Assignment 3

Elizabeth Stoner

10/23/2021

Collaborators: Carmen Avery and Halle Wasser.

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?

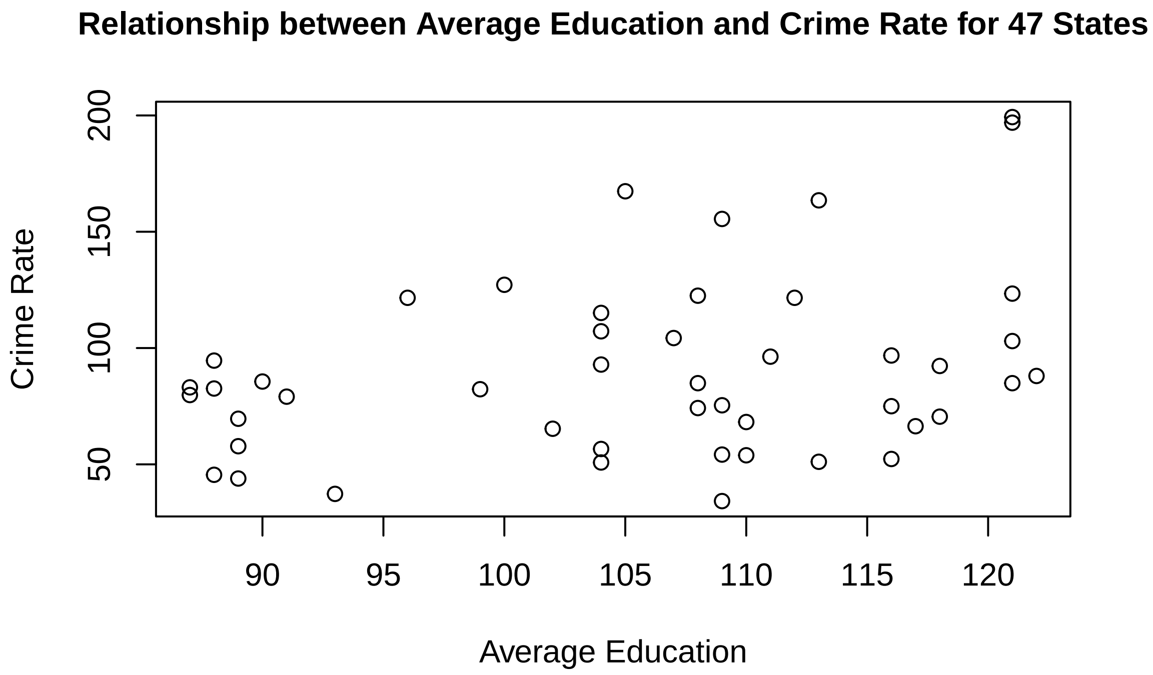
```
summary(dat.crime)
```

```
##        R               Age              S                Ed
##  Min.   : 34.20   Min.   :119.0   Min.   :0.0000   Min.   : 87.0
##  1st Qu.: 65.85   1st Qu.:130.0   1st Qu.:0.0000   1st Qu.: 97.5
##  Median : 83.10   Median :136.0   Median :0.0000   Median :108.0
##  Mean   : 90.51   Mean   :138.6   Mean   :0.3404   Mean   :105.6
##  3rd Qu.:105.75   3rd Qu.:146.0   3rd Qu.:1.0000   3rd Qu.:114.5
##  Max.   :199.30   Max.   :177.0   Max.   :1.0000   Max.   :122.0
##       Ex0              Ex1               LF              M
##  Min.   : 45.0   Min.   : 41.00   Min.   :480.0   Min.   : 934.0
##  1st Qu.: 62.5   1st Qu.: 58.50   1st Qu.:530.5   1st Qu.: 964.5
##  Median : 78.0   Median : 73.00   Median :560.0   Median : 977.0
##  Mean   : 85.0   Mean   : 80.23   Mean   :561.2   Mean   : 983.0
##  3rd Qu.:104.5   3rd Qu.: 97.00   3rd Qu.:593.0   3rd Qu.: 992.0
##  Max.   :166.0   Max.   :157.00   Max.   :641.0   Max.   :1071.0
##        N               NW              U1               U2
##  Min.   :  3.00   Min.   :  2.0   Min.   : 70.00   Min.   :20.00
##  1st Qu.: 10.00   1st Qu.: 24.0   1st Qu.: 80.50   1st Qu.:27.50
##  Median : 25.00   Median : 76.0   Median : 92.00   Median :34.00
##  Mean   : 36.62   Mean   :101.1   Mean   : 95.47   Mean   :33.98
##  3rd Qu.: 41.50   3rd Qu.:132.5   3rd Qu.:104.00   3rd Qu.:38.50
##  Max.   :168.00   Max.   :423.0   Max.   :142.00   Max.   :58.00
##        W               X
##  Min.   :288.0   Min.   :126.0
##  1st Qu.:459.5   1st Qu.:165.5
##  Median :537.0   Median :176.0
##  Mean   :525.4   Mean   :194.0
##  3rd Qu.:591.5   3rd Qu.:227.5
##  Max.   :689.0   Max.   :276.0
```

There are 47 observations in the dataset, which correspond to 47 of the USA states and their 1960 crime data. There are 14 columns of variables providing information for each observation.

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?

```
plot(dat.crime$Ed, dat.crime$R, main="Relationship between Average Education and Crime Rate for 47 Stat
```



```
cor(dat.crime$Ed, dat.crime$R)
```

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

The correlation between the two variables is 0.3228349, which is a positive correlation. According to the codebook, R, the crime rate, is the number of offenses reported to police per million population, and Ed is the average years of schooling multiplied by ten for people of 25 years or older. So, as the average amount of schooling increases, so does the crime rate. A possible explanation for this relationship is that areas where people have higher levels of education may be better targets for those committing crimes. Education level is often correlated with better jobs and financial success, so perhaps areas with these people may be home to more criminal activity. Also, because R is specifically number of crimes reported, then maybe people with higher education levels are also more likely to report crimes to the police; however, this hypothesis cannot be proven by this data alone.

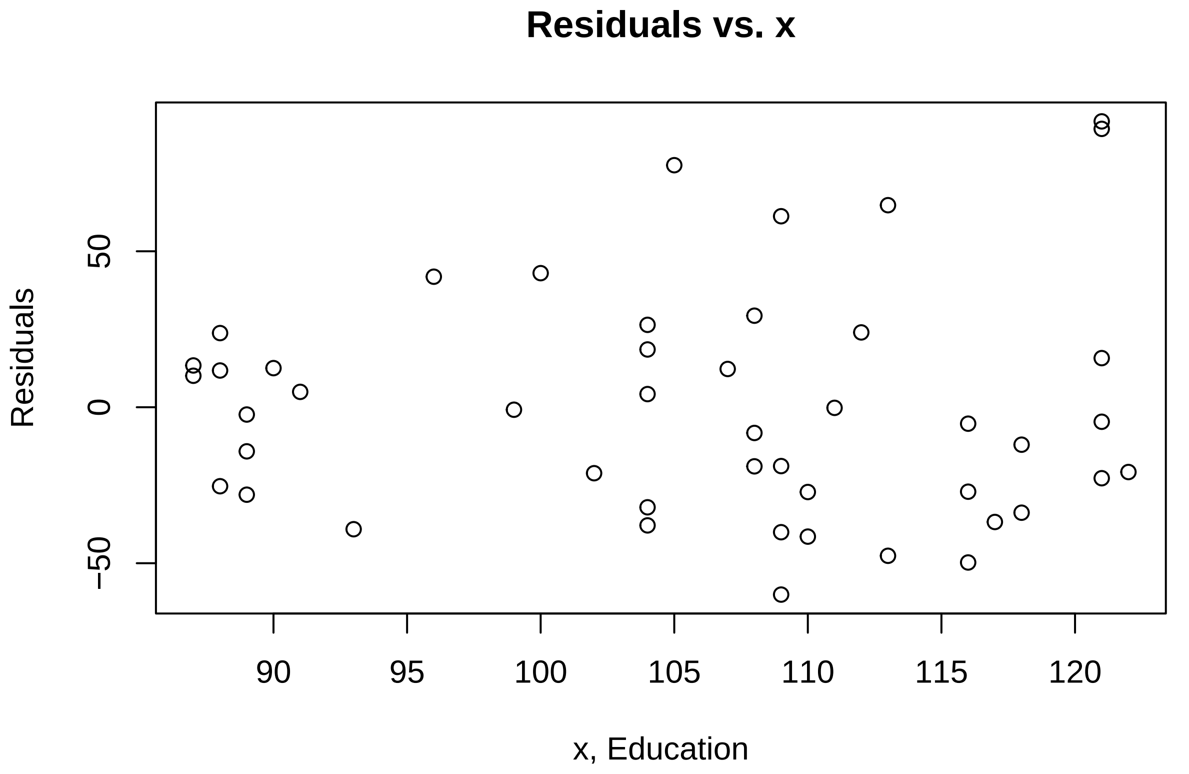
3. Regress reported crime rate (y) on average education (x) and call this linear model `crime.lm` and write the summary of the regression by using this code, which makes it look a little nicer `{r, eval=FALSE} kable(summary(crime.lm)$coef, digits = 2)`.

```
#install.packages("kableExtra")
crime.lm <- lm(formula=R~Ed, data=dat.crime)
# Remember to remove eval=FALSE above!
summary(crime.lm)
```

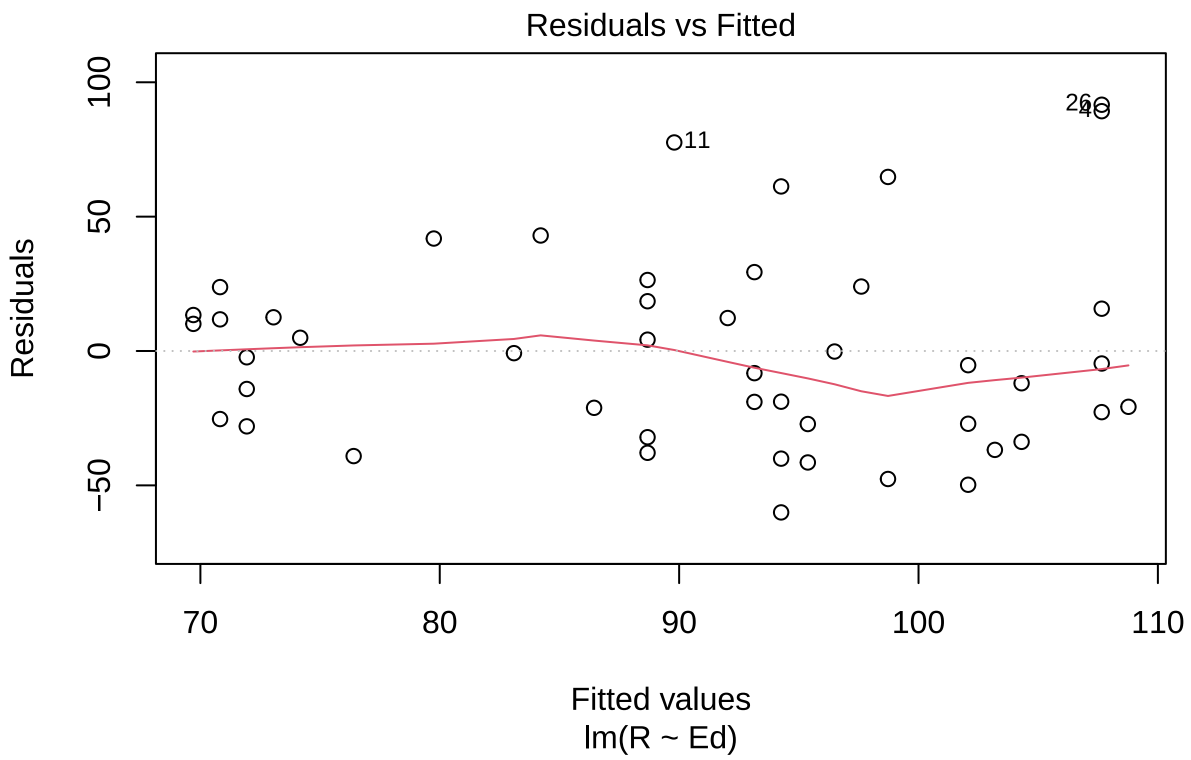
```
##
## Call:
## lm(formula = R ~ Ed, data = dat.crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.061 -27.125  -4.654  17.133  91.646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.3967    51.8104  -0.529   0.5996
## Ed            1.1161     0.4878   2.288   0.0269 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 37.01 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(dat.crime$Ed, crime.lm$residuals, main="Residuals vs. x", xlab="x, Education", ylab="Residuals") #
```

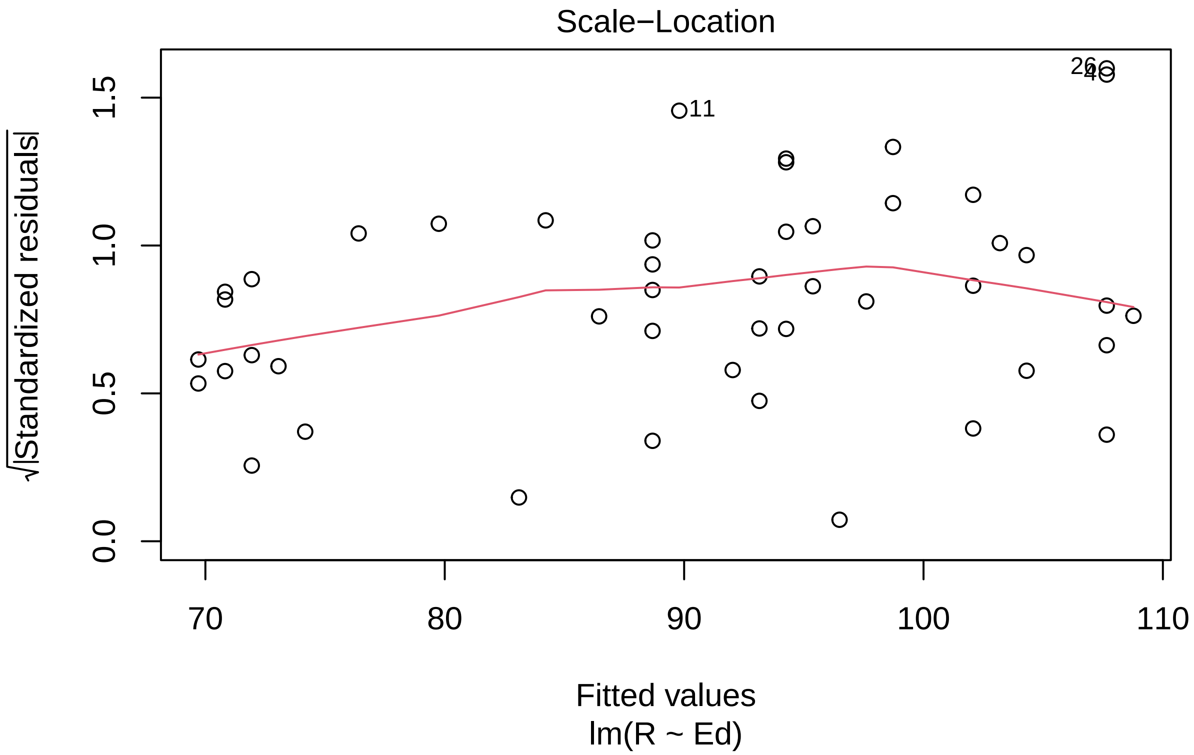


```
plot(crime.lm, which=1) #Plot for the linearity assumption and independence assumption
```



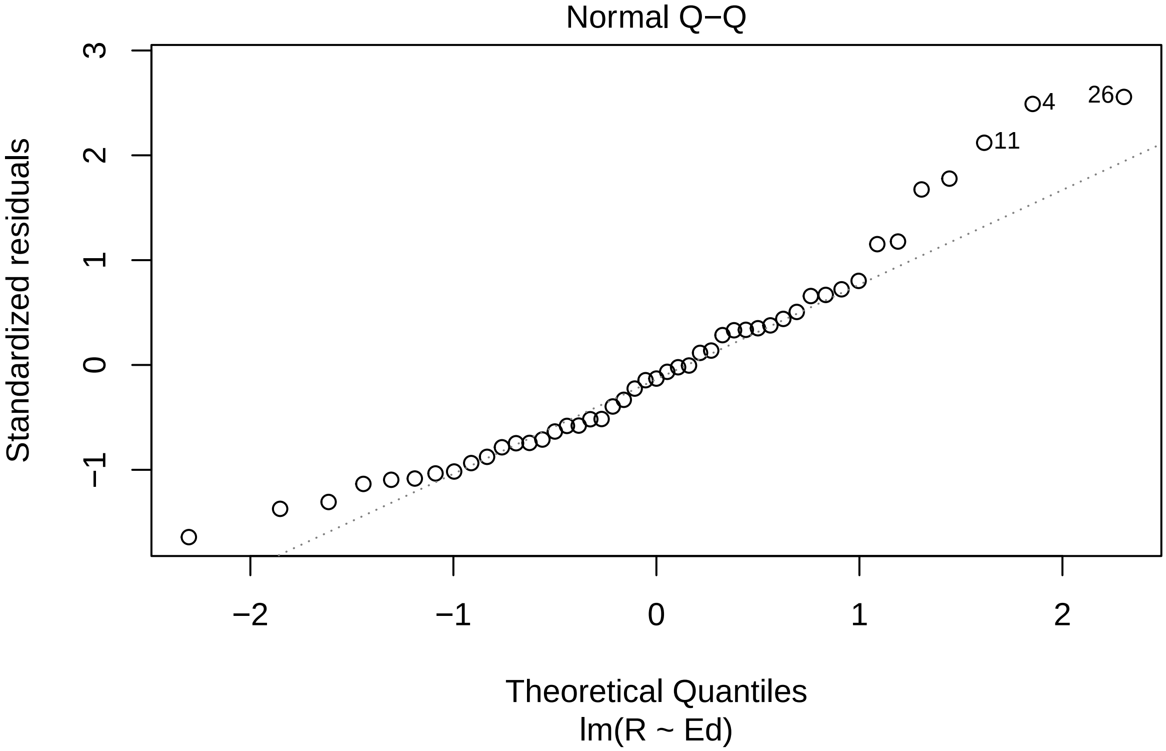
Based on the plot, the linearity assumption does not look to be met as the red line appears to have a varying pattern that is non-linear. The same can be said for the independence assumption; there is some sort of pattern, so the independence assumption is not met as well as we would like it to be.

```
plot(crime.lm, which=3) #Plot to test for the equal variance assumption
```



Using the scale-location plot, it appears as though the data does not perfectly satisfy the equal variance assumption as the line curves upward then decreases around the x-value of 100, and the distribution of the data points increase around the middle but is thinner near the beginning, suggesting a lack of homoscedasticity.

```
plot(crime.lm, which=2) #Plot to test the normal population assumption
```



Based on the normal qq plot there are issues with the normal population assumption, as the plot is light tailed on both ends, meaning that those data points could be outliers, which no longer follow the slope of the line.

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 of the slope is 1.1161, which indicates the effect that education has on the reported crime rate. The standard error is 0.4878, which indicates the potential variance of crimes actually reported. The p-value is 0.0269 for the slope, which indicates that there is only a weak relationship between reported crime and average education. This means that the relationship is not statistically significant. A statistically significant relationship would be such that the relationship is unlikely to occur randomly, or unlikely to occur given the null hypothesis. Therefore, this relationship is more likely to have occurred due to randomness or some other factor.

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?

Based on the data, for every increase in one unit of average education, number of offenses reported to the police increases by 1.1161 per million per state.

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

Using the data provided we cannot conclude that if individuals were to receive more education, then crime would be reported more often. The statistical relationship is weak, at best. And, even so, that relationship does not automatically indicate causality. There could be other factors that have yet to be measured that could influence the perceived relationship. For example, community relations with local law enforcement is an important factor in a willingness to report crimes, but is not observably related to education levels. So, in order to infer any further relationship, more information and more data are needed.