Stats101A, Spring 2023 - Homework 6

Luke Villanueva - 206039397

05/12/23

Problem 1 (Ch.3 #1b)

Yes, based on the data and graph, we can see multiple signs that the model fits the data well. One sign is that the standardized residuals do not exceed past 2 standard deviations. This implies that all the data points generally follow the model. Another sign would be that the p-values of the intercept and slope are close to 0, which implies that the intercept and slope are very likely to not be sampled randomly.

Problem 2 (Ch. 3 #3 A, B, C)

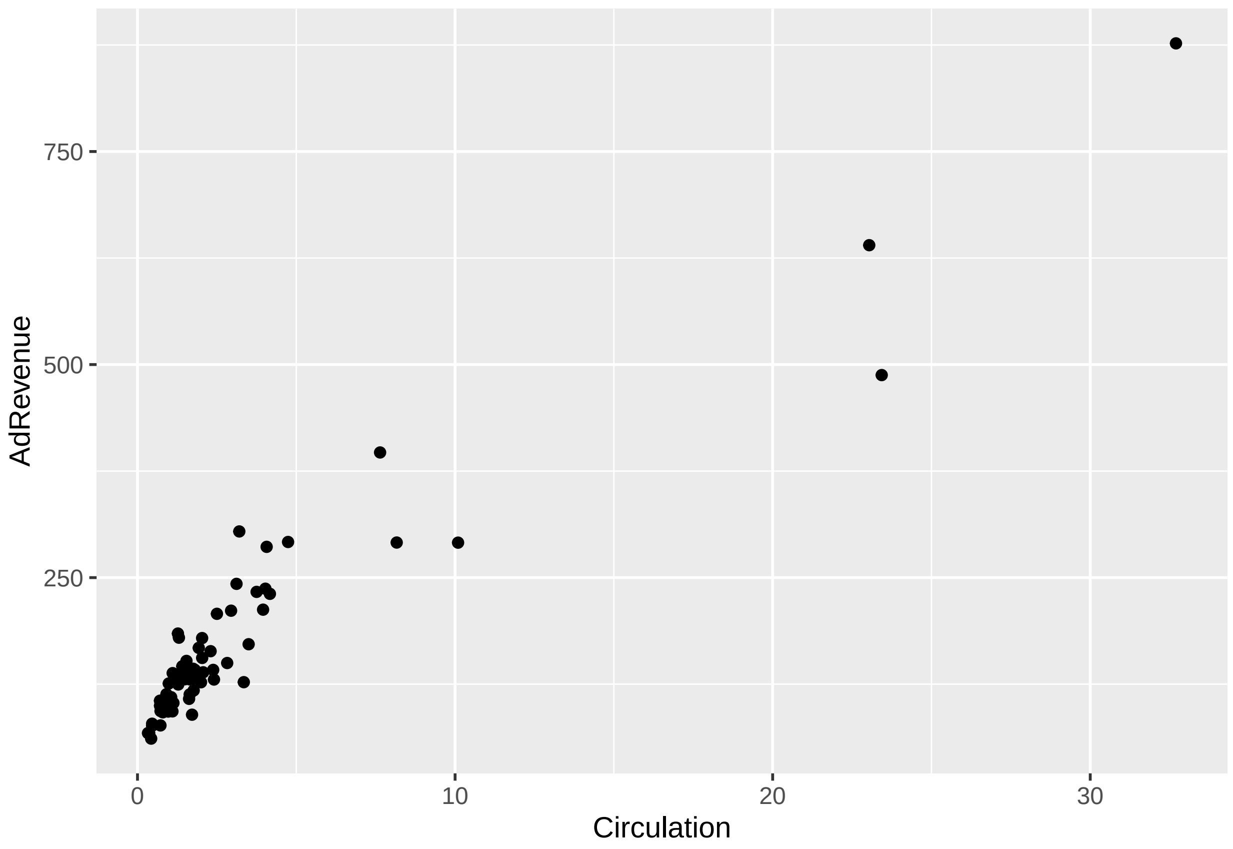
```r
library(tidyverse)
ad <- read.csv(paste0(getwd(),"/AdRevenue.csv"), header = TRUE)
glimpse(ad)
```

```
## Rows: 70
## Columns: 4
## $ Magazine                    <chr> "People", "Better Homes and Gardens", "Time~
## $ PARENT.COMPANY..SUBSIDIARY <chr> "Time Warner, (Time Inc.)", "Meredith Corp.~
## $ AdRevenue                   <dbl> 233.259, 396.865, 286.108, 876.907, 304.185~
## $ Circulation                 <dbl> 3.751, 7.639, 4.067, 32.700, 3.205, 4.741, ~
```

A

a.

```r
# plot to check linearity
ggplot(ad) + geom_point(aes(Circulation, AdRevenue))
```



```r
# linear model of ad rev based on circulation
m1 <- lm(AdRevenue~Circulation, ad)

# log models
m2 <- lm(log(AdRevenue)~Circulation, ad)

m3 <- lm(AdRevenue~log(Circulation), ad)

m4 <- lm(log(AdRevenue)~log(Circulation), ad)

# sqrt models

m5 <- lm(sqrt(AdRevenue)~Circulation, ad)

m6 <- lm(AdRevenue~sqrt(Circulation), ad)

m7 <- lm(sqrt(AdRevenue)~sqrt(Circulation), ad)
```

```r
# plot checks

plot(m1)
plot(m2)
plot(m3)
plot(m4)
plot(m5)
plot(m6)
plot(m7)
```

From eyeing the residual plots, the model with the most randomly scattered residual plots, most constant variance, and least amount of potential bad leverage points would be model 4, which is the model taking the log of both variables.

b.

```
coef(m4)
```

```
##      (Intercept) log(Circulation)
##         4.674734         0.528758
```

```
# i.
predict(m4, list(Circulation = 0.5), interval = "p", level = 0.95)
```

```
##        fit      lwr    upr
## 1 4.308227 3.947855 4.6686
```

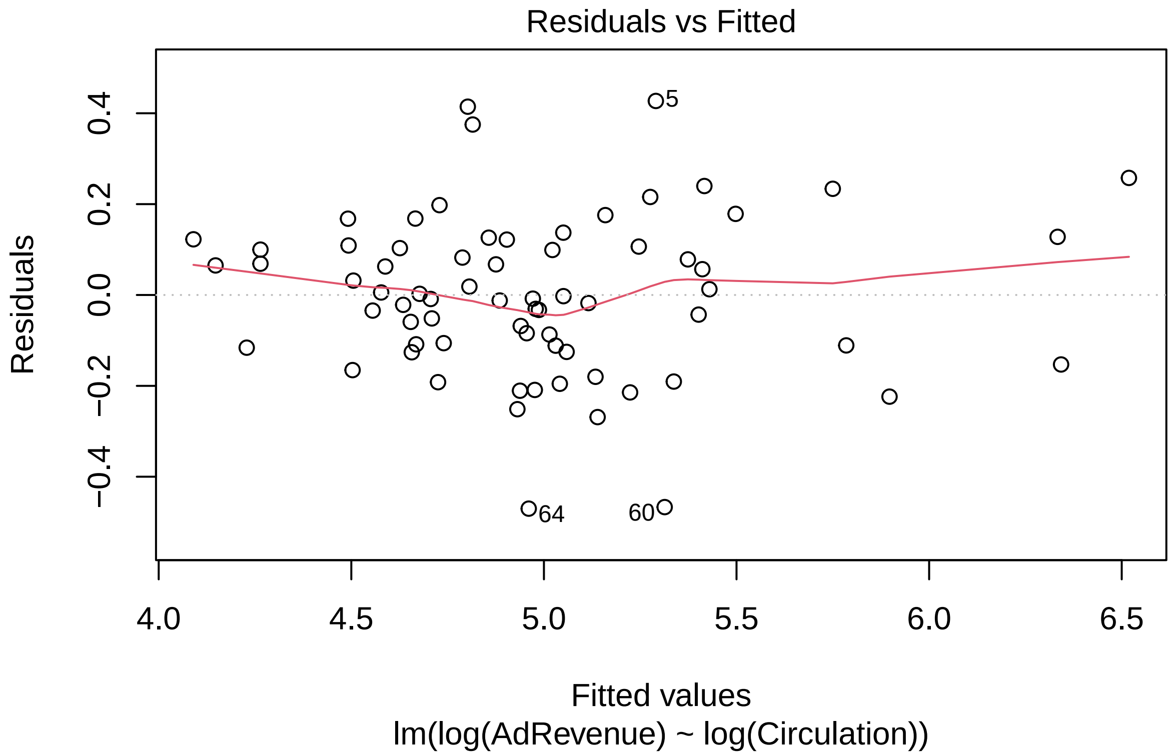
```
# ii.
predict(m4, list(Circulation = 20), interval = "p", level = 0.95)
```

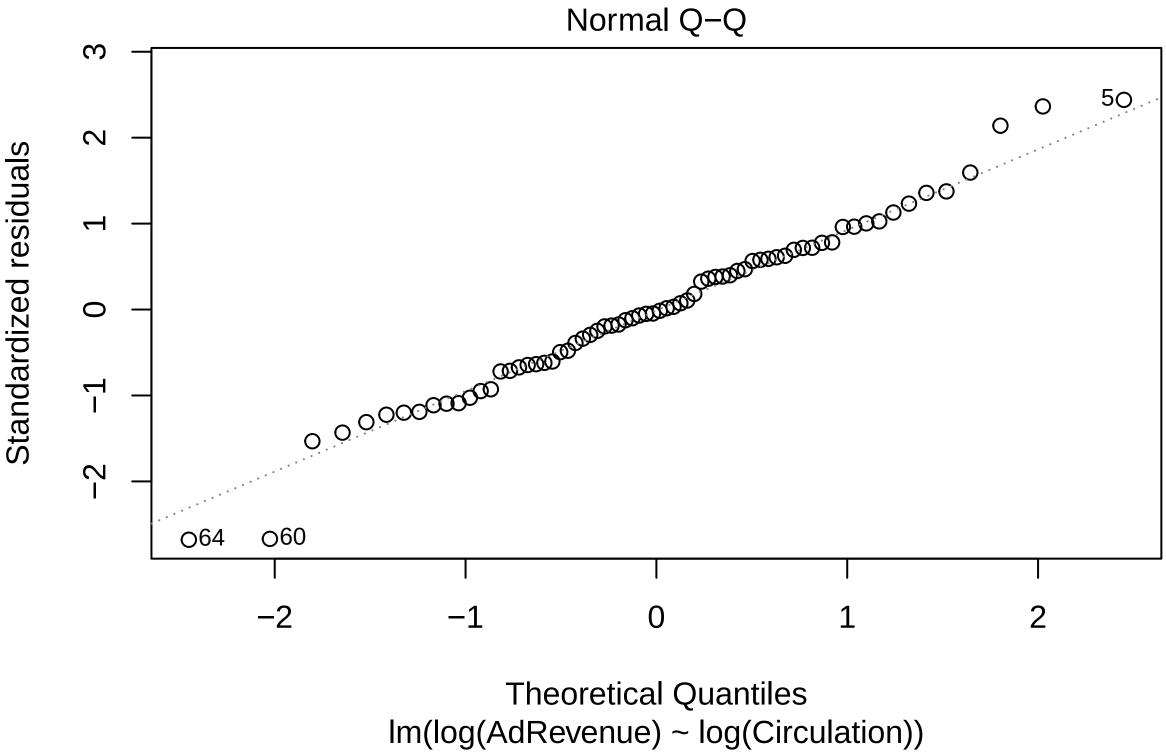
```
##        fit      lwr      upr
## 1 6.258752 5.885815 6.631689
```

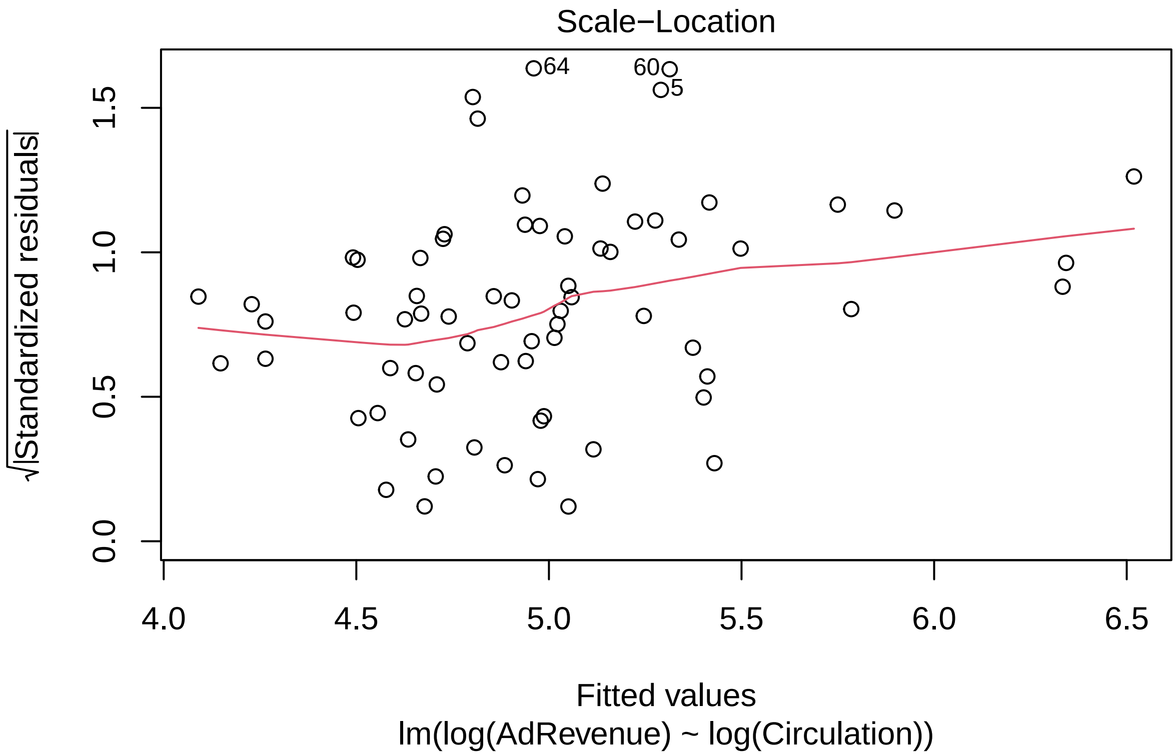
When circulation is 0.5 million, the predicted revenue is around 4.3 thousand dollars. When circulation is 20 million, the predicted revenue is around 6.3 thousand dollars.

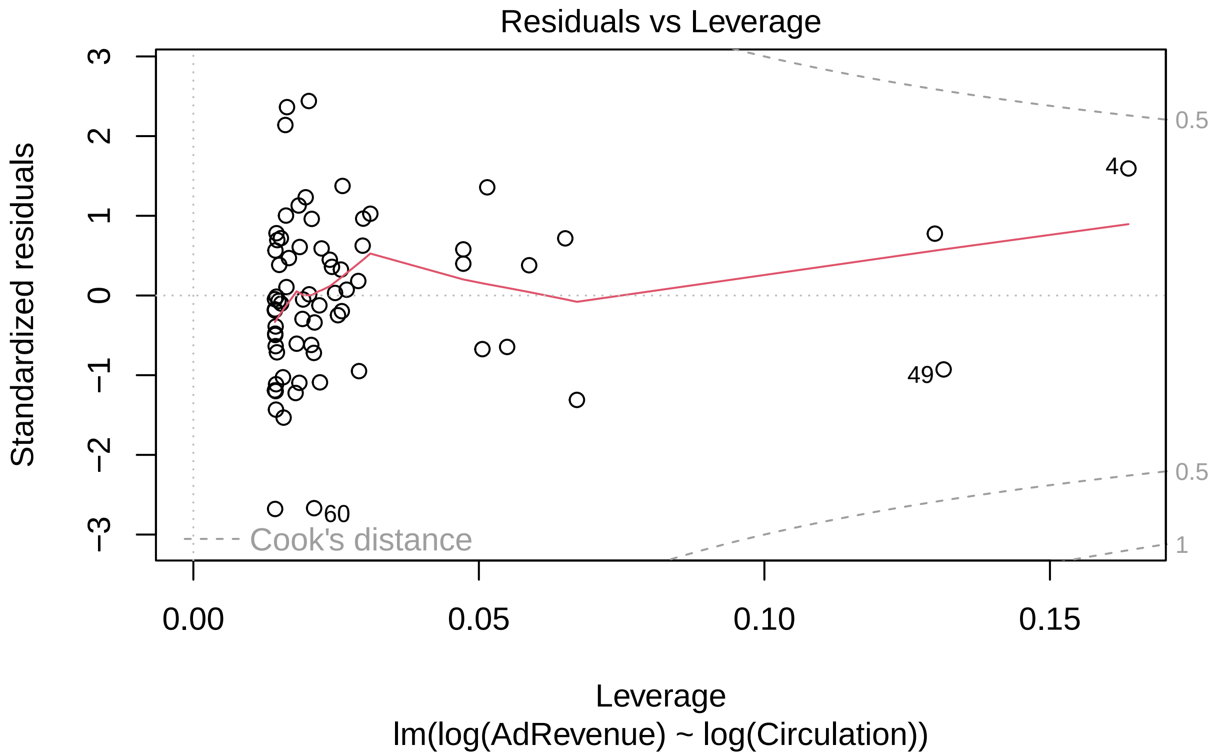
c.

```
plot(m4)
```

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


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


Scale−Location
√|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(log(AdRevenue) ~ log(Circulation))
64
60
5



Residual: There seems to be some trend and clumping in the residuals, which implies the data might have inconsistent variance and/or a non-linear behavior.

QQ: Generally, most points follow a normal distribution. However, there are a handful of outliers that do not, implying there is at least some but not solid normality to the data.

Scale Location: Not only is there clumping, but there is also a very vague positive trend in the data. This implies the data does not have constant variance.

Std. Residuals Vs Leverage:

```r
4/nrow(ad)
```

```
## [1] 0.05714286
```

Since the "big" leverage amount is 0.05, there doesn't seem to be any observations with a standard deviation over 2 that has a leverage over 0.05. This implies that there aren't reasonable outliers that would heavily bias the model. This validates the lack of bad leverage points.

B.

a.

```r
# 2nd order
secondOrderModel <- lm(AdRevenue~Circulation + I(Circulation^2), ad)
```

```r
# 3rd order
thirdOrderModel <- lm(AdRevenue~Circulation + I(Circulation^2) + I(Circulation^3), ad)
```

```r
# check residual and leverages
plot(secondOrderModel)
plot(thirdOrderModel)
```

Because the second order polynomial has the most constant variances and linearity comapared to the third order, we will use the second order model.

b.

```r
# i.
# 2nd order
predict(secondOrderModel, list(Circulation = 0.5), interval = "p", level = 0.95)
```

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

```r
# ii.
predict(secondOrderModel, list(Circulation = 20), interval = "p", level = 0.95)
```

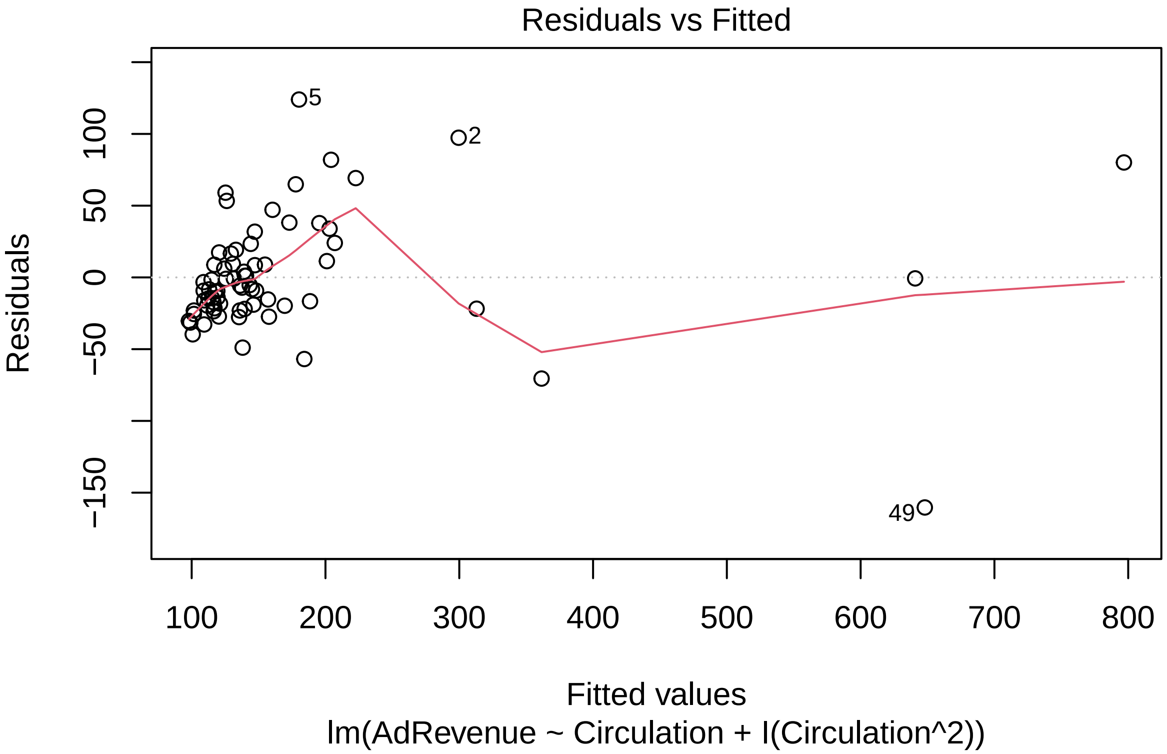
```
##        fit      lwr     upr
## 1 582.3869 490.5858 674.188
```

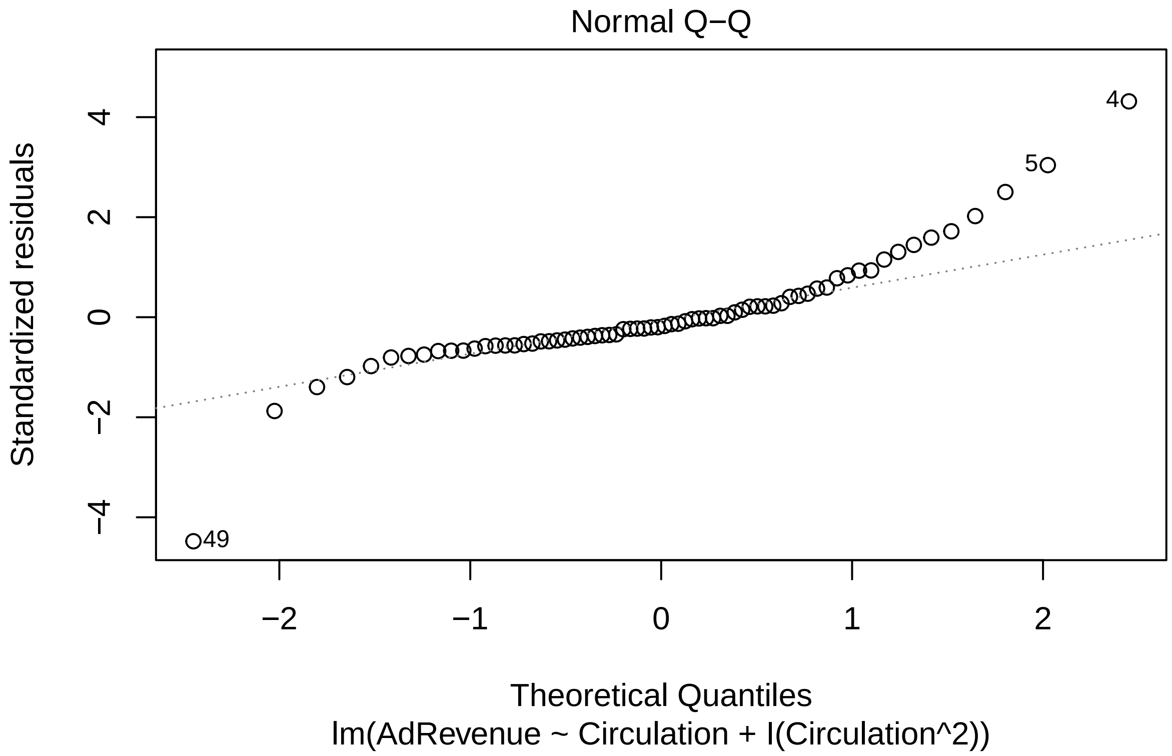
When circulation hits 0.5 million, ad revenue is predicted to be around 102.83 thousand dollars with the order-2 polynomial.

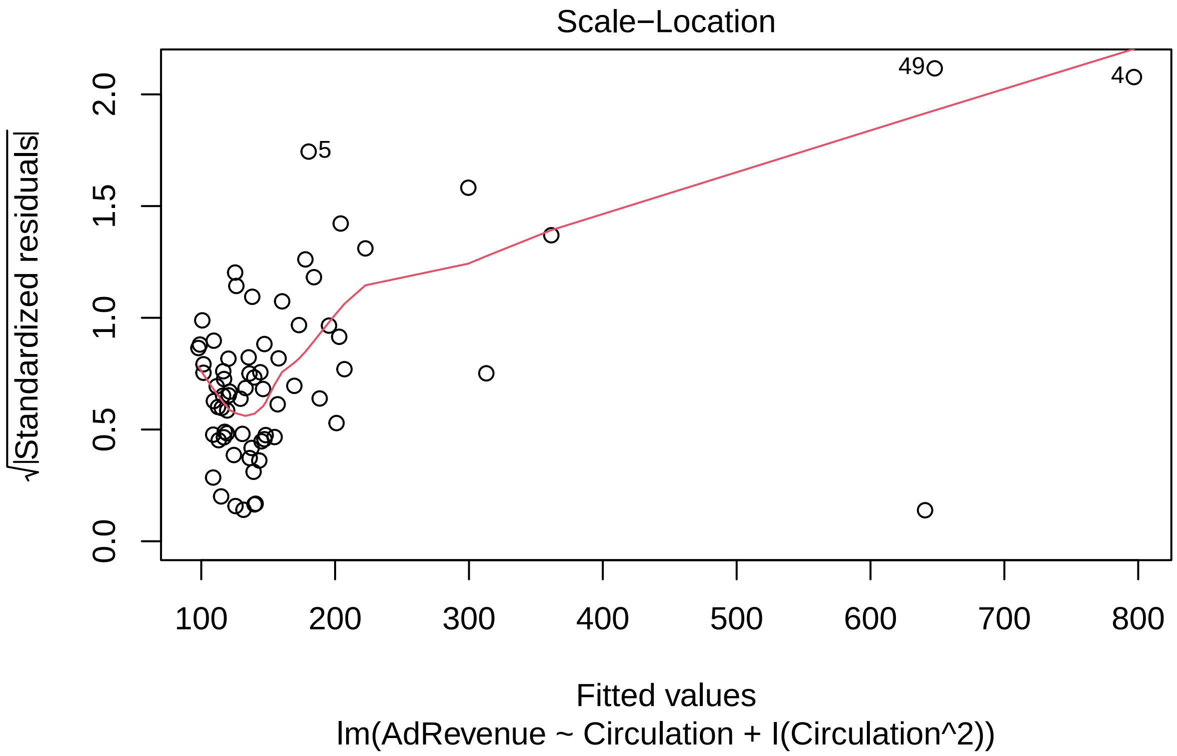
When circulation hits 20 million, ad revenue is predicted to be around 582.39 thousand dollars with the order-2 polynomial.

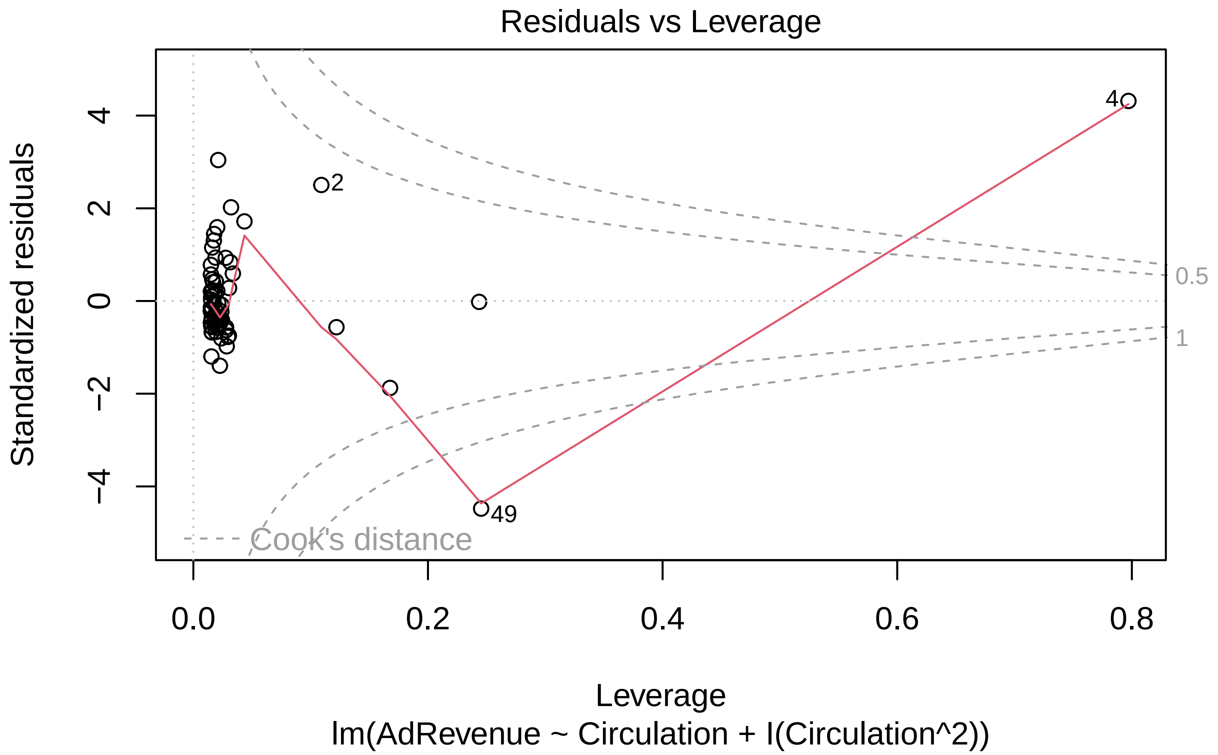
c.

```r
plot(secondOrderModel)
```


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


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


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
200
300
400
500
600
700
800
Fitted values
lm(AdRevenue ~ Circulation + I(Circulation^2))
5
49
4



Residuals: There is a lot of clumping and a trend in the outliers. This implies there is inconsistent variance and non-linearity.

QQ: There seems to be a good handful of data points that deviate from the normal line. This implies a lot of observations do not follow a normal curve.

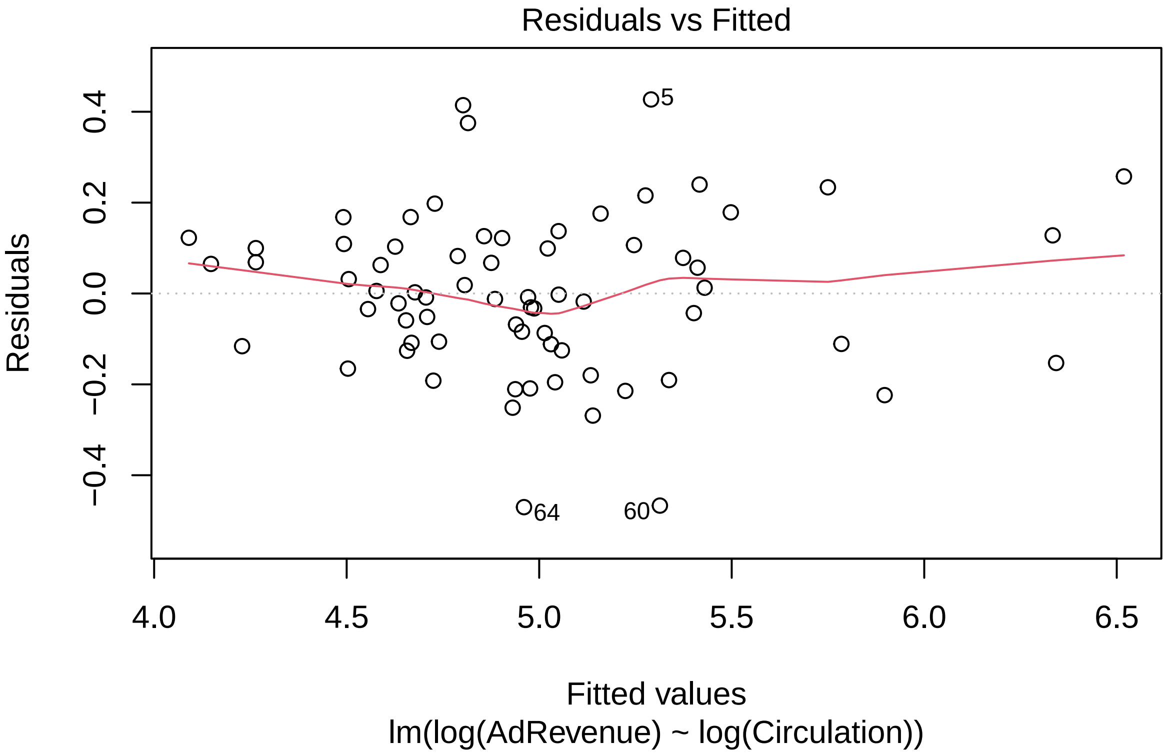
Scale Location: There is a lot of clumping and a positive trend. This implies that there is inconsistent variance.

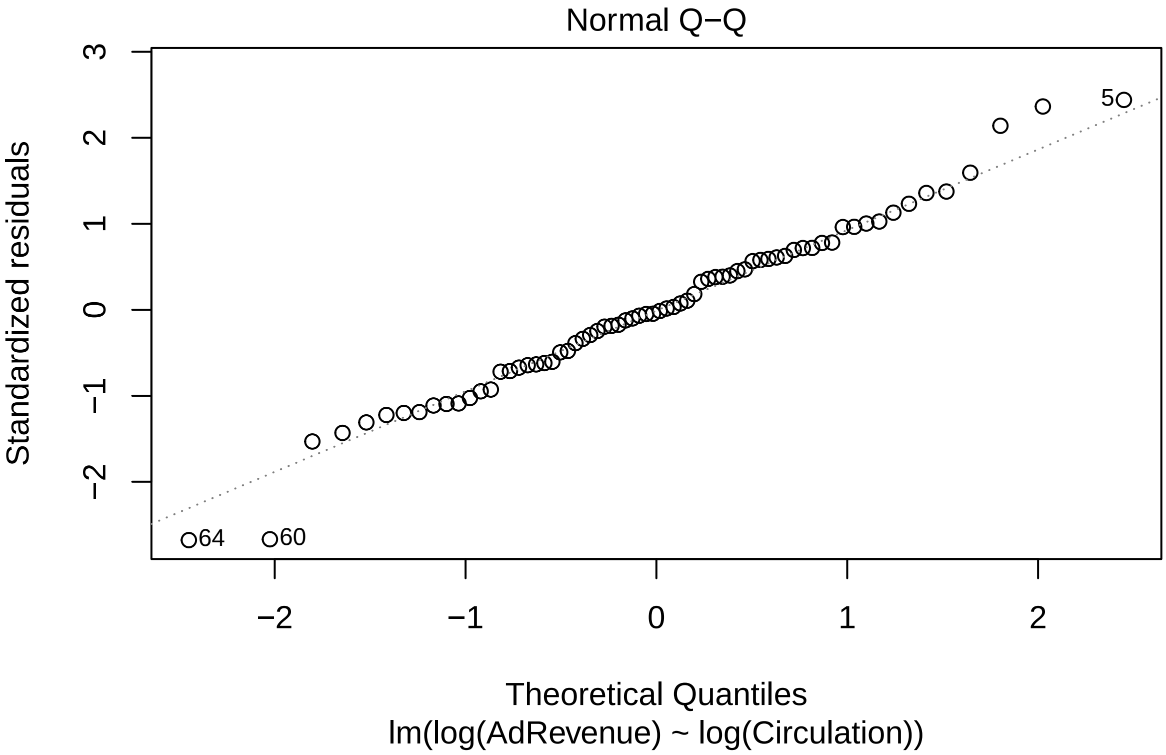
Std. Residuals vs Leverage: There is 1 point that exceeds the “big” 0.05 leverage amount and exceeds past 2 standard deviations. This implies the point might be a bad leverage point, possibly biasing the regression model.

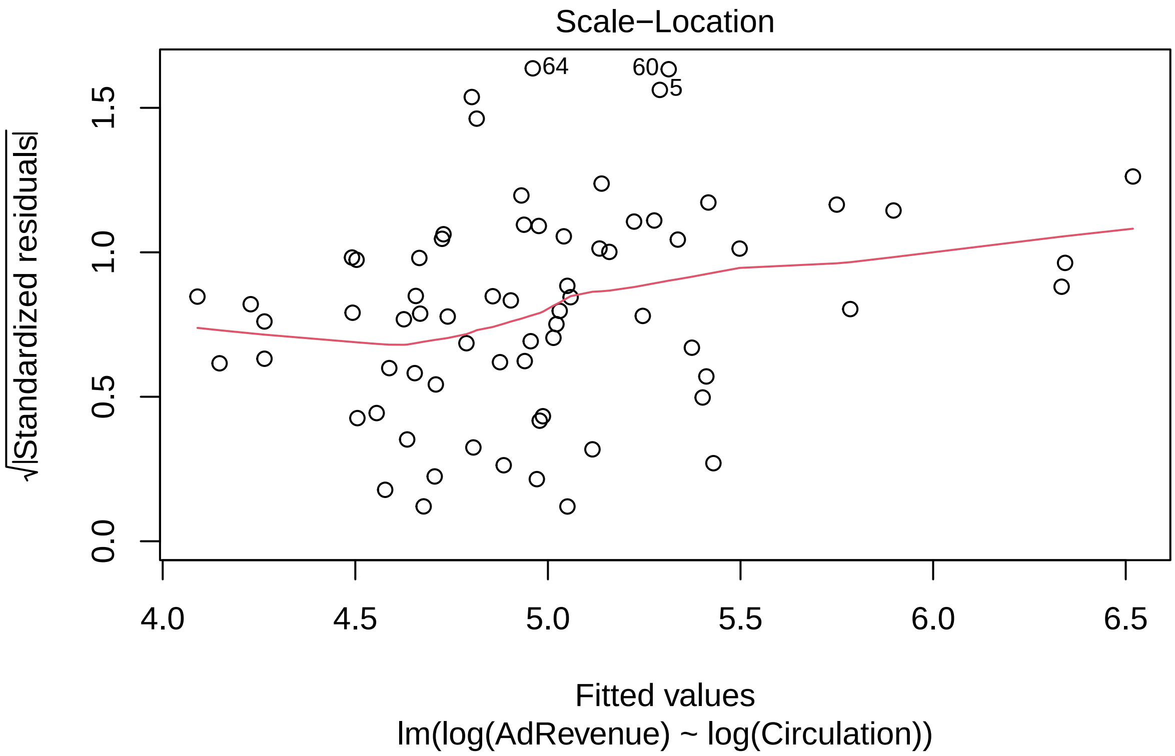
C.

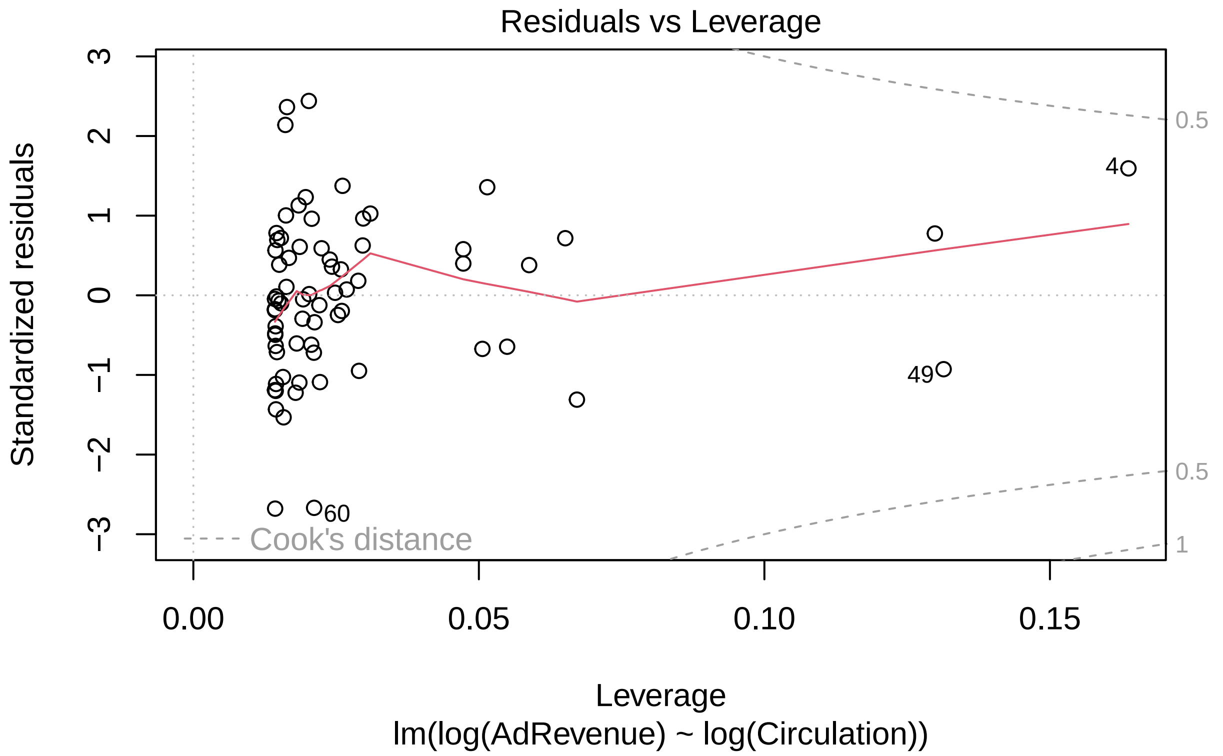
a.

```r
plot(m4)
```

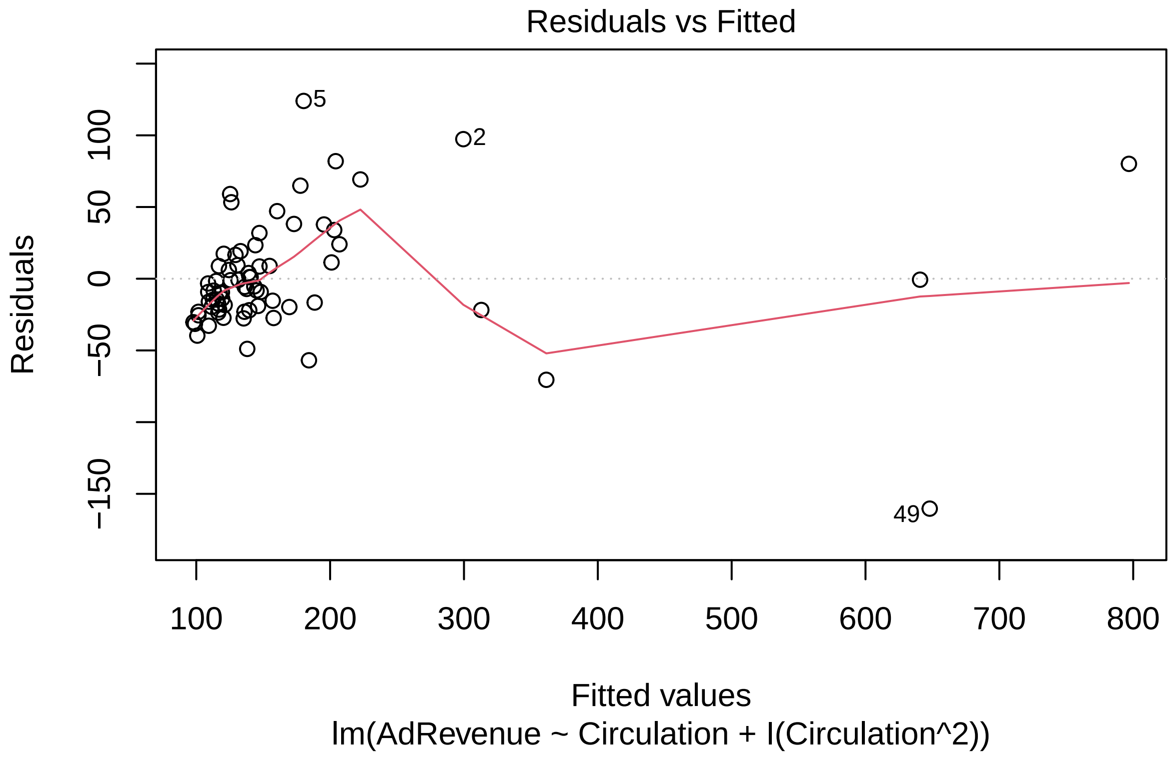

Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
4.0
4.5
5.0
5.5
6.0
6.5
Fitted values
lm(log(AdRevenue) ~ log(Circulation))
5
64
60

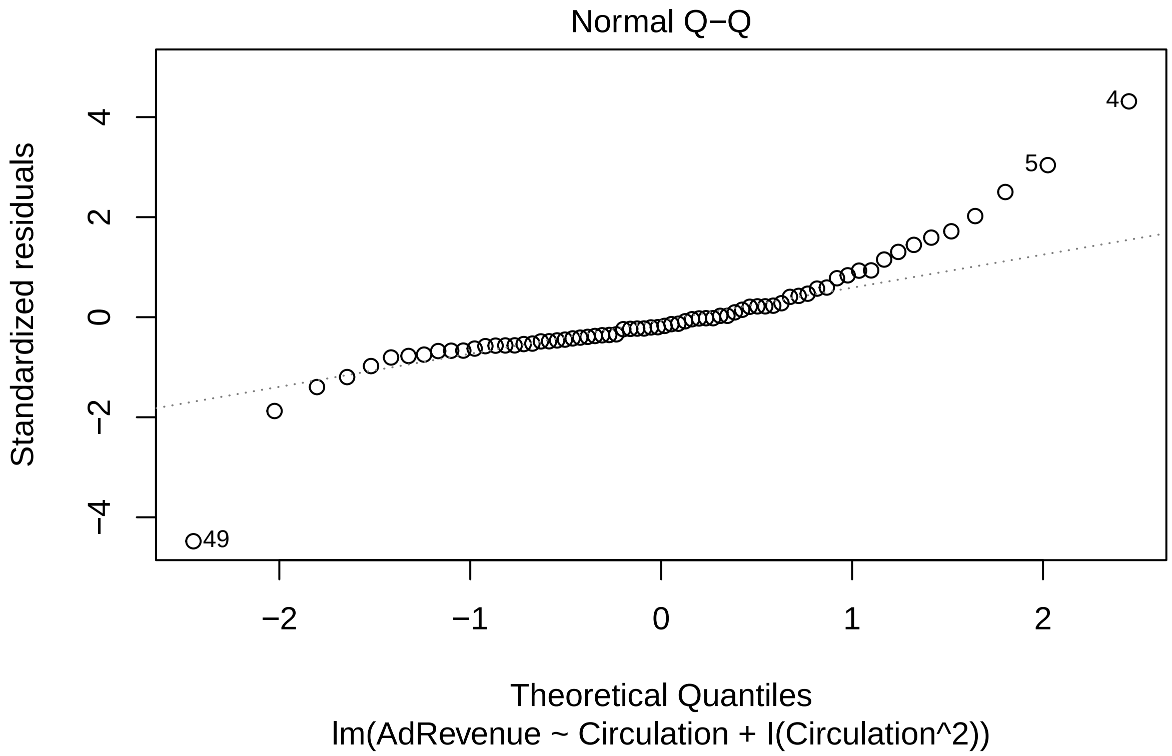

Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(log(AdRevenue) ~ log(Circulation))
64
60
5

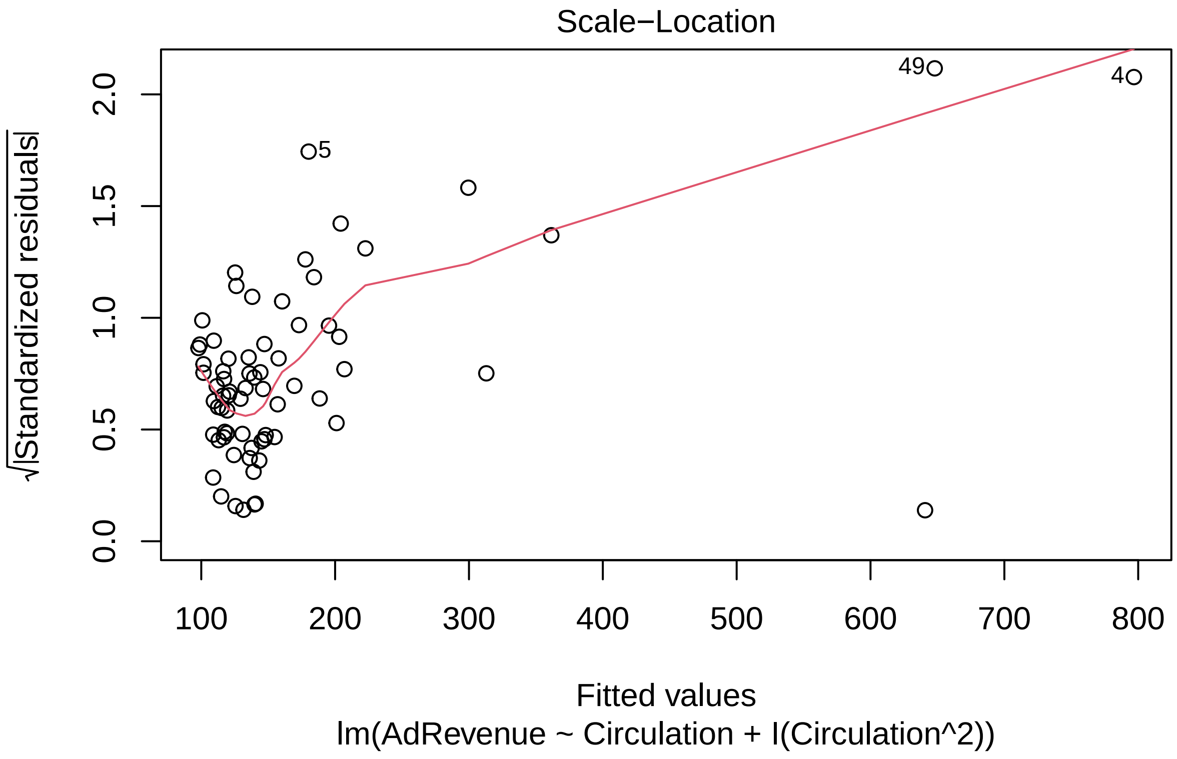

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5
6.0
6.5
64
60
5
Fitted values
lm(log(AdRevenue) ~ log(Circulation))

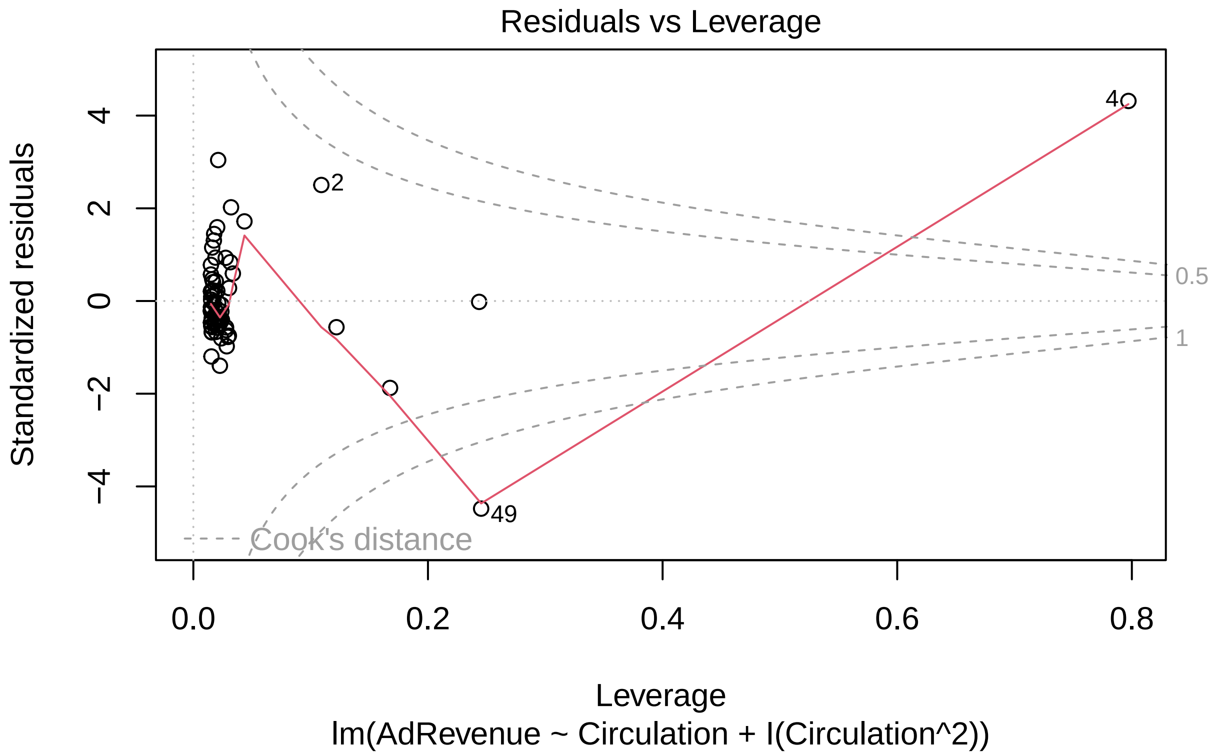


```r
plot(secondOrderModel)
```


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


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


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
200
300
400
500
600
700
800
Fitted values
lm(AdRevenue ~ Circulation + I(Circulation^2))
5
49
4



Based on the residual and leverage plots, clearly model 4 has the better linearity, normality, and consistent variance compared to the polynomial model. Model 4 has a more randomly scattered residual plot, straighter QQ plot, and no potential bad leverage points.

b.

I would recommend the prediction interval of model 4. This is because the standard error for model 4 is smaller than the standard error of the second order polynomial. This lets the prediction for model 4 be more accurate for the same amount of confidence level as the second order polynomial.

Problem 3

a.

For each X, we would take the standard deviation of all the Y's that would correspond to the predictor variable. In other words, if there were $(y_0, ..., y_n)$ response variables for x_1, then the standard deviation of Y at x_1 would be $\sqrt{\frac{\sum_{i=0}^{n}(y_i-\bar{y})^2}{N-1}}$ where N is the number of observations and $\bar{y}$ would be the mean of the y values at x_1.

b.

This is a special case because in usual plots, there is a single Y value corresponding to an X value. This special case has multiple Y values for a single X value, so each X value can have a standard deviation.

Usually, it’d only be possible to take the standard deviation of all X values.