Stats_101A_hw_4_anna_piskun

Anna Piskun

1/31/2020

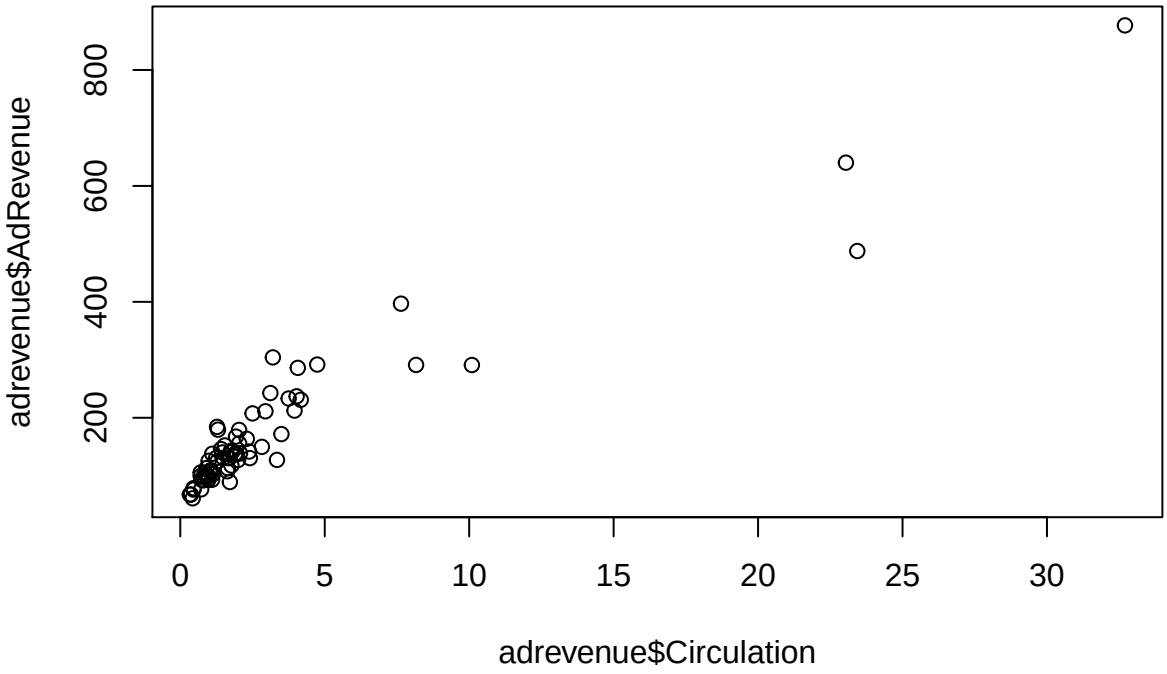
Part A: Chapter 3 Question 1

a) Looking at the plot of Model 1, we can say that there exists a strong, positive, linear relationship. However, looking at the residual plot there is a clear pattern (inverted parabola) that suggests a better fit with a nonlinear model. Therefore there is a better model that can be used to predict fare given the distance.

b) Given the output from R we get that the ordinary regression model follows the following formula: fare = 48.9718 + 0.2197 (distance) with r squared = 0.994 which correlates to 99.4% of the variation in fare being explained by the model. A high rsquared value indicates a good model since it shows that almost all of the variation can be explained by our model. Likewise, since both pvalues of the intercept and slope are less than 0.05, the regression model is signficant and fits the data well.

Part B: Chapter 3 Question 3

```
setwd("~/Desktop")
adrevenue <- read.csv("AdRevenue.csv")
plot(adrevenue$AdRevenue~adrevenue$Circulation)
```

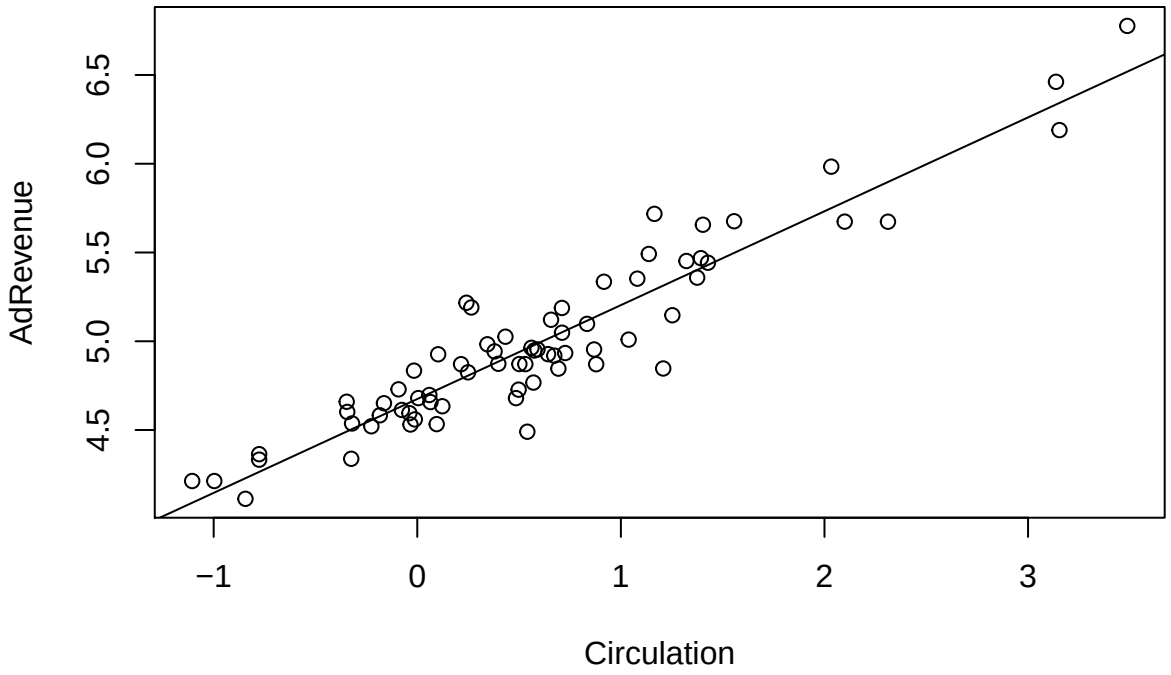


Part A:

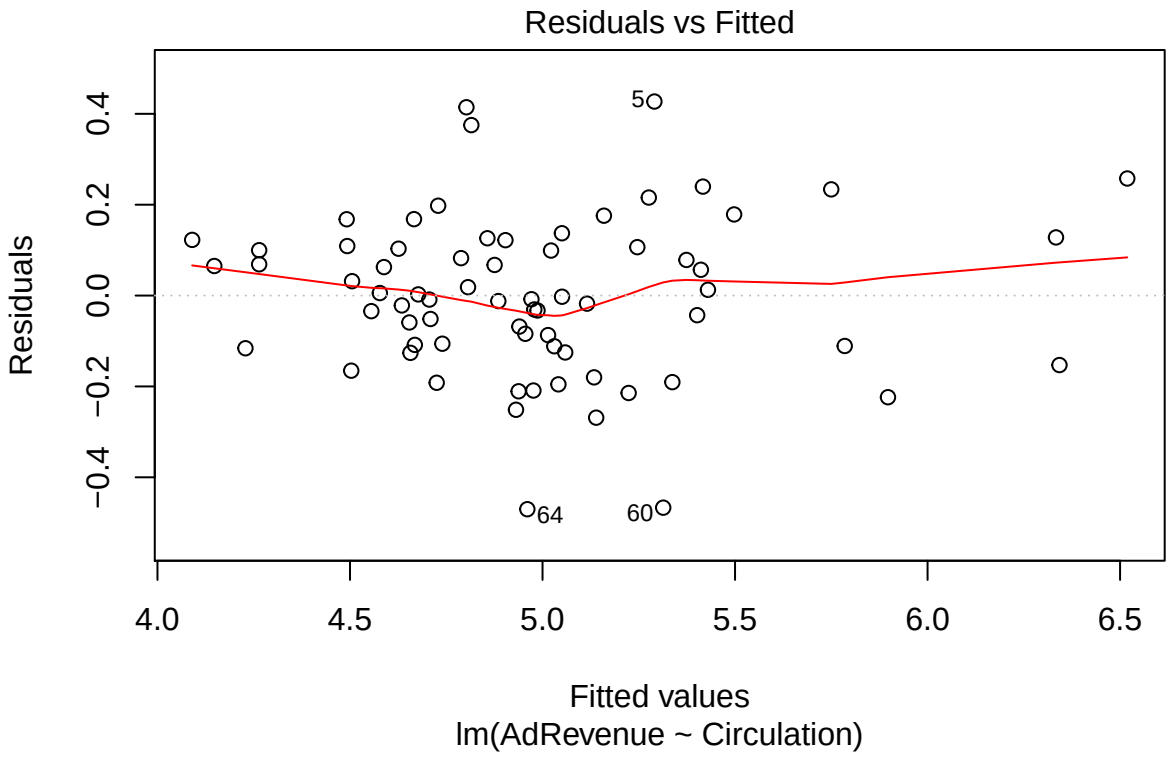
a) Develop a simple linear regression model based on least squares that predicts advertising revenue per

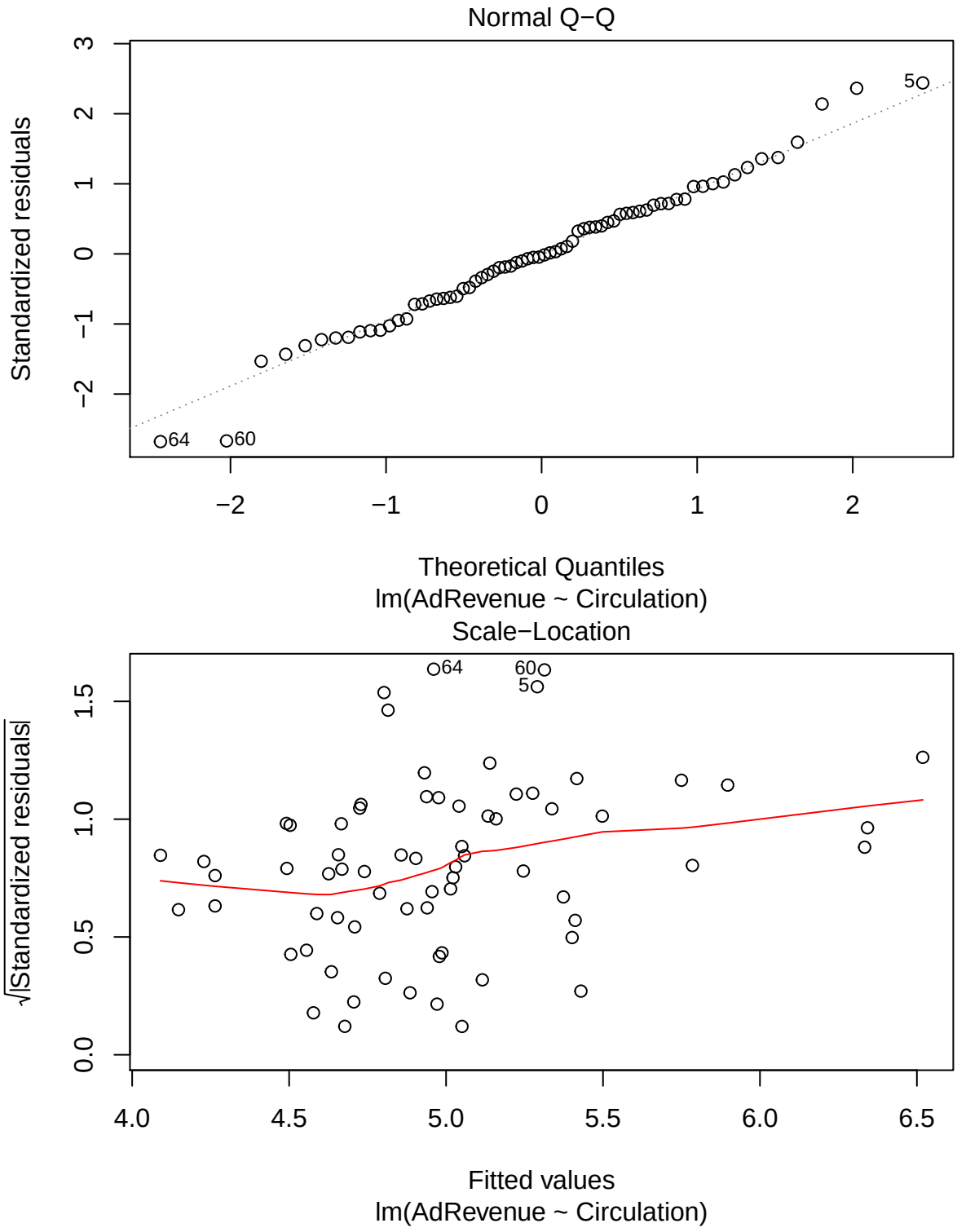
page from circulation (i.e., feel free to transform either the predictor or the response variable or both variables). Ensure that you provide justification for your choice of model.

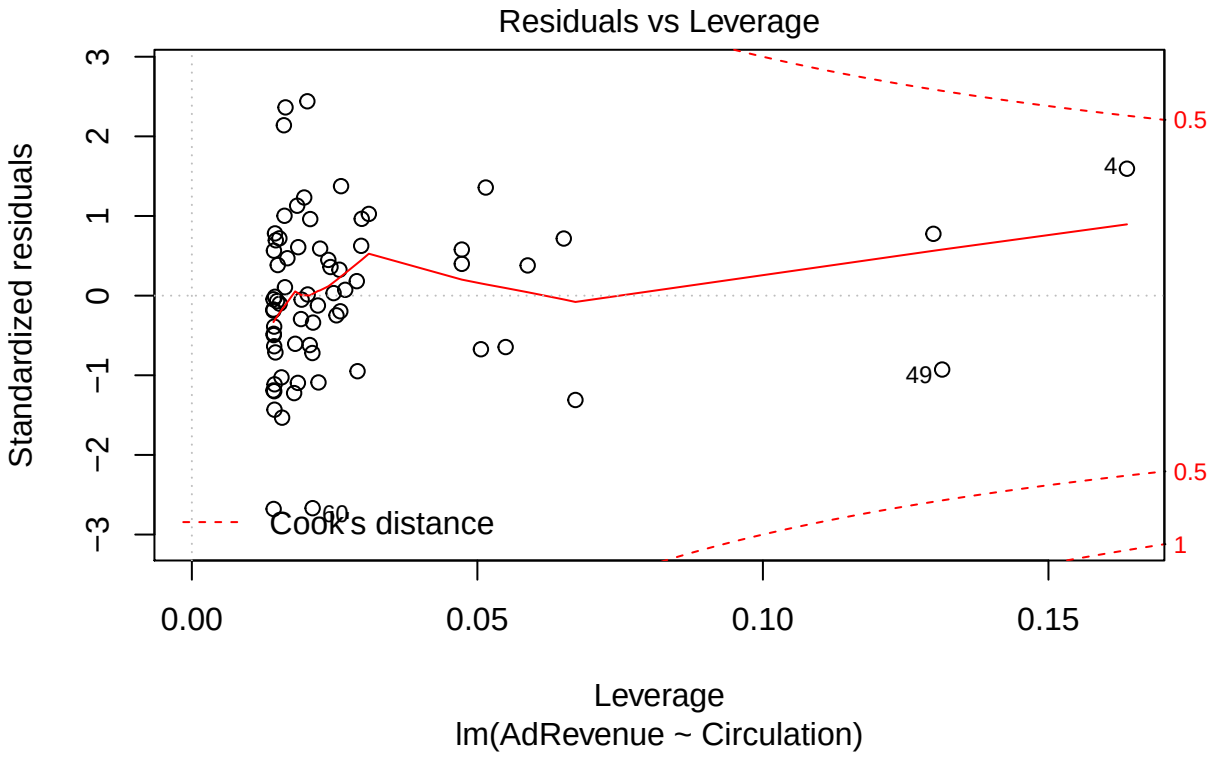
```
adrevenue_log <- transform(adrevenue, AdRevenue = log(AdRevenue), Circulation = log(Circulation))
model1 <- lm(AdRevenue~Circulation, data = adrevenue_log)
plot(AdRevenue~Circulation, data = adrevenue_log)
abline(model1)
```




```
plot(model1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation)
Scale−Location
√|Standardized residuals|
Fitted values
lm(AdRevenue ~ Circulation)



For AdRevenue, values range from 1000's to 100,000's (two orders of magnitude increase), so using a logarithmic transformation will allow us to fit the data better. Looking at the ordinary regression plot, the data seems to fit the model well with a strong, positive, linear relationship and there is no clear pattern in the residual plot indicating that our linear model (with a log transformation) is a good fit.

(b) Find a 95% prediction interval for the advertising revenue per page for magazines with the following circulations: i) 0.5 million ii) 20 million

```
exp(predict(model1, data.frame("Circulation" = log(.5)), interval = "prediction"))
```

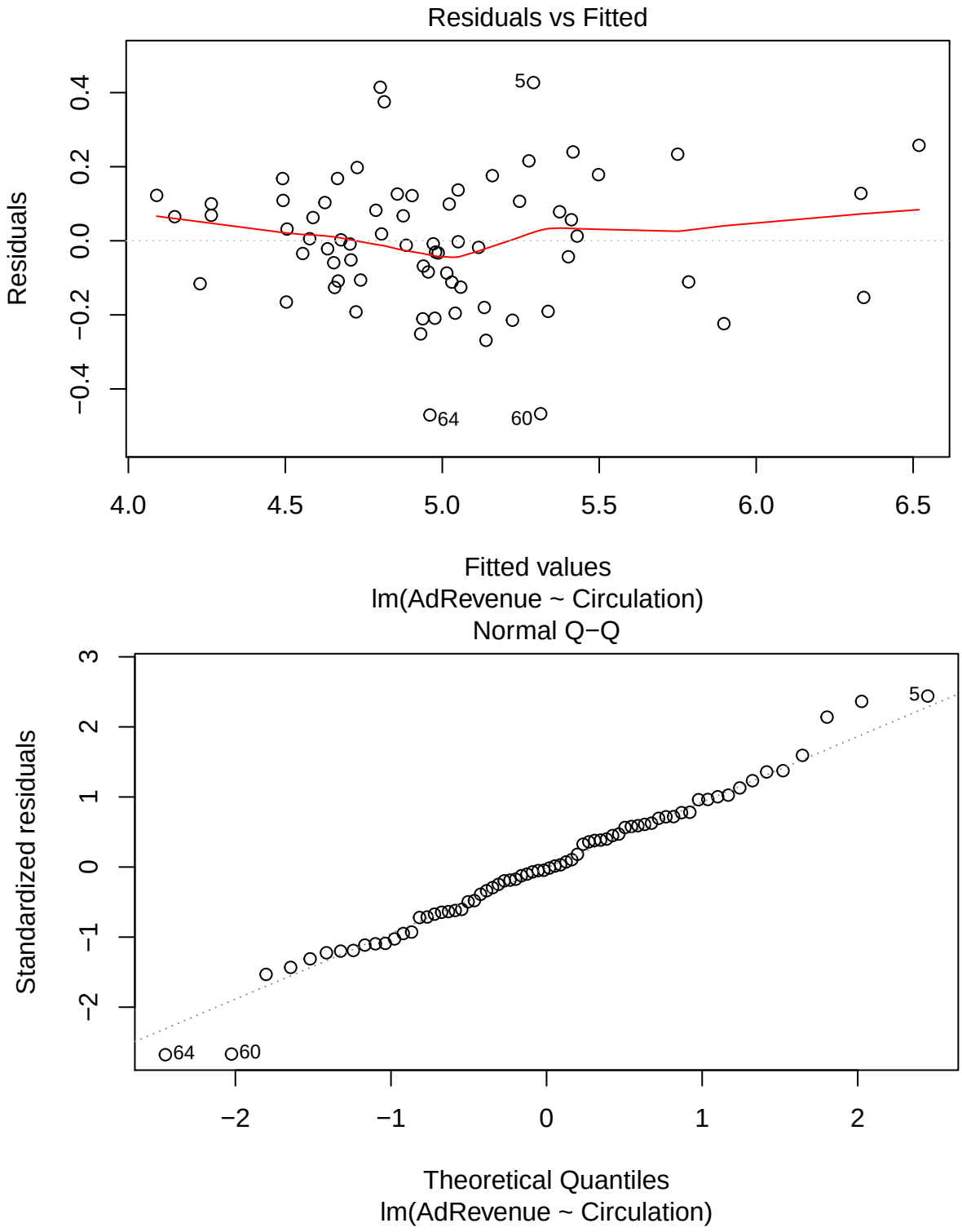
```
##        fit      lwr      upr
## 1 74.30864 51.82406 106.5485
```

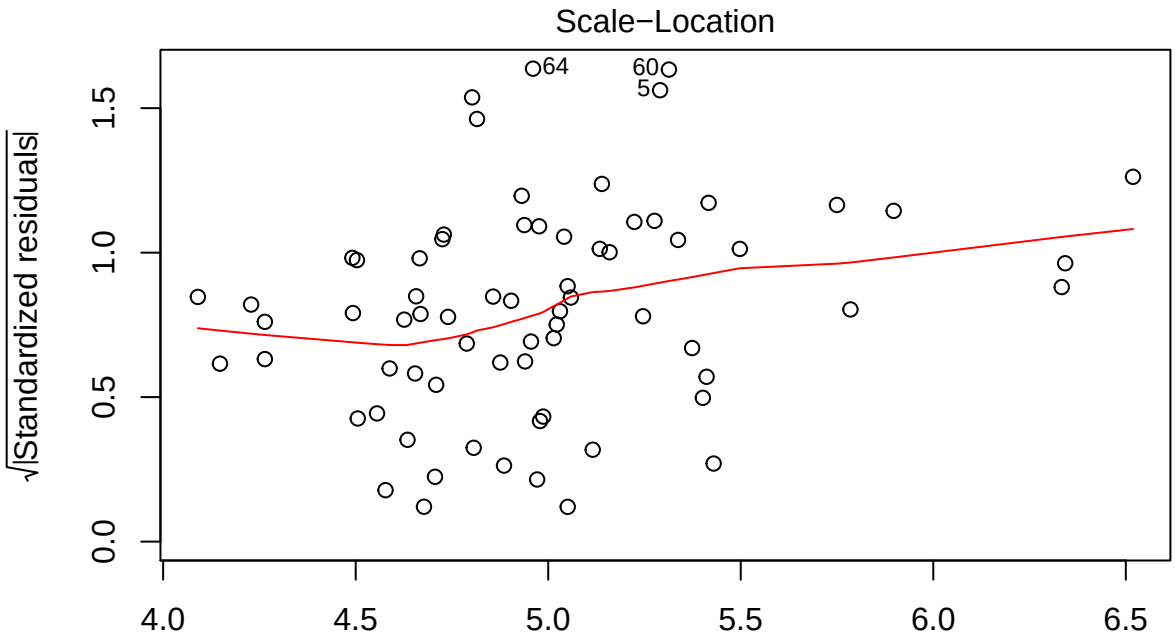
```
exp(predict(model1, data.frame("Circulation" = log(20)), interval = "prediction"))
```

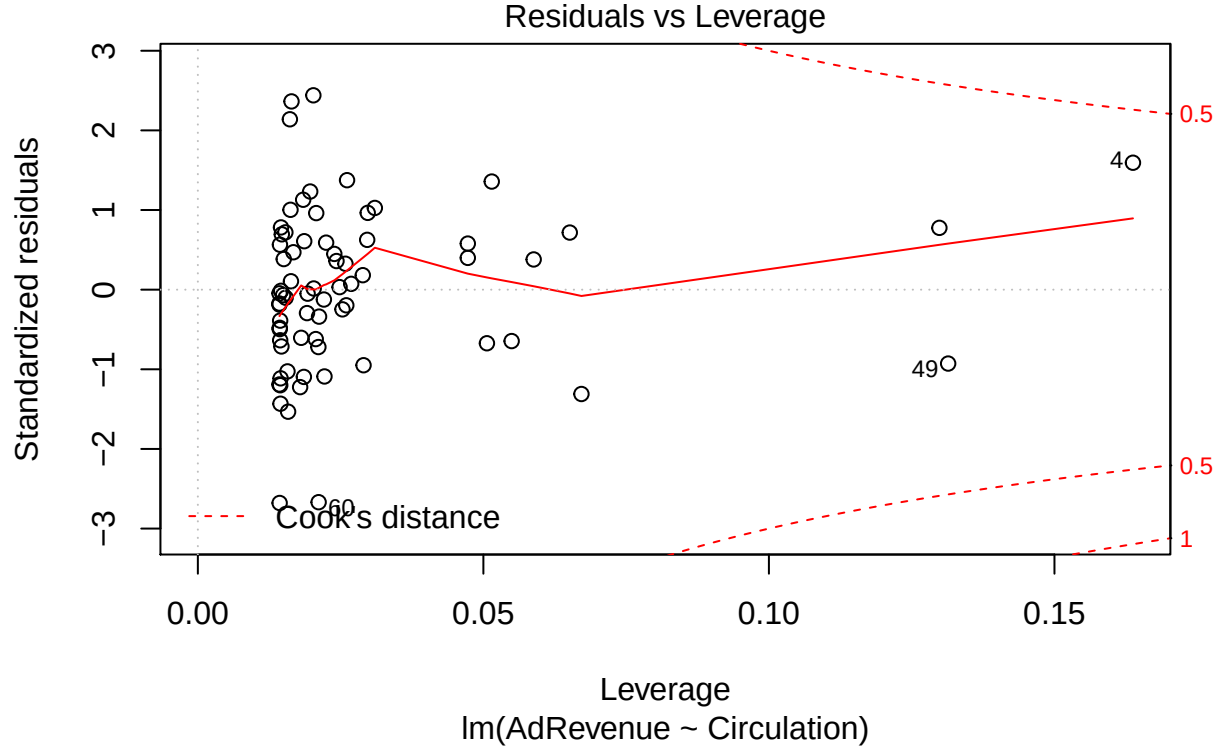
```
##        fit      lwr      upr
## 1 522.5663 359.8958 758.7626
```

(c) Describe any weaknesses in your model.

```
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ Circulation)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ Circulation)
5
64
60



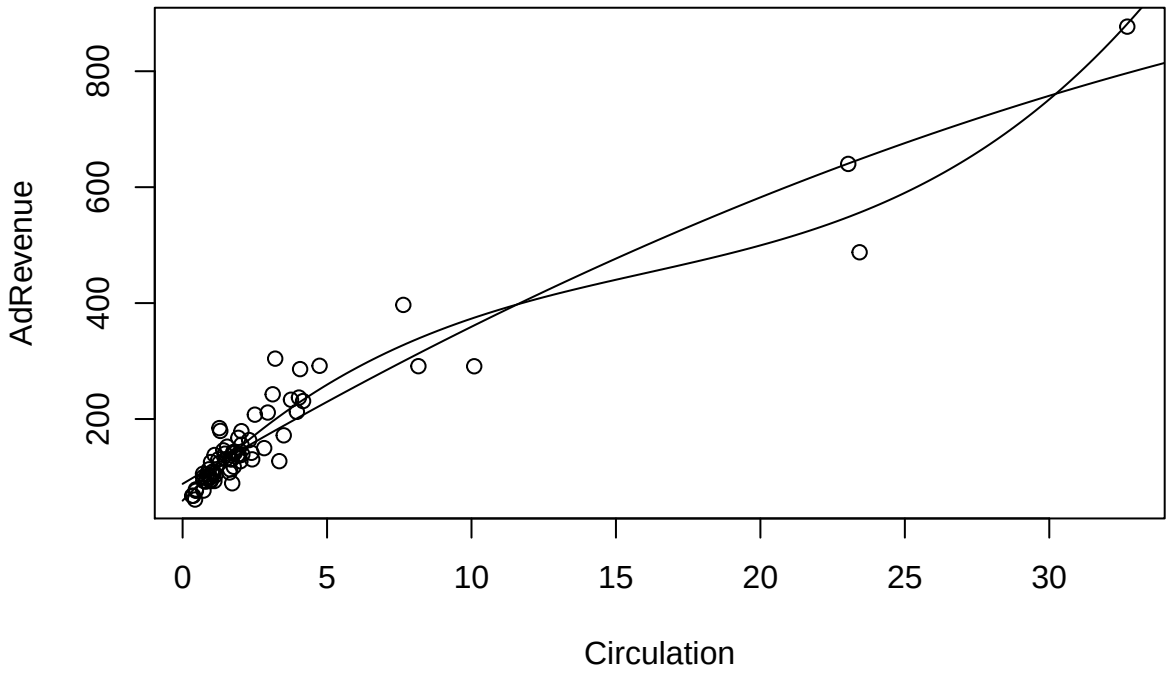


Looking at the normal Q-Q plot we see that it is not completely straight, which may potentially indicate non-normality, and thus illustrate a weakness in our model. Other than that, the residual plot shows no clear pattern, there is no trend in the scale-location plot (indicating constant variance), and there are no points with substantially high leverage.

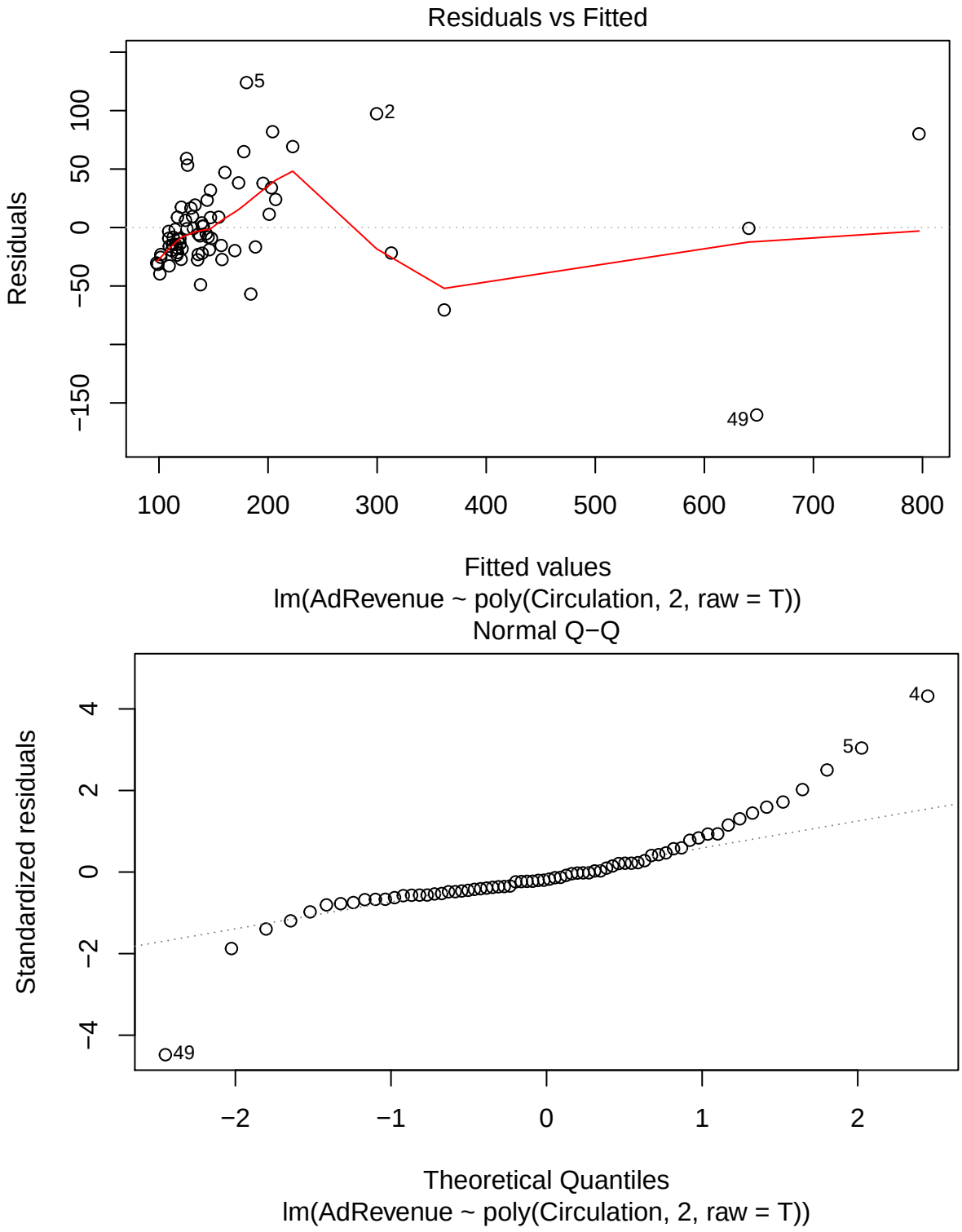
Part B:

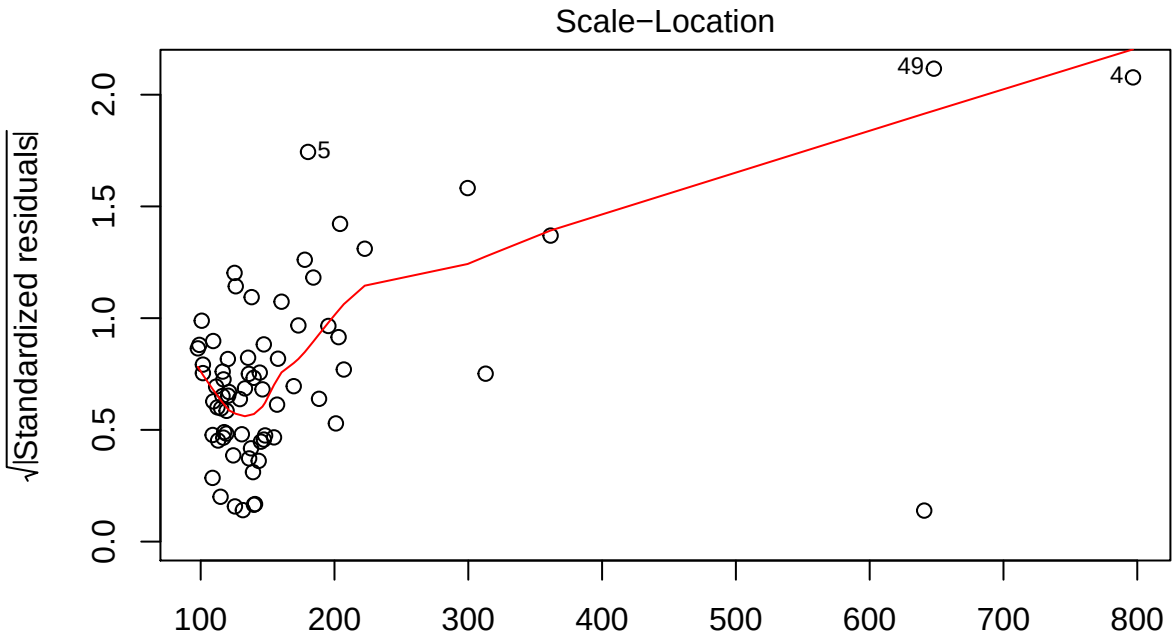
(a) Develop a polynomial regression model based on least squares that directly predicts the effect on advertising revenue per page of an increase in circula- tion of 1 million people (i.e., do not transform either the predictor nor the response variable). Ensure that you provide detailed justification for your choice of model. [Hint: Consider polynomial models of order up to 3.]

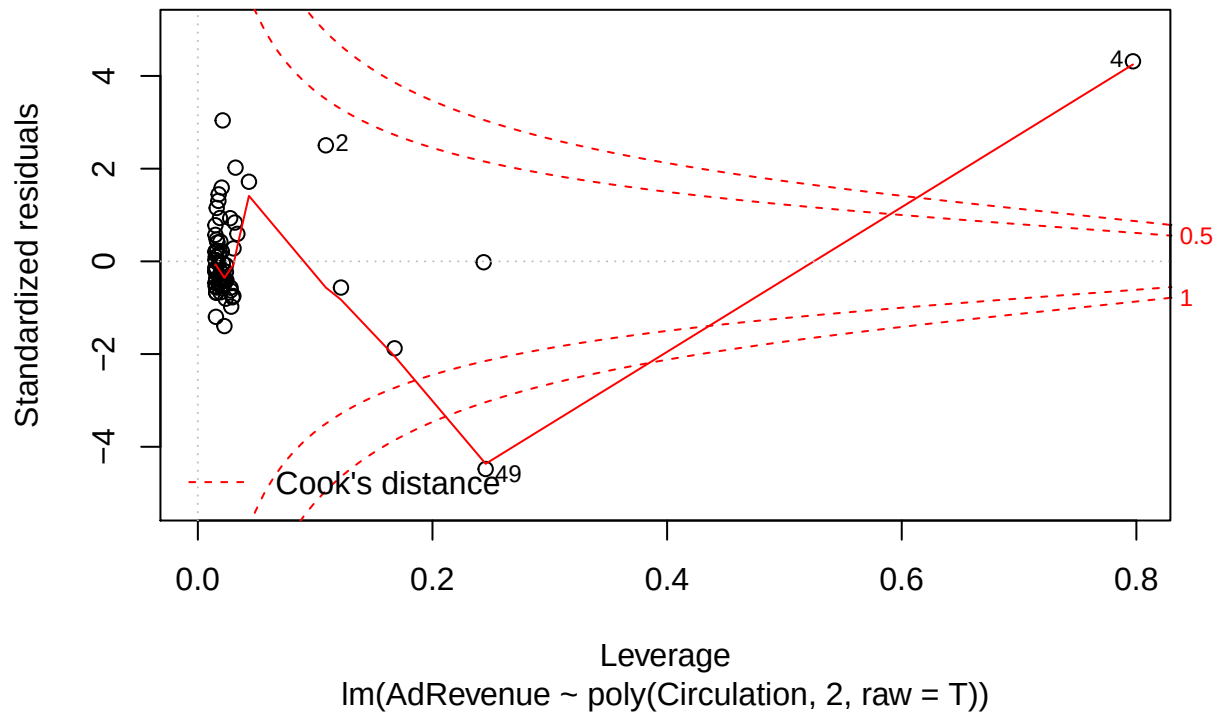
```
plot(AdRevenue~Circulation, data = adrevenue)
model2 <- lm(AdRevenue~poly(Circulation,2,raw=T),data=adrevenue)
xs <- seq(0,40,length=1000)
ys <- predict(model2, data.frame(Circulation=xs))
lines(xs, ys)

model3 <- lm(AdRevenue~poly(Circulation,3),data=adrevenue)
x <- seq(0,40,length=1000)
y <- predict(model3, data.frame(Circulation=x))
lines(x, y)
```



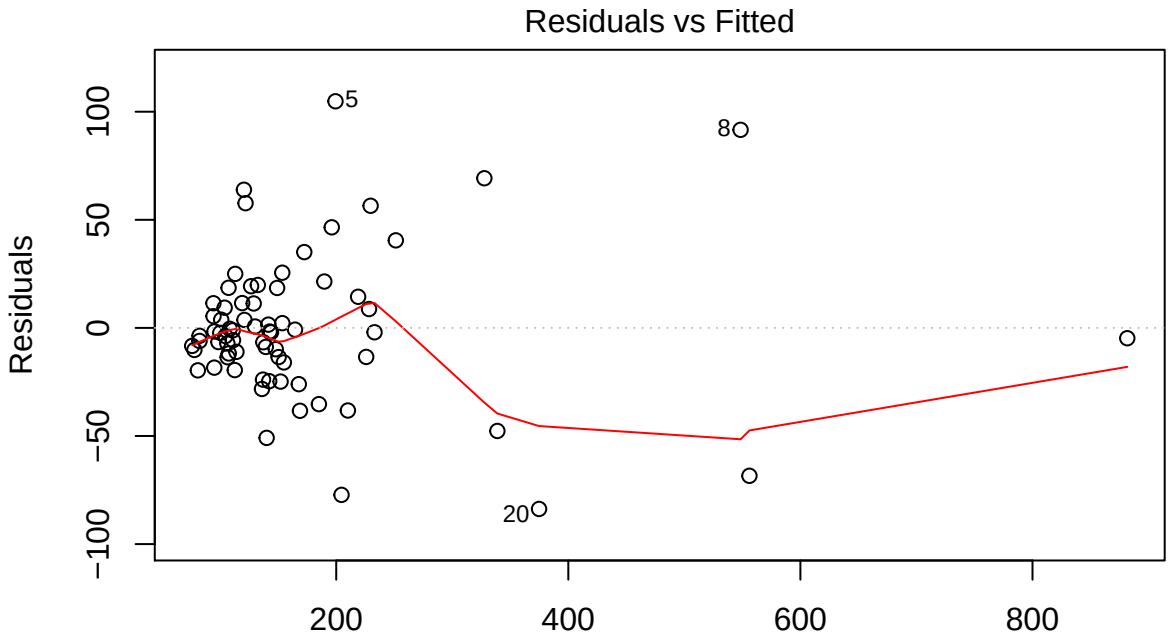

```
plot(model2)
```

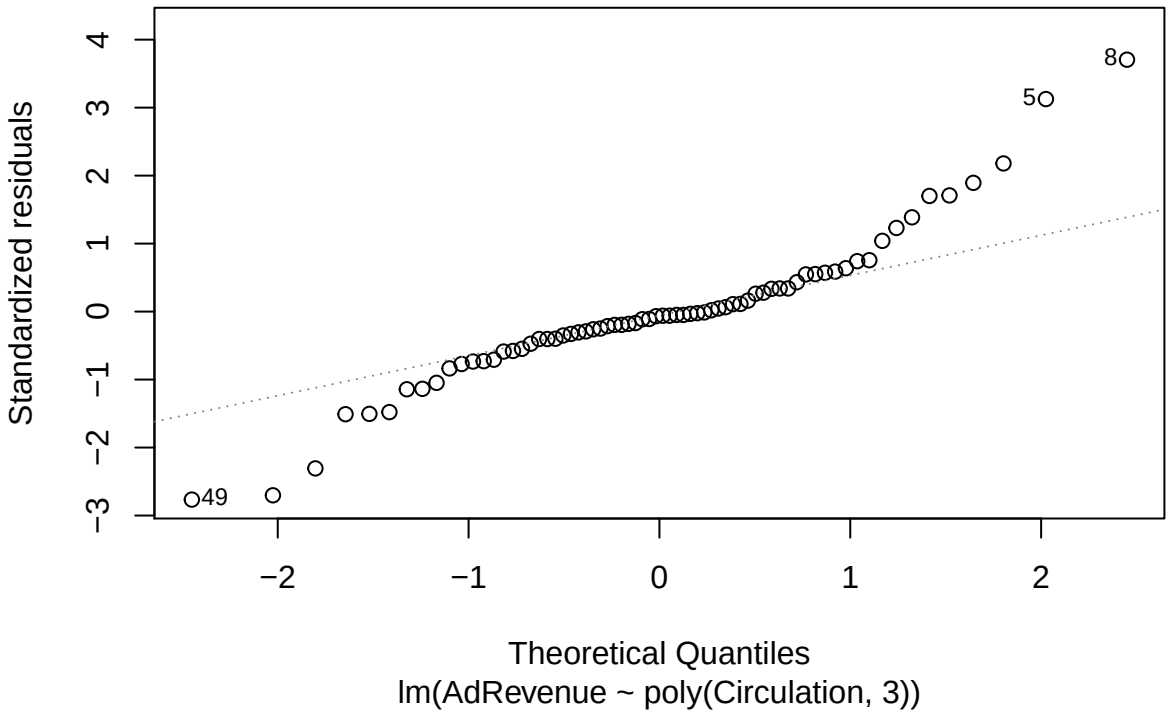

Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ poly(Circulation, 2, raw = T))
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ poly(Circulation, 2, raw = T))

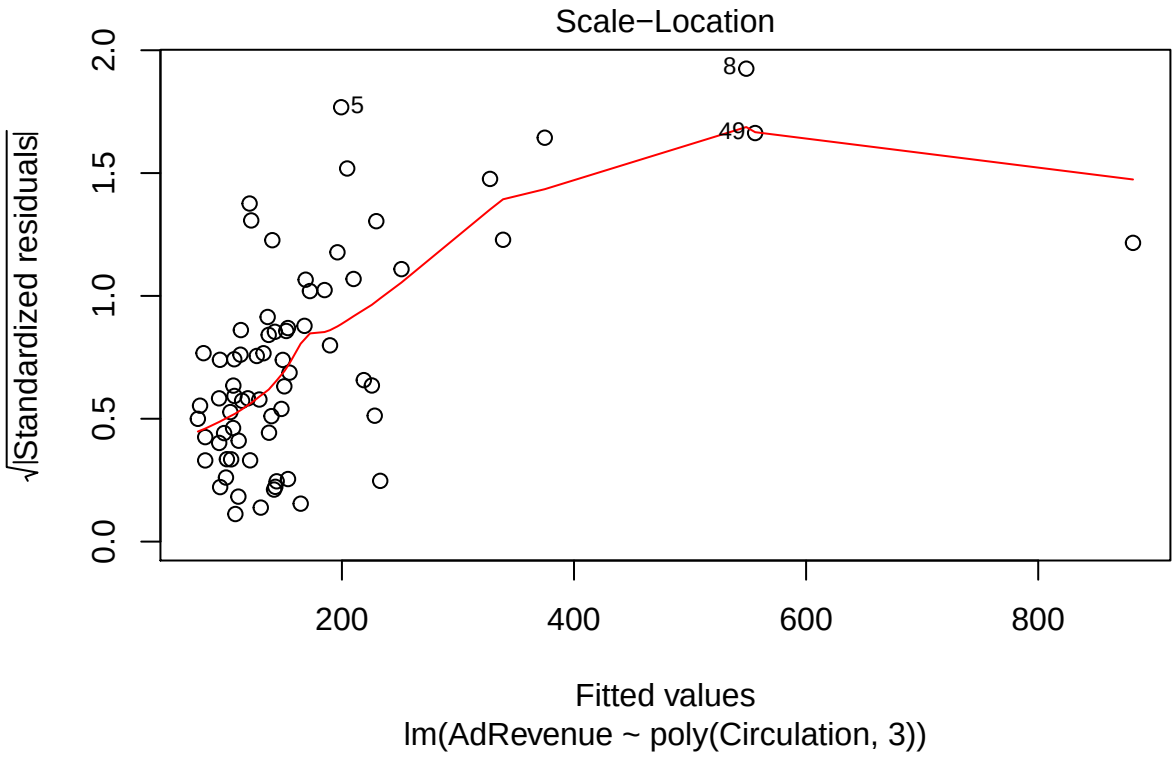

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
200
300
400
500
600
700
800
Fitted values
lm(AdRevenue ~ poly(Circulation, 2, raw = T))
5
49
40


Residuals vs Leverage
Standardized residuals
−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
Leverage
lm(AdRevenue ~ poly(Circulation, 2, raw = T))
Cook's distance
0.5
1
2
49
40

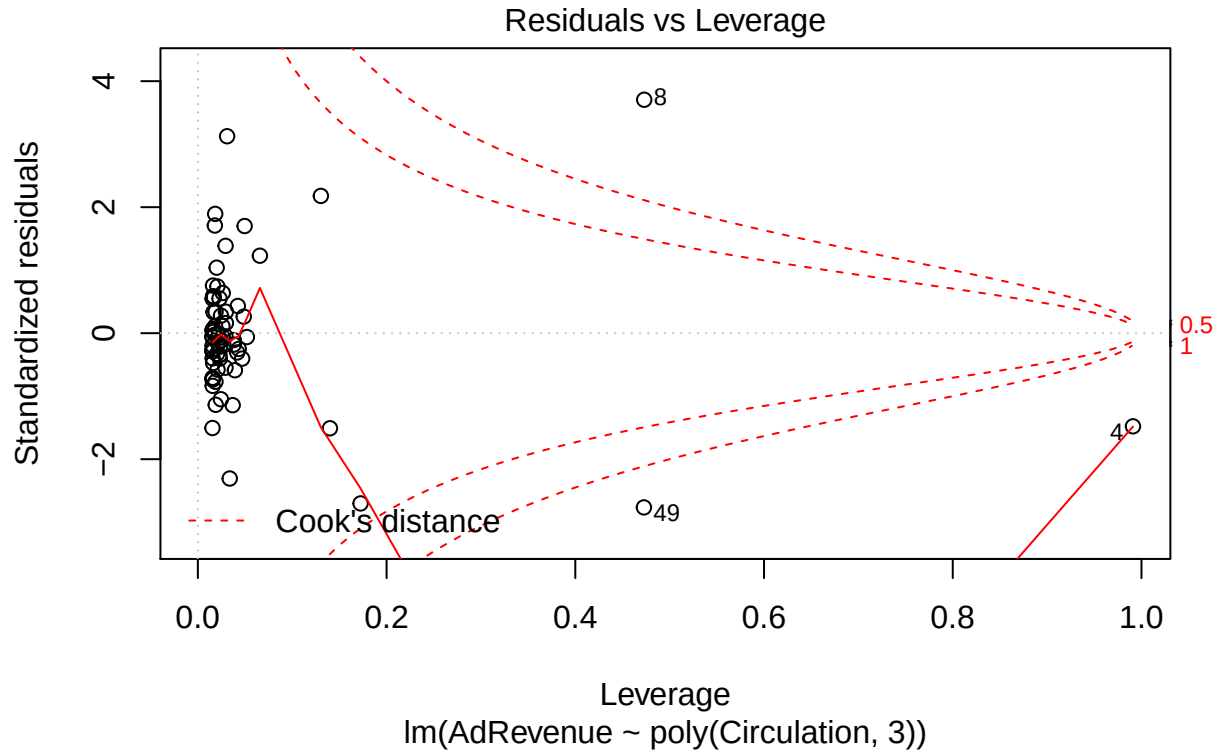
```
plot(model3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(AdRevenue ~ poly(Circulation, 3))
5
8
20


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AdRevenue ~ poly(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
```



(b) Find a 95% prediction interval for the advertising page cost for magazines with the following circulations: (i) 0.5 million (ii) 20 million

```
predict(model2, data.frame("Circulation" = 0.5), interval = "prediction")
```

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

```
predict(model2, data.frame("Circulation" = 20), interval = "prediction")
```

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

(c) Describe any weaknesses in your model.

Looking at the quadratic model first, its residual plot shows no clear trend but there is a cluster with a few outliers. The QQ plot does not follow a straight line (in fact it almost looks like it follows a cubic trend), showing potential non-normality of our data. The scale-location plot shows no upward trend, allowing the constant variation condition to still hold. The leverage plot shows two points with high leverage (4 and 49) with both having bad leverage since neither follow the linear trend of the data. Looking at the diagnostic plots for the cubic model next, the resdual plot shows a fanshape indicating nonconstant variance. Again, the QQ plot does not follow a straight line indicating non-normality and the scale-location plot shows a definite upward trend confirming the failure of the constant variance condition. There are three high leverage points for this model, with all being bad leverage points because none follow the linear trend of the data.

Part C: (a) Compare the model in Part A with that in Part B. Decide which provides a better model. Give reasons to justify your choice.

The logarithmic model in Part A is better than both the quadratic and cubic models in part B. The only weakness of the model in part A was that the QQ plot was not completely straight, however, it was straighter than both models provided in part B. Additionally, the quadratic and cubic models had issues with nonconstant variance, errors in normality, and bad, high leverage points. Intuitively it makes sense that the log model would serve as a better representation of our data given that the data itself ranges over two orders of magnitude indicating that a log tranformation would allow a model to better fit the data.

(b) Compare the prediction intervals in Part A with those in Part B. In each case, decide which interval you would recommend. Give reasons to justify each choice.

I would recommend choosing the log prediction interval, since the log model is the better model for our data. If a model is invalid, then its resulting prediction intervals would also be invalid, thus showing how the prediction intervals in Part B would be less accurate than those in Part A.

Part C

Load the housescrapeWW1.txt data into a dataframe. This includes characteristics of houses/condos in Westwood from three years ago. Create a new data frame that includes only listings for which sqft>0.

```
ww1 <- read.table("housescrapeWW1.txt", header = T, sep = "\t", fill = FALSE)
ww1_clean <- subset(ww1, sqft > 0)
ww1_clean
```

```
##        city      type bed bath garage sqft pool spa   price
## 1  Westwood       SFR   5 3.50     NA 3656       NA 1995000
## 2  Westwood       SFR   3 3.00      2 1870       NA 1350000
## 3  Westwood       SFR   3 2.50     NA 2372    Y  NA 1250000
## 4  Westwood       SFR   3 3.00     NA 1488       NA 1198000
## 5  Westwood Condo/Twh   2 3.00     NA 2310       NA 1145000
## 6  Westwood Condo/Twh   3 2.50     NA 2800       NA 1075000
## 7  Westwood       SFR   3 2.00      2 1579       NA 1034000
## 8  Westwood Condo/Twh   2 2.50     NA 1900       NA  995000
## 9  Westwood Condo/Twh   3 3.50     NA 2650       NA  995000
```

```
## 10 Westwood       SFR   3 1.25     NA 1594       NA  949000
## 11 Westwood Condo/Twh   3 2.50     NA 2662       NA  925000
## 12 Westwood Condo/Twh   3 2.50     NA 1534       NA  925000
## 13 Westwood Condo/Twh   3 2.50     NA 1847       NA  899000
## 14 Westwood       SFR   2 1.00     NA 1240       NA  895000
## 15 Westwood Condo/Twh   3 2.50      2 1900       NA  879900
## 16 Westwood Condo/Twh   3 2.50      2 2118       NA  849000
## 17 Westwood Condo/Twh   2 2.50     NA 1900       NA  849000
## 18 Westwood Condo/Twh   3 3.00     NA 2200       NA  834500
## 19 Westwood Condo/Twh   3 3.00     NA 1870       NA  799000
## 20 Westwood Condo/Twh   3 2.50     NA 1812       NA  799000
## 21 Westwood Condo/Twh   3 2.50     NA 1832       NA  799000
## 22 Westwood Condo/Twh   2 2.50     NA 1900       NA  795000
## 23 Westwood       SFR   2 1.00     NA  968       NA  789000
## 24 Westwood Condo/Twh   2 2.50     NA 1549       NA  780000
## 25 Westwood Condo/Twh   3 2.50      2 1733       NA  779000
## 26 Westwood Condo/Twh   2 2.50     NA 1700       NA  777000
## 27 Westwood Condo/Twh   3 2.50     NA 1832       NA  759000
## 28 Westwood Condo/Twh   4 2.50      2 1830       NA  750000
## 29 Westwood Condo/Twh   3 2.50     NA 1703       NA  749000
## 30 Westwood Condo/Twh   2 3.00      2 1643       NA  749000
## 31 Westwood Condo/Twh   3 2.00     NA 1800       NA  744000
## 32 Westwood Condo/Twh   3 2.50     NA 1887       NA  739000
## 33 Westwood Condo/Twh   2 2.50      2 1828       NA  739000
## 34 Westwood Condo/Twh   3 2.00     NA 1697       NA  735000
## 36 Westwood Condo/Twh   2 2.00     NA 1331       NA  699900
## 37 Westwood Condo/Twh   2 2.50     NA 2031       NA  699000
## 38 Westwood Condo/Twh   2 2.50      2 1526       NA  699000
## 39 Westwood Condo/Twh   3 2.50     NA 1620       NA  699000
## 40 Westwood Condo/Twh   3 3.00      2 1487       NA  699000
## 41 Westwood Condo/Twh   2 2.50     NA 1774       NA  699000
## 42 Westwood Condo/Twh   3 2.50     NA 1536       NA  699000
## 44 Westwood Condo/Twh   3 2.00      2 1913       NA  695000
## 45 Westwood Condo/Twh   2 2.50     NA 1516       NA  695000
## 46 Westwood Condo/Twh   2 2.50     NA 1380       NA  689000
## 47 Westwood Condo/Twh   2 2.50     NA 1475       NA  680000
## 48 Westwood Condo/Twh   2 2.00     NA 1517       NA  679000
## 49 Westwood Condo/Twh   2 2.00      2 1511       NA  679000
## 50 Westwood Condo/Twh   3 2.00     NA 1549       NA  675000
## 51 Westwood       SFR   2 1.00     NA  810       NA  675000
## 52 Westwood Condo/Twh   1 1.50     NA 1167       NA  675000
## 53 Westwood Condo/Twh   2 2.50     NA 1774       NA  675000
## 54 Westwood Condo/Twh   2 2.50     NA 1334       NA  659000
## 55 Westwood Condo/Twh   3 2.00     NA 1662       NA  650000
## 56 Westwood Condo/Twh   2 3.00     NA 1620       NA  649000
## 57 Westwood Condo/Twh   2 3.00     NA 1800       NA  649000
## 58 Westwood Condo/Twh   2 2.00      2 1640       NA  649000
## 59 Westwood Condo/Twh   2 2.50     NA 1403       NA  649000
## 60 Westwood Condo/Twh   3 2.00      2 1603       NA  640000
## 61 Westwood Condo/Twh   2 2.00     NA 1532       NA  626400
## 62 Westwood Condo/Twh   2 3.00      2 1540       NA  625000
## 63 Westwood Condo/Twh   3 3.00     NA 1693       NA  599000
## 64 Westwood Condo/Twh   2 2.00     NA 1508       NA  589000
## 65 Westwood Condo/Twh   2 2.00      2 1694       NA  579000
```

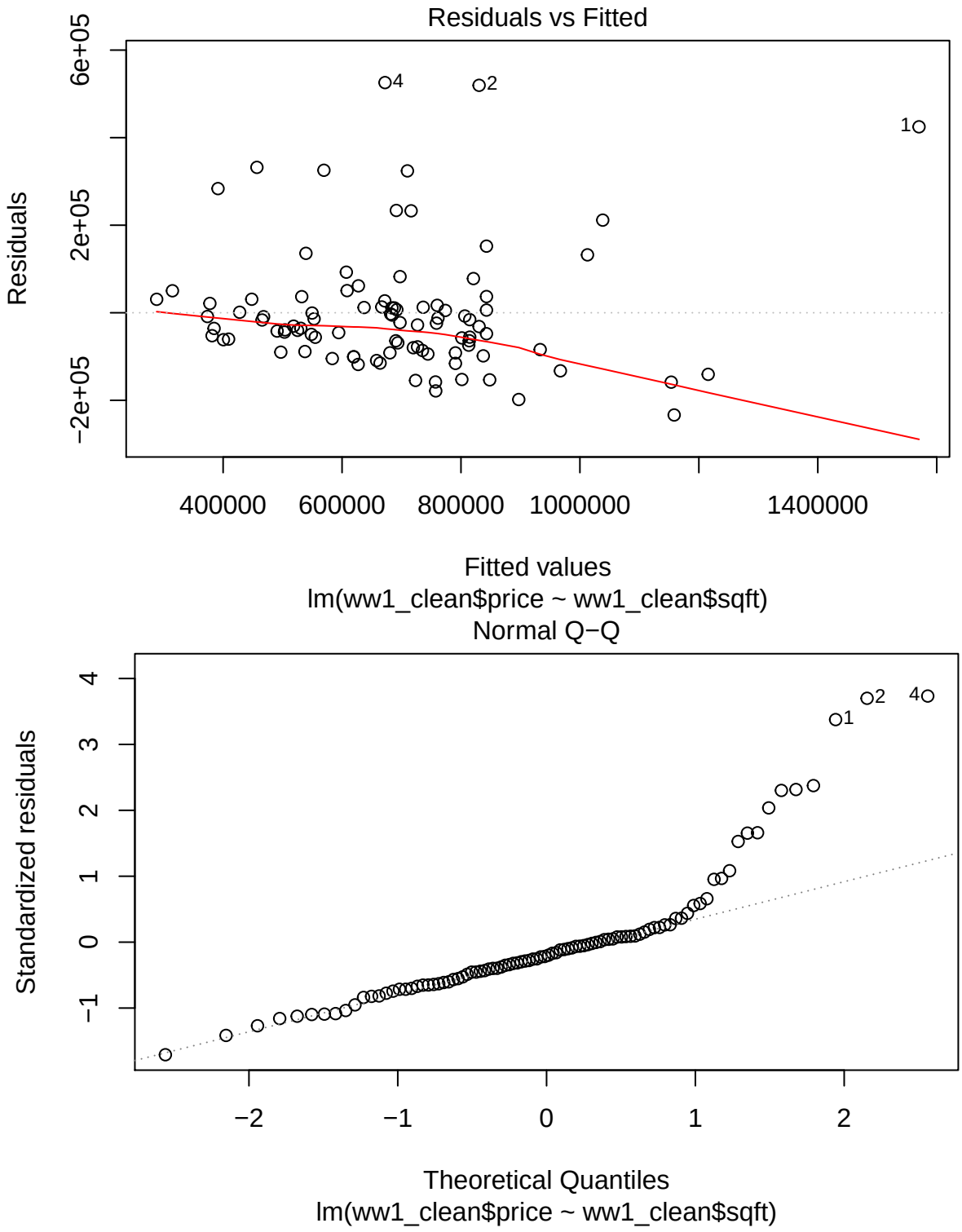
```
## 66 Westwood Condo/Twh   2 2.50     NA 1612       NA  569000
## 67 Westwood Condo/Twh   2 2.00     NA 1150       NA  569000
## 68 Westwood Condo/Twh   2 2.00     NA 1454       NA  549000
## 69 Westwood Condo/Twh   2 2.50      2 1192       NA  549000
## 70 Westwood Condo/Twh   3 3.00     NA 1468       NA  549000
## 71 Westwood Condo/Twh   3 2.00     NA 1300       NA  549000
## 72 Westwood Condo/Twh   2 2.50     NA 1200       NA  539000
## 73 Westwood Condo/Twh   3 3.50     NA 1361       NA  519000
## 74 Westwood Condo/Twh   3 3.50     NA 1361       NA  519000
## 75 Westwood Condo/Twh   2 2.00     NA 1379       NA  509000
## 76 Westwood Condo/Twh   2 2.00     NA 1189       NA  499000
## 77 Westwood Condo/Twh   2 2.00      2 1205       NA  499000
## 78 Westwood Condo/Twh   2 2.00     NA 1146       NA  495000
## 79 Westwood Condo/Twh   2 2.00     NA 1117       NA  488000
## 80 Westwood Condo/Twh   1 1.50      2 1133       NA  485000
## 81 Westwood Condo/Twh   2 2.50     NA 1274       NA  479000
## 82 Westwood Condo/Twh   1 1.50     NA  947       NA  479000
## 83 Westwood Condo/Twh   2 2.00     NA 1082       NA  465000
## 84 Westwood Condo/Twh   2 2.00     NA 1080       NA  459000
## 85 Westwood Condo/Twh   2 1.75     NA  995       NA  459000
## 86 Westwood Condo/Twh   2 2.00     NA 1050       NA  449000
## 87 Westwood Condo/Twh   2 2.00     NA 1163       NA  449000
## 88 Westwood Condo/Twh   2 1.75     NA  989       NA  449000
## 89 Westwood Condo/Twh   2 2.00      2  898       NA  429000
## 90 Westwood Condo/Twh   1 2.00     NA 1065       NA  407000
## 91 Westwood Condo/Twh   2 2.00      2  777       NA  399000
## 92 Westwood Condo/Twh   1 1.00     NA  767       NA  365000
## 93 Westwood Condo/Twh   1 1.00     NA  625       NA  365000
## 94 Westwood Condo/Twh   2 2.00      2  853       NA  349000
## 95 Westwood Condo/Twh   2 1.00     NA  794       NA  349000
## 96 Westwood Condo/Twh   1 1.00      2  832       NA  339000
## 97 Westwood Condo/Twh   1 1.00     NA  786       NA  329000
## 98 Westwood Condo/Twh   0 1.00     NA  561       NA  319000
```

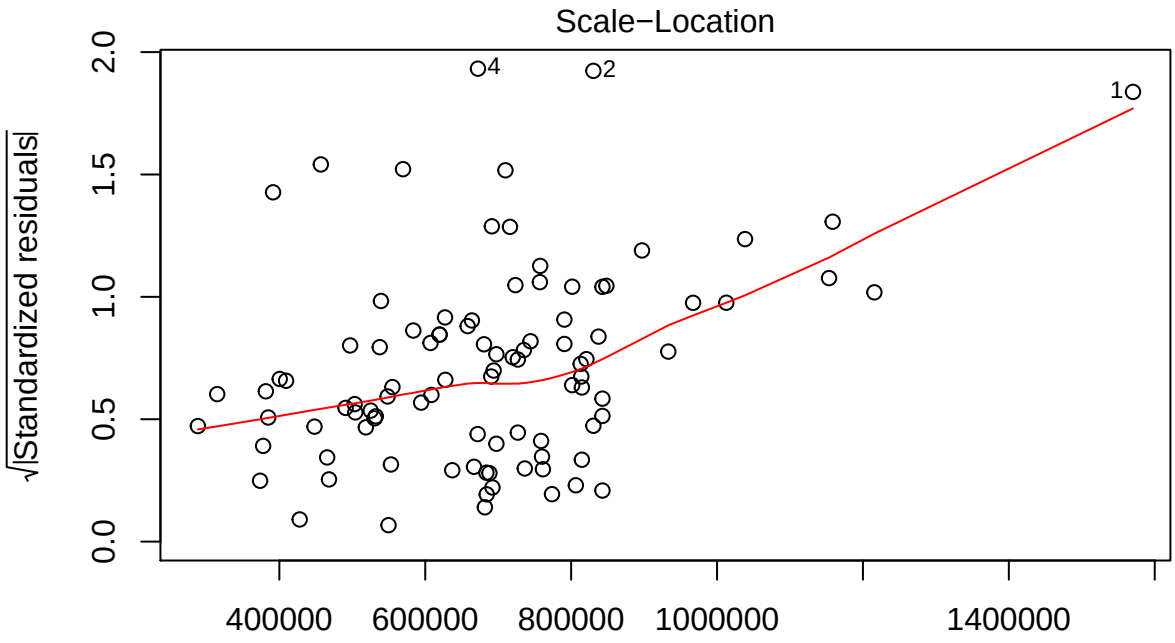
a) Fit the model price= b_0 + b_1 size(size is measured in square-feet and is the variable sqft). Report the model and iterpret the intercept and slope.

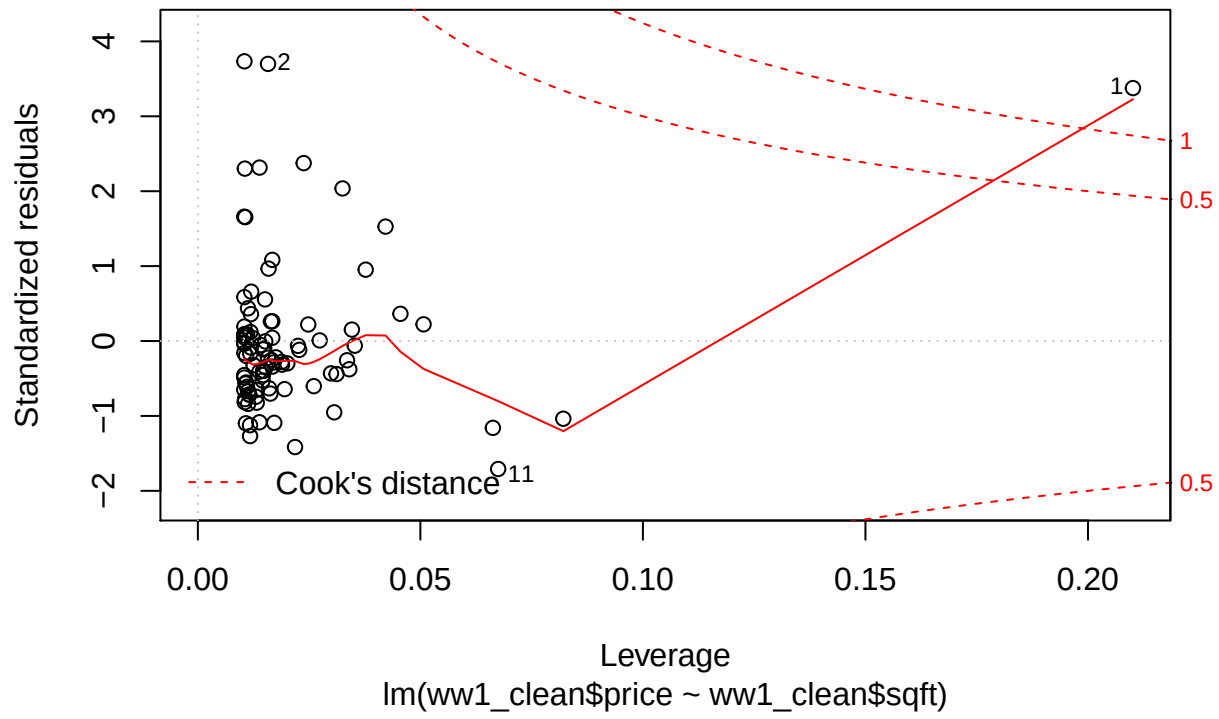
```
ww1_clean_lm <- lm(ww1_clean$price~ww1_clean$sqft)
ww1_clean_lm
```

```
##
## Call:
## lm(formula = ww1_clean$price ~ ww1_clean$sqft)
##
## Coefficients:
##    (Intercept)  ww1_clean$sqft
##        55902.6           414.2
```

```
plot(ww1_clean_lm)
```


Residuals vs Fitted
Residuals
6e+05
2e+05
−2e+05
4
2
1
400000
600000
800000
1000000
1400000
Fitted values
lm(ww1_clean$price ~ ww1_clean$sqft)
Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
2
4
1
−2
−1
0
1
2
Theoretical Quantiles
lm(ww1_clean$price ~ ww1_clean$sqft)






```
# this gives us the following model: price = 55902.6 + 414.2size with the starting price for a house be
```

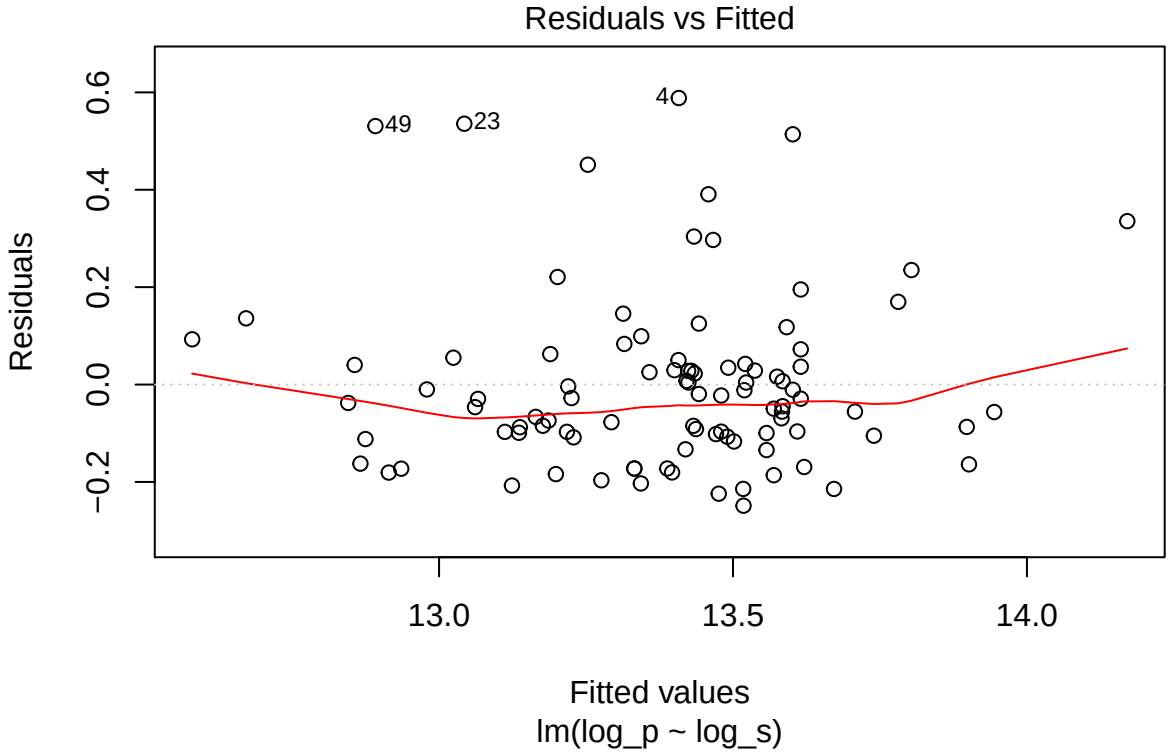
b) Fit the model log(price)=b_0+b_1 log(size). Report the model and Interpret the slope.

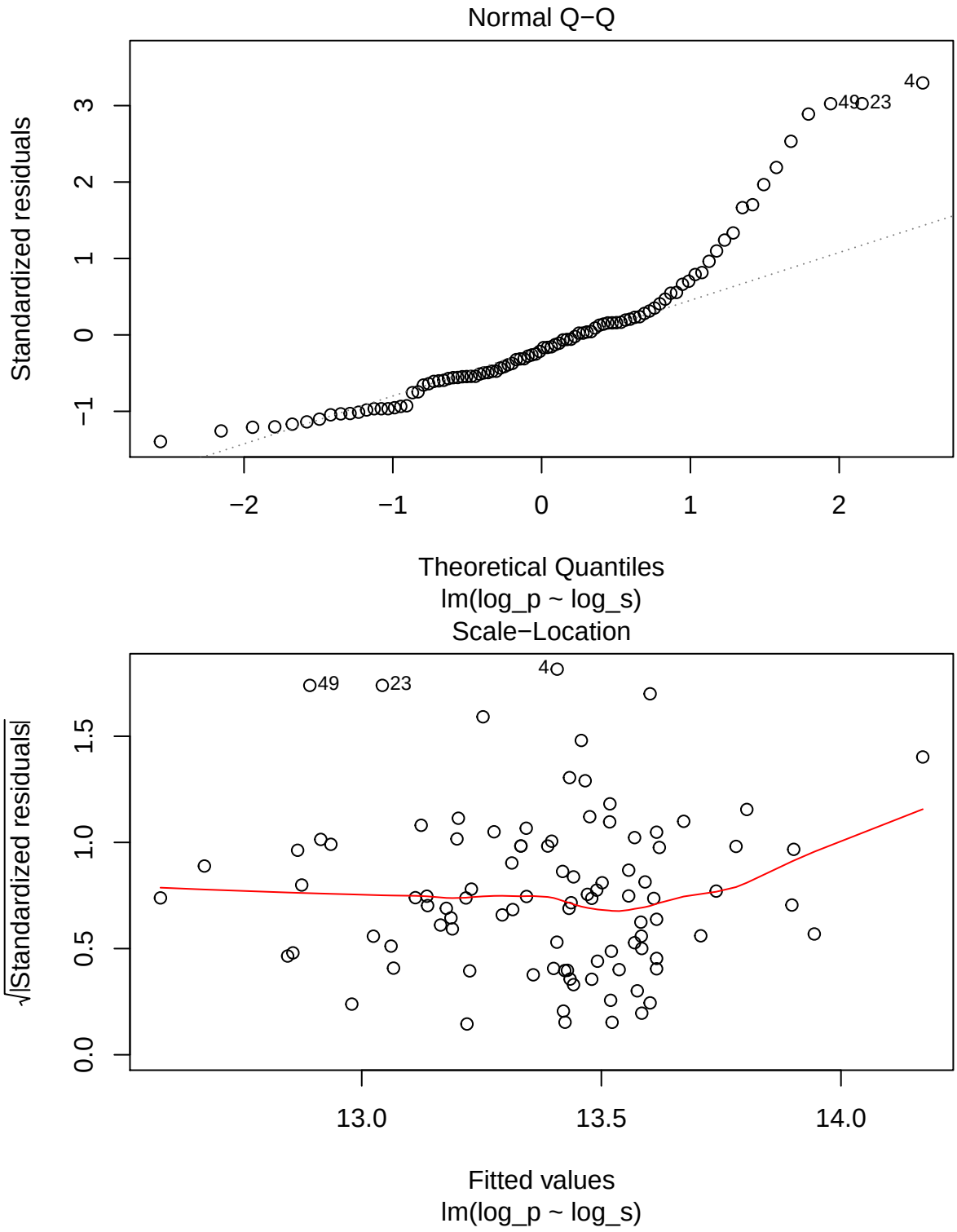
```
log_p <- log(ww1_clean$price)
log_s <- log(ww1_clean$sqft)
```

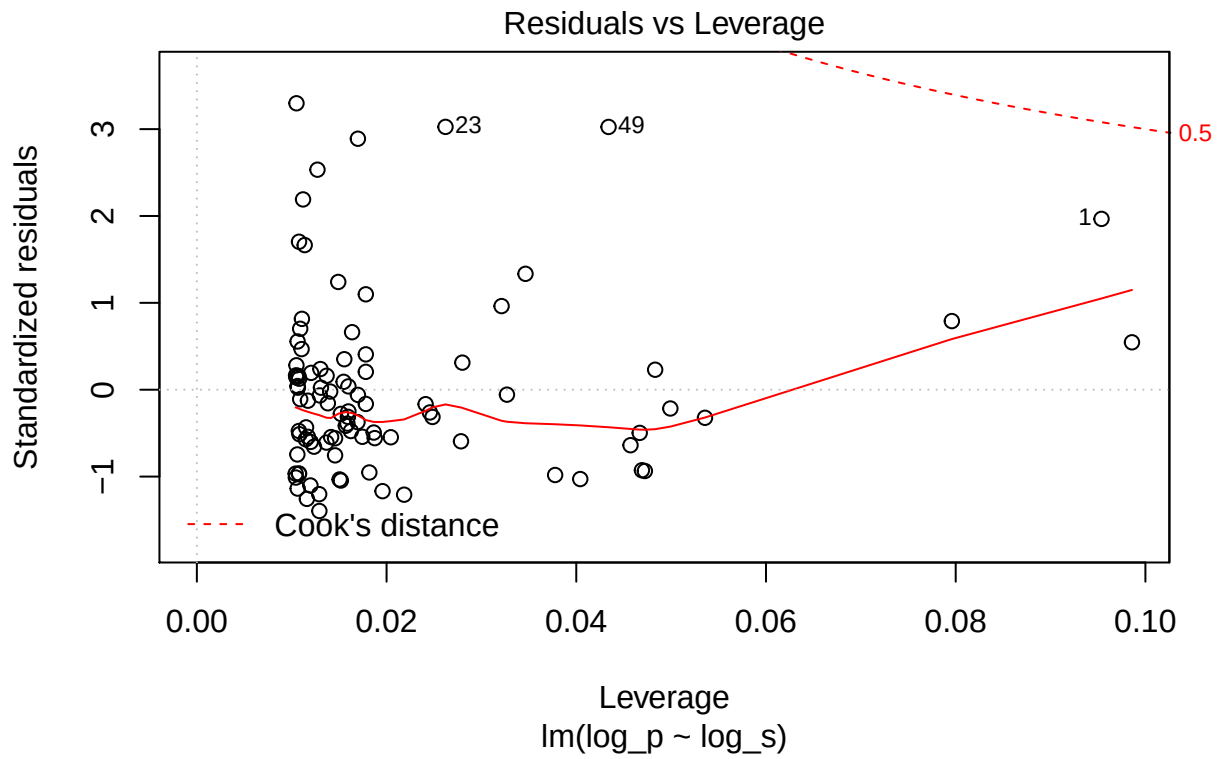
```r
ww1_clean_log <- lm(log_p~log_s)
ww1_clean_log
```

```
##
## Call:
## lm(formula = log_p ~ log_s)
##
## Coefficients:
## (Intercept)        log_s
##      7.2086       0.8486
```

```r
plot(ww1_clean_log)
```



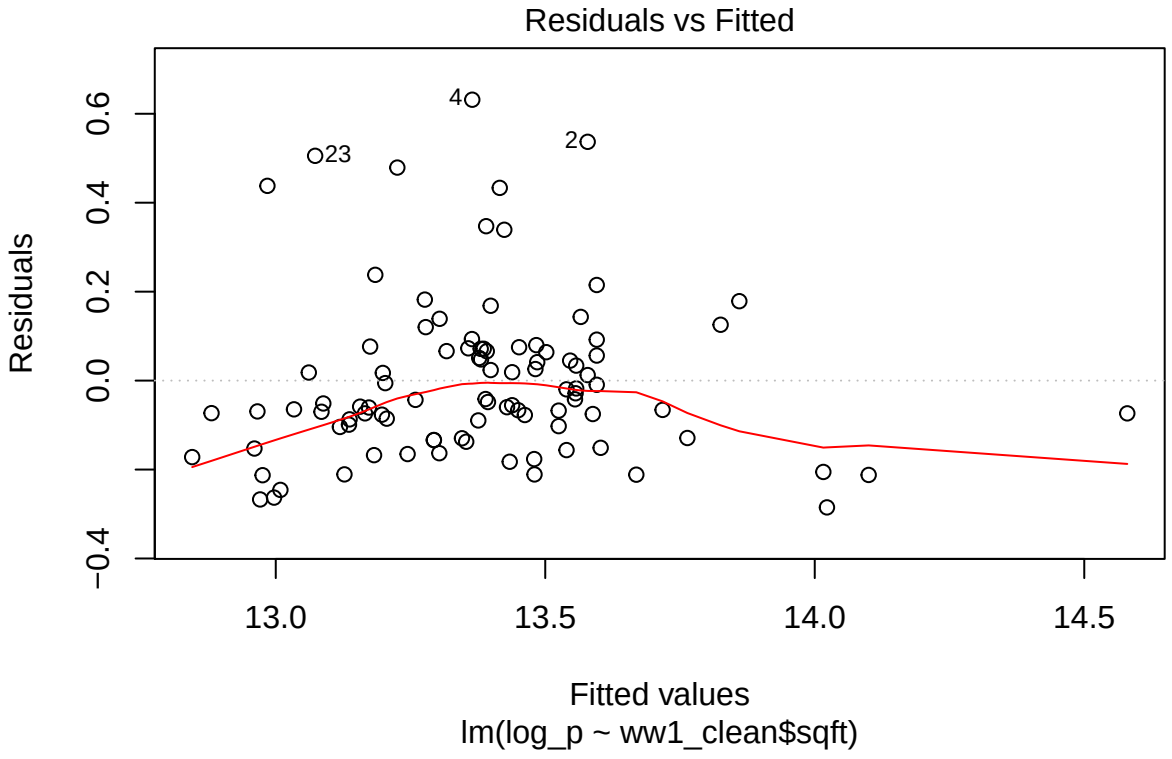

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_p ~ log_s)
Scale−Location
√|Standardized residuals|
Fitted values
lm(log_p ~ log_s)
4
49
23

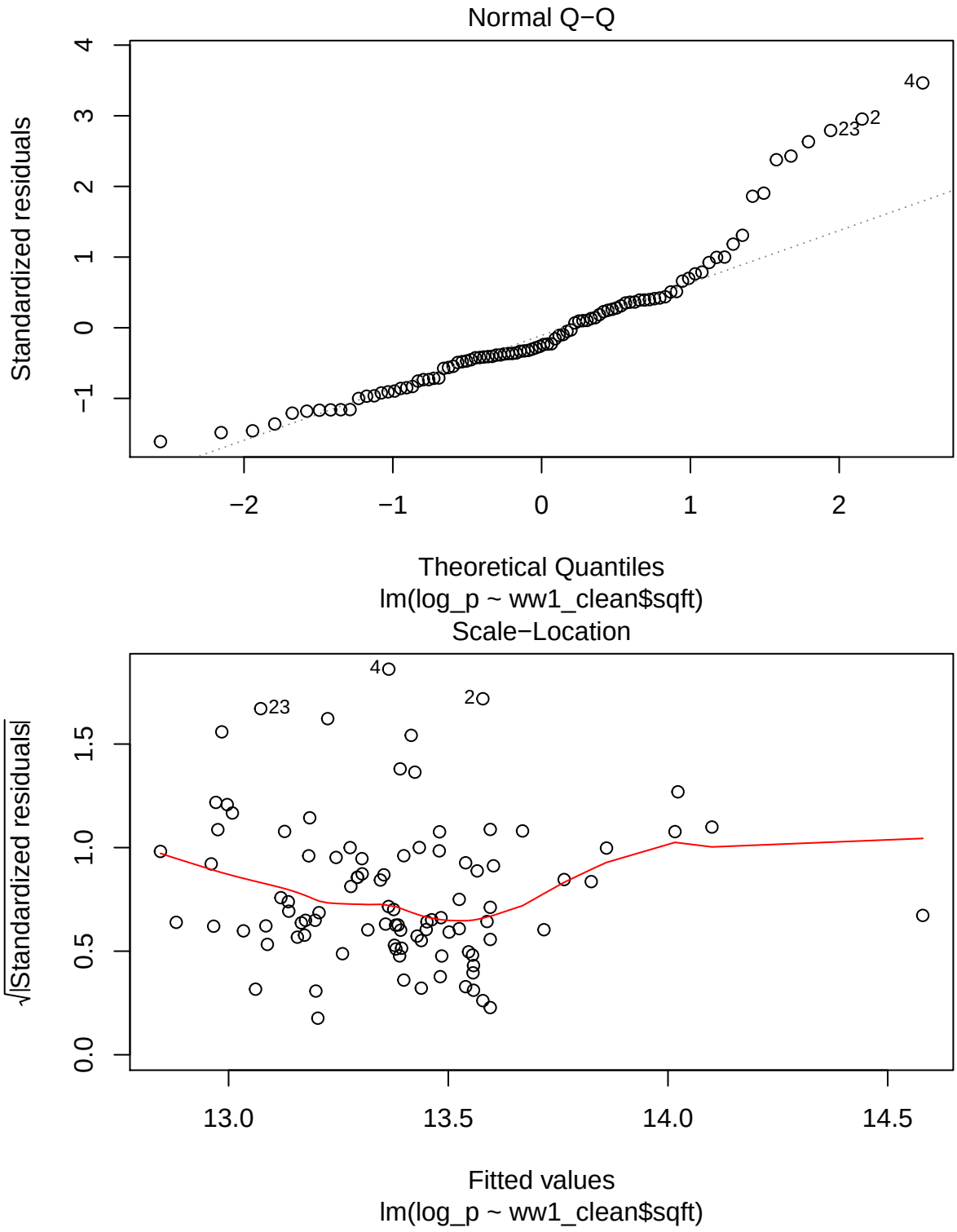


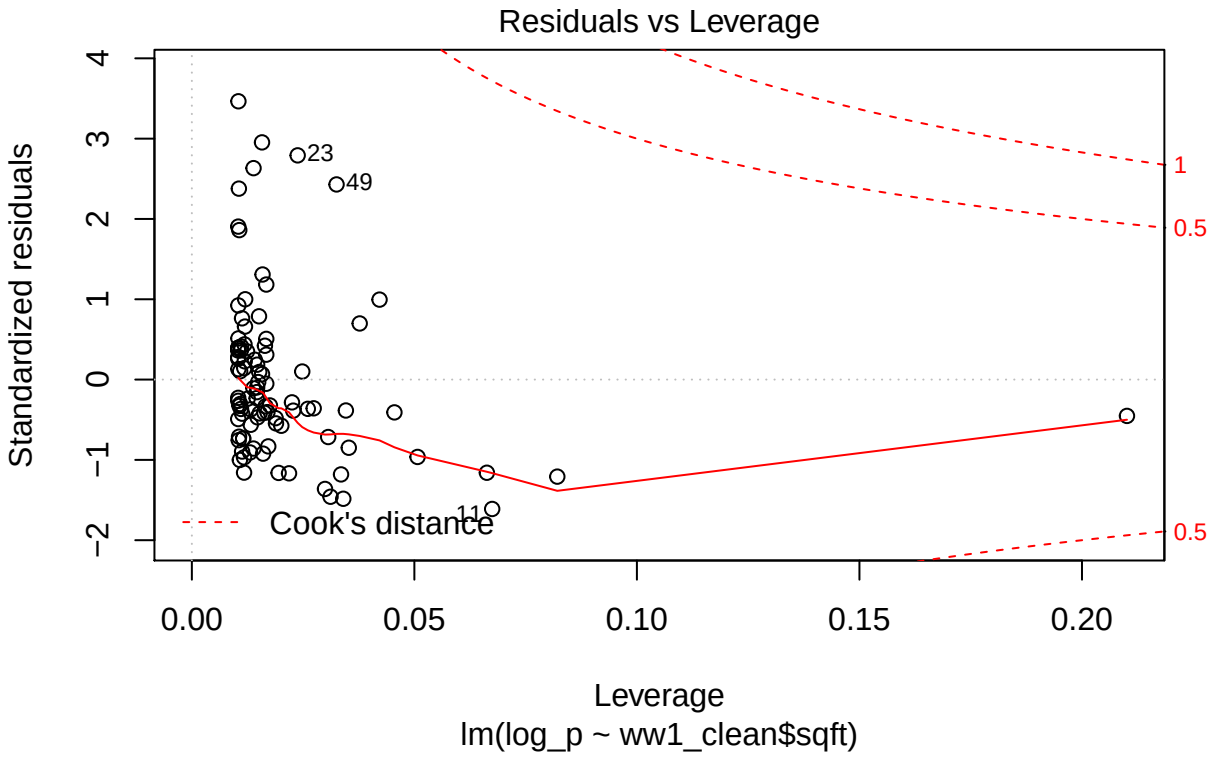

```
# this gives us the following model: price = 7.2086 + 0.8486size where the slope tells us that a one pe
```

c) Fit the model log(price)=b_0+b_1 *size. Report the model

```
ww1_clean_both <- lm(log_p~ww1_clean$sqft)
plot(ww1_clean_both)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_p ~ ww1_clean$sqft)
Scale−Location
√|Standardized residuals|
Fitted values
lm(log_p ~ ww1_clean$sqft)




```
ww1_clean_both
```

```
##
## Call:
## lm(formula = log_p ~ ww1_clean$sqft)
##
## Coefficients:
##    (Intercept)  ww1_clean$sqft
##      1.253e+01       5.605e-04
```

```
# gives the following model : price = 12.53 + 0.00056size
```

d) Which model fits better, in terms of model validity? Comment on all ways in which the better model is better, and the ways in which it is not better (and maybe even worse.)

The model in part B (log/log) fits better. Looking at the residual plots for all three models first, the models in part A and C resemble a fanshape indicating nonconstant variance. This is confirmed by their scale location plots which for the model in part A shows an upward trend and a downward trend for the model in part B. The model in part A has a bad, high leverage point while the leverage plot for part B shows the presence of a potential influential point. While all three models dont have a QQ plot that follows a very straight line, the model in B is better due to it satisfying the constant variance condition and having no high leverage points (as well as more varied points in the residuals vs. leverage plot).