Assignment 3

Sophie Faircloth

Today's date: 10/27/2021

This assignment is due on Canvas on Wednesday 10/27/2021 before class, at 10:15 am. Include the name of anyone with whom you collaborated at the top of the assignment.

Submit your responses as either an HTML file or a PDF file on Canvas. Also, please upload it to your website.

Save the file (found on Canvas) crime_simple.txt to the same folder as this file (your Rmd file for Assignment 3).

Load the data.

```
library(readr)
library(knitr)
dat.crime <- read_delim("crime_simple.txt", delim = "\t")
```

```
## Rows: 47 Columns: 14
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## dbl (14): R, Age, S, Ed, Ex0, Ex1, LF, M, N, NW, U1, U2, W, X
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

This is a dataset from a textbook by Brian S. Everitt about crime in the US in 1960. The data originate from the Uniform Crime Report of the FBI and other government sources. The data for 47 states of the USA are given.

Here is the codebook:

R: Crime rate: # of offenses reported to police per million population

Age: The number of males of age 14-24 per 1000 population

S: Indicator variable for Southern states (0 = No, 1 = Yes)

Ed: Mean of years of schooling x 10 for persons of age 25 or older

Ex0: 1960 per capita expenditure on police by state and local government

Ex1: 1959 per capita expenditure on police by state and local government

LF: Labor force participation rate per 1000 civilian urban males age 14-24

M: The number of males per 1000 females

N: State population size in hundred thousands

NW: The number of non-whites per 1000 population

U1: Unemployment rate of urban males per 1000 of age 14-24

U2: Unemployment rate of urban males per 1000 of age 35-39

W: Median value of transferable goods and assets or family income in tens of $

X: The number of families per 1000 earning below 1/2 the median income

We are interested in checking whether the reported crime rate (# of offenses reported to police per million population) and the average education (mean number of years of schooling for persons of age 25 or older) are related.

1. How many observations are there in the dataset? To what does each observation correspond?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

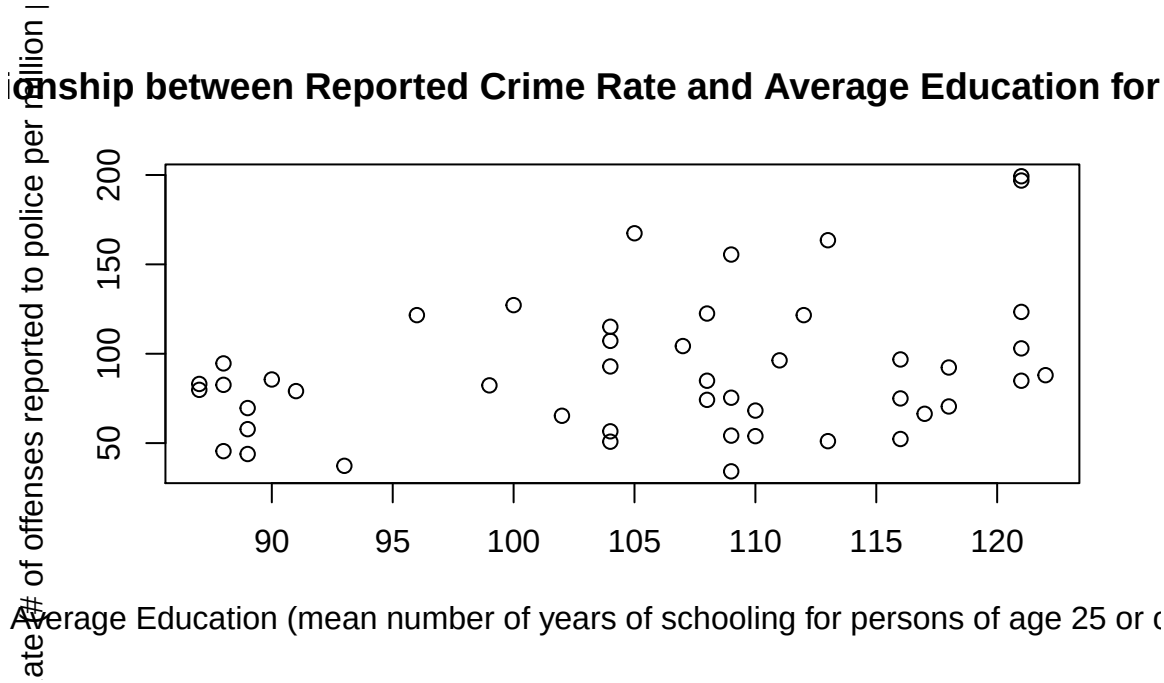
```
count(dat.crime)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    47
```

There are 47 observations in the dataset. The observations correspond to the different rows, aka different states.

2. Draw a scatterplot of the two variables. Calculate the correlation between the two variables. Can you come up with an explanation for this relationship?

```
library(datasets)
plot(dat.crime$Ed,dat.crime$R,  main="Relationship between Reported Crime Rate and Average Education fo
    xlab="Average Education (mean number of years of schooling for persons of age 25 or older)", ylab="
```




```r
x <- cor(dat.crime$R, dat.crime$Ed)
x
```

```
## [1] 0.3228349
```

I cannot come up with an explanation for this relationship on first impression. The correlation of these two variables is 0.3228349, meaning that there is a fairly weak correlation. Because there are so many other factors in this dataset, we can assume that there is likely another factor that correlates more highly to the reported crime rate, but this low of a correlation does not allow us to make any causal inferences.

3. Regress reported crime rate (y) on average education (x) and call this linear model `crime.lm` and write the summary of the regression by using this code, which makes it look a little nicer `{r, eval=FALSE}` `kable(summary(crime.lm)$coef, digits = 2)`.

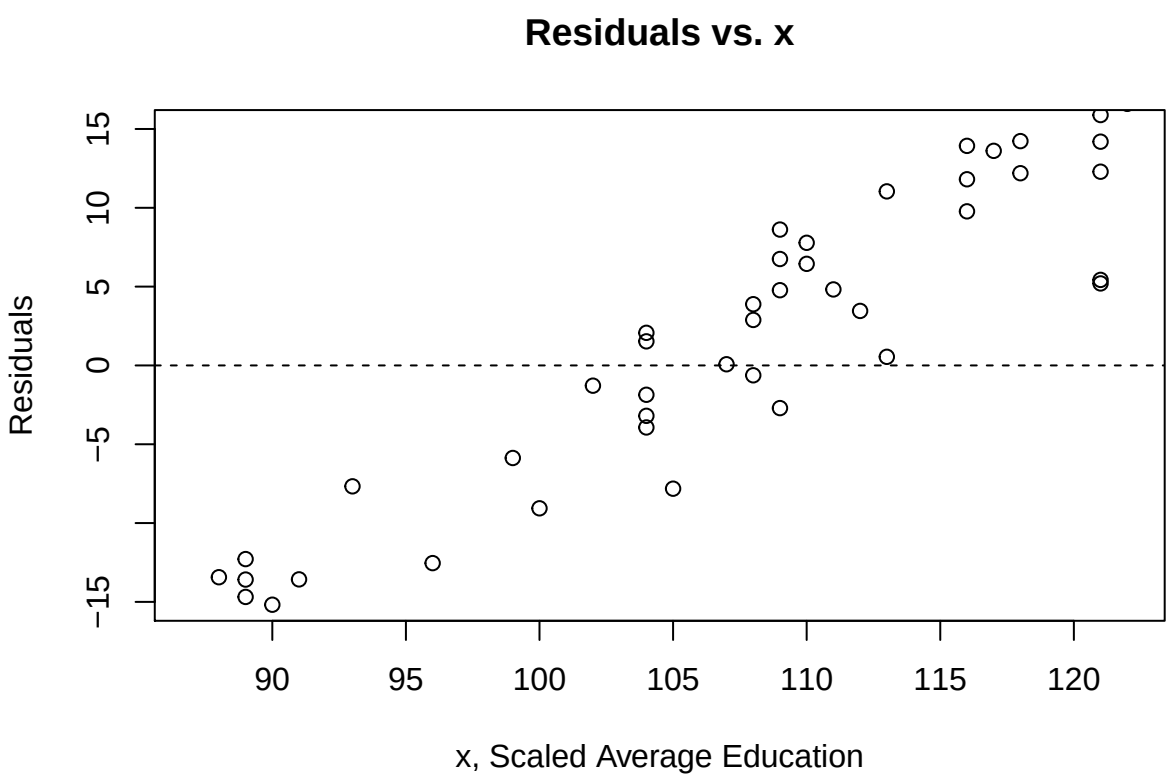
```r
# Remember to remove eval=FALSE above!

dat.crime$R.c = scale(dat.crime$R, center=TRUE, scale=FALSE)
crime.lm <- lm(formula = Ed ~ R.c, data = dat.crime)
# kable(summary(crime.lm)$coef, digits = 2)
summary(crime.lm)
```

```
##
## Call:
## lm(formula = Ed ~ R.c, data = dat.crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.020  -8.441   1.528   8.200  16.596
##
```

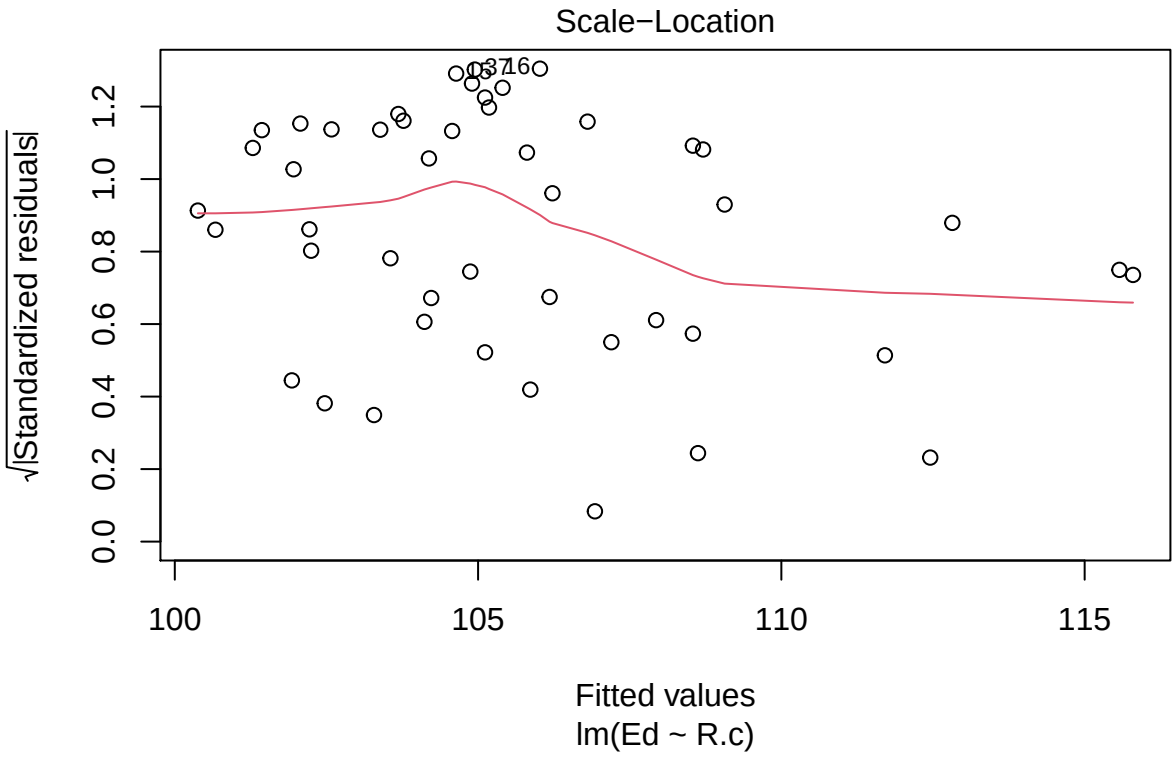
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 105.63830    1.56148  67.653   <2e-16 ***
## R.c           0.09338    0.04081   2.288   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 45 degrees of freedom
## Multiple R-squared:  0.1042, Adjusted R-squared:  0.08432
## F-statistic: 5.236 on 1 and 45 DF,  p-value: 0.02688
```

4. Are the four assumptions of linear regression satisfied? To answer this, draw the relevant plots. (Write a maximum of one sentence per assumption.)

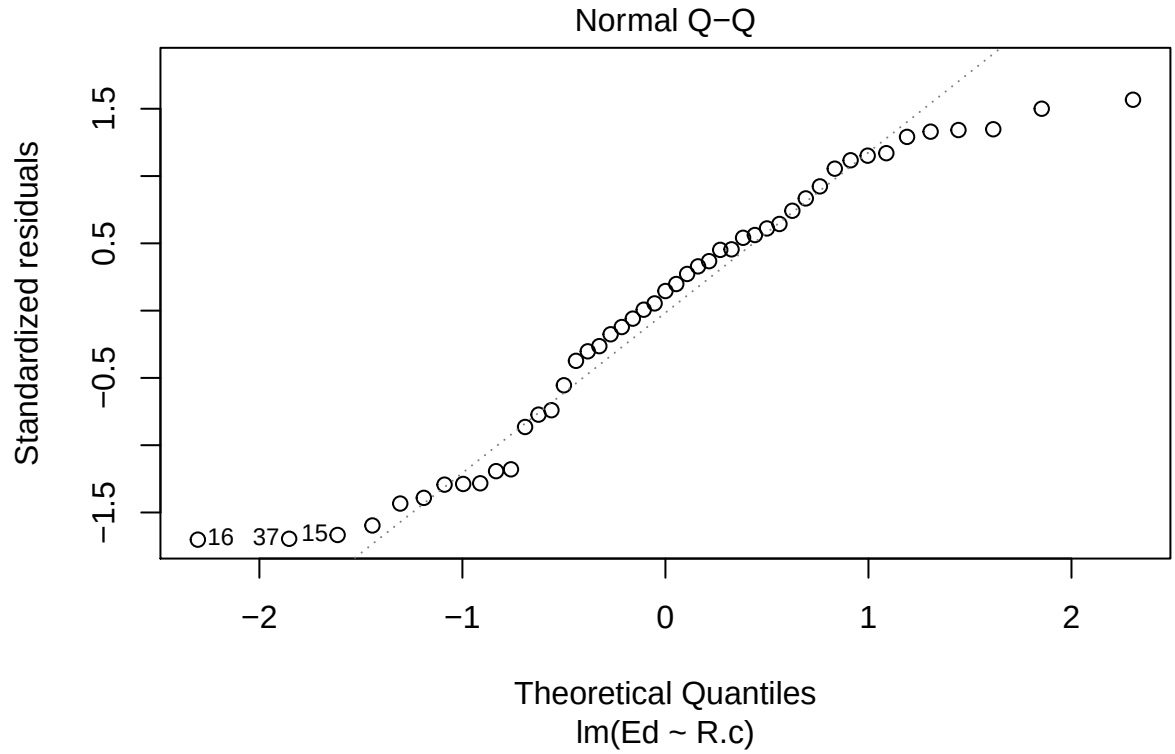
```
# plot 1 & plot 2, residuals vs x
plot(dat.crime$Ed, crime.lm$residuals, ylim=c(-15,15), main="Residuals vs. x", xlab="x, Scaled Average E
abline(h = 0, lty="dashed")
```



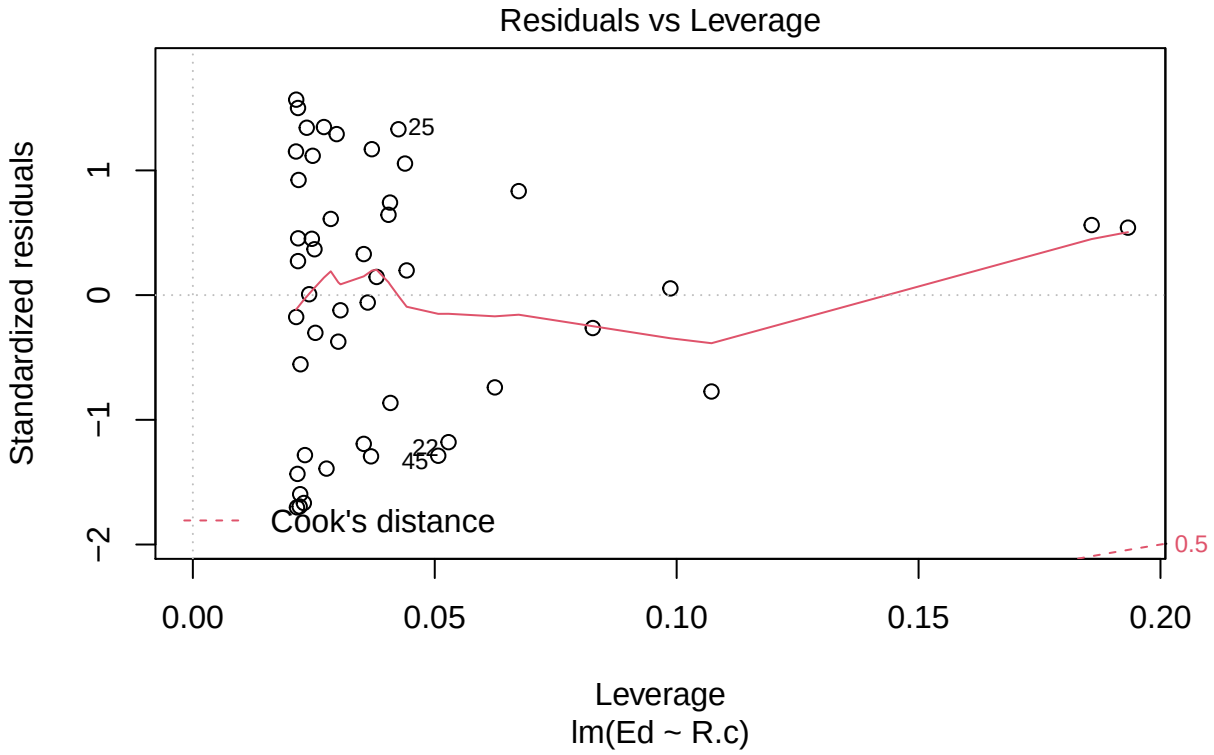

```
# plot 3
plot(crime.lm, which=3)
```




```
# plot 4, qq and outlier condition
plot(crime.lm, which=2)
```



```
plot(crime.lm, which=5)
```



The four assumptions are as follows: linearity, independence, equal variance, and normal population. The first assumption - linearity - is satisfied because the first plot of the residuals vs x has a straight line of plots. The second assumption - independence - is satisfied because the plot of residuals vs x does not display any distinct patterns. The third assumption - equal variance - is satisfied because the flat line shows that the errors are of constant variance. The fourth assumption - normal population - is debatably not satisfied because the qq plot is heavily tailed, meaning it does not follow the normal model well.

5. Is the relationship between reported crime and average education statistically significant? Report the estimated coefficient of the slope, the standard error, and the p-value. What does it mean for the relationship to be statistically significant?

The estimated coefficient is 0.09. The standard error is 0.04. The p-value is 0.02688. This means that this relationship is statistically significant, as the p-value is less than 0.05. If the relationship is statistically significant, this means that there is a high likelihood that the observance is not wrong or random (I use this word carefully) chance.

6. How are reported crime and average education related? In other words, for every unit increase in average education, how does reported crime rate change (per million) per state?

For every unit increase in average education, the reported crime rate increases by 0.09 units, meaning that the reported crime rate increases by 0.09 offenses reported to police per million population.

7. Can you conclude that if individuals were to receive more education, then crime will be reported more often? Why or why not?

We cannot conclude this from the dataset. While it appears there is a statistically significant correlation, this does not mean we can assume causation. While the variables correlate to each other, this is not conducive of a causal relationship.