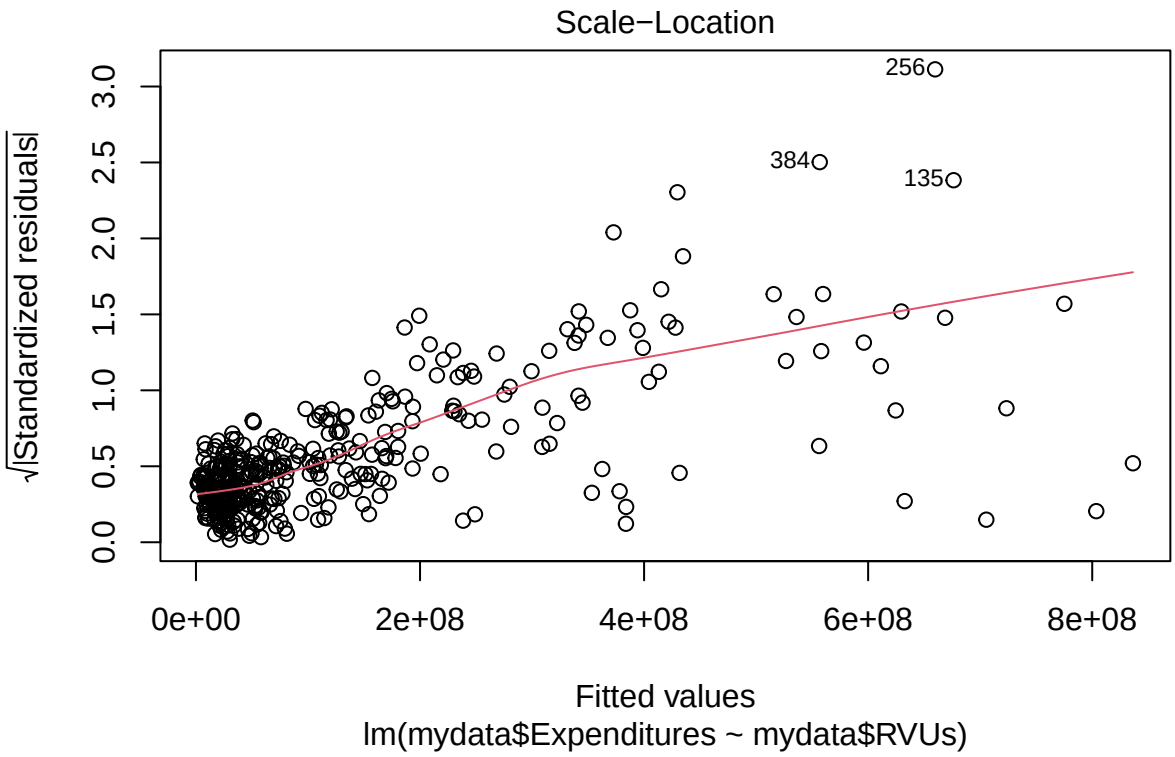
```r
plot(mydata$Expenditures~mydata$RVUs)
abline(model1) # create a scatter plot with the slope
```

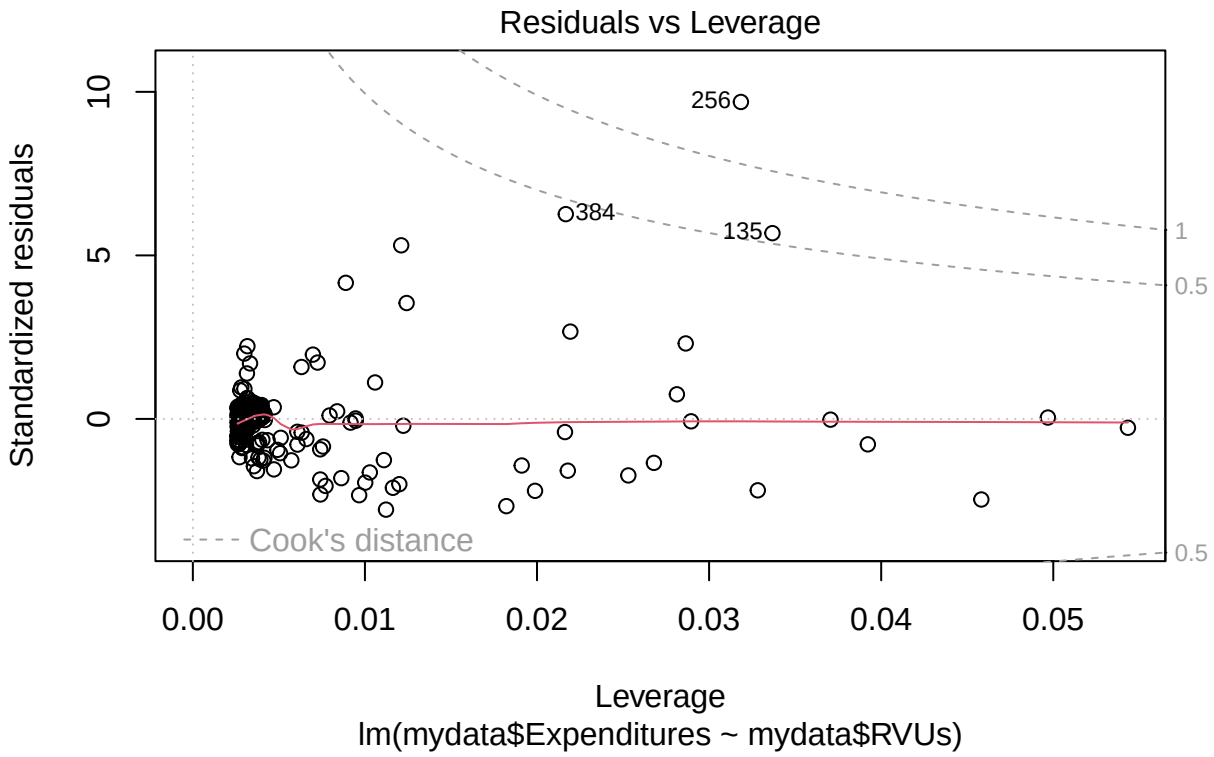



```r
plot(model1)
```

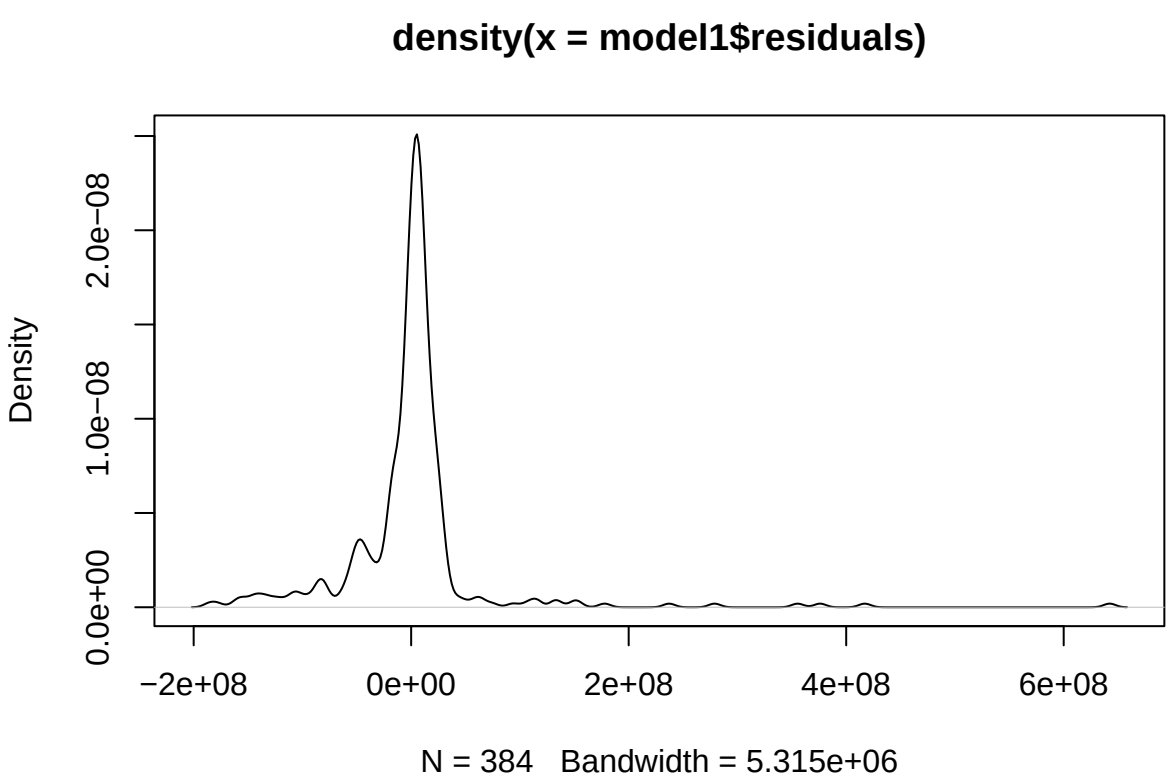

Residuals vs Fitted
Residuals
6e+08
2e+08
−2e+08
256
384
135
0e+00
2e+08
4e+08
6e+08
8e+08
Fitted values
lm(mydata$Expenditures ~ mydata$RVUs)


Q−Q Residuals
Standardized residuals
10
8
6
4
2
0
−2
256
384
135
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mydata$Expenditures ~ mydata$RVUs)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
256
384
135
0e+00
2e+08
4e+08
6e+08
8e+08
Fitted values
lm(mydata$Expenditures ~ mydata$RVUs)




```
plot(density(model1$residuals))
```



Interpretation:

1)With the median of the residuals being far away from 0, we know that the distribution is not symmetrical and it is slightly left-skewed.

2)With the estimate of intercept being negative, it means that the model is overestimating on average the y values.

3)With a coefficient (2.351e+02) larger than the standard error of coefficient (5.061e+00), we can conclude that the coefficient is significant to this model.

4)With Pr(>|t|) of < 0.05, we can conclude that the coefficient is significant to this model. It means that expenditures has a statistically significant relationship with RVUs

5) Residuals Standard Erros shows that the average amount of the actual values of Y differ from the predictions in units of Y is large with 6735000 which means that the precision of the model's prediction is not low.

6) R-squared value of this model is 0.8496 which indicates that 84.96% of the variance is explained by this model. With $0.8496 > 0.5$, we can conclude that this model fits the data well.

Assumptions :

1. Level of Measurement: Two variables are measured on a ratio or scale level
2. Linear Relationship: The relationship between two variables is linear
3. Normality: According to the density plot, the data is not normally distributed.

4. Homogeneity of residuals variance: According to “Scale-Location” plot, we can conclude that there is heteroscedasticity.

5. No Outliers: According to “Residuals vs Leverage” plot, there are extreme outliers.

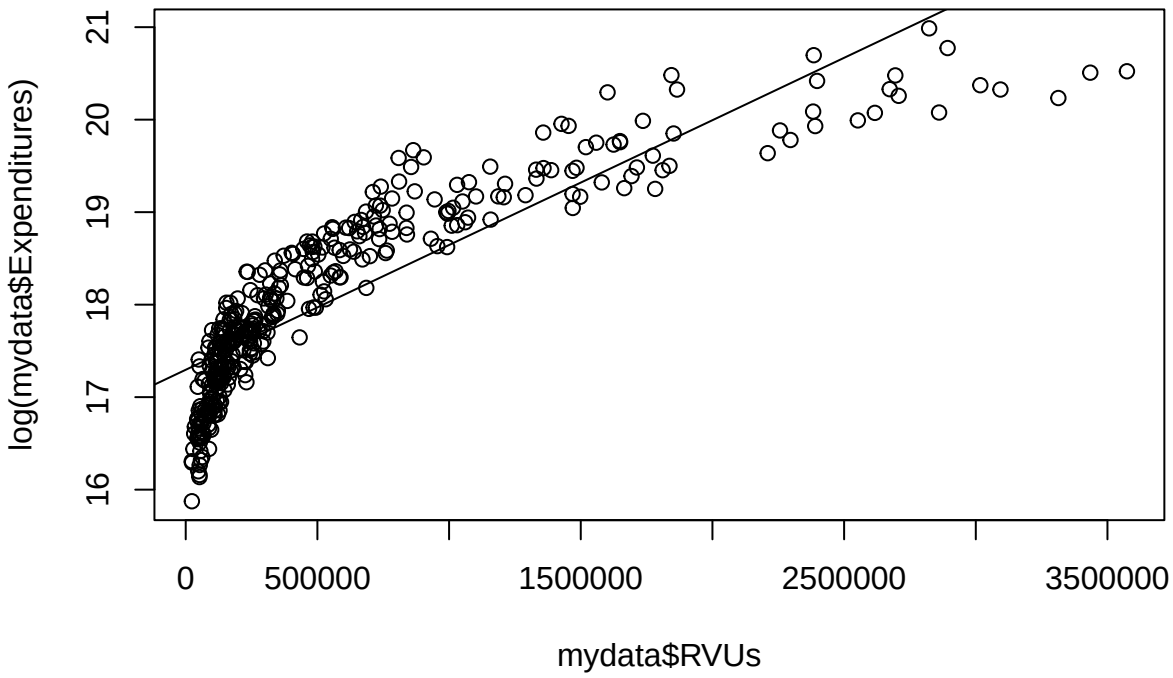
Conclusion: Overall the assumptions do not hold.

Linear model 2 : ln(Expensitures) ~ RVUs

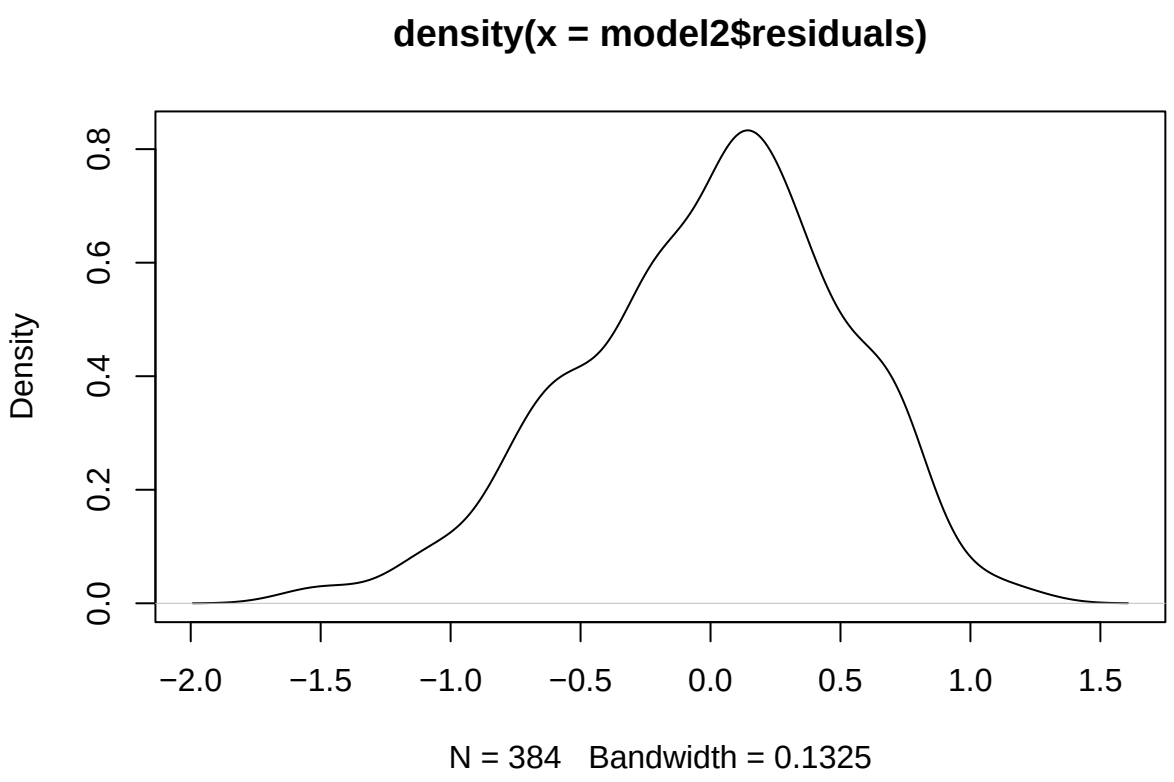
```
model2 <- lm(log(mydata$Expenditures)~ mydata$RVUs)
summary(model2)
```

```
##
## Call:
## lm(formula = log(mydata$Expenditures) ~ mydata$RVUs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59439 -0.29504  0.06135  0.35333  1.20871
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.730e+01  3.325e-02  520.11   <2e-16 ***
## mydata$RVUs 1.349e-06  3.814e-08   35.38   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5076 on 382 degrees of freedom
## Multiple R-squared:  0.7661, Adjusted R-squared:  0.7655
## F-statistic:  1251 on 1 and 382 DF,  p-value: < 2.2e-16
```

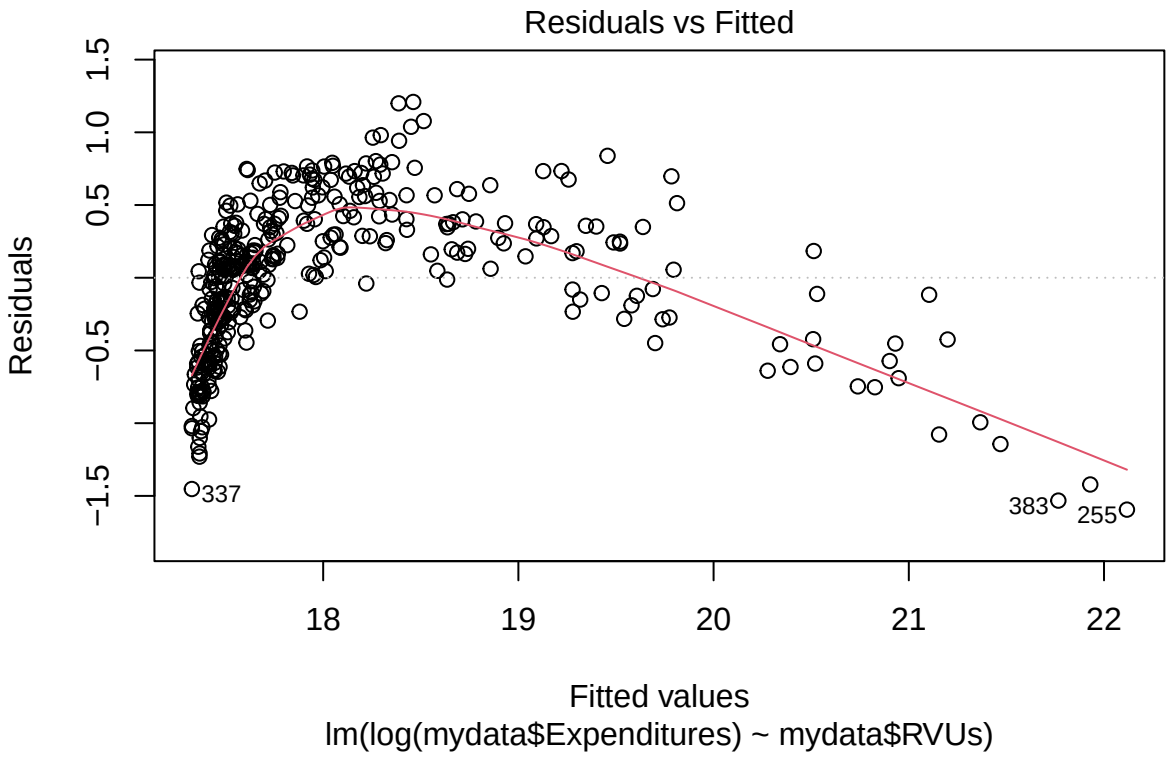
```
plot(log(mydata$Expenditures)~ mydata$RVUs)
abline(model2)
```

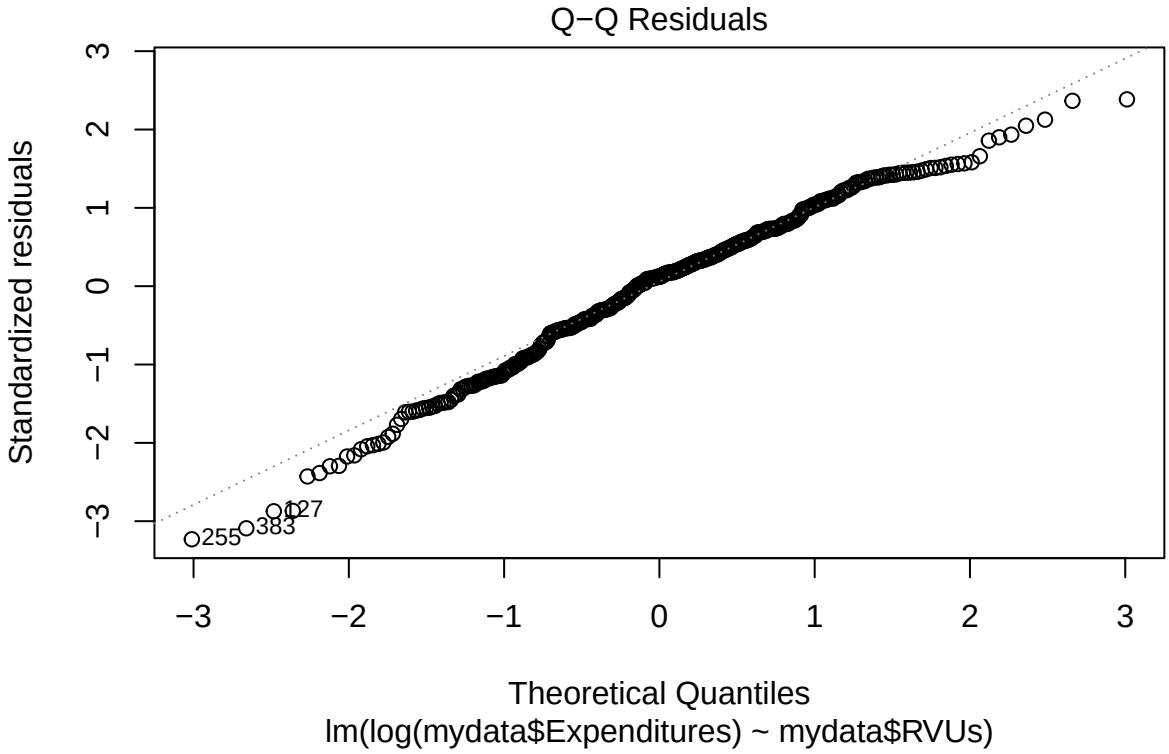


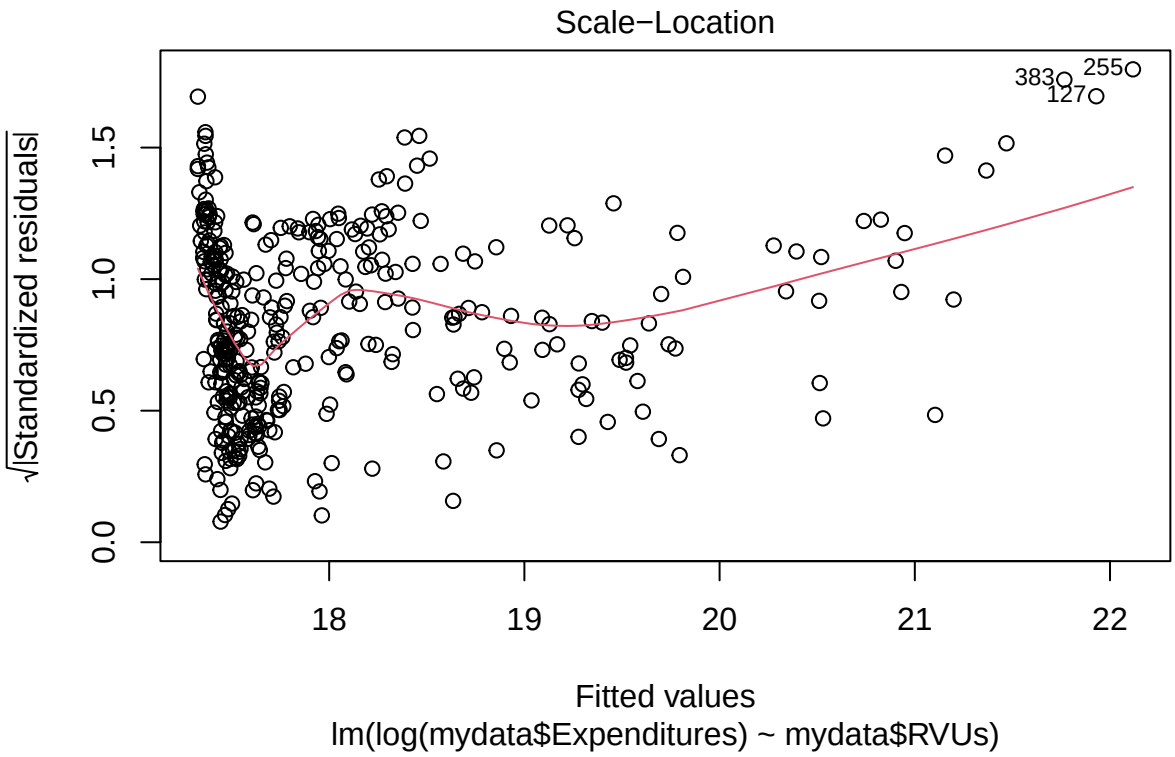

```
plot(density(model2$residuals))
```

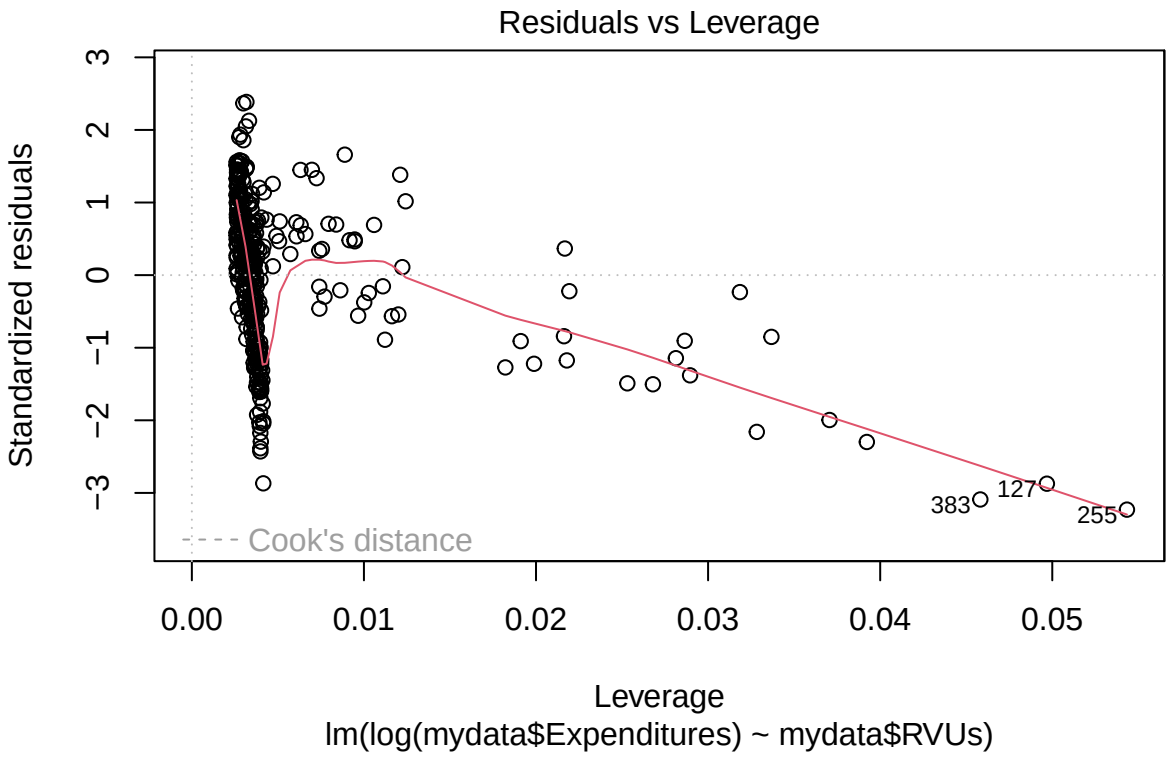

density(x = model2$residuals)
Density
0.0
0.2
0.4
0.6
0.8
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
N = 384 Bandwidth = 0.1325

```
plot(model2)
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
−0.5
−1.5
337
383
255
18
19
20
21
22
Fitted values
lm(log(mydata$Expenditures) ~ mydata$RVUs)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(log(mydata$Expenditures) ~ mydata$RVUs)
255
383
127


Scale−Location
383
255
127
√|Standardized residuals|
0.0
0.5
1.0
1.5
18
19
20
21
22
Fitted values
lm(log(mydata$Expenditures) ~ mydata$RVUs)



Interpretation:

1)With the median of the residuals being closer to 0, we know that the distribution is symmetrical.

2)With a coefficient (1.349e-06) larger than the standard error of coefficient (3.814e-08), we can conclude that the coefficient is significant to this model.

3)With $Pr(>|t|) < 0.05$, we can conclude that the coefficient is significant to this model.

4) Residuals Standard Erros shows that the average amount of the actual values of Y differ from the predictions in units of Y is small with a 0.5076 which means that the precision of the model's prediction is considered high.

5)R-squared value of this model is 0.7661 which indicates that 76.61% of the variance is explained by this model. With $0.7661 > 0.5$, we can conclude that this model fits the data well.

Assumptions:

1. Level of Measurement: Two variables are measured on a ratio or scale level
2. Linear Relationship: The relationship between two variables is non linear
3. Normality: According to the density plot, the data is overall normally distributed.
4. Homogeneity of residuals variance: According to "Scale-Location" plot, we can conclude that there is homoscedasticity.
5. No Outliers: According to "Residuals vs Leverage" plot, there is no extreme outliers.

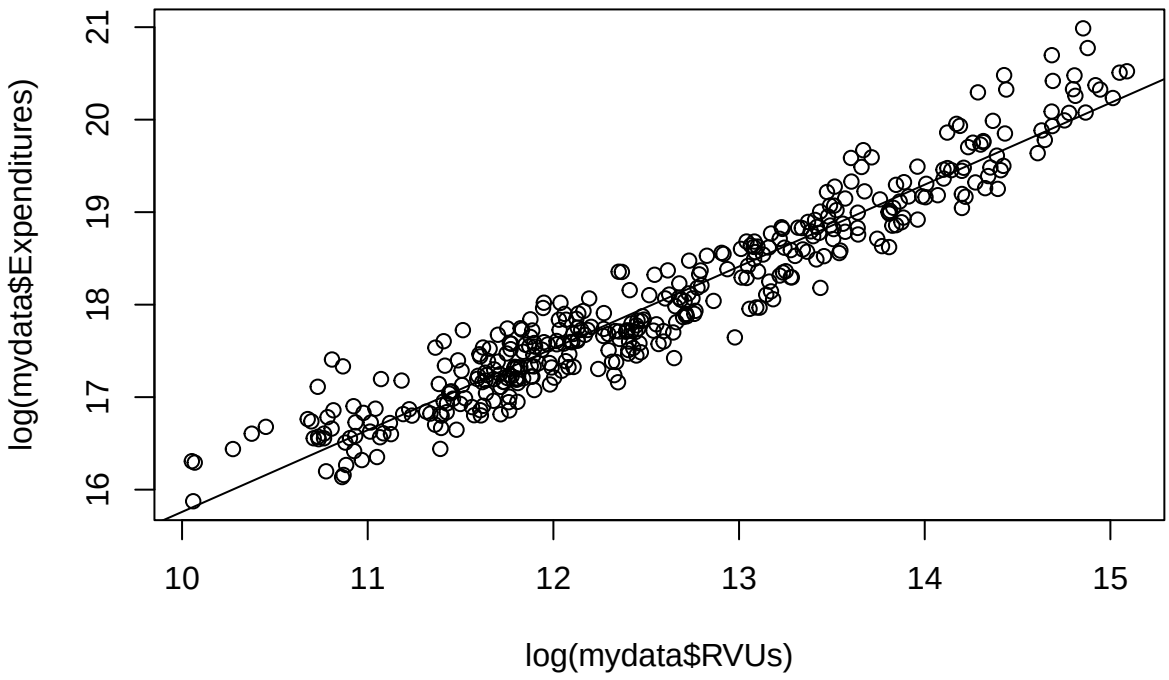
Conclusion: Overall the assumptions do not hold.

Linear model 3: ln(Expenditures)~ln(RVUs)

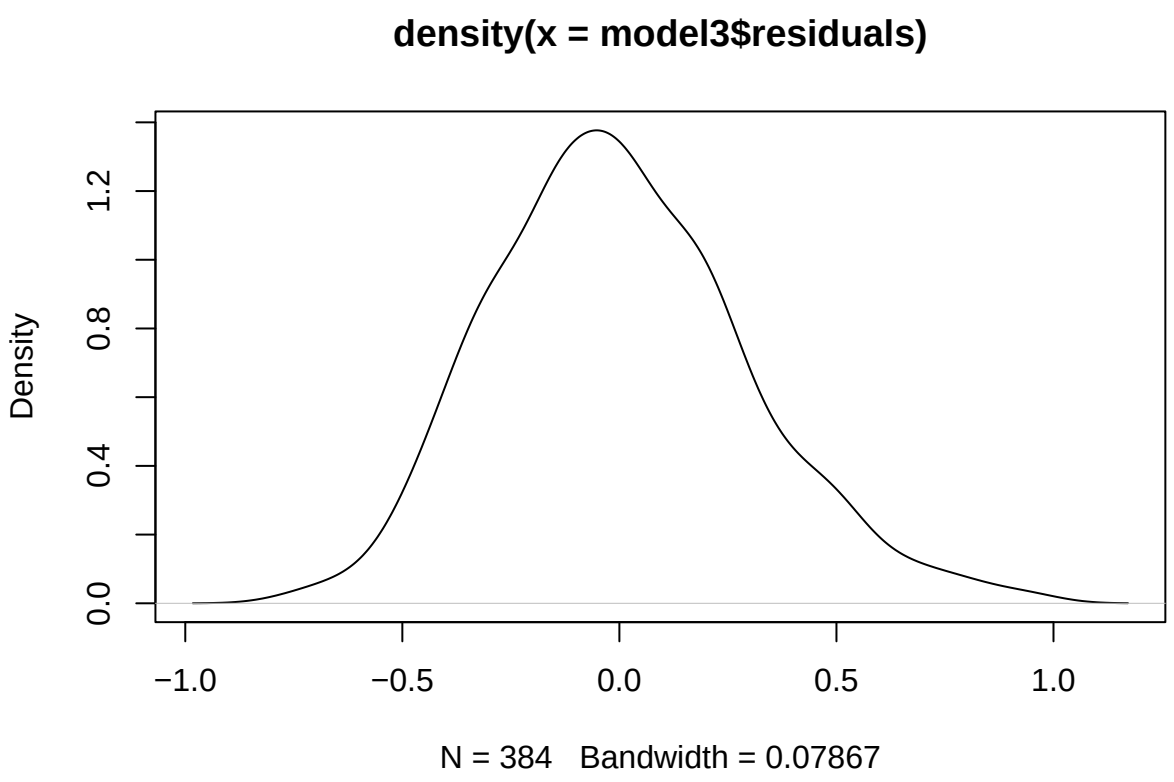
```
model3 <- lm(log(mydata$Expenditures)~log(mydata$RVUs))
summary(model3)
```

```
##
## Call:
## lm(formula = log(mydata$Expenditures) ~ log(mydata$RVUs))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74657 -0.19864 -0.02431  0.18642  0.93551
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.91487    0.16621   41.60   <2e-16 ***
## log(mydata$RVUs)  0.88444    0.01317   67.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2932 on 382 degrees of freedom
## Multiple R-squared:  0.9219, Adjusted R-squared:  0.9217
## F-statistic:  4512 on 1 and 382 DF,  p-value: < 2.2e-16
```

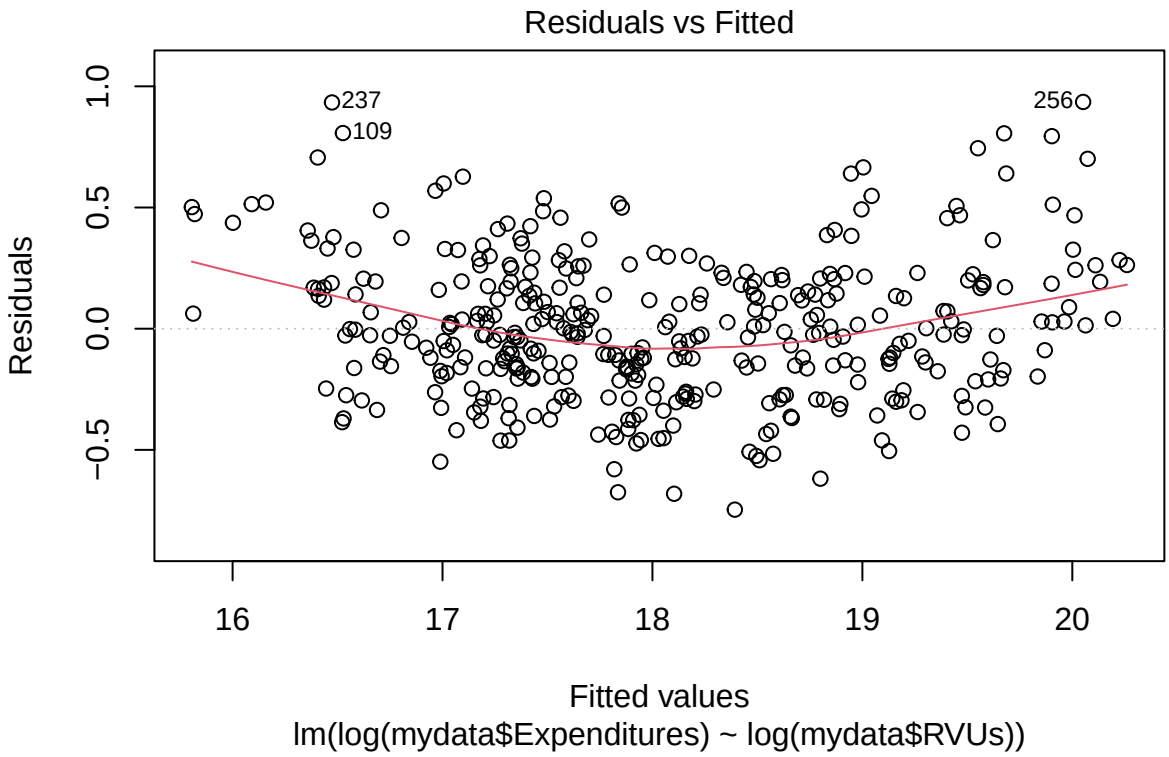
```
plot(log(mydata$Expenditures)~log(mydata$RVUs))
abline(model3)
```

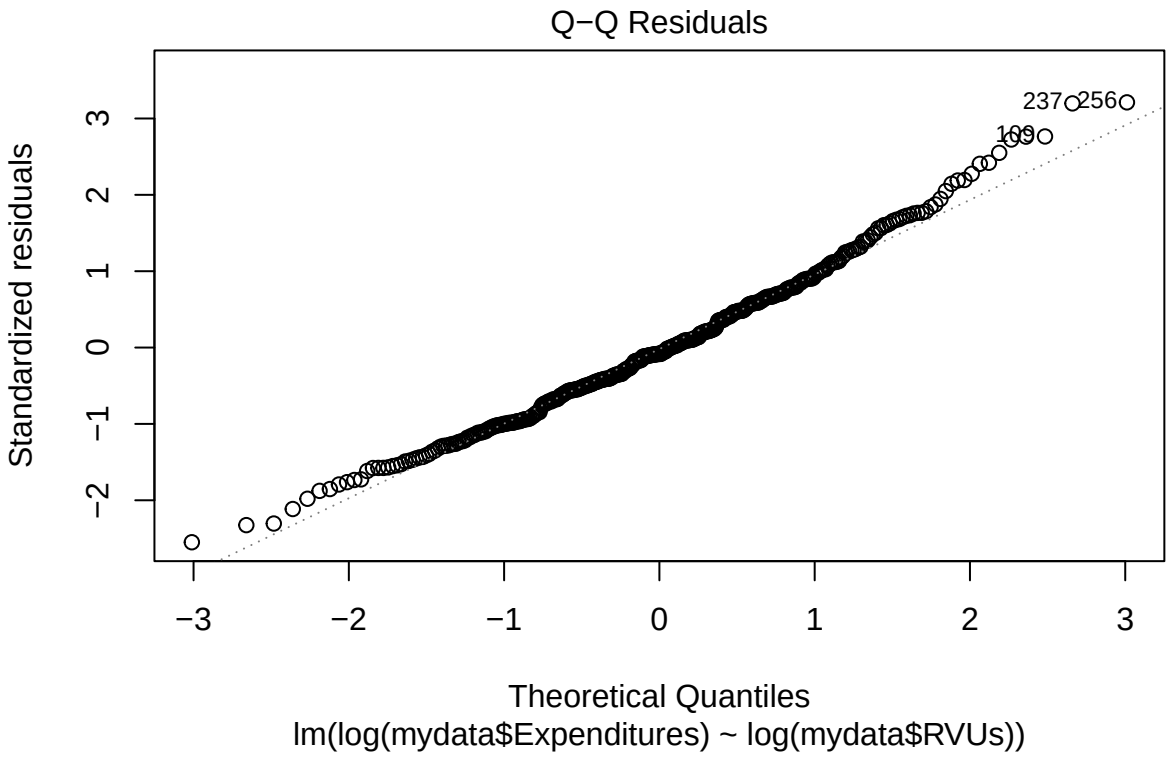


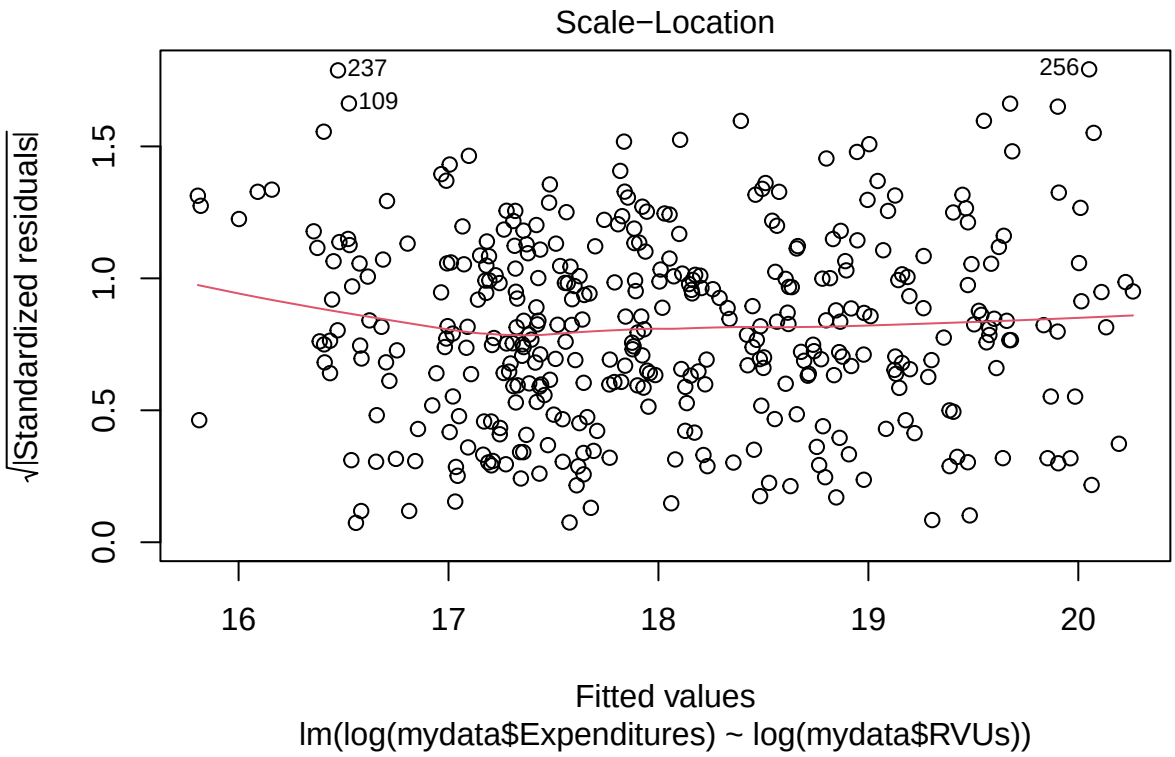

```
plot(density(model3$residuals))
```

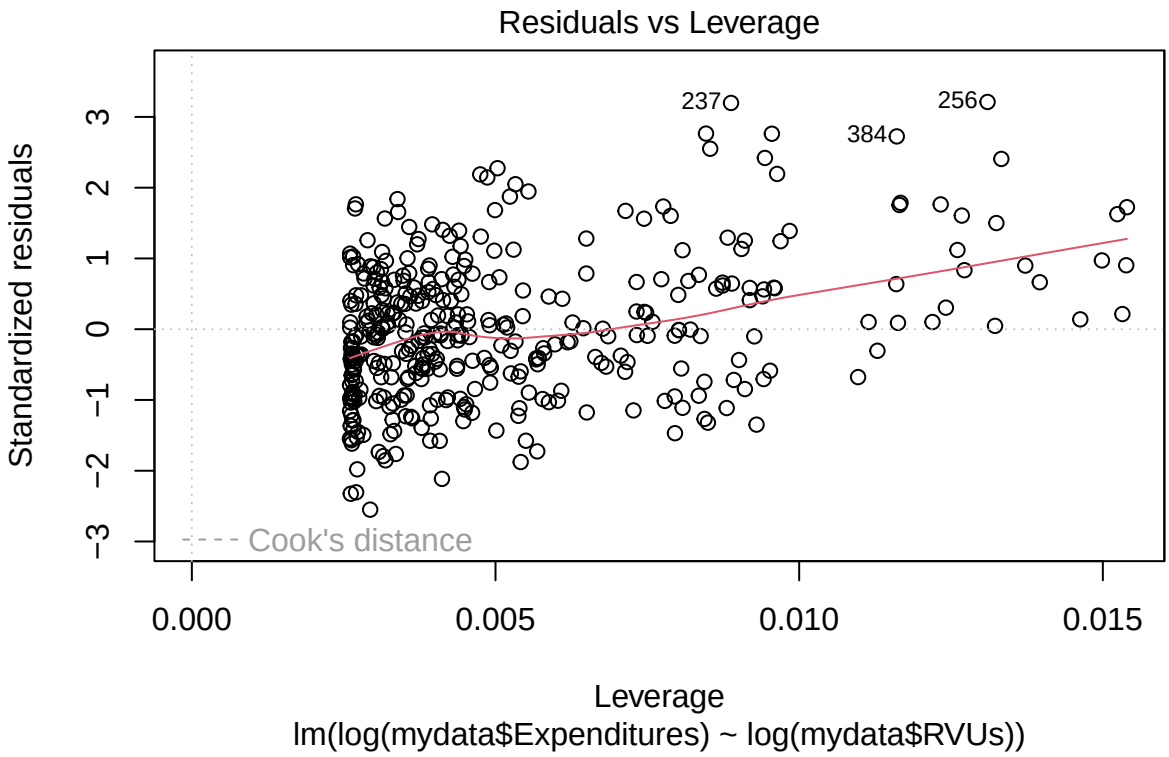

density(x = model3$residuals)
Density
0.0
0.4
0.8
1.2
−1.0
−0.5
0.0
0.5
1.0
N = 384 Bandwidth = 0.07867

```
plot(model3)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
237
109
256
16
17
18
19
20
Fitted values
lm(log(mydata$Expenditures) ~ log(mydata$RVUs))


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(log(mydata$Expenditures) ~ log(mydata$RVUs))
237
256
−3
−2
−1
0
1
2
3


Scale−Location
237
109
256
√|Standardized residuals|
0.0
0.5
1.0
1.5
16
17
18
19
20
Fitted values
lm(log(mydata$Expenditures) ~ log(mydata$RVUs))



Effect of Transformation:

1)By comparing the density plot of each model, we can conclude that the normality of the distribution is increasingly closer to normal.

2) The effect of Transformation increases the value of R-squared, which means that the fitness of the model increased.

3) By transforming the dependent variable, the linear relationship changed to non-linear relationship. On the other hand, transforming both dependent and independent variables, the relationship changed from non linear to closer to linear.

4) The residual standard error increased from the transformation which indicates that the precision of the model's prediction increased.

5) Extreme outliers do not exist after the transformation.