Exam 2

Your name

2024-11-11

Review exercises

```
library(tidyverse)
```

Codebook:

The Boston Housing Dataset

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

CRIM - per capita crime rate by town ZN - proportion of residential land zoned for lots over 25,000 sq.ft. INDUS - proportion of non-retail business acres per town. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise) NOX - nitric oxides concentration (parts per 10 million) RM - average number of rooms per dwelling AGE - proportion of owner-occupied units built prior to 1940 DIS - weighted distances to five Boston employment centres RAD - index of accessibility to radial highways TAX - full-value property-tax rate per $10,000 PTRATIO - pupil-teacher ratio by town B - 1000(Bk - 0.63)^2 where Bk is the proportion of Black individuals by town LSTAT - % lower status of the population MEDV - Median value of owner-occupied homes in $1000's

1. Load the data.

```
dat <- read_csv("data/bostonhousing.csv")
dat
```

```
## # A tibble: 506 x 14
##       crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio     b
##      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
##  1 0.00632  18    2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  397.
##  2 0.0273    0    7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  397.
##  3 0.0273    0    7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  393.
##  4 0.0324    0    2.18     0 0.458  7.00  45.8  6.06     3   222    18.7  395.
##  5 0.0690    0    2.18     0 0.458  7.15  54.2  6.06     3   222    18.7  397.
##  6 0.0298    0    2.18     0 0.458  6.43  58.7  6.06     3   222    18.7  394.
##  7 0.0883   12.5  7.87     0 0.524  6.01  66.6  5.56     5   311    15.2  396.
##  8 0.145    12.5  7.87     0 0.524  6.17  96.1  5.95     5   311    15.2  397.
##  9 0.211    12.5  7.87     0 0.524  5.63 100    6.08     5   311    15.2  387.
## 10 0.170    12.5  7.87     0 0.524  6.00  85.9  6.59     5   311    15.2  387.
## # i 496 more rows
## # i 2 more variables: lstat <dbl>, medv <dbl>
```

2. Read the codebook. Your task today is to answer the question, is crime associated with the value of homes? State your null hypothesis and alternative hypothesis.

Answer:

3. Do EDA for crim and medv.

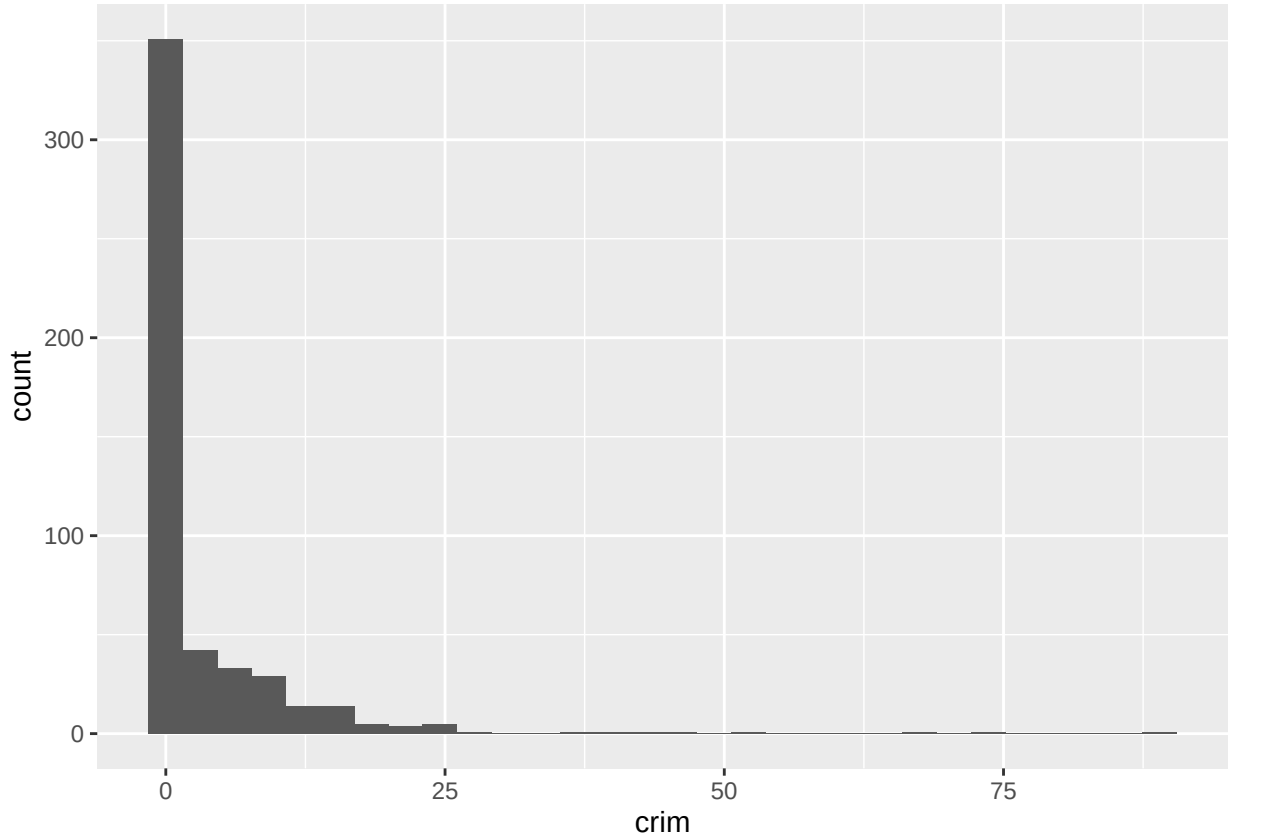
Answer:

4. For crim, what transformation makes the distribution more symmetric and spread out?

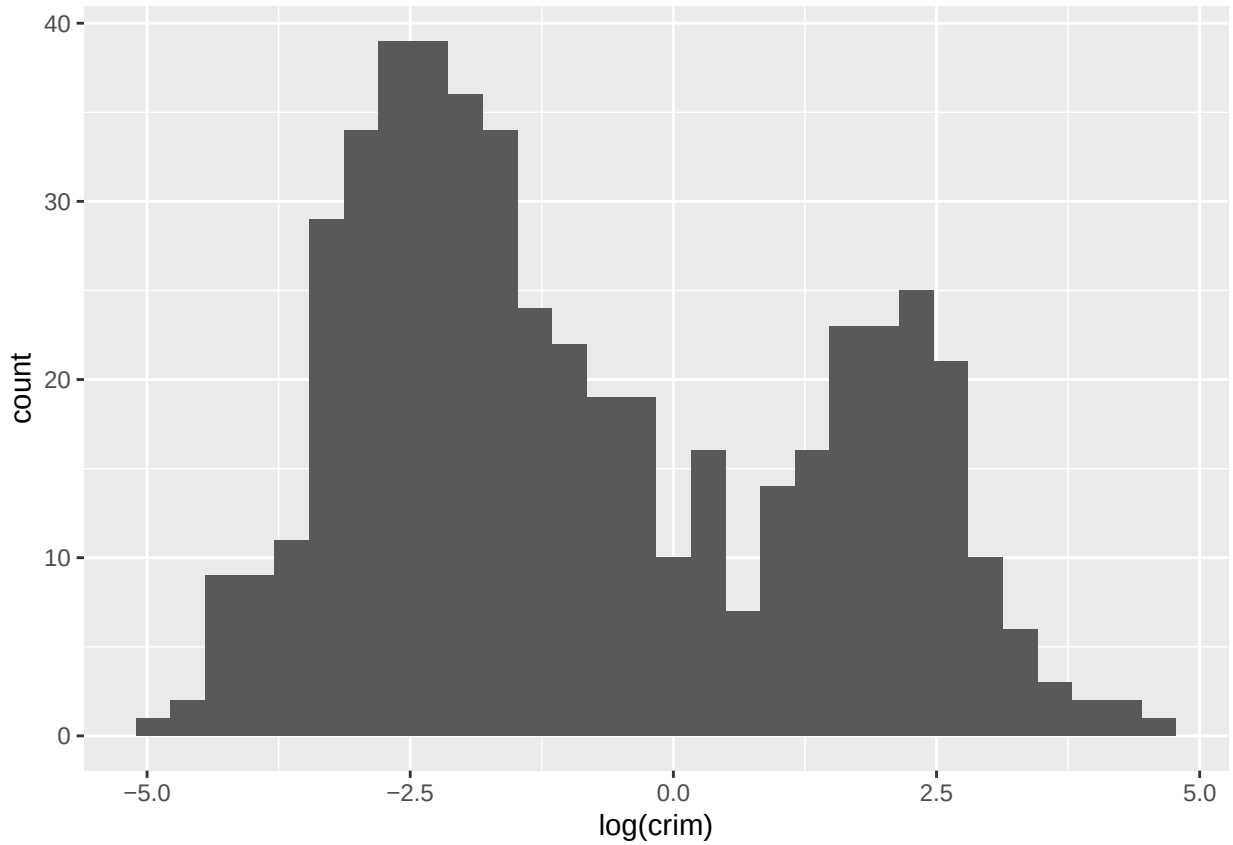
Answer:

5. For medv, filter to keep only the cases for which medv (the Median value of owner-occupied homes) is less than 50. The dataset was truncated at medv>=50, so we'll just keep anything below it since the distribution is artificially truncated. Make a new dataset called data.edited that only keeps the cases for which medv<50.

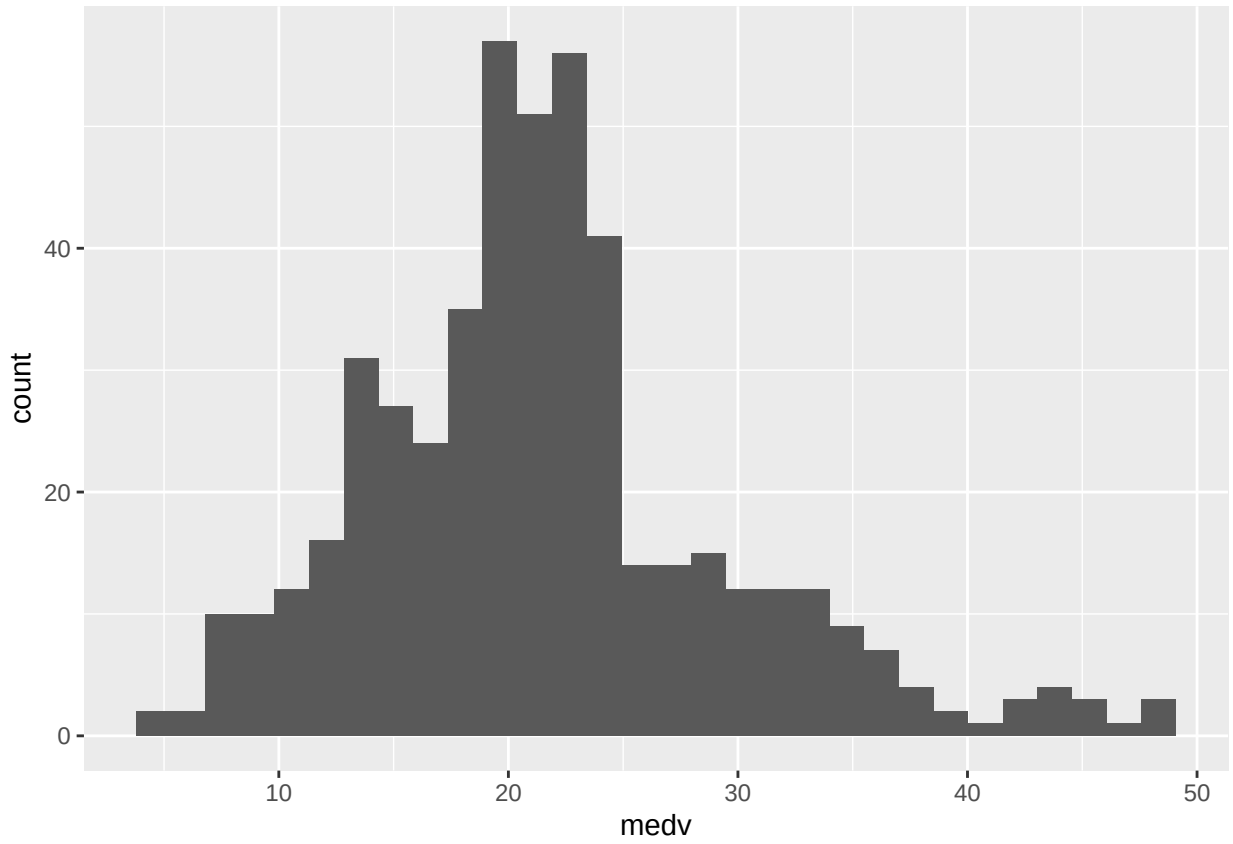
```
dat %>% ggplot(aes(x=crim)) + geom_histogram()
```



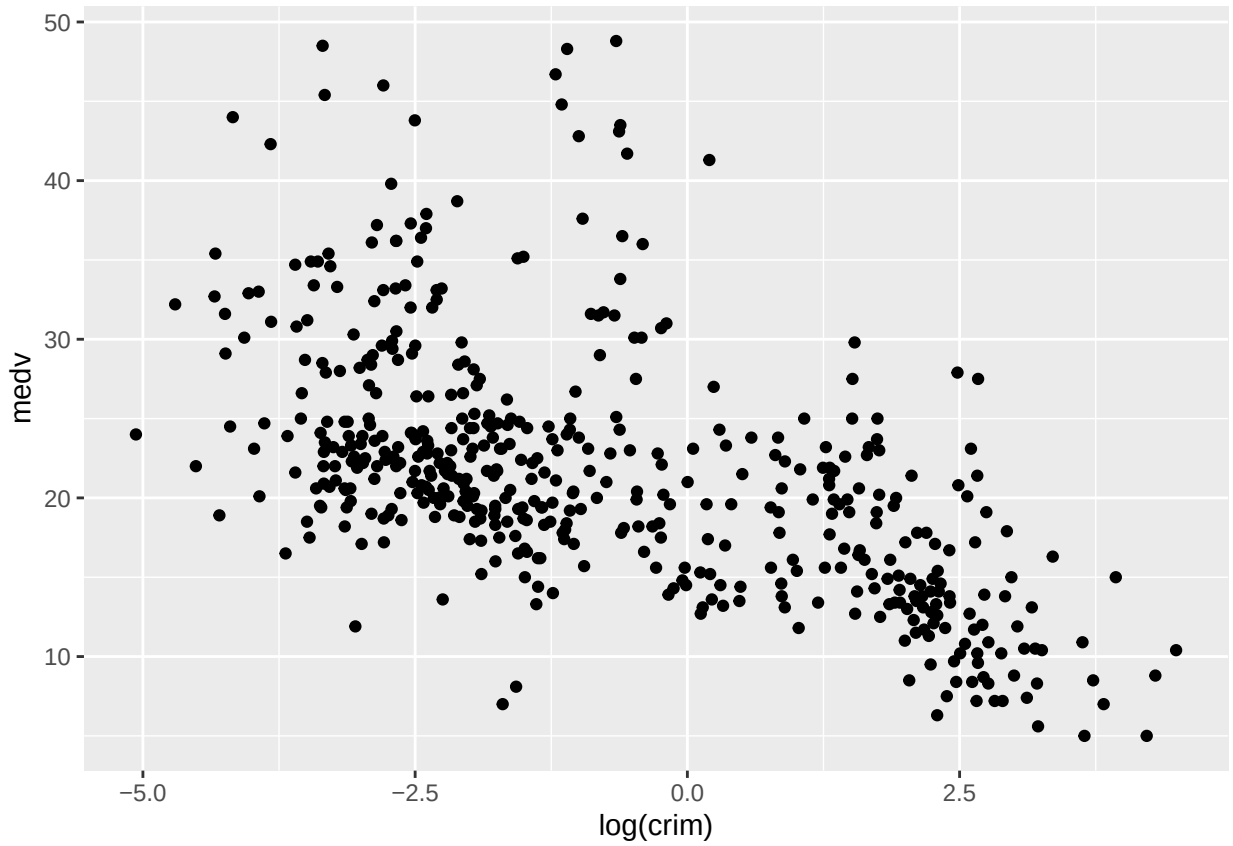

```
dat %>% ggplot(aes(x=log(crim))) + geom_histogram()
```




```
dat %>% filter(medv<50) %>% ggplot(aes(x=medv)) + geom_histogram()
```




```
dat %>% filter(medv<50) %>% ggplot(aes(x=log(crim), y=medv)) + geom_point()
```

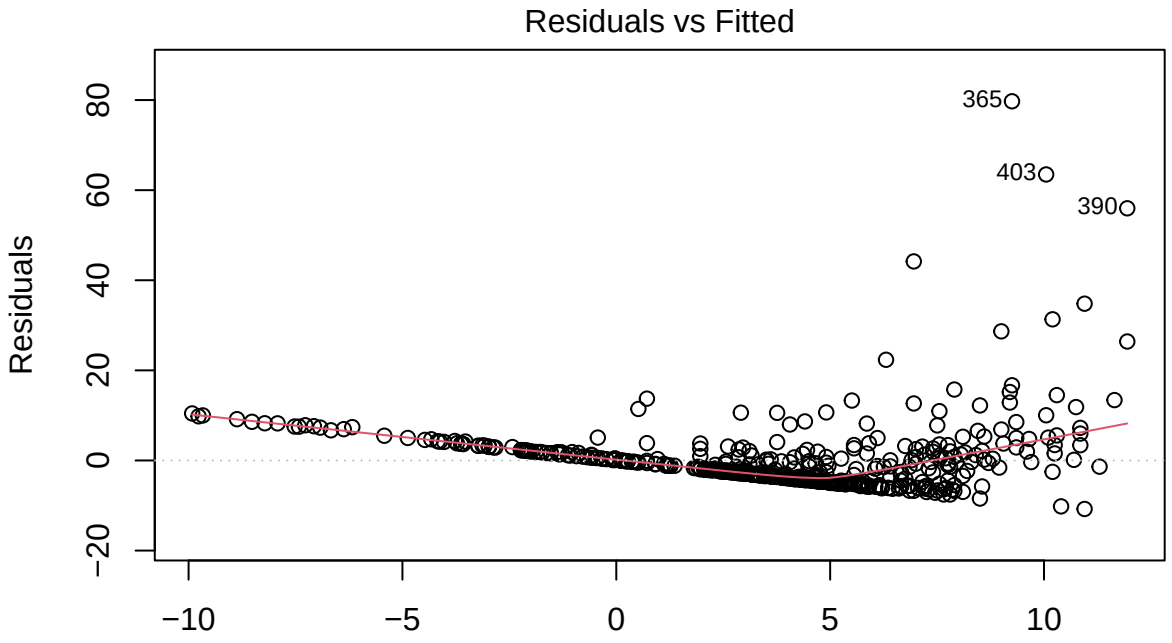


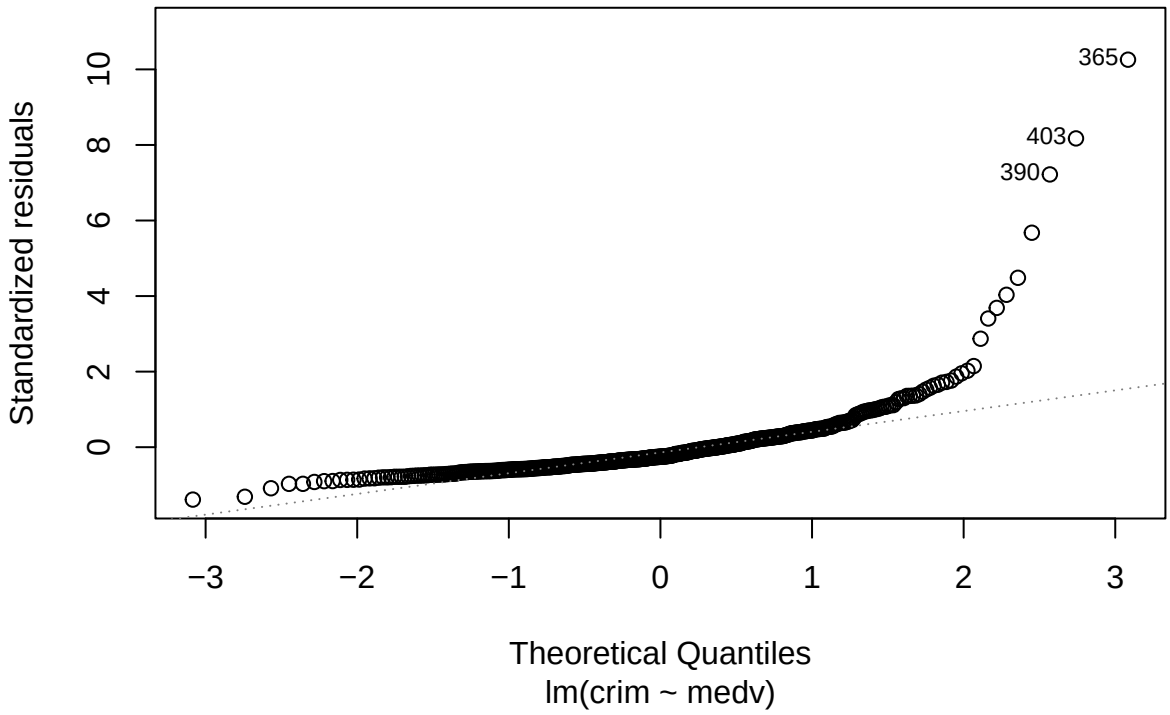

```
dat.edited <- dat %>% filter(medv<50)
```

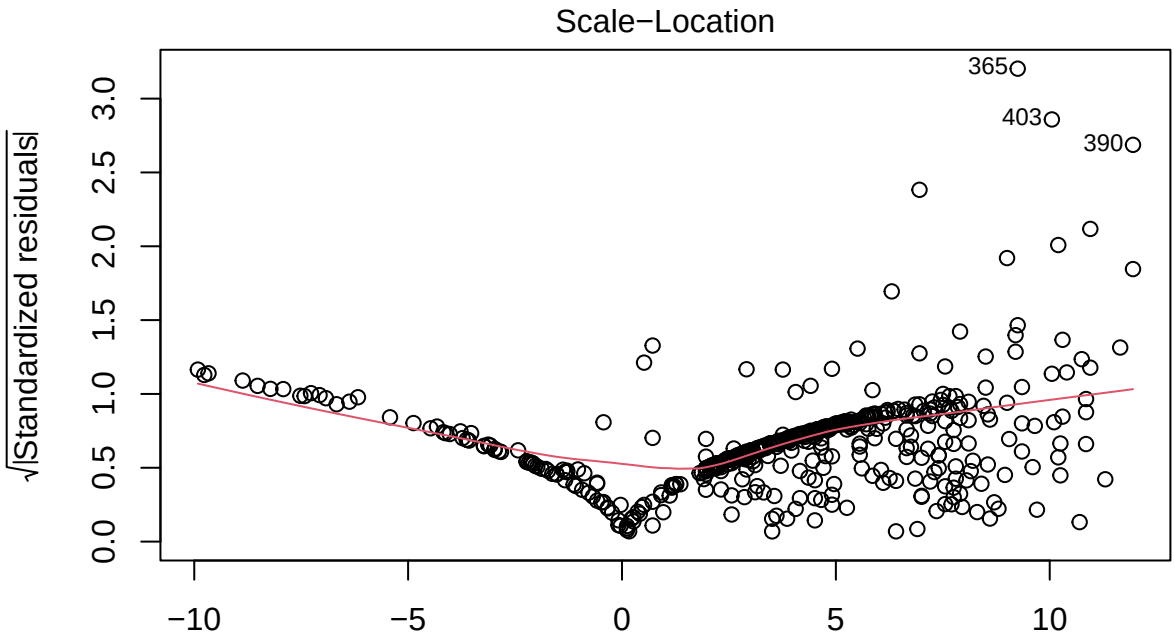
Answer:

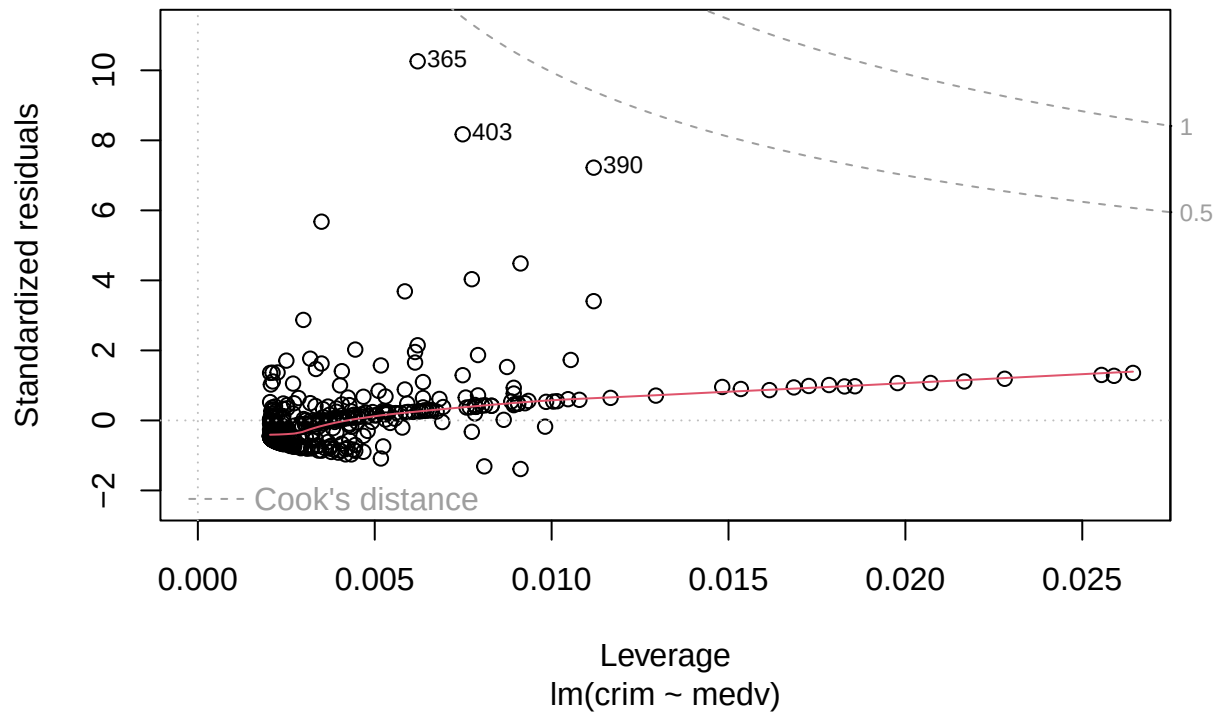
6. Fit a linear regression model on crim and medv called out, using the dat.edited dataset. Using a Box-Cox transformation, find out which transformation would help make the linear model fit better.

```
out <- lm(crim ~ medv, data=dat.edited)
plot(out)
```

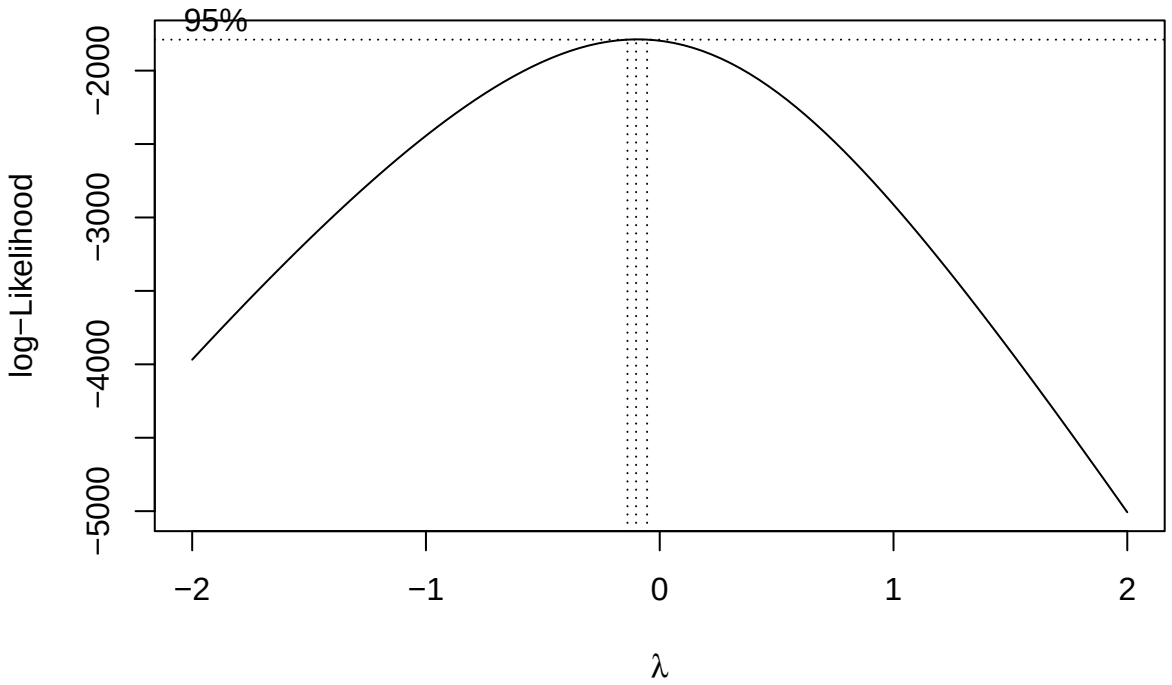

Residuals vs Fitted
Residuals
Fitted values
lm(crim ~ medv)
365
403
390
−20
0
20
40
60
80
−10
−5
0
5
10


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(crim ~ medv)
365
403
390
0
2
4
6
8
10
−3
−2
−1
0
1
2
3





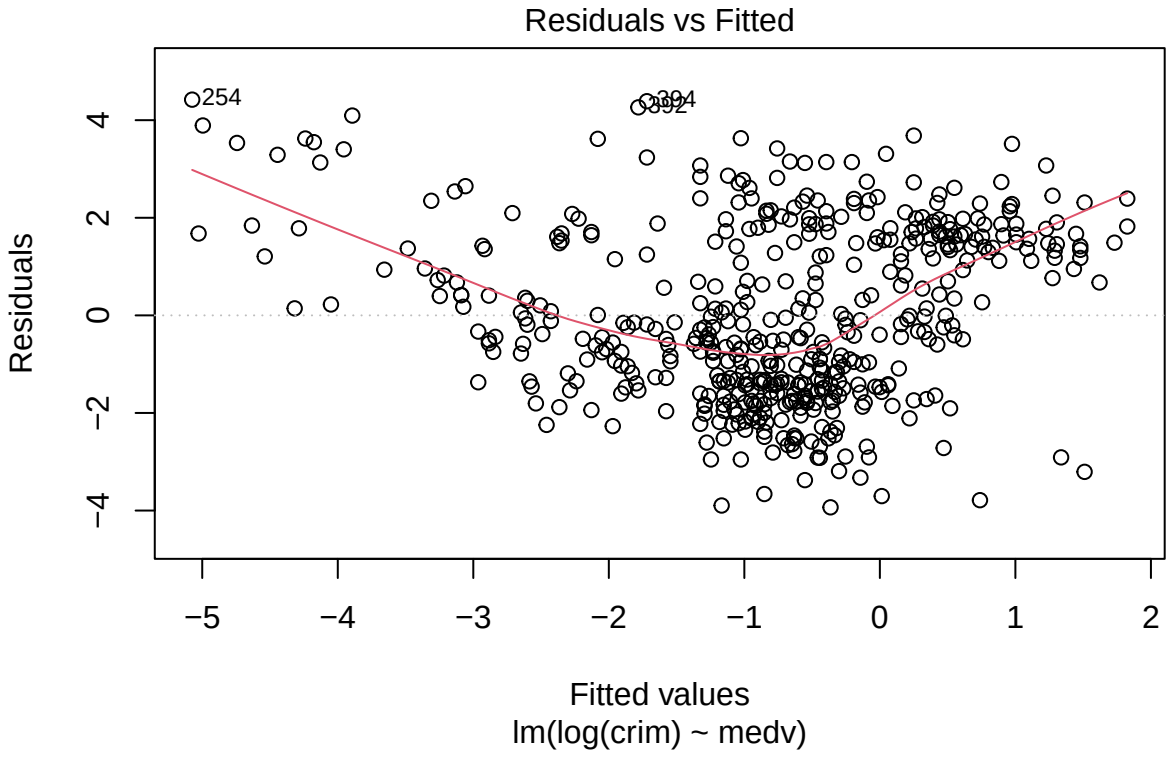

```
library(MASS)
boxcox(out)
```

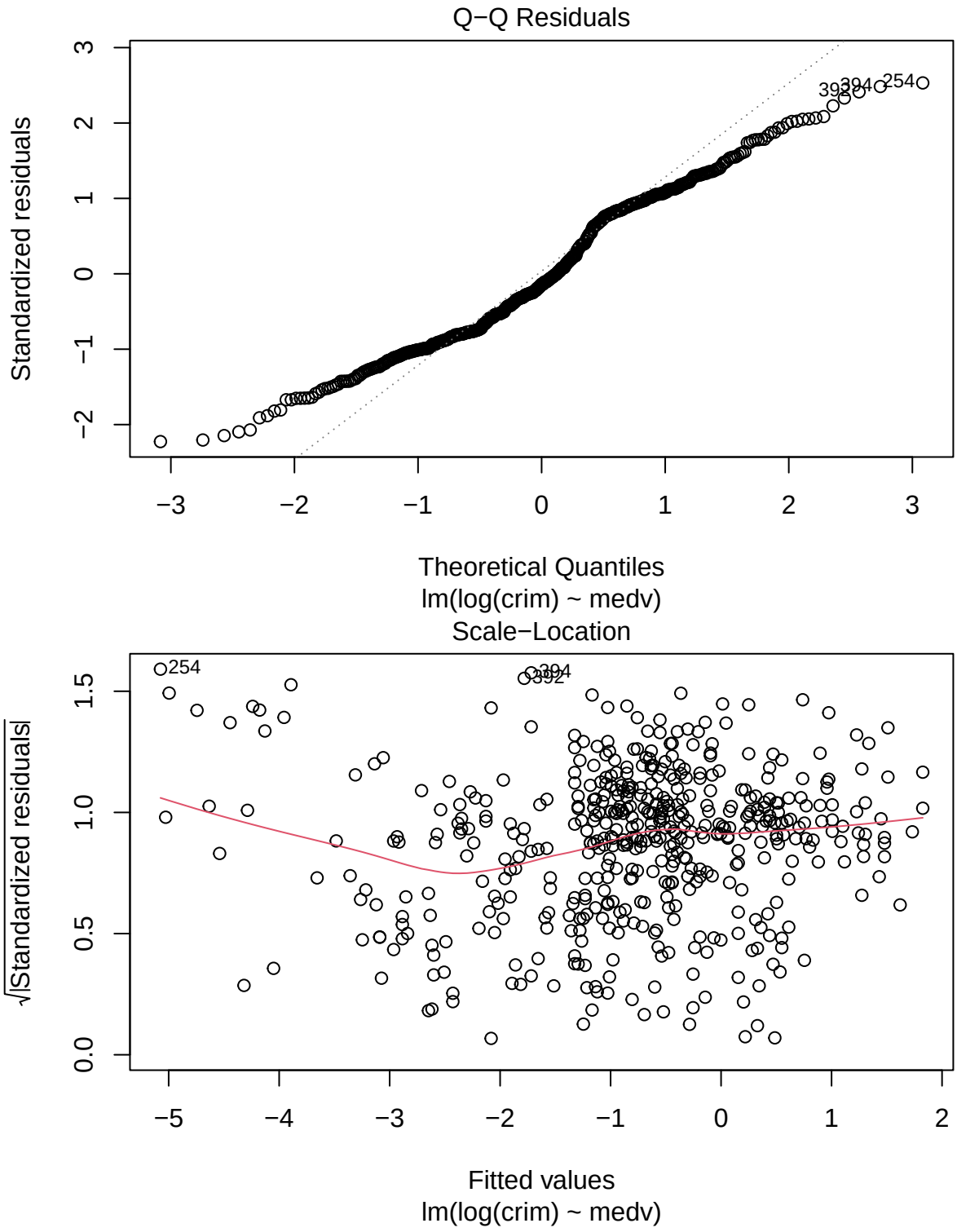


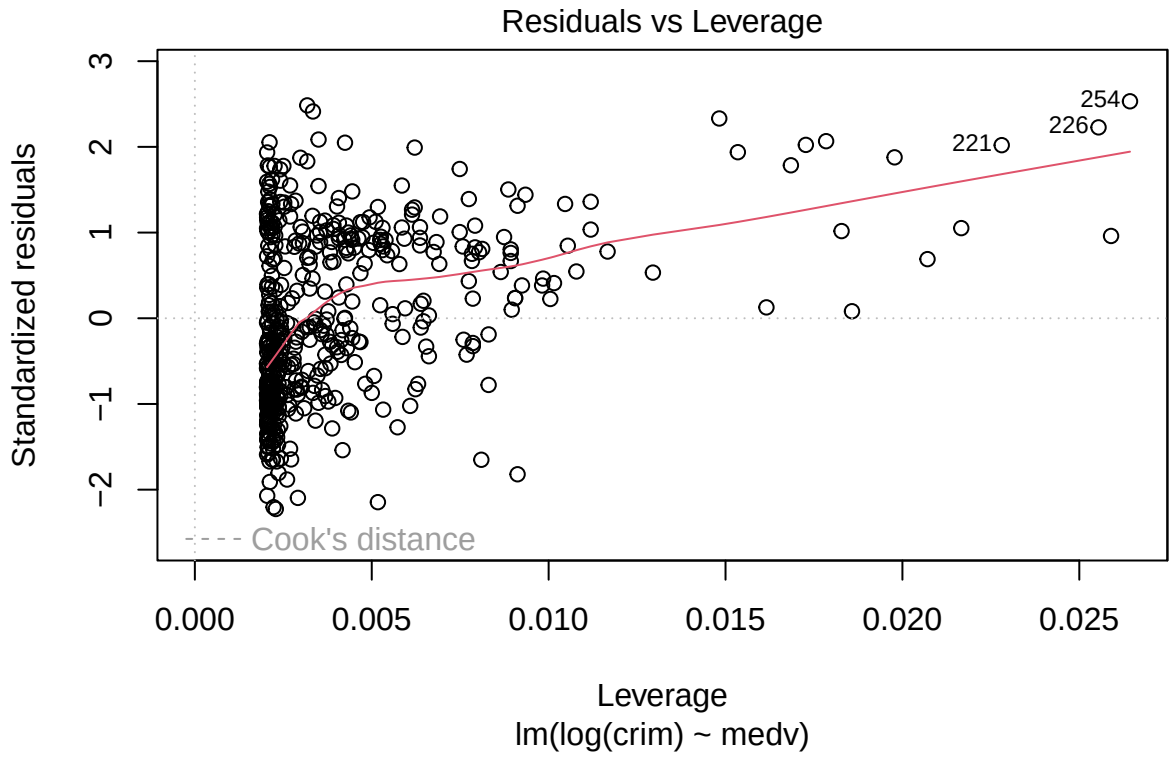
Answer:

7. Fit a linear model on the transformed variable(s) called out.transformed. Do the diagnostics look better? In your opinion, which assumptions are satisfied and which are not?

```
out.transformed <- lm(log(crim) ~ medv, data=dat.edited)
plot(out.transformed)
```



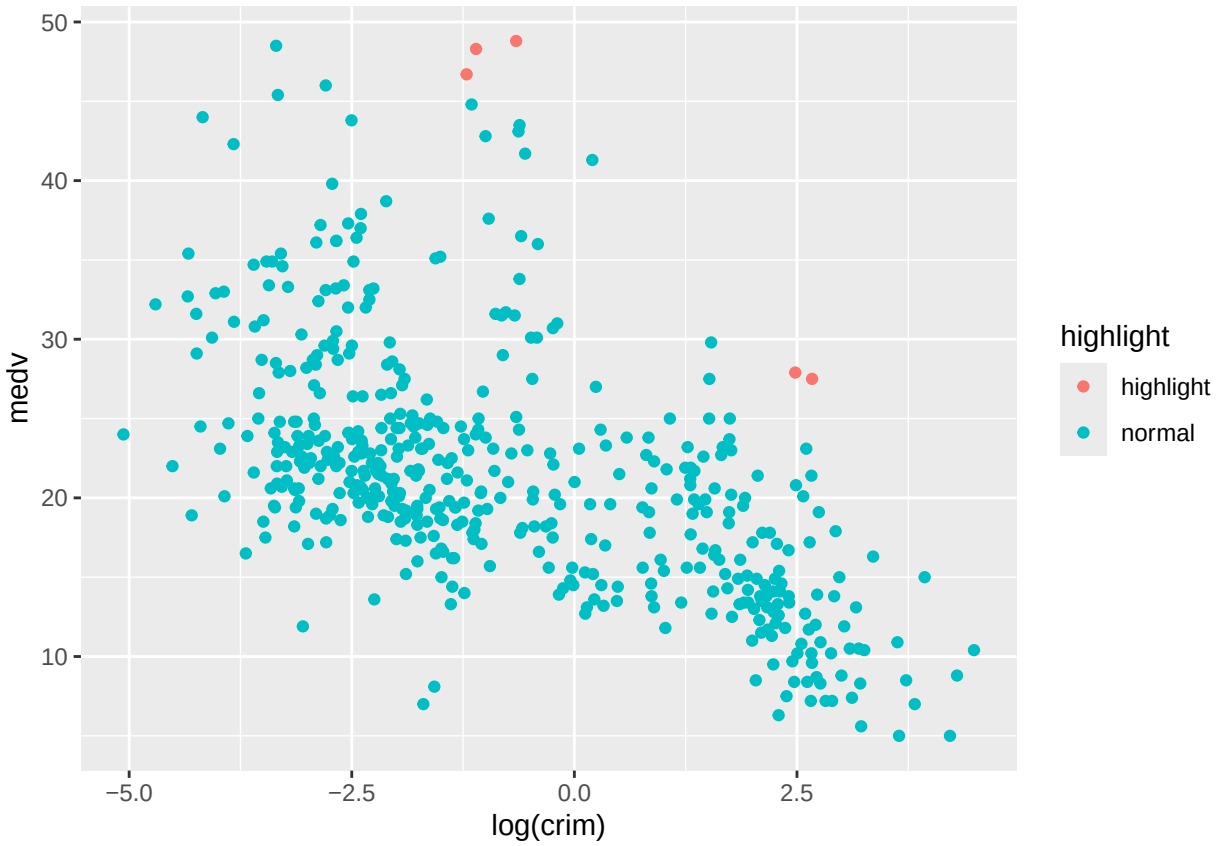

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(log(crim) ~ medv)
Scale−Location
√|Standardized residuals|
Fitted values
lm(log(crim) ~ medv)



Answer:

8. In the diagnostic plots can you see any outliers? If so, draw a scatterplot that shows these outliers. Describe where they are in the scatterplot.

```
dat.edited %>% mutate(highlight = ifelse(row_number() %in% c(221, 226, 254, 392, 394), "highlight", "no
  ggplot(aes(x=log(crim), y=medv)) + geom_point(aes(colour = highlight))
```



Answer:

9. Are these outliers problematic? What would you do with these outliers in fitting your model?

Answer: I would do nothing with the outliers because they are not very influential.

10. Is crime associated with the value of homes? If so, by how much? Note: Fill in the blanks.

Answer: A decrease in median home value by \$1000 is associated with an 16% decrease in crime rate, and this is statistically significantly different from zero at the 0.05 level.