705604096_stats101a_hw6

Jade Gregory

2023-05-11

Question 1

Yes, the ordinary straight line regression model does seem to fit the data well. This is evident in the residuals plot, as there is no obvious observable pattern in the data points. The data in the residual plot is horizontally plotted, from which we can infer constant variance. The correlation coefficient, denoted by the r squared value, is fairly high as well, being roughly 99.4%. This is relatively high, meaning that the two variables are very correlated to one another. Also, in our summary we can observe the p-value of 2e-16 which is smaller than our chosen significance level of 0.05. From this we can conclude that the p-value supports the idea that the two variables are associated. From these statistics provided in our summary, we can conclude that the ordinary straight line regression model does fit the data well.

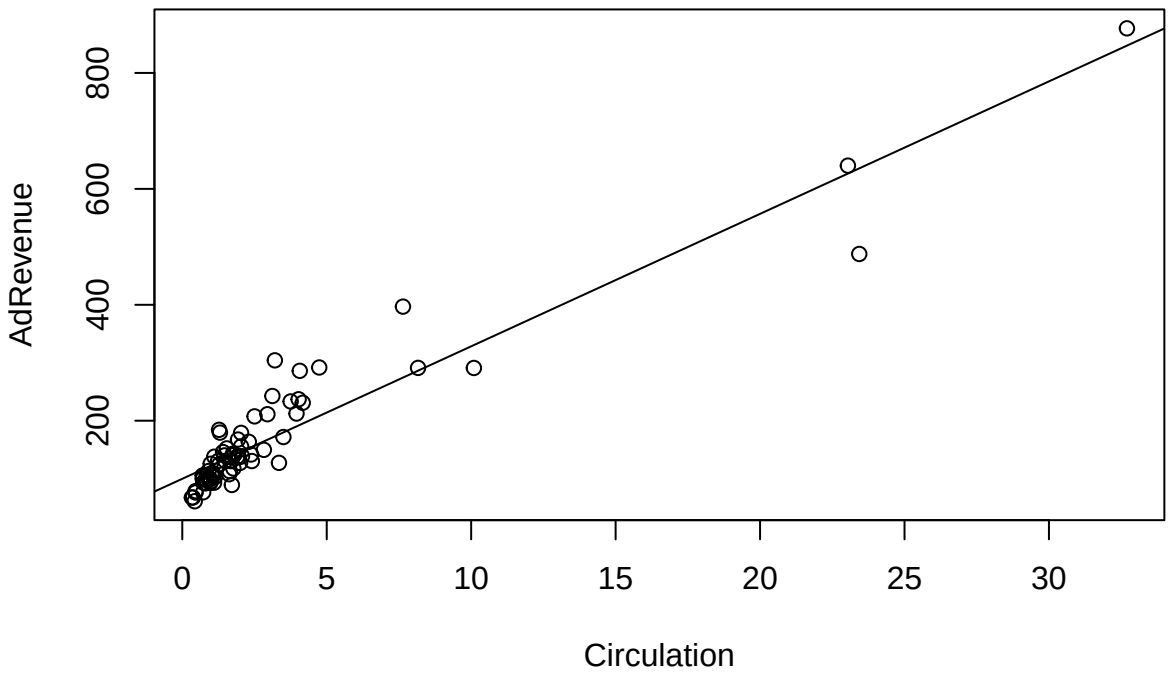
Question 2

```
revenue <- read.csv('AdRevenue.csv')
revenue1 <- revenue
```

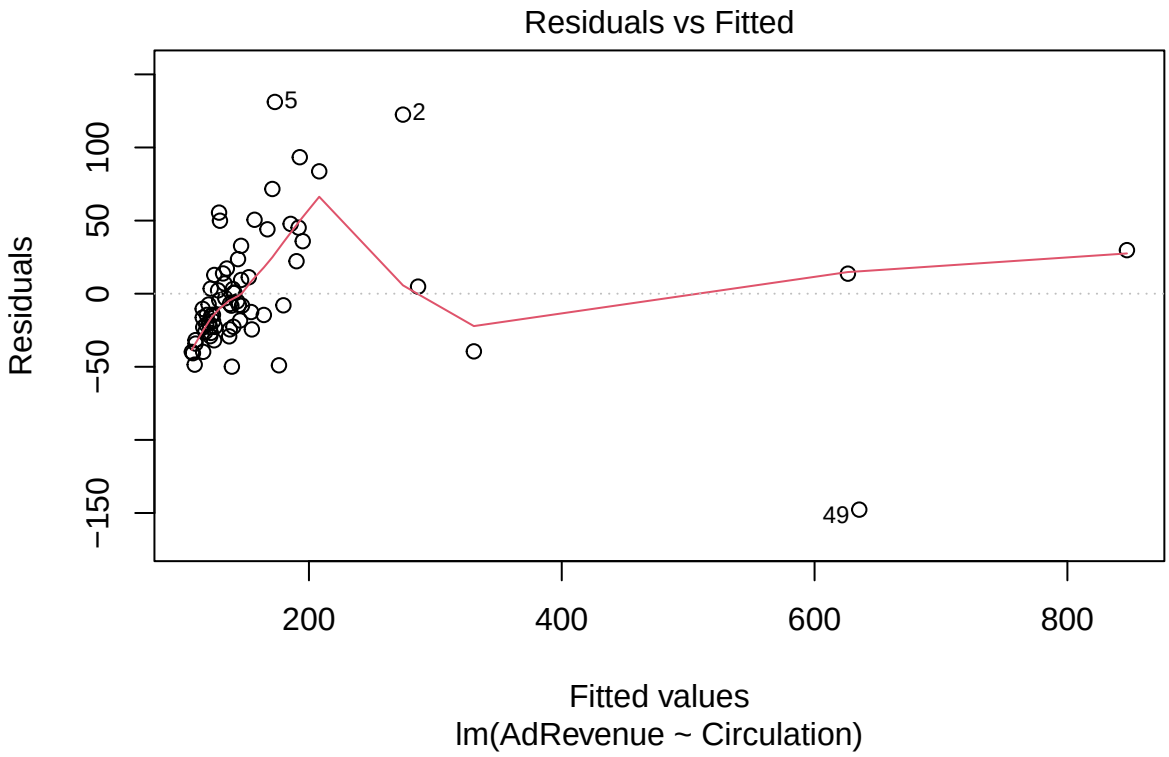
Part A

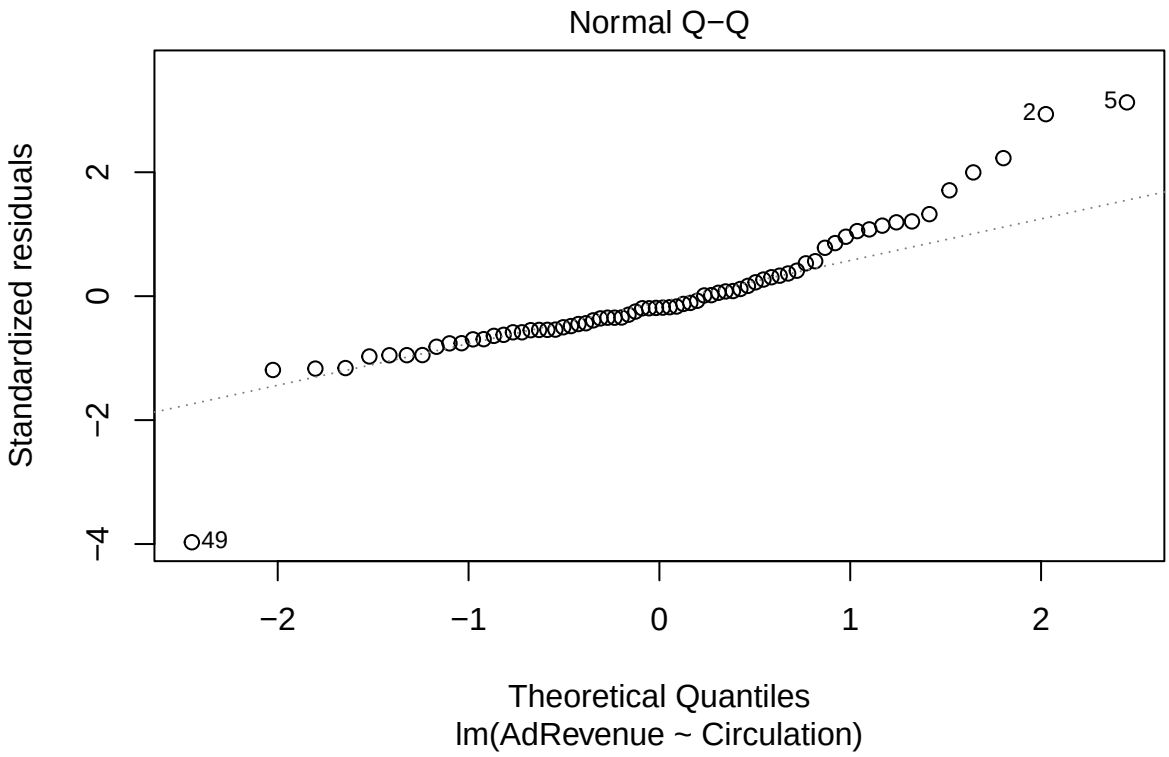
a)

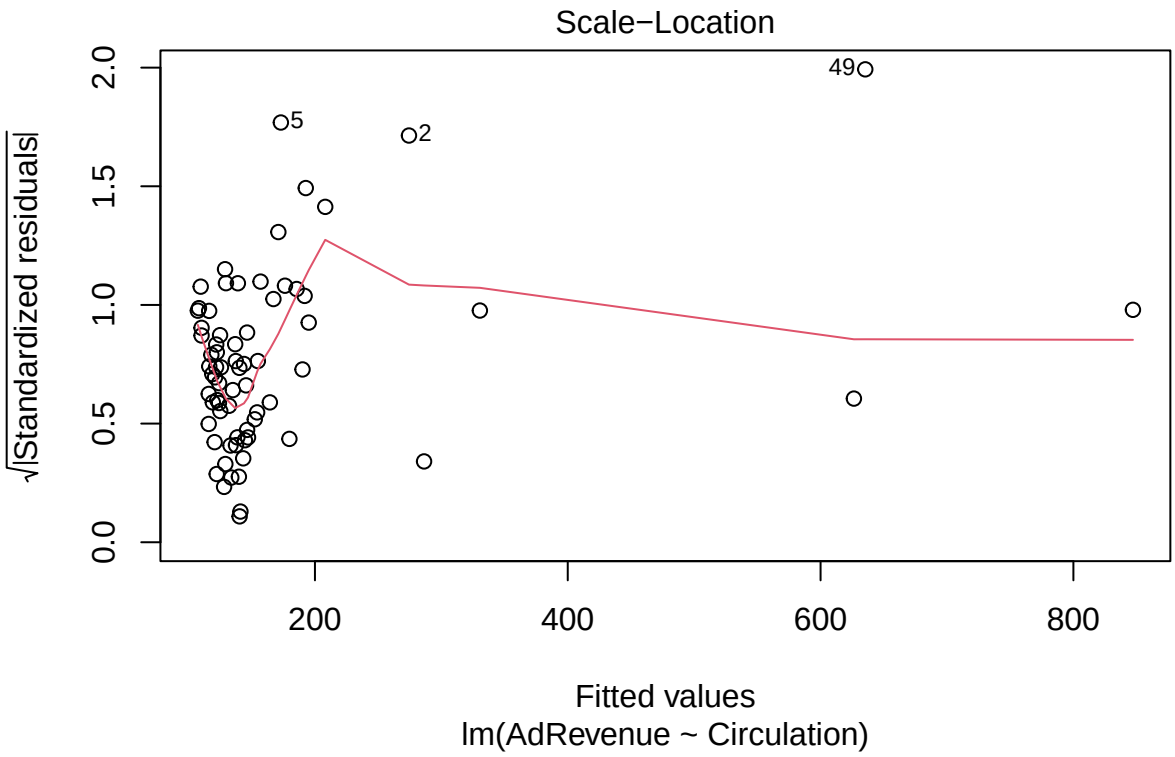
```
l.revenue <- transform(revenue1, lCirculation = log(Circulation),lAdRevenue = log(AdRevenue))
m.log1 <- lm(AdRevenue ~ Circulation, data = revenue1)
plot(AdRevenue ~ Circulation, data = revenue1)
abline(m.log1)
```

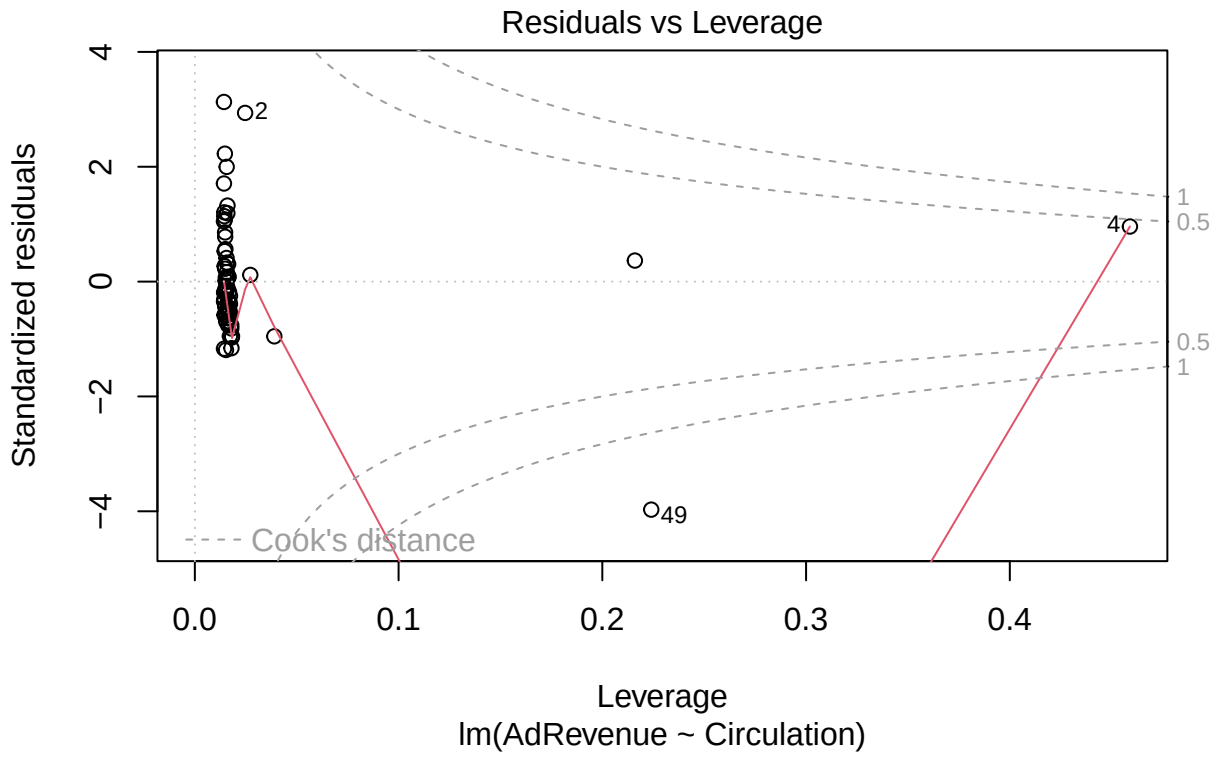



```
plot(m.log1)
```

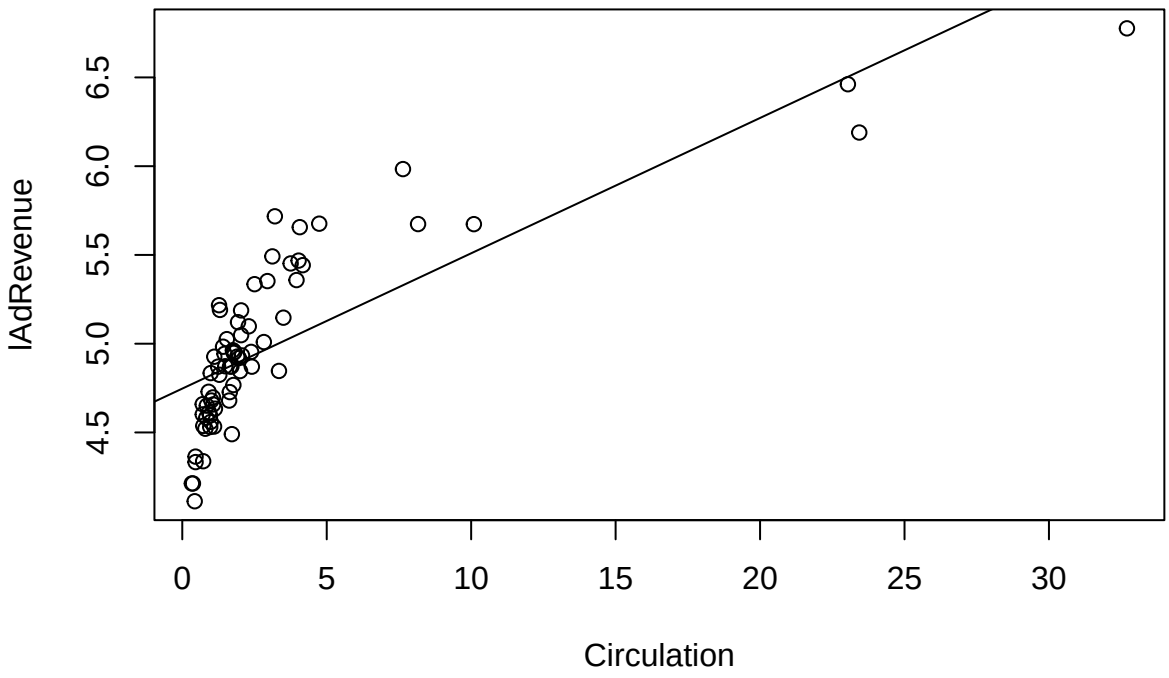

Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ Circulation)
5
2
49
100
50
0
−50
−150
200
400
600
800


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation)
2
5
49
−4
−2
0
2
−2
−1
0
1
2

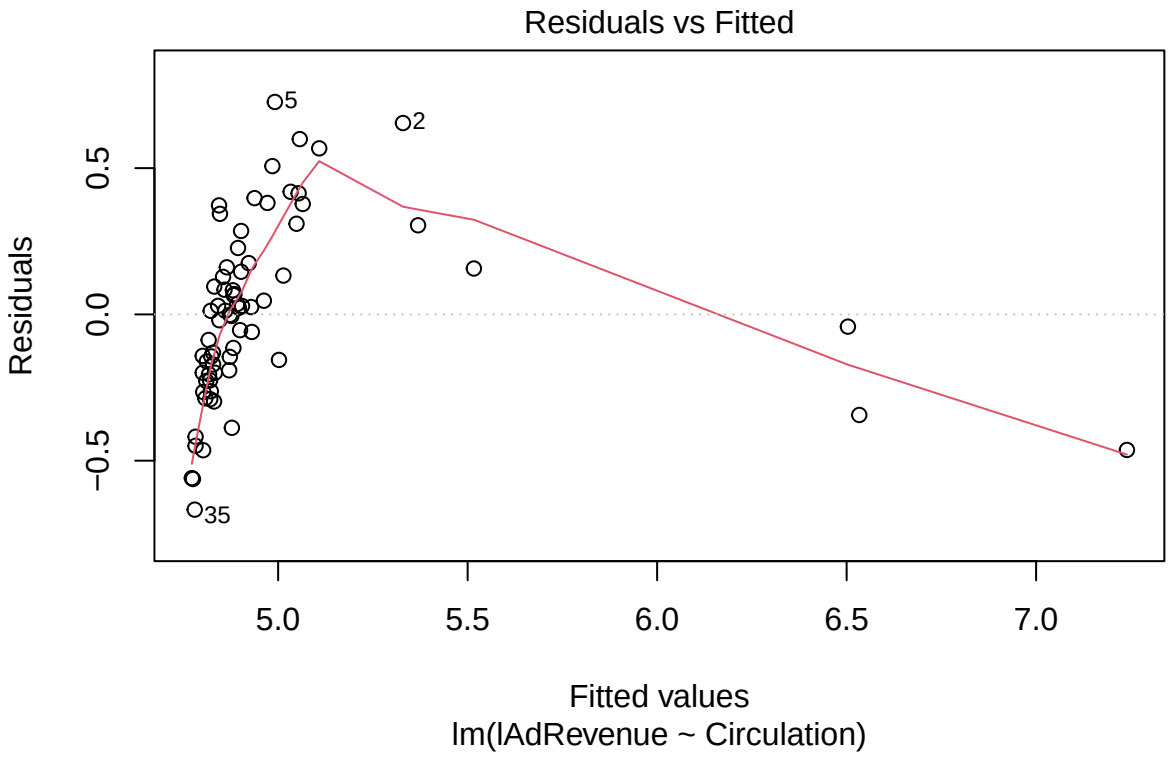

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
49
5
2
200
400
600
800
Fitted values
lm(AdRevenue ~ Circulation)

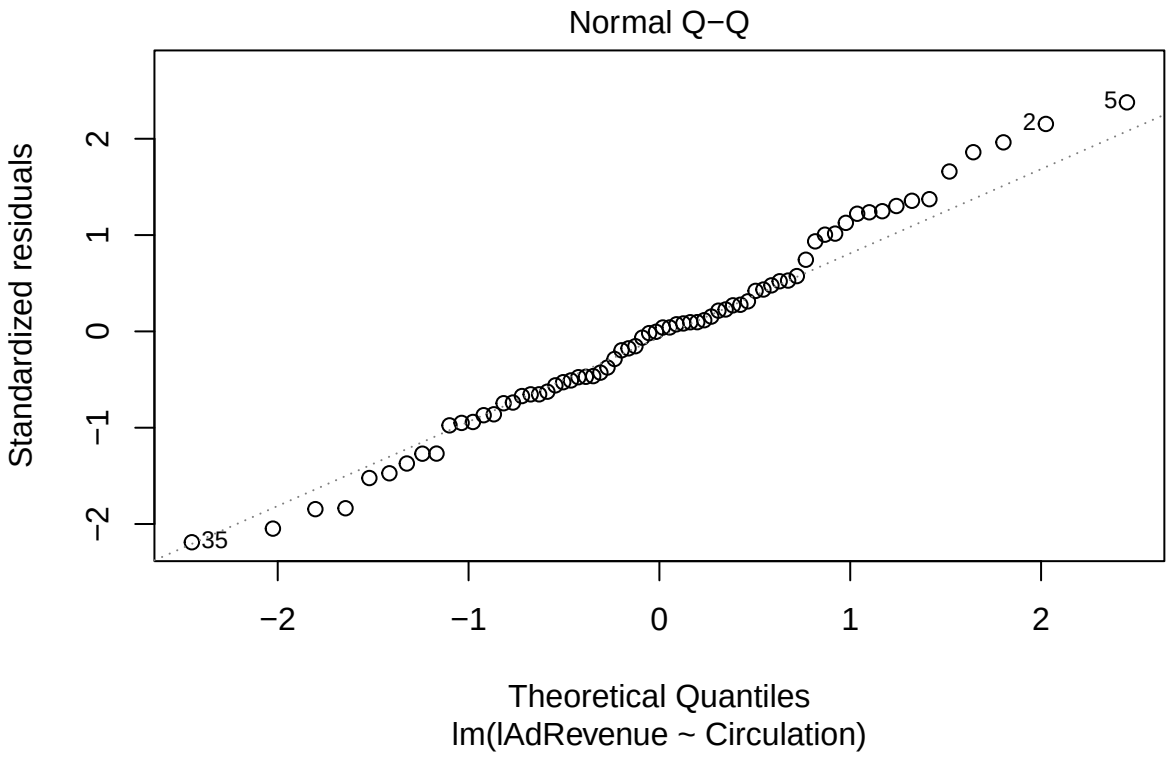


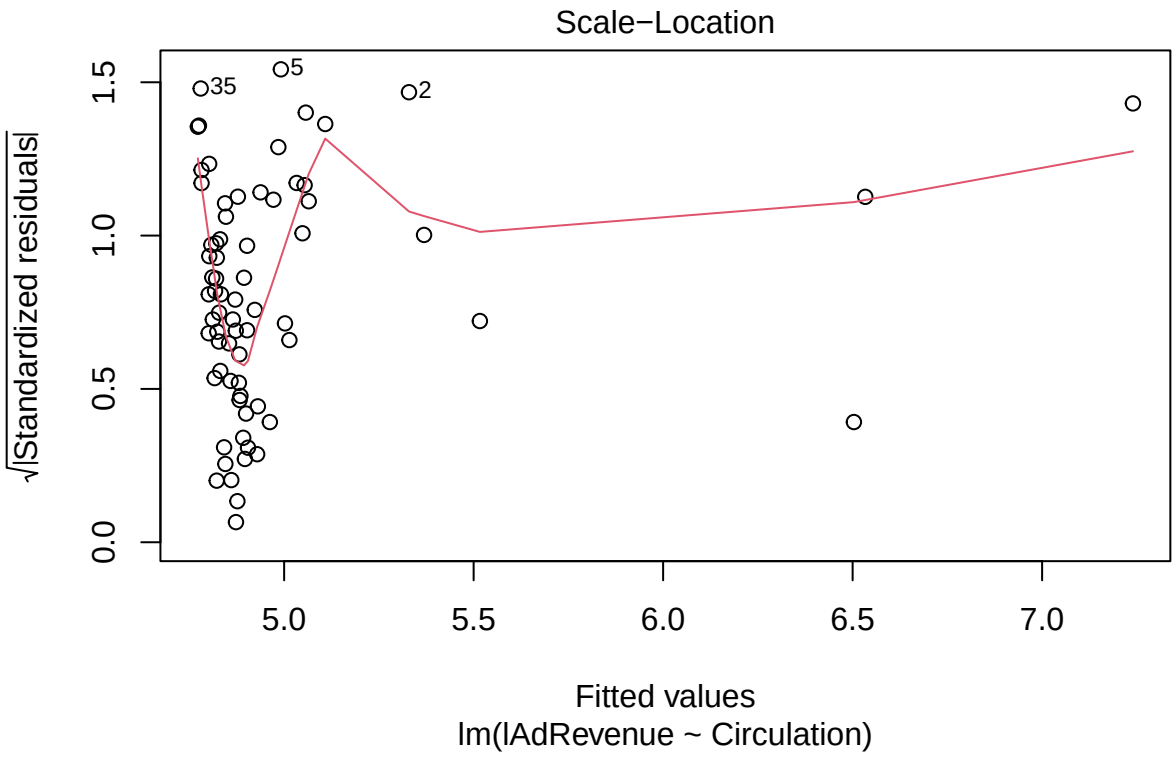

```
m.log2 <- lm(lAdRevenue ~ Circulation, data = l.revenue)
plot(lAdRevenue ~ Circulation, data = l.revenue)
abline(m.log2)
```

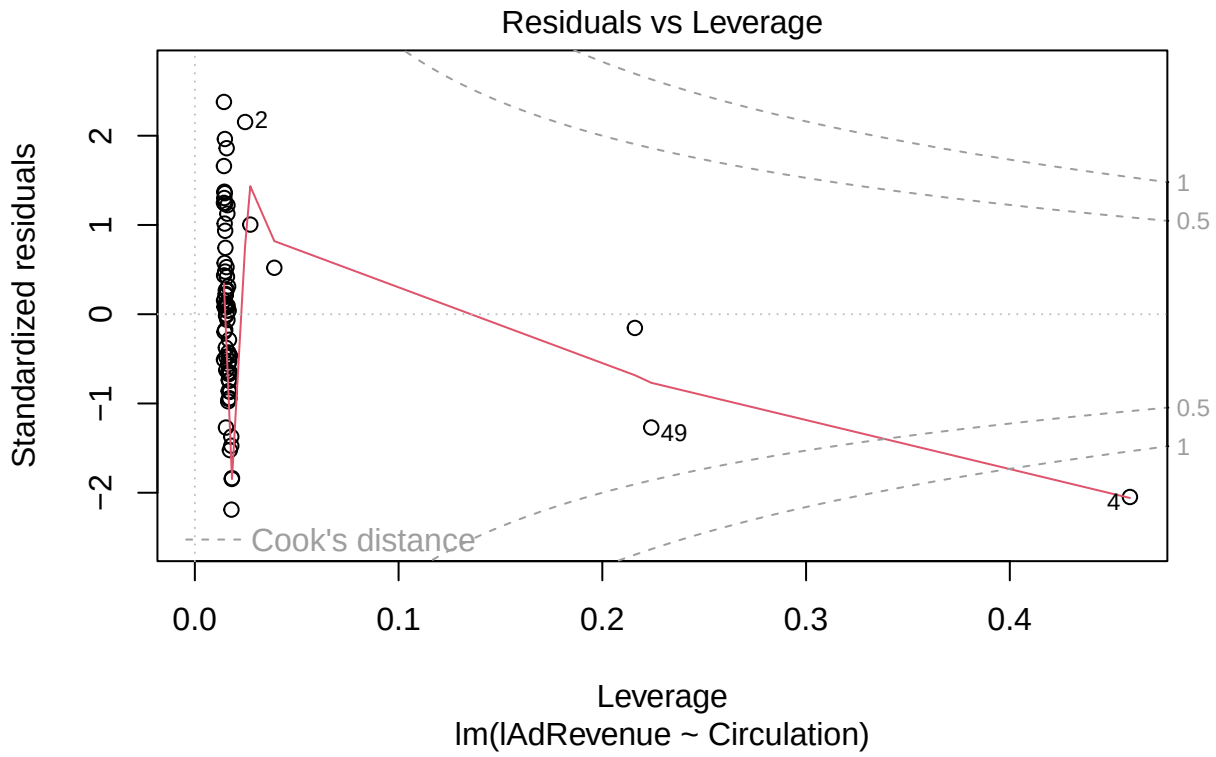



```
plot(m.log2)
```

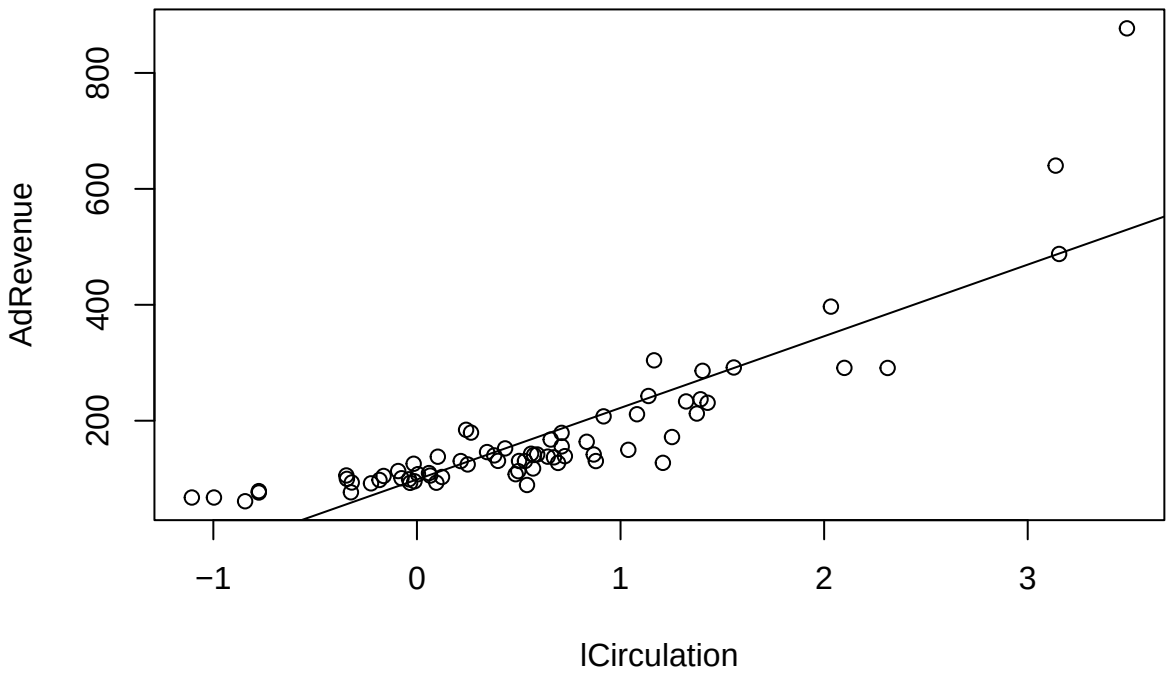

Residuals vs Fitted
Residuals
0.5
0.0
−0.5
5
2
35
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(lAdRevenue ~ Circulation)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(lAdRevenue ~ Circulation)
35
2
5
−2
−1
0
1
2

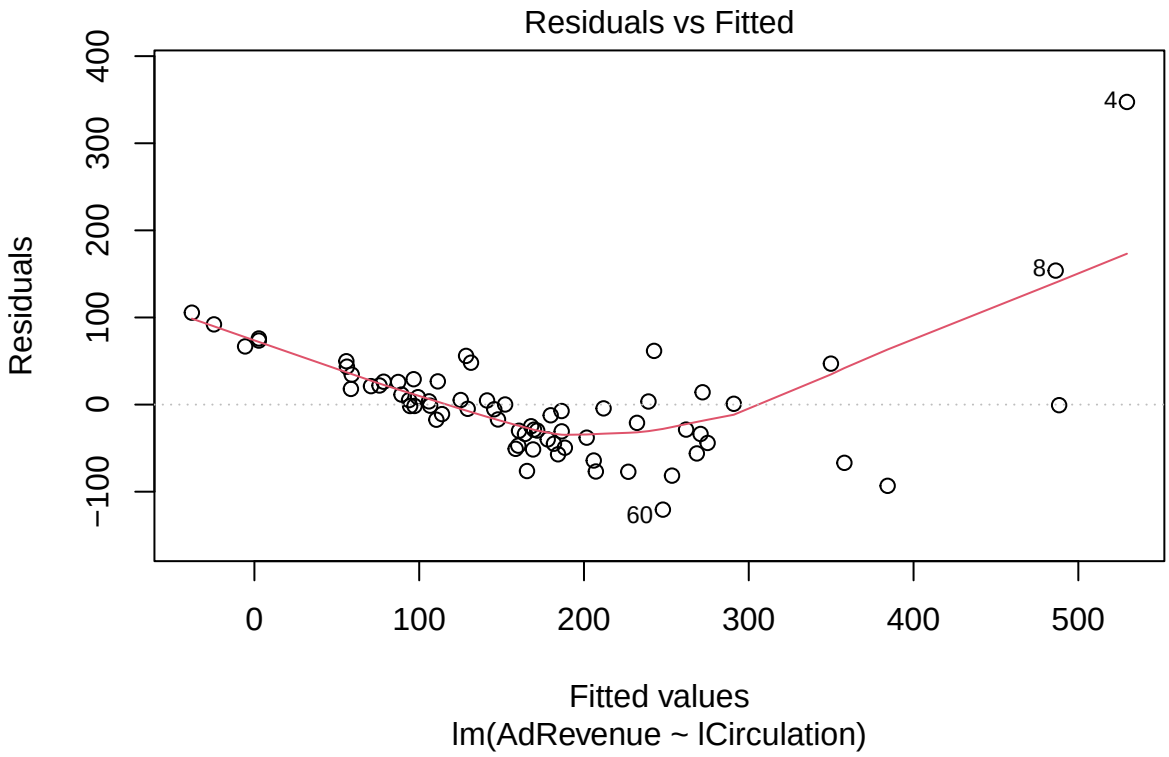

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
35
5
2
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(lAdRevenue ~ Circulation)

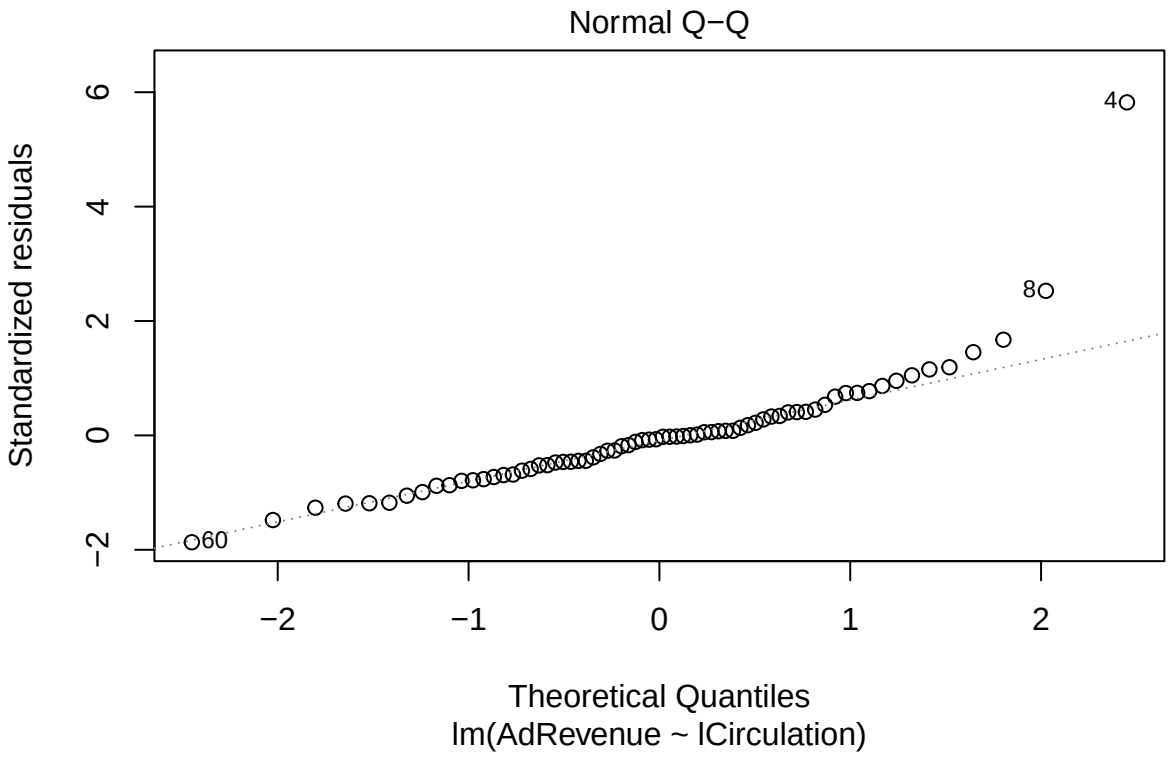


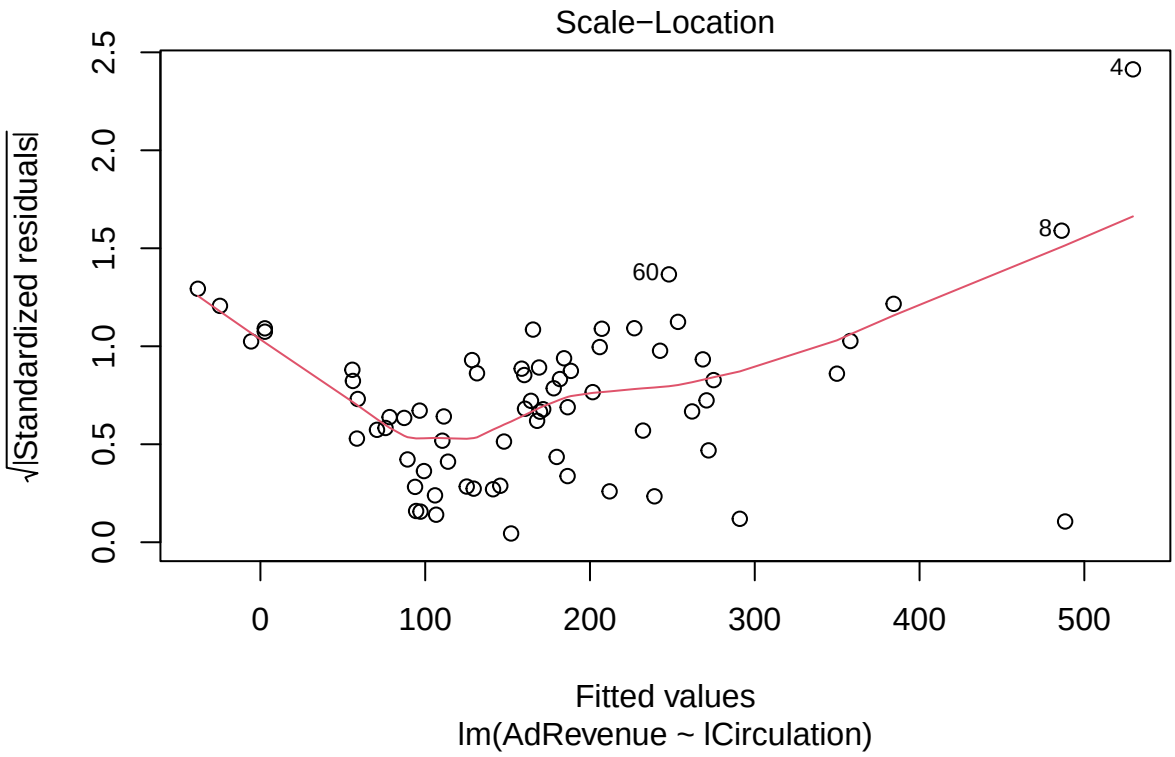

```
m.log3 <- lm(AdRevenue ~ lCirculation, data = l.revenue)
plot(AdRevenue ~ lCirculation, data = l.revenue)
abline(m.log3)
```

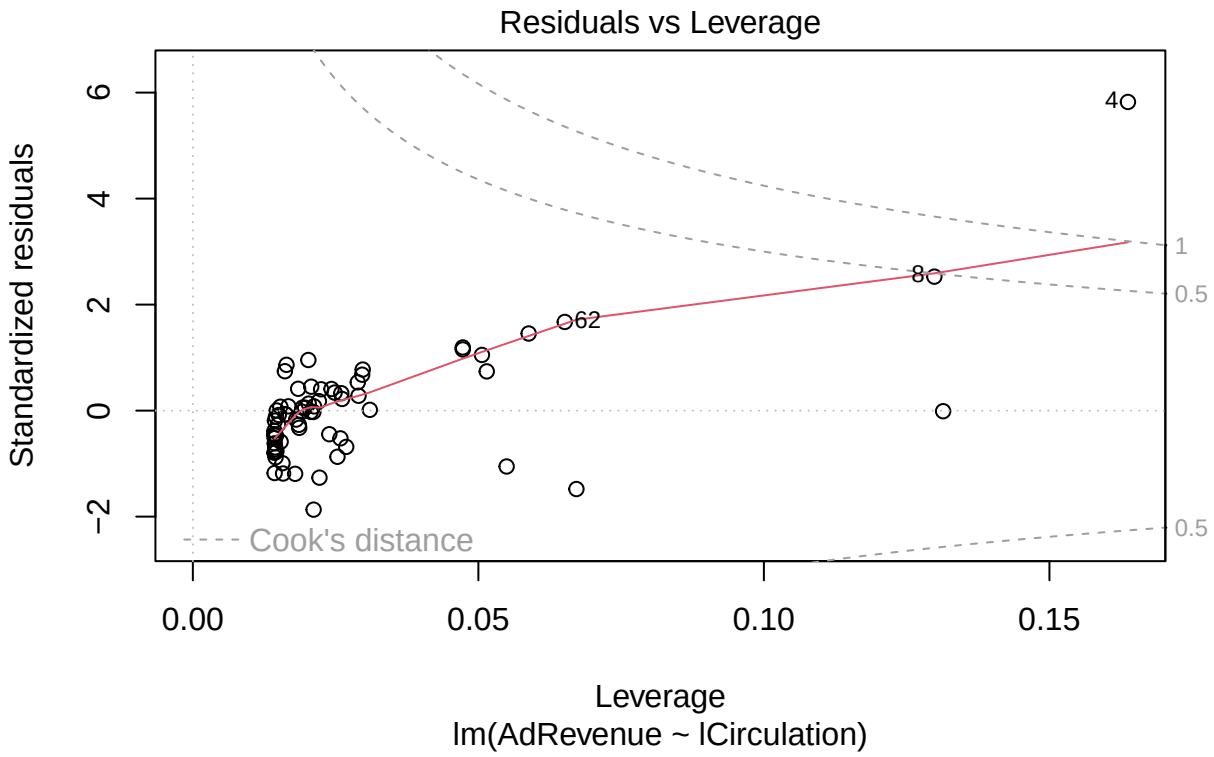



```
plot(m.log3)
```

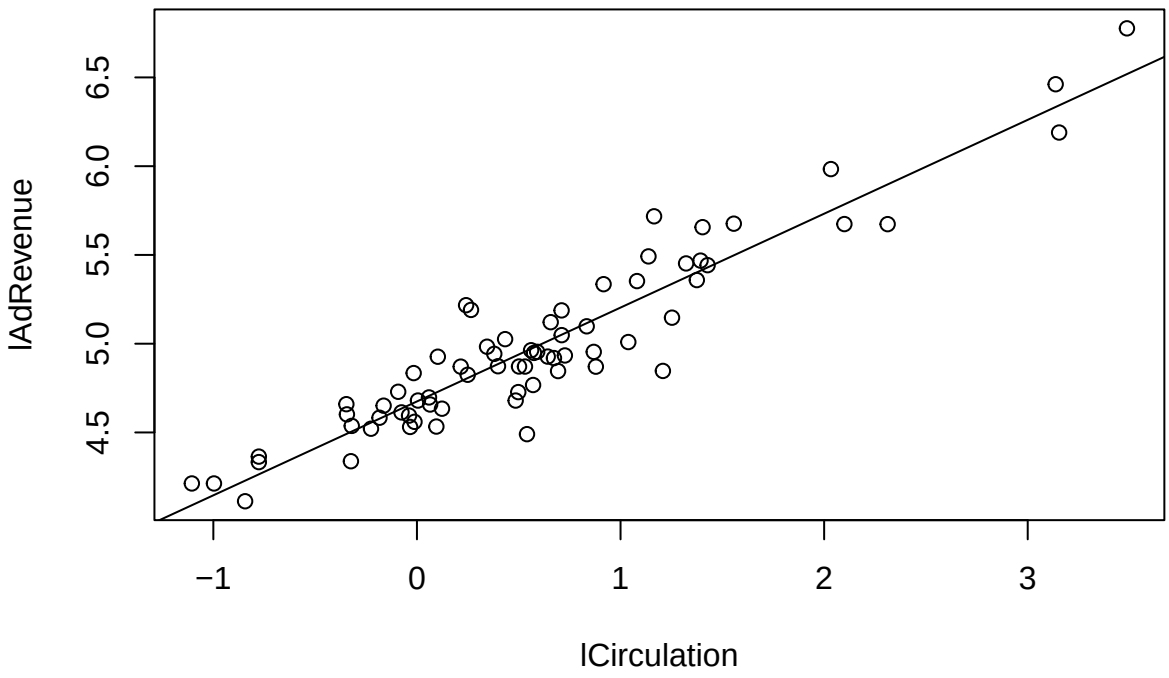

Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ lCirculation)
400
300
200
100
0
−100
0
100
200
300
400
500
4
8
60


Normal Q−Q
Standardized residuals
6
4
2
0
−2
4
8
60
−2
−1
0
1
2
Theoretical Quantiles
lm(AdRevenue ~ lCirculation)

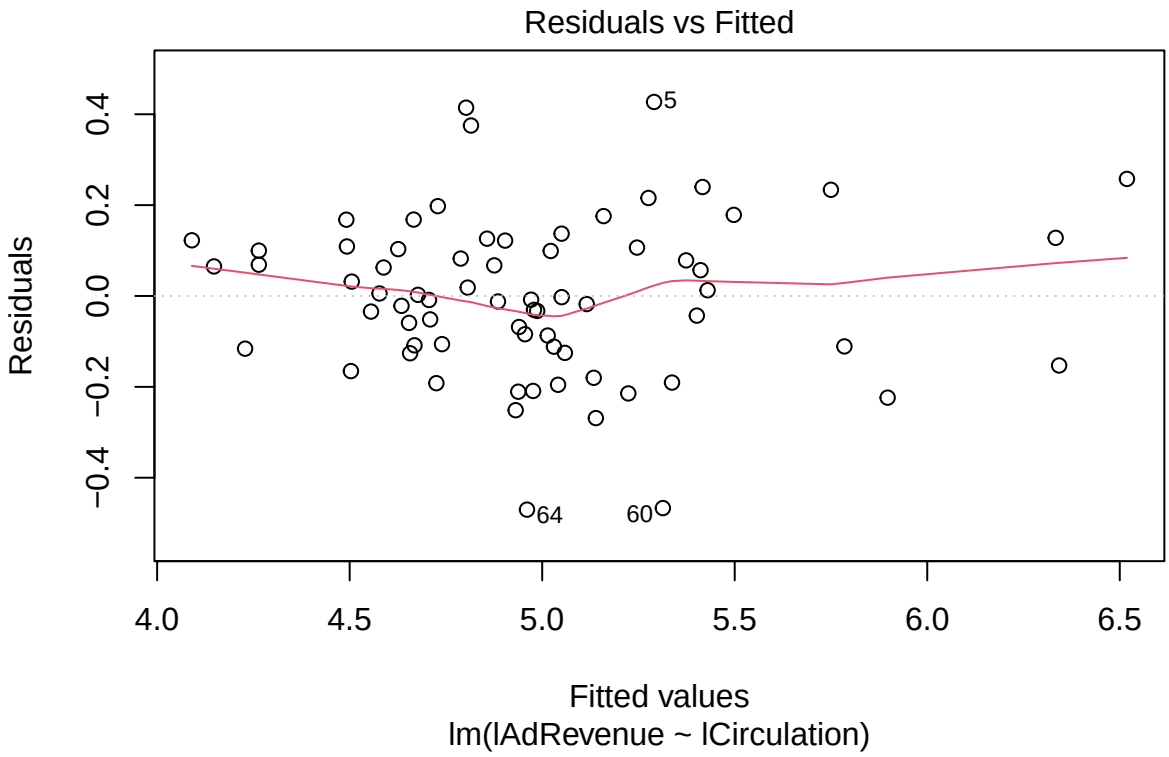

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
100
200
300
400
500
Fitted values
lm(AdRevenue ~ lCirculation)
4
8
60

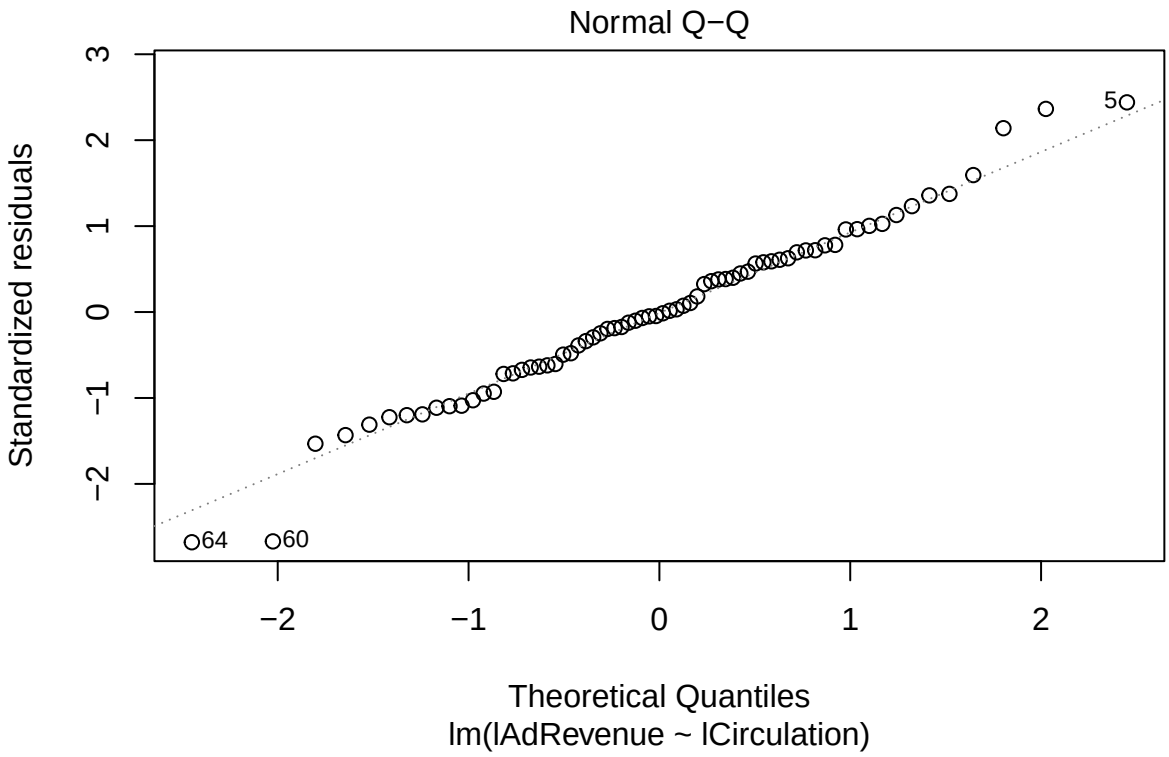


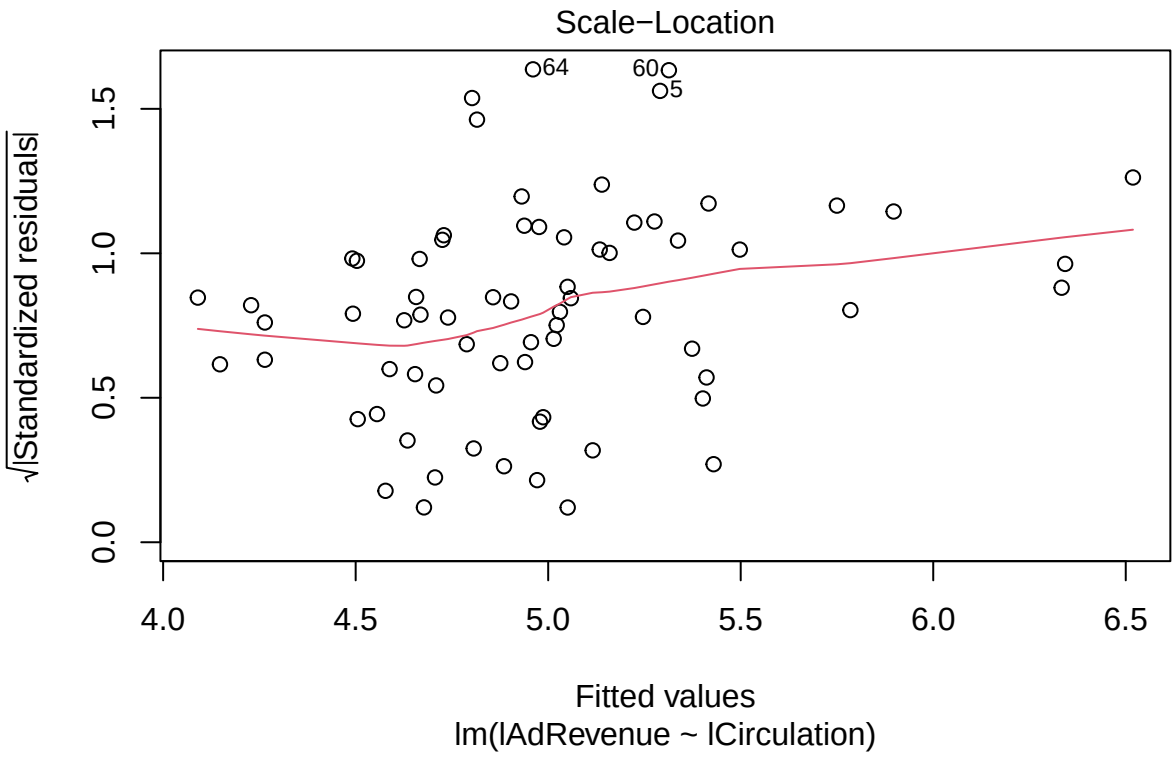

```
m.log4 <- lm(lAdRevenue ~ lCirculation, data = l.revenue)
plot(lAdRevenue ~ lCirculation, data = l.revenue)
abline(m.log4)
```

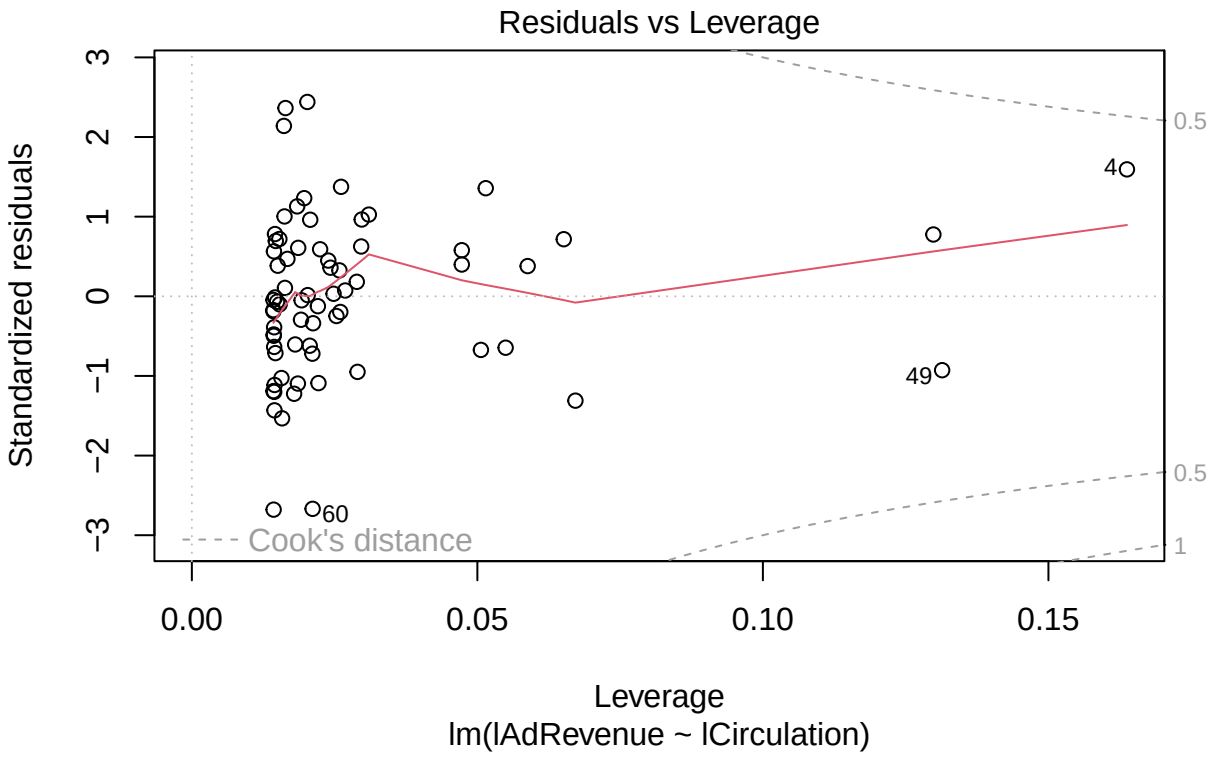



```
plot(m.log4)
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
5
64
60
4.0
4.5
5.0
5.5
6.0
6.5
Fitted values
lm(lAdRevenue ~ lCirculation)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
64
60
5
−2
−1
0
1
2
Theoretical Quantiles
lm(lAdRevenue ~ lCirculation)


Scale−Location
64
60
5
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5
6.0
6.5
Fitted values
lm(lAdRevenue ~ lCirculation)



First, I fit our data to a simple linear regression model. When observing the residuals plot, it is evident that this model is not the best model to fit to this data due to the points in residual plot not being horizontally graphed, breaking the constant variance assumption. Also, in our leverage plot, we can discern at least one bad leverage point in our data. In our QQ-norm plot, we can see that the data points greatly stray from the dotted line. Because of these factors, I decided to do three more regressions where I transform the variables using logarithms in order to better fit the data. In our regression in which the AdRevenue variable is transformed using a logarithm, we observe that our residual plot still is not plotted horizontally, from which we can infer that the constant variance assumption is still violated. Our QQ-norm plot seems to follow the dashed line closer than in our previous regression, but our leverage plot still displays bad leverage points. So, in my next regression I transformed the Circulation variable with a logarithm. Compared to our last two regressions, the residuals plot for this regression does seem to be more horizontally graphed but we can observe a slight parabolic shape in the points, still violating the constant variance assumption. Again, our QQ-norm plot for this regression seems to follow the dotted line closer than the previous two regressions, but our leverage plot still displays bad leverage points. This is not optimal for a regression model. Finally, I made a regression model that transformed both of the variables, AdRevenue and Circulation, with logarithms. In this regression's residual plot, we can see that the points are graphed horizontally with no discernible pattern. Therefore, our constant variance assumption holds. In our QQ-norm plot, the data points follow the dashed line tightly. Finally, in our residual plot, we have no bad leverage points. Therefore, we can conclude that our last regression is the best fit for this model when using a simple linear regression method.

b)

i.

```
exp(1)^predict(m.log4, newdata = data.frame(lCirculation = log(0.5)), interval = 'predict', level = 0.95
```

```
##        fit      lwr      upr
## 1 74.30864 51.82406 106.5485
```

The 95% prediction interval for the advertising revenue per page given a circulation of 0.5 million is (51.82406, 106.5485)

ii.

```
exp(1)^predict(m.log4, newdata = data.frame(lCirculation = log(20)), interval = 'predict', level = 0.95)
```

```
##        fit      lwr      upr
## 1 522.5663 359.8958 758.7626
```

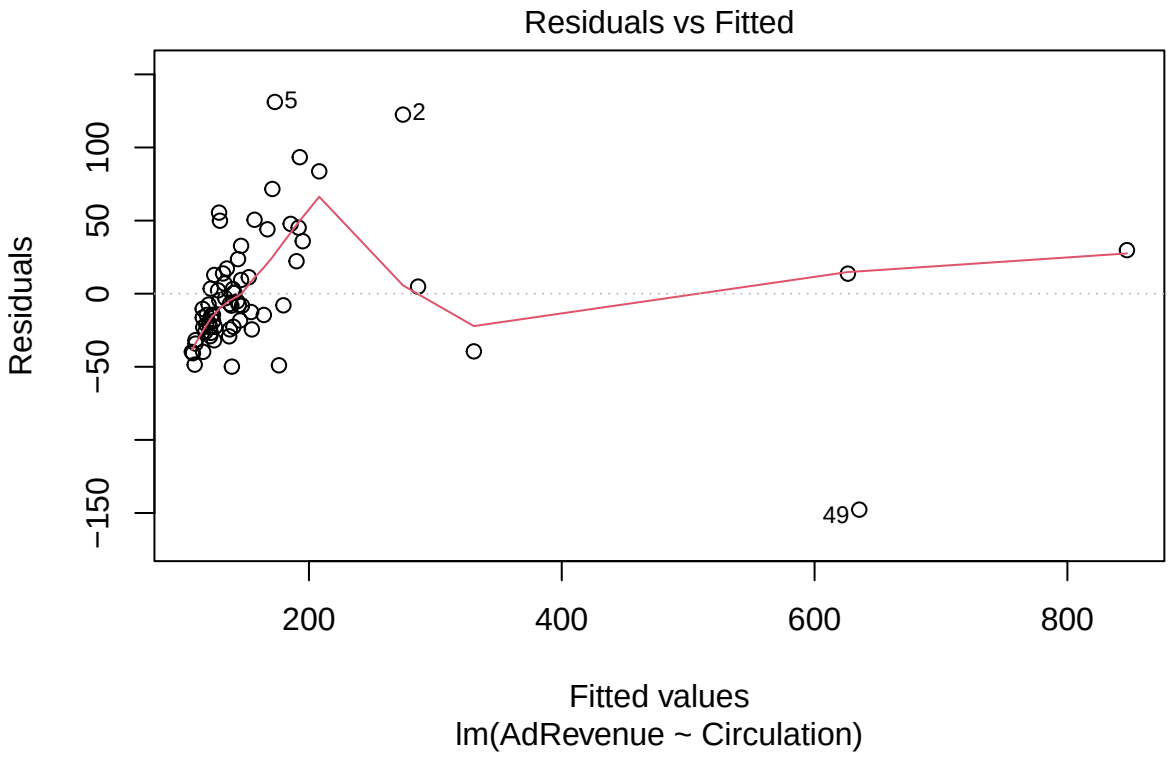
The 95% prediction interval for the advertising revenue per page given a circulation of 20 million is (359.8958, 758.7626)

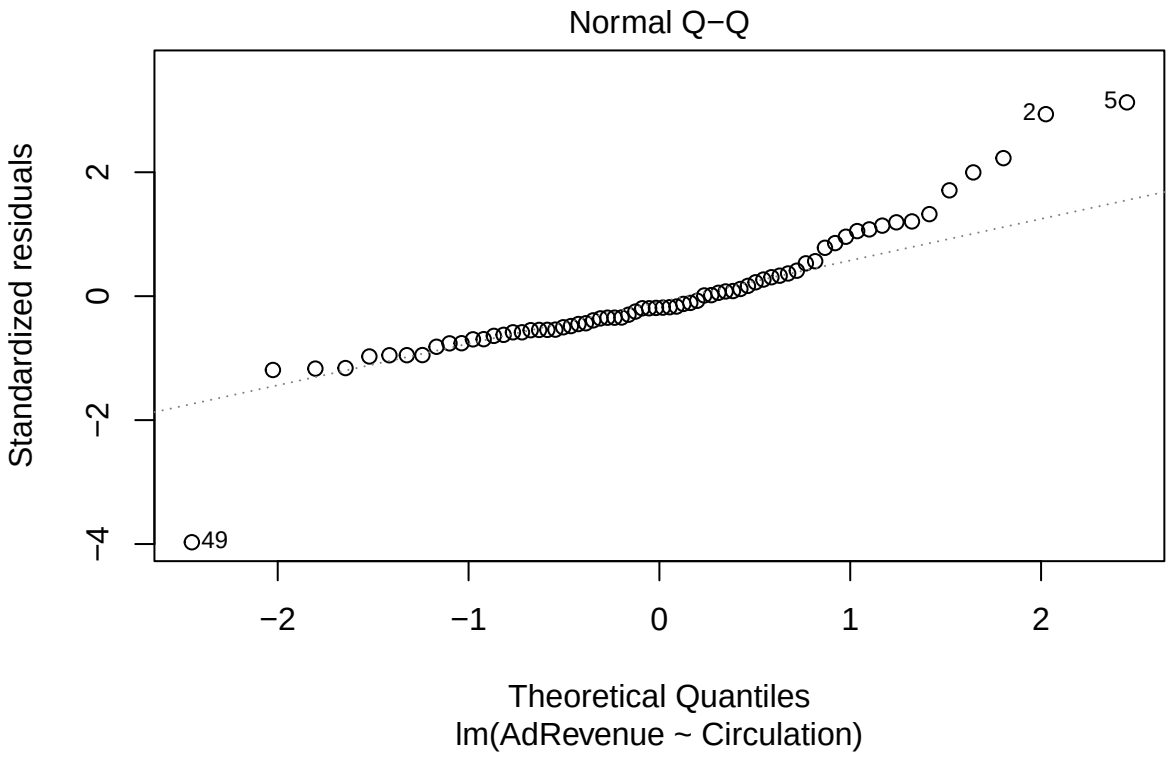
c) Some weaknesses in our model are seem in our QQ-norm plot and our scale-location plot. In our QQ-norm plot, the data points do stray from the dashed line even though it is the closest fit in our four models. There are two outliers pointed out by this plot as well, 64 and 60. This could insinuate that our normality assumption does not hold. Also, our scale-location plot can be interpreted as having a slightly positive trend, potentially violating our constant variance assumption.

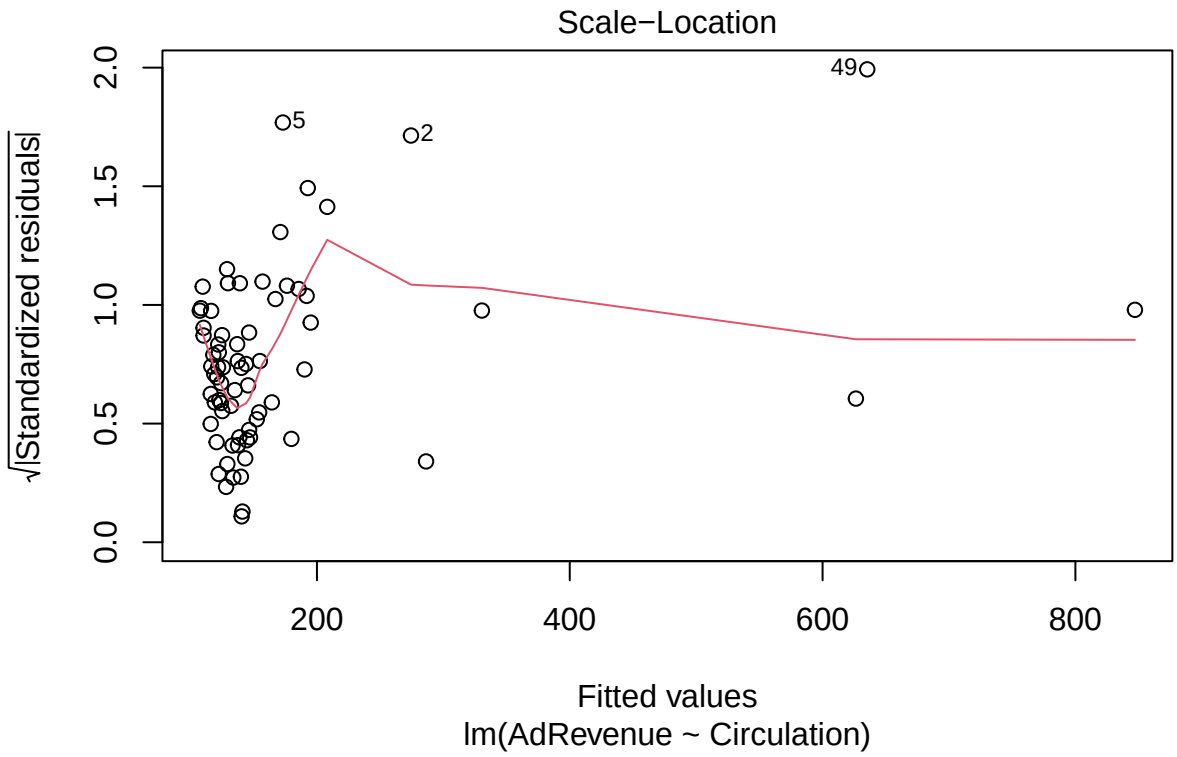
Part B

a)

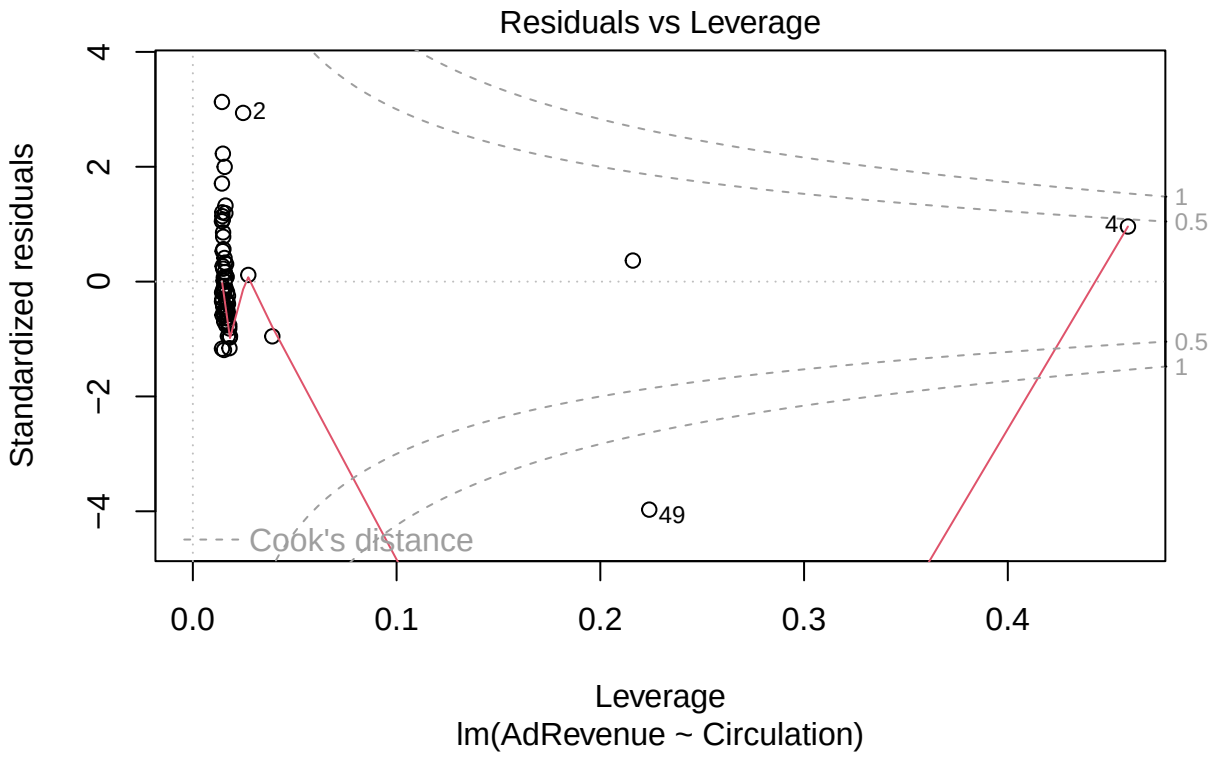
```
m.firstorder <- lm(AdRevenue ~ Circulation, data = revenue)
plot(m.firstorder)
```


Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ Circulation)
5
2
49
100
50
0
−50
−150
200
400
600
800

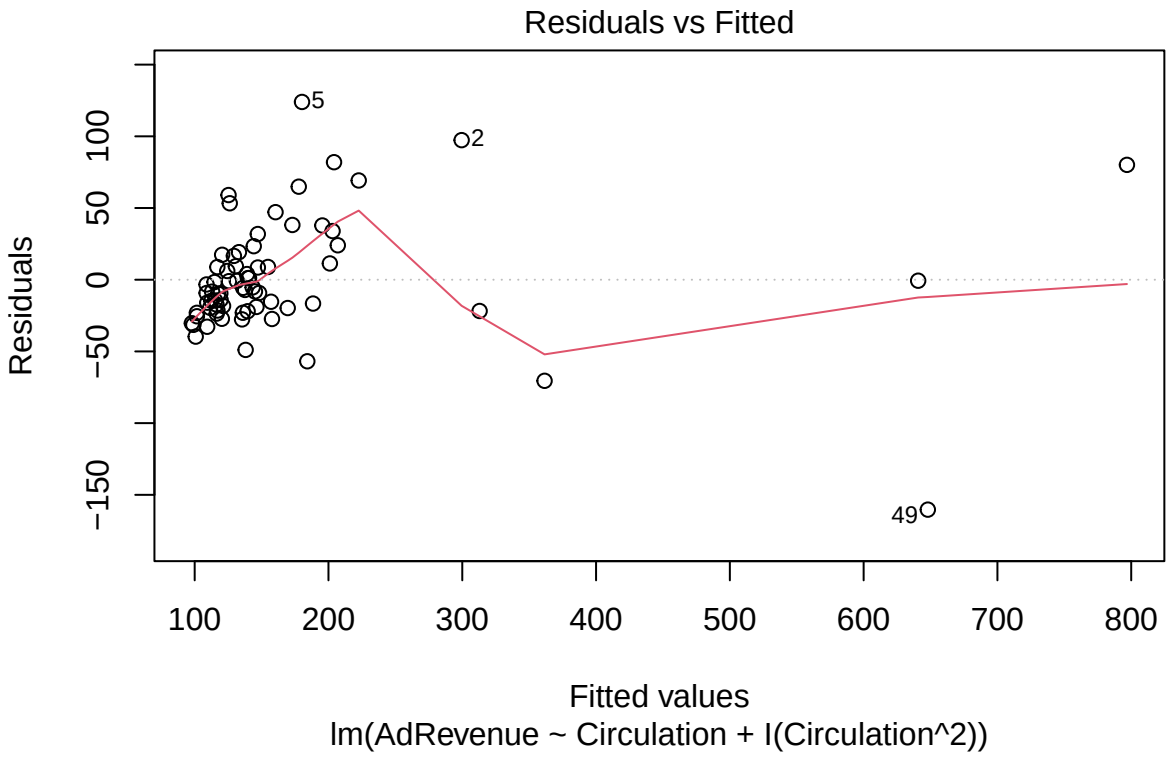

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation)
2
5
49
−4
−2
0
2
−2
−1
0
1
2

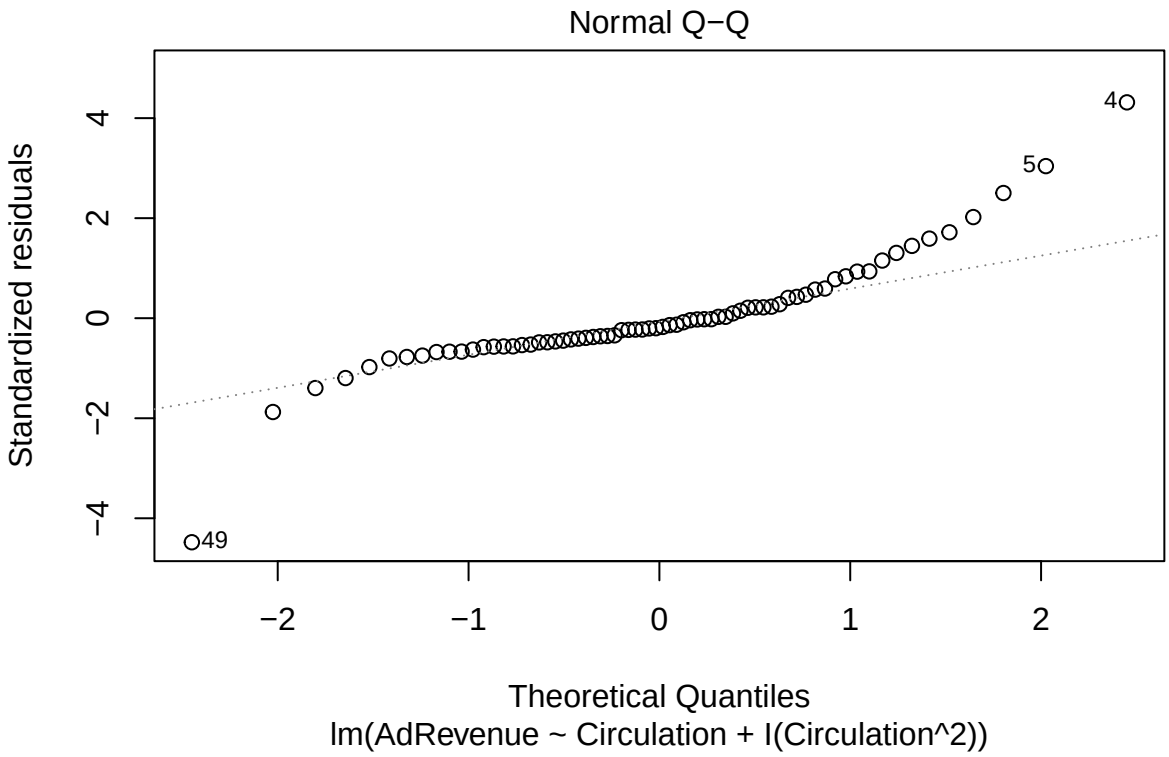


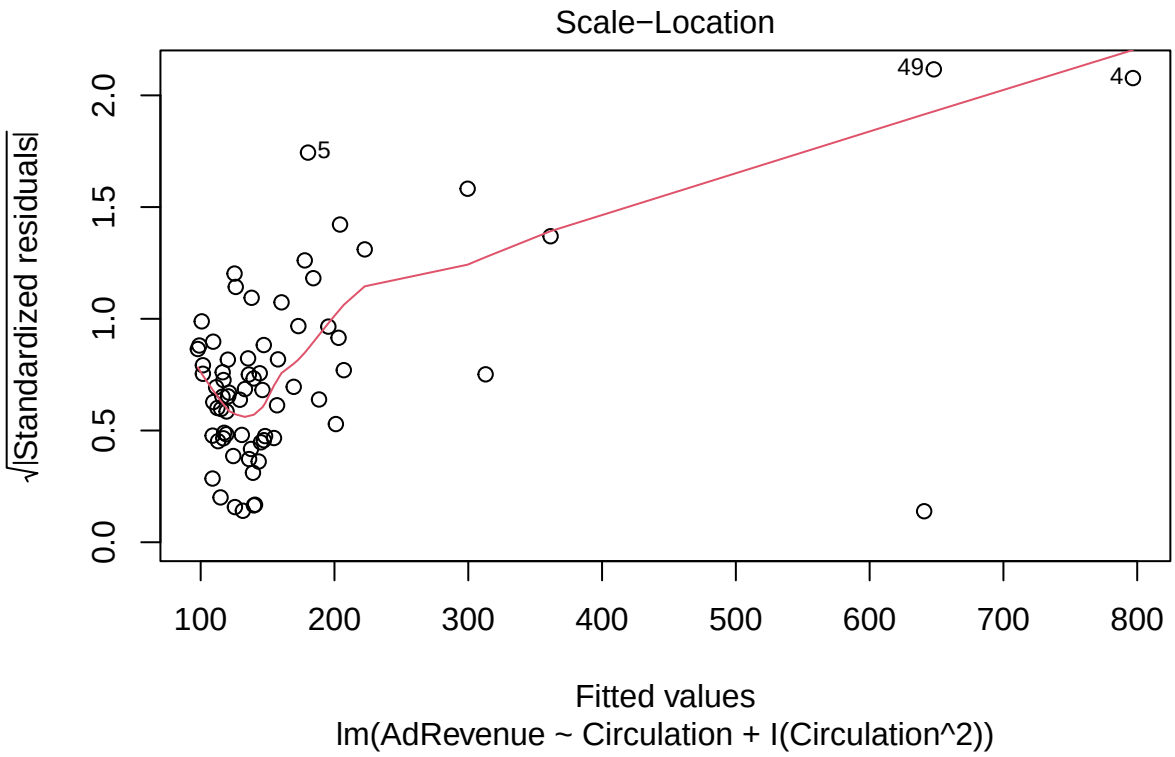

```
abline(m.firstorder)
```




```
m.secondorder <- lm(AdRevenue ~ Circulation + I(Circulation^2), data = revenue)
plot(m.secondorder)
```

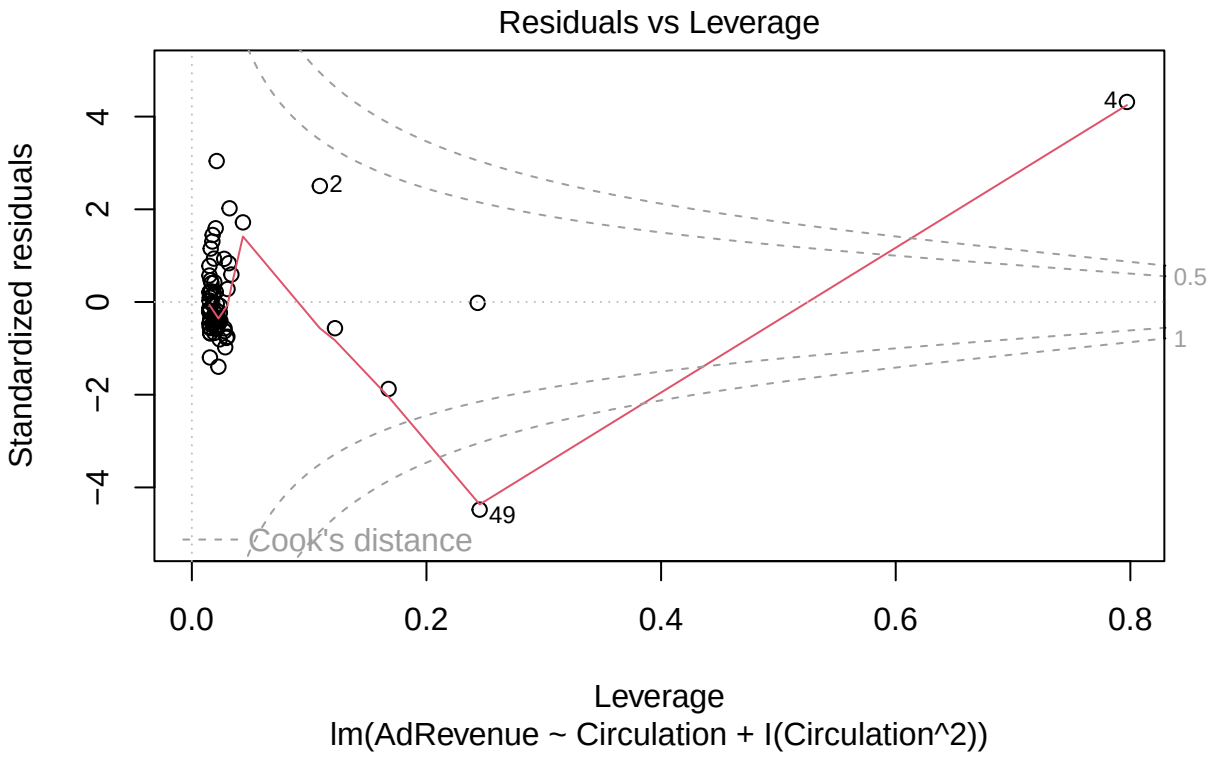

Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ Circulation + I(Circulation^2))
5
2
49
100
50
0
−50
−150
100
200
300
400
500
600
700
800


Normal Q−Q
Standardized residuals
4
5
49
−4
−2
0
2
4
−2
−1
0
1
2
Theoretical Quantiles
lm(AdRevenue ~ Circulation + I(Circulation^2))




```
abline(m.secondorder)
```

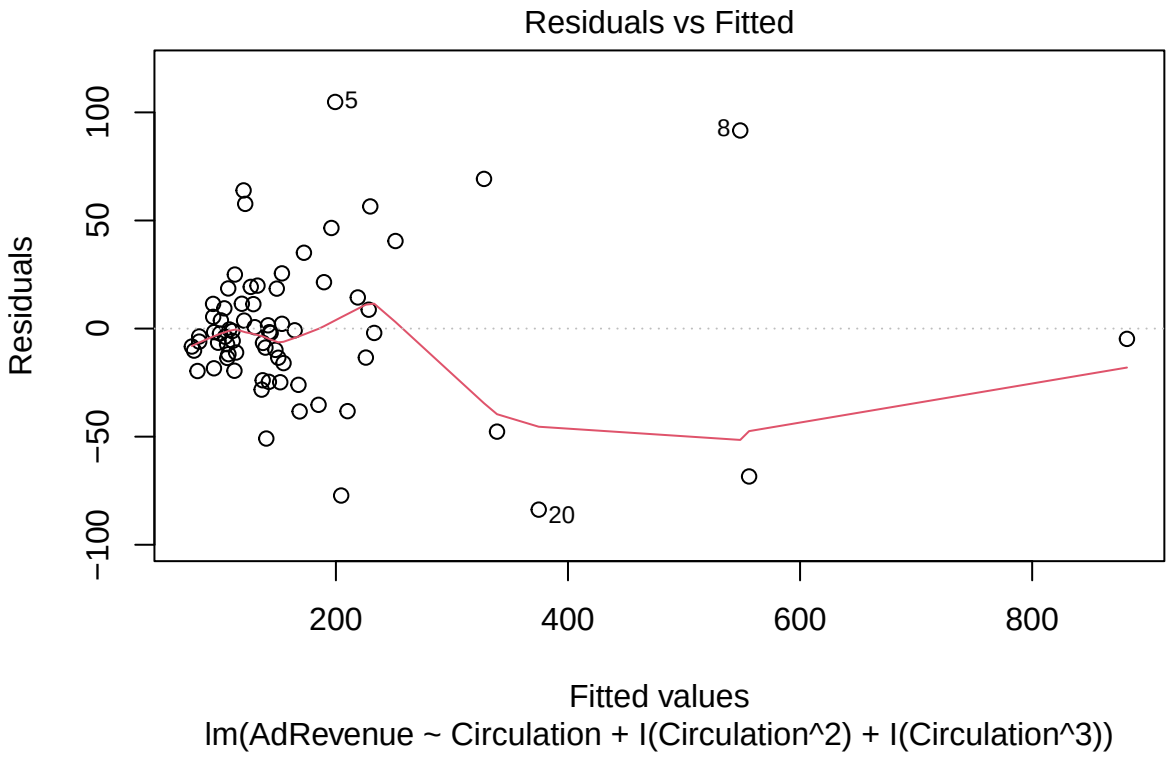
```
## Warning in abline(m.secondorder): only using the first two of 3 regression
## coefficients
```

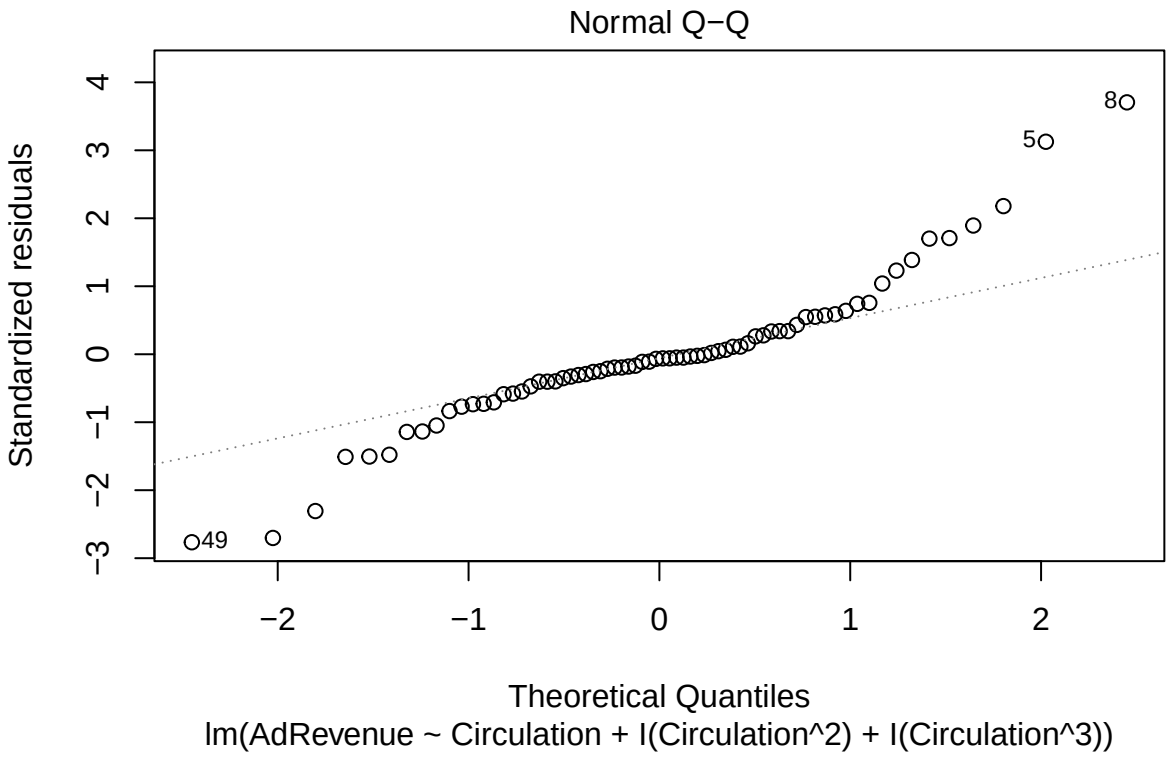


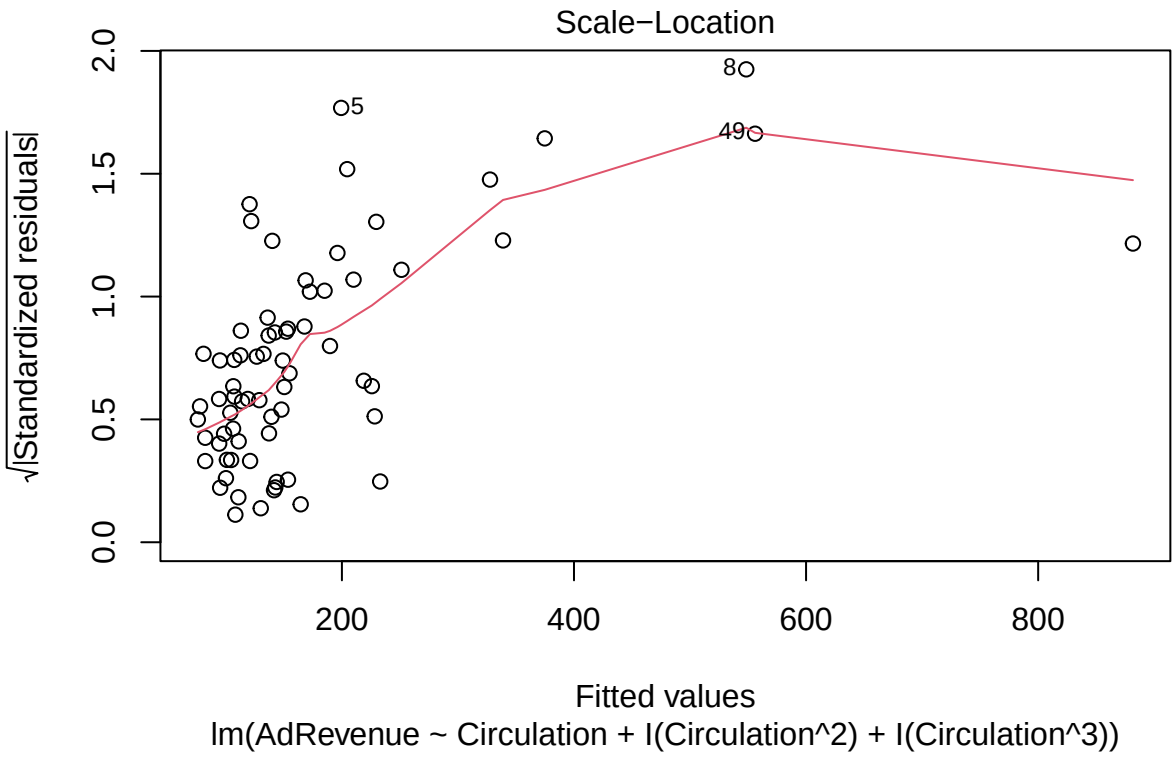

```
summary(m.secondorder)
```

```
##
## Call:
## lm(formula = AdRevenue ~ Circulation + I(Circulation^2), data = revenue)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -160.32  -21.05   -7.65   15.30  123.96
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       88.1390     7.9708  11.058  < 2e-16 ***
## Circulation       29.5006     3.2992   8.942 4.87e-13 ***
## I(Circulation^2)  -0.2394     0.1140  -2.100   0.0395 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.2 on 67 degrees of freedom
## Multiple R-squared:  0.901,  Adjusted R-squared:  0.8981
## F-statistic: 304.9 on 2 and 67 DF,  p-value: < 2.2e-16
```

```
m.thirdorder <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3), data = revenue)
plot(m.thirdorder)
```


Residuals vs Fitted
Residuals
100
50
0
−50
−100
200
400
600
800
5
8
20
Fitted values
lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3))


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3))
8
5
49

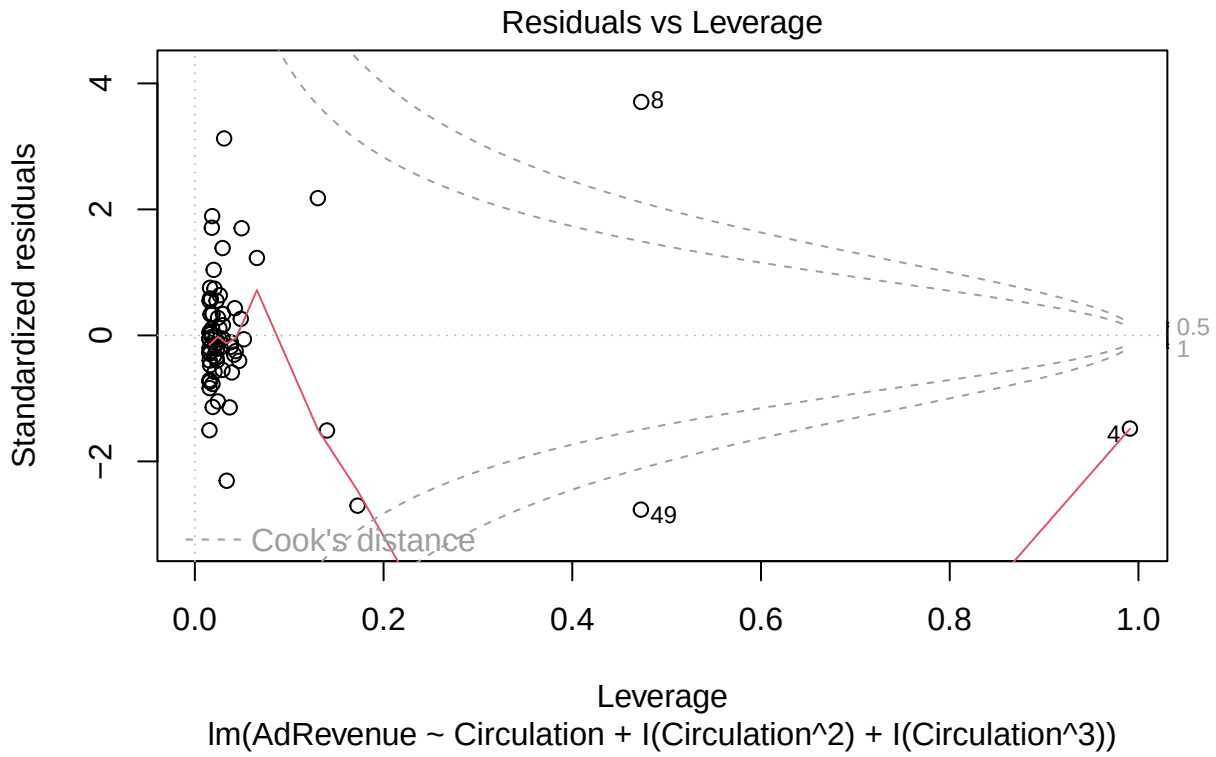



```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
abline(m.thirdorder)
```

```
## Warning in abline(m.thirdorder): only using the first two of 4 regression
## coefficients
```




```
summary(m.thirdorder)
```

```
##
## Call:
## lm(formula = AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3),
##     data = revenue)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83.75 -13.56  -2.16  11.46 104.82
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      59.17037    8.34505   7.090 1.12e-09 ***
## Circulation      51.23582    4.71123  10.875 2.33e-16 ***
## I(Circulation^2) -2.50538    0.41141  -6.090 6.48e-08 ***
## I(Circulation^3)  0.05223    0.00923   5.658 3.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.06 on 66 degrees of freedom
## Multiple R-squared:  0.9333, Adjusted R-squared:  0.9303
## F-statistic: 308.1 on 3 and 66 DF,  p-value: < 2.2e-16
```

Between the second order regression and the third order regression, we can conclude that the third order regression is the better fit for this data. When comparing the residual plots, we can see that the third order

regression residual plot, though still densely plotted, has more of a horizontal dispersion of the data points. When comparing the QQ-norm plots of the two regressions, we observe that both plots point out a number of outliers in the data, with both of the sets of data points straying away from the dashed line. When comparing our scale-location plots, we can see that for the third order regression the data is roughly more horizontally plotted than in our second order regression, with our data points equally dispersed about the red line. There are a handful of bad leverage points discernible in both leverage plots for the regressions. When observing our summary statistics, we can see that our correlation coefficient for our third order regression has a higher value than for our second order regression, 93.3% and 90.1% respectively. This supports the claim that the third order regression is a better fit for this data. Also, our third order regression has smaller p-values than our second order regression, further supporting the claim that our third order regression is a better fit than our second order regression. From this analysis, we can conclude that our third order regression model is a better fit for our data than our second order regression model.

b)

i.

```
predict(m.thirdorder, newdata = data.frame(Circulation = 0.5), interval = 'predict', level = 0.95)
```

```
##        fit      lwr      upr
## 1 84.16846 14.92314 153.4138
```

The 95% prediction interval for advertising revenue per page given a circulation of 0.5 million is (14.92314, 153.4138)

ii.

```
predict(m.thirdorder, newdata = data.frame(Circulation = 20), interval = 'predict', level = 0.95)
```

```
##        fit     lwr      upr
## 1 499.5334 418.179 580.8878
```

The 95% prediction interval for advertising revenue per page given a circulation of 20 million is (418.179, 580.8878)

c) In our model, we can spot many potential weaknesses. For example, in our QQ-norm plot we can see that our data points do not closely follow the dashed line, suggesting a violation of the normality assumption. In our residual plot, we can see a fan shape appear, which suggests a violation of the constant variance assumption. Also, in our scale-location plot, we can argue that there is an observable positive trend in our data points, further suggesting a violation in the constant variance assumption. There are also bad leverage points in this model, as seen in our leverage plot.

Part C

a) I believe that our model in Part A is a better fit for this data. This is due to many reasons, such as the QQ-norm plot's data points in Part A following the dashed line more closely than in Part B, meaning that the normality assumption is more likely to be held in Part A's model. In Part A's residual plot, we do not observe any noticeable pattern such as the fan shape in Part B. This means that the constant variance assumption is more likely to be held in Part A than in Part B. Also, when comparing the scale location plots, we can easily observe that in Part A the data points are plotted horizontally while in Part B they are not. Also, there are no bad leverage points in Part A. Because of these observations, we can conclude that the model in Part A is a better fit for the data when compared to the model in Part B.

b)

i. Part A's model creates a prediction interval of (51.82406, 106.5485) when the circulation is 0.5 million. The model in Part B creates a prediction interval of (14.92314, 153.4138) when the circulation is 0.5 million. We can conclude that the prediction interval produced by Part A's model is narrower and therefore more precise, meaning it would the preferred prediction interval to use.

ii. The model in Part A creates a prediction interval of (359.8958, 758.7626) when the circulation is 20 million, while the model in Part B creates a prediction interval of (418.179, 580.8878) when the circulation is 20 million. We can conclude that Part B's model produces the preferred interval since its interval is narrower and therefore more precise when compared to Part A's interval.

Question 3

a) For each discrete value of x, you would have to calculate the standard deviation of y individually. Since there are multiple observations of the y variable at each x value, this is possible.

b) In this special case, we know that the x variable is discrete. We also know there are also multiple y value measurements at each discrete value of x. Because of this, we are able to calculate the standard deviation of y at each discrete value of x directly. This is not like the usual case of calculating standard deviations because we do not usually have multiple observations of the y variable associated with one value of the x variable. If we did try to calculate the standard deviation of the y variable at each x variable when there was only one y value associated with each x value, it would equal zero. This is because the mean of the y values at each x value would equal the observed value of y.