Homework 6

Jun Ryu, UID: 605574052

2023-05-12

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

Question 1

Although the model does seem to fit the data well since most data points lie close to the line, looking at the standardized residuals vs. distance plot, we see that there is a problem with non-constant variance. This is noted through a quadratic shape of the residuals and also an overall decreasing trend. Moreover, we notice some outliers in the residual plot that might be affecting the model. Thus, some changes we could make are removing these outliers and/or fitting a quadratic model instead.

Question 2 Part A

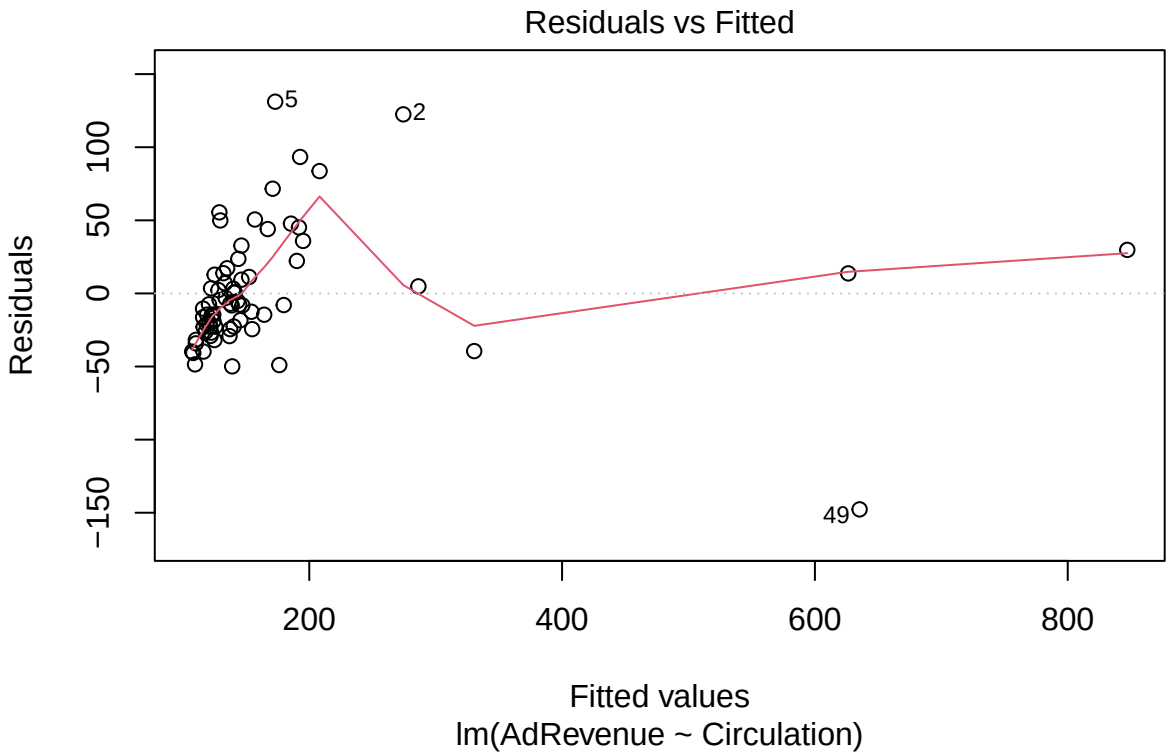
```
rev <- read.csv("AdRevenue.csv")
head(rev)
```

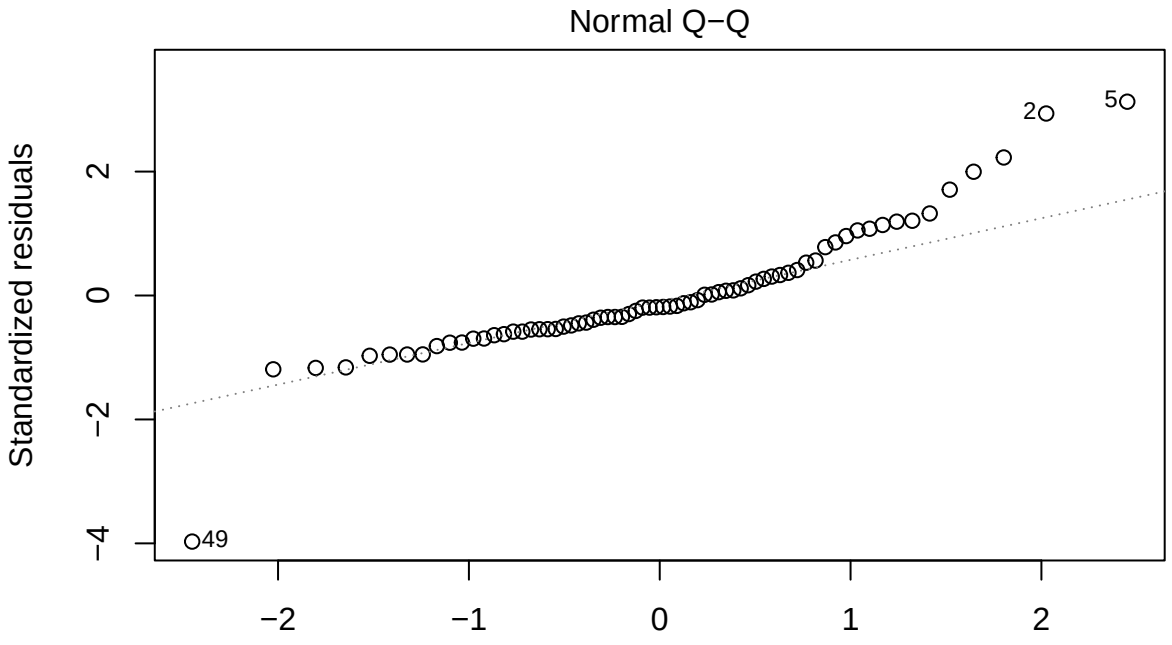
```
##                   Magazine                   PARENT.COMPANY..SUBSIDIARY
## 1                   People                     Time Warner, (Time Inc.)
## 2 Better Homes and Gardens                                Meredith Corp.
## 3                     Time                     Time Warner, (Time Inc.)
## 4                Parade (1) Advance Publications, (Parade Publications)
## 5       Sports Illustrated                     Time Warner, (Time Inc.)
## 6        Good Housekeeping           Hearst Corp., (Magazine Division)
##   AdRevenue Circulation
## 1   233.259       3.751
```

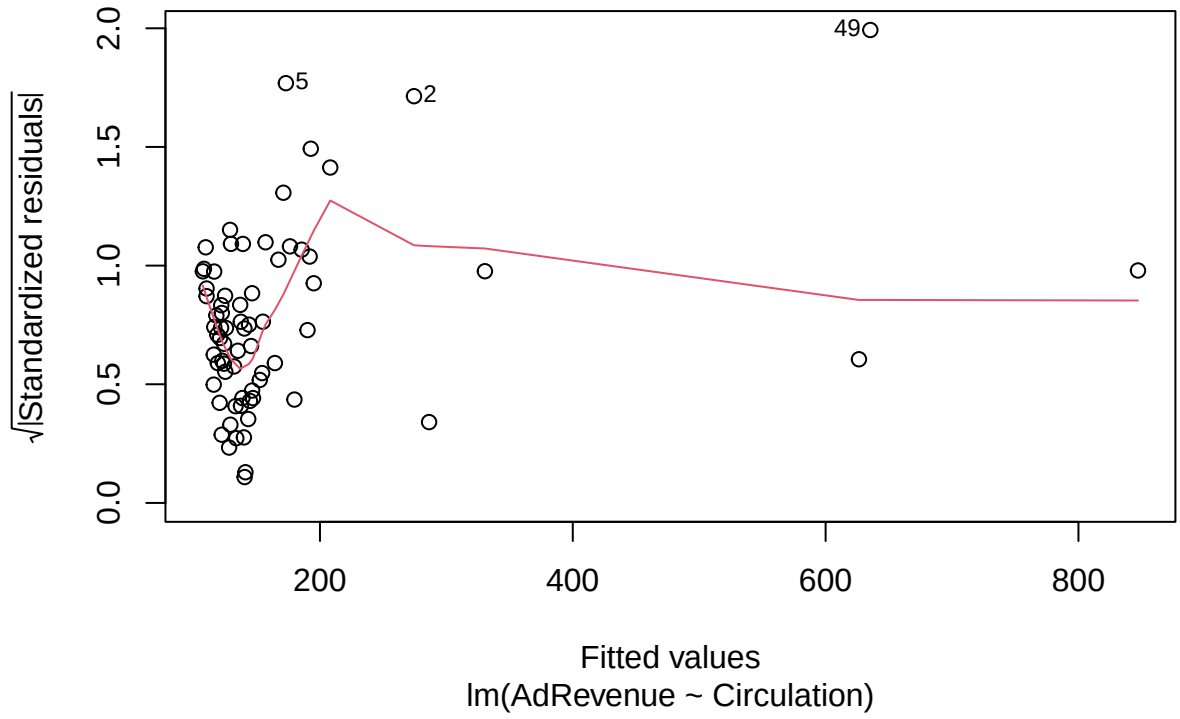
```
## 2   396.865       7.639
## 3   286.108       4.067
## 4   876.907      32.700
## 5   304.185       3.205
## 6   291.829       4.741
```

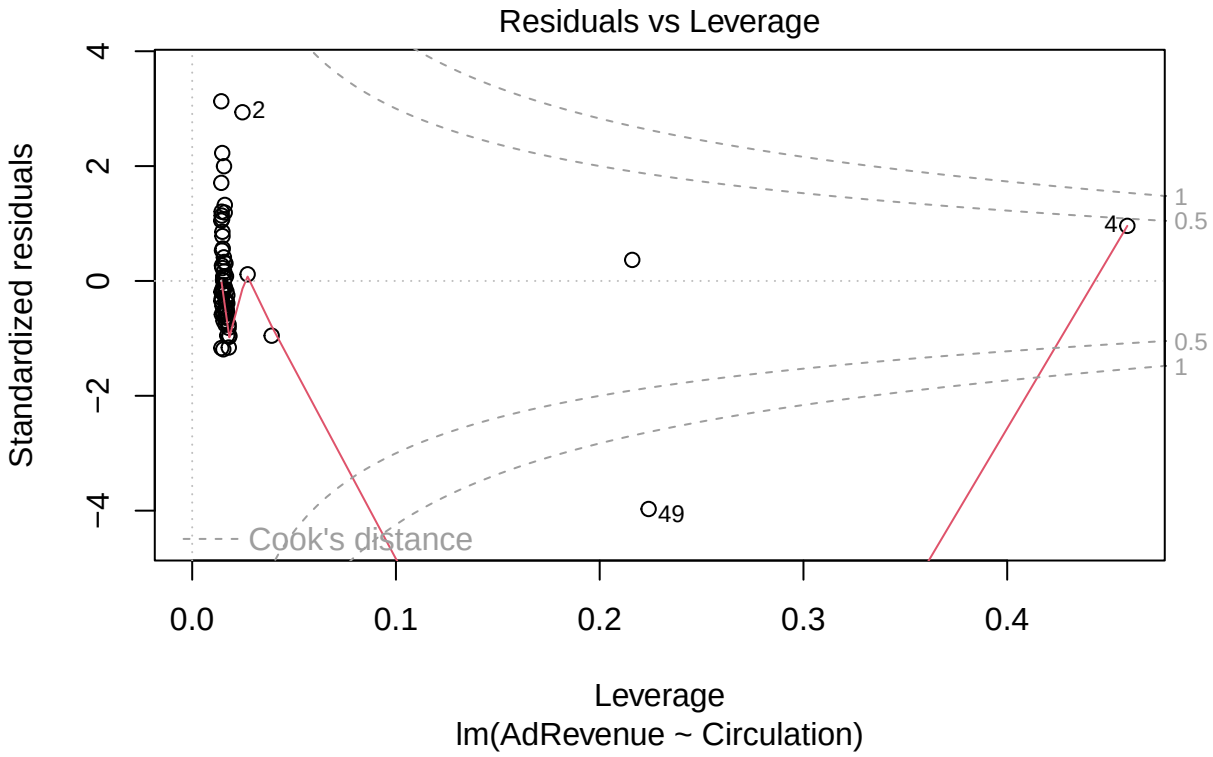
a)

```
model <- lm(AdRevenue ~ Circulation, data = rev)
plot(model)
```



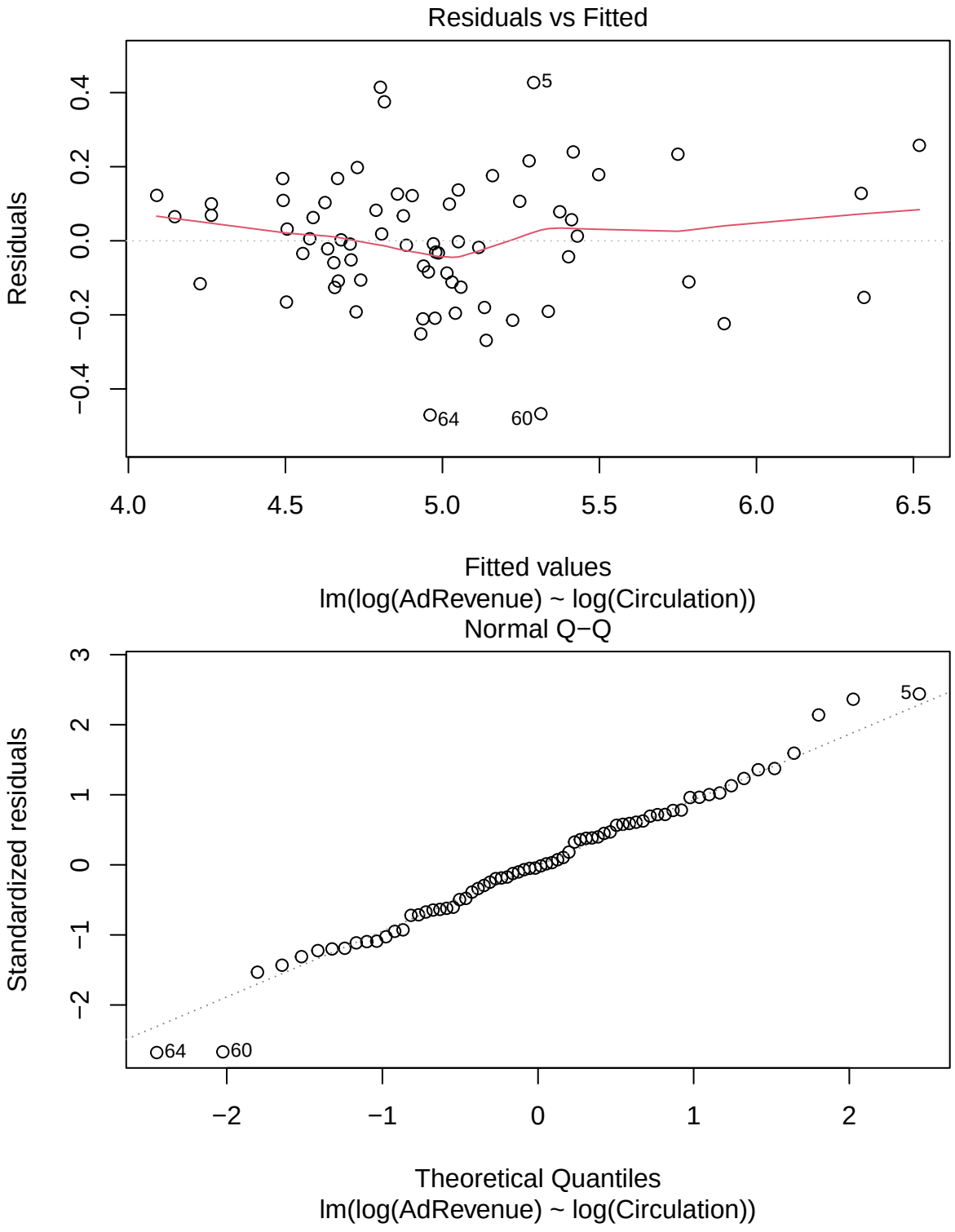

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation)
49
2
5

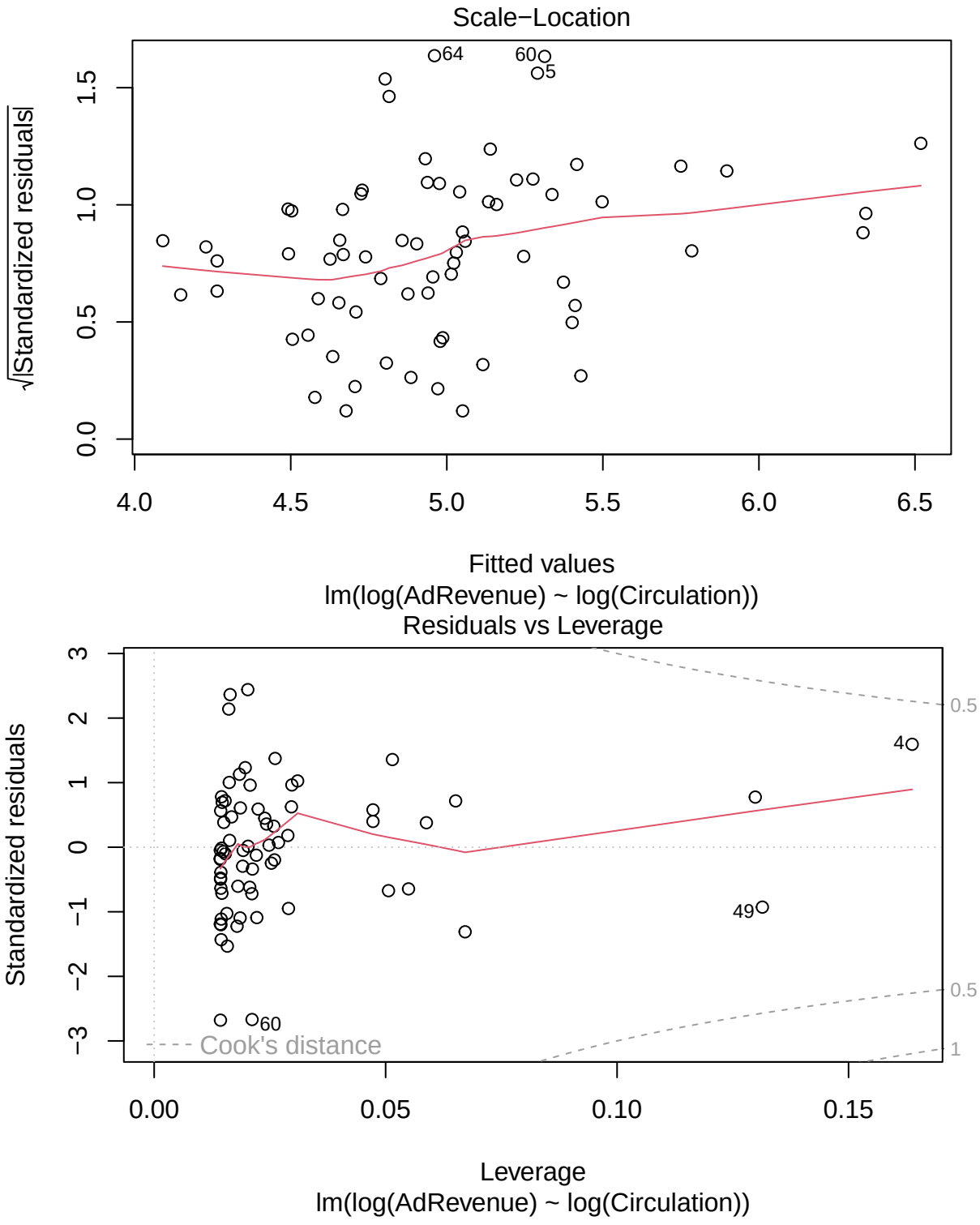

Scale−Location
√|Standardized residuals|
Fitted values
lm(AdRevenue ~ Circulation)
49
5
2



Looking at this fit, we notice that the model is not valid because even from just looking at the qq-plot, we notice that the plot is not straight and thus shows non-normality. In the other plots, we observe that we have an issue dealing with range more than one order of magnitude as some fitted values are detached really far away from the rest. Thus, we can adjust this by taking the log of both variables and seeing what happens.

```
log_model <- lm(log(AdRevenue) ~ log(Circulation), data = rev)
plot(log_model)
```


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(log(AdRevenue) ~ log(Circulation))
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
5
64
60
−2
−1
0
1
2
Theoretical Quantiles
lm(log(AdRevenue) ~ log(Circulation))



We notice that this is a much improved model since not only have we achieved straightness in the qq-plot, but we notice a better scaling of our observations in the other plots.

b)

```
# we have to to e^(interval) because the returned interval would represent log(y) values
# intervals represent the cost in thousands of dollars

exp(1)^predict(log_model, data.frame(Circulation = 0.5), interval="prediction")
```

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

```
exp(1)^predict(log_model, data.frame(Circulation = 20), interval="prediction")
```

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

c)

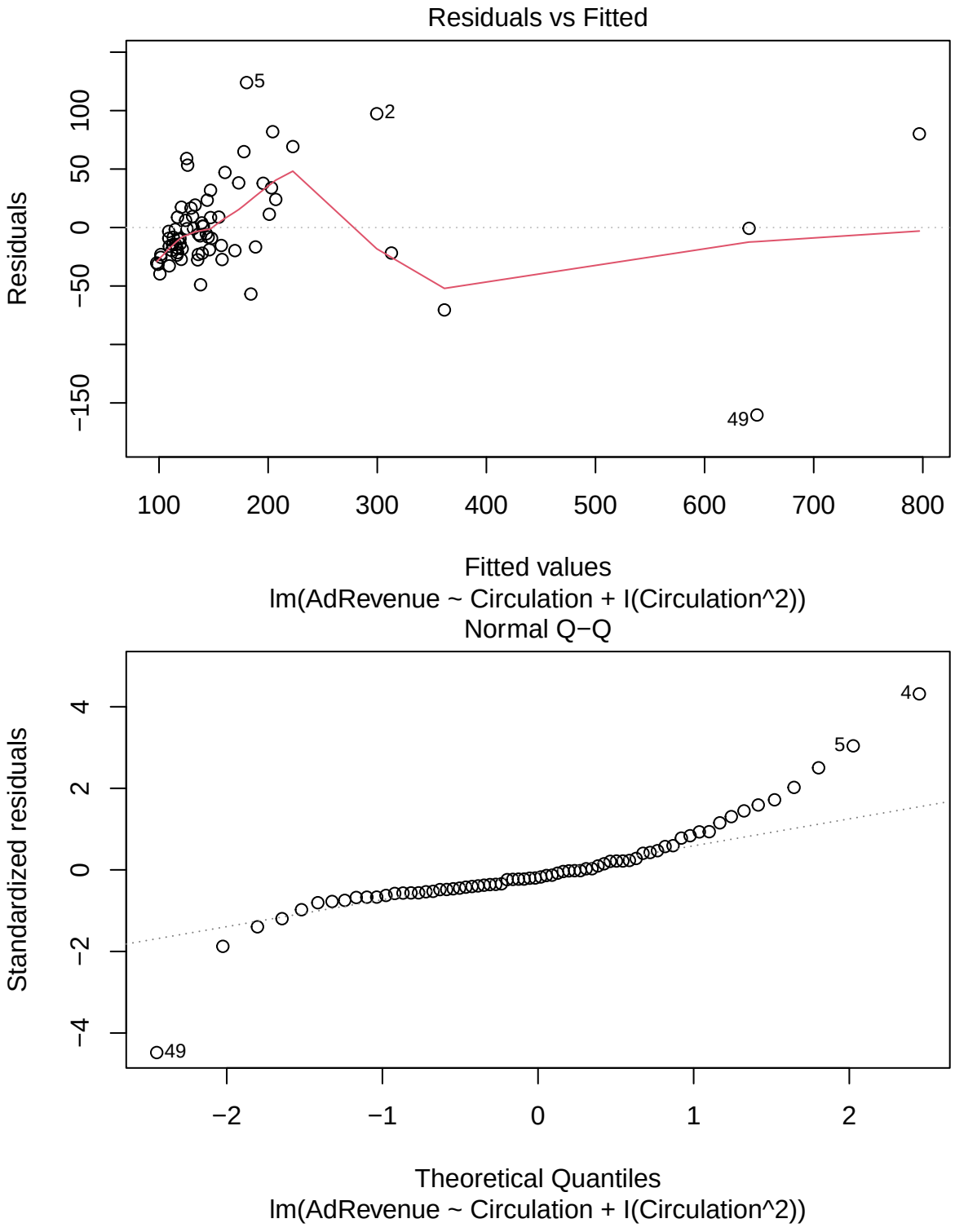
Referring to the plots from part a) of our revised log model, we do notice a good normality through the straight line in the qq-plot. In the leverage plot, we don't really see bad leverage points, which is also a strength of our model. However, there is a slight issue with constant variance as we see an increasing trend in the scale-location graph and this is a certainly a weakness of our model.

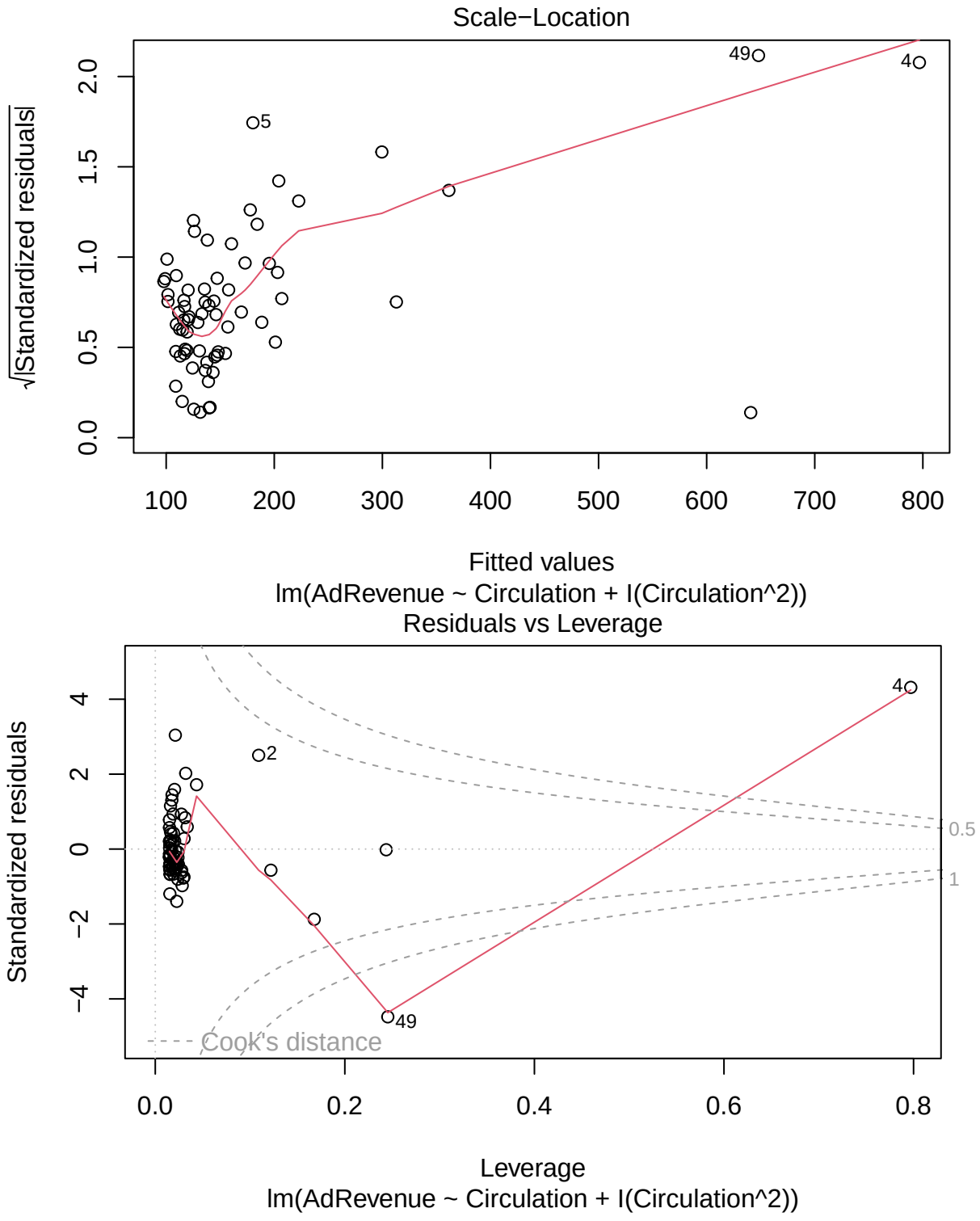
Question 2 Part B

a)

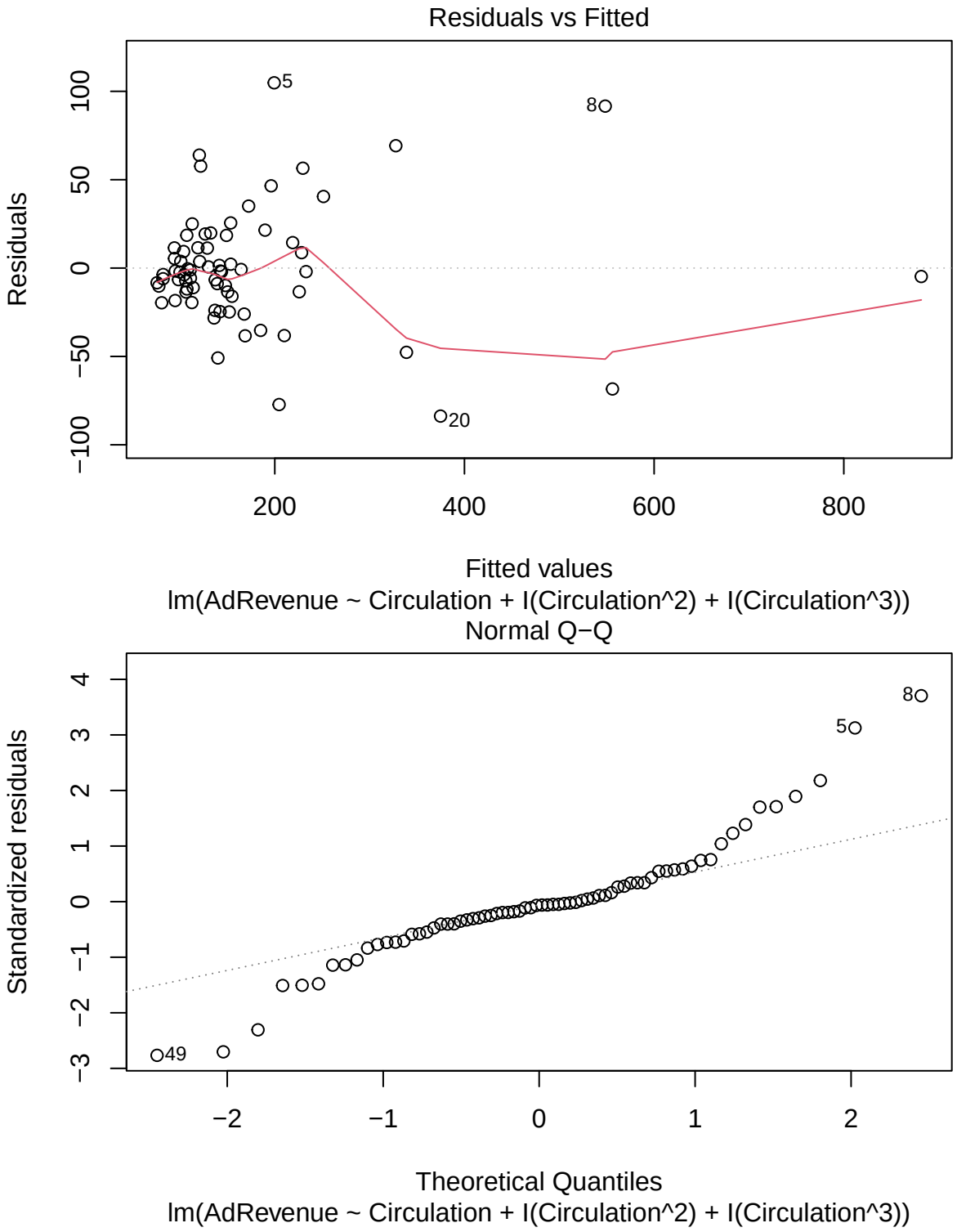
Now, we will try to fit a quadratic and a cubic model to see if a polynomial regression model will better fit our data.

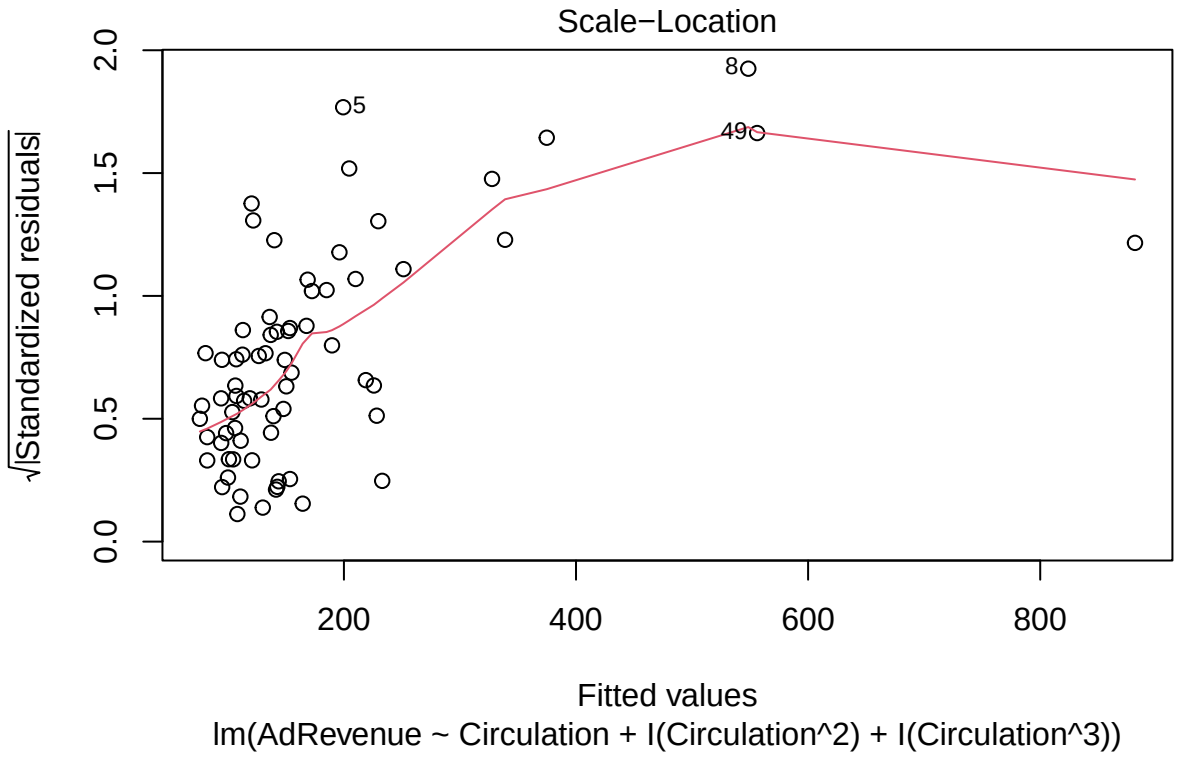
```
# fit a quadratic model (no data transformation)
secondorder <- lm(AdRevenue ~ Circulation + I(Circulation^2), data = rev)
plot(secondorder)
```


Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ Circulation + I(Circulation^2))
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation + I(Circulation^2))



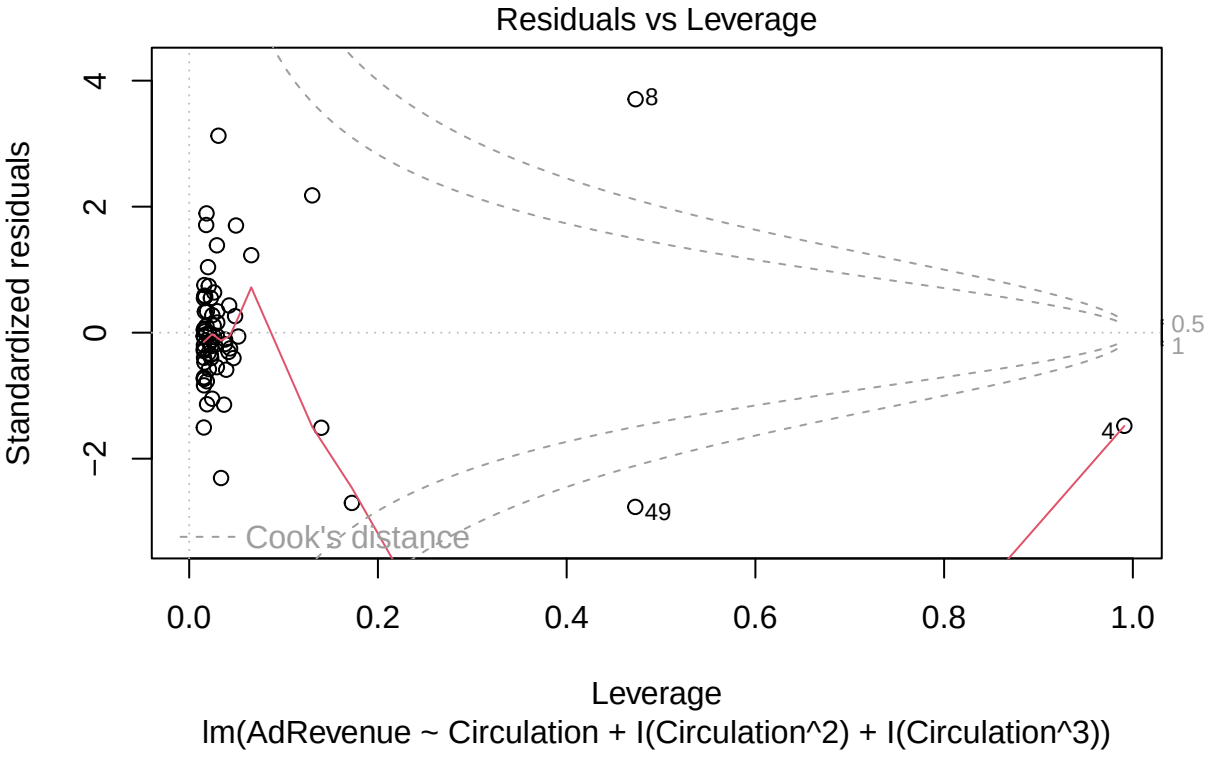

```
# fit a cubic model
thirdorder <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3), data = rev)
plot(thirdorder)
```


Residuals vs Fitted
Residuals
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))




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

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



b)

```
# intervals represent the cost in thousands of dollars
predict(secondorder, data.frame(Circulation = 0.5), interval="prediction")
```

```
##        fit      lwr      upr
## 1 102.8294 19.47858 186.1802
```

```
predict(secondorder, data.frame(Circulation = 20), interval="prediction")
```

```
##        fit      lwr     upr
## 1 582.3869 490.5858 674.188
```

```
predict(thirdorder, data.frame(Circulation = 0.5), interval="prediction")
```

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

```
predict(thirdorder, data.frame(Circulation = 20), interval="prediction")
```

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

c)

We notice a lot of weaknesses in our quadratic and cubic models. First, for both the quadratic and the cubic model, the qq-line is clearly not straight and trails off dramatically on the ends of each plot, indicating non-normality. Also, looking at the scale-location plots, both models show an increasing trend of the standardized residuals, indicating non-constant variance. Also, the leverage plot shows the existence of high influence points in both models having Cook's distances greater than 1.

Question 2 Part C

a)

Comparing all three models, the linear model with the data transformed using the log function is the best. This is because we were able to achieve normality of the data, maintain an appropriate scaling of fitted values, and have no bad leverage points/high influence points. All of these are violated in the quadratic/cubic models.

b)

For the prediction intervals, I would recommend the intervals posed by the linear model as I have mentioned in part a) about how it is the most accurate model out of the three.

Question 3

a)

For each discrete value of x, calculate the variance by computing $\sigma^2 = \sum(x_i - \bar{x})/n$, where x_i represents the different observations under a specific discrete x value and $\bar{x}$ represents the average of these observation quantities. Then, derive the standard deviation by taking the square root of the variance.

b)

The method above only works with discrete values of x and not continuous sets because looking at the formula, it can only account for calculating the standard deviation of one variable (which, in this case, is the y-variable representing the number of rooms cleaned). Thus, if x were to be continuous, we would essentially have to calculate an infinite number of standard deviations since there would be infinite x-values in any non-empty continuous interval.