Assignment 3

Akshat Shah

Today's date here: 10/22/2021

Collaborators: .

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?

```
names(dat.crime)
```

```
##  [1] "R"   "Age" "S"   "Ed"  "Ex0" "Ex1" "LF"  "M"   "N"   "NW"  "U1"  "U2"
## [13] "W"   "X"
```

```
dim(dat.crime)
```

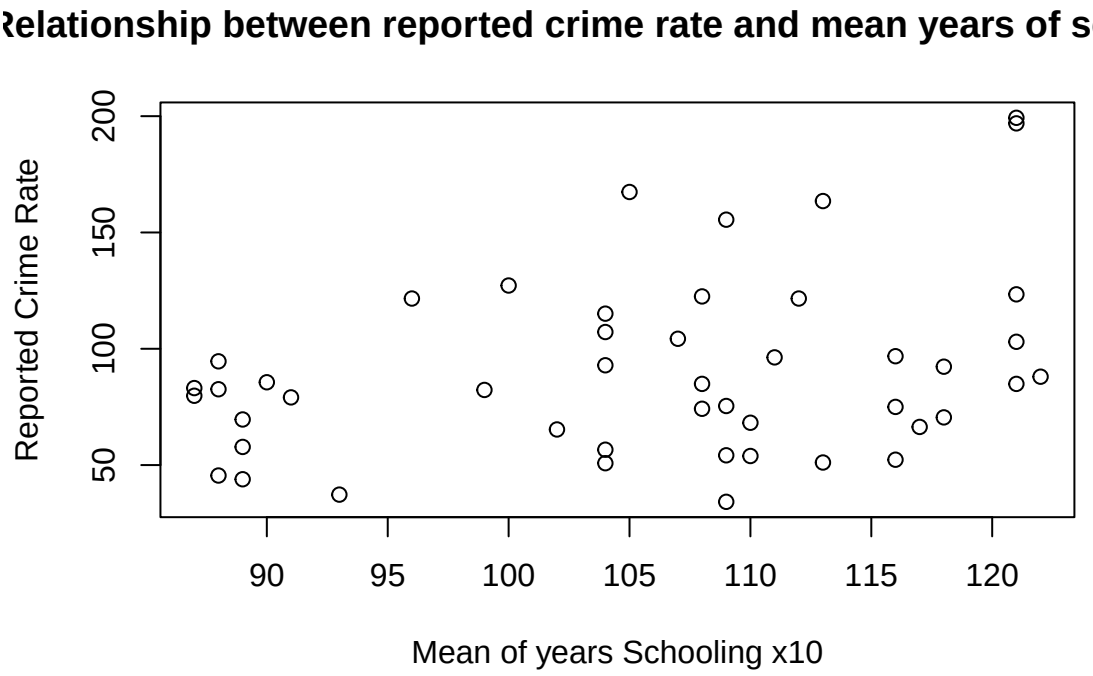
```
## [1] 47 14
```

Excluding the first row (because it tells us the names of the columns), there are 46 observations. Each column corresponds to “R” “Age” “S” “Ed” “Ex0” “Ex1” “LF” “M” “N” “NW” “U1” “U2” “W” “X” respectively. In our codebook we see that each of the observations is a different data that was recorded.

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?

R ED

```
plot(dat.crime$Ed, dat.crime$R, main="Relationship between reported crime rate and mean years of schooli
```




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

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

The correlation between these two variables is 0.3328349. I don't think this correlation is high enough for us to draw a particular conclusion yet.

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)`.

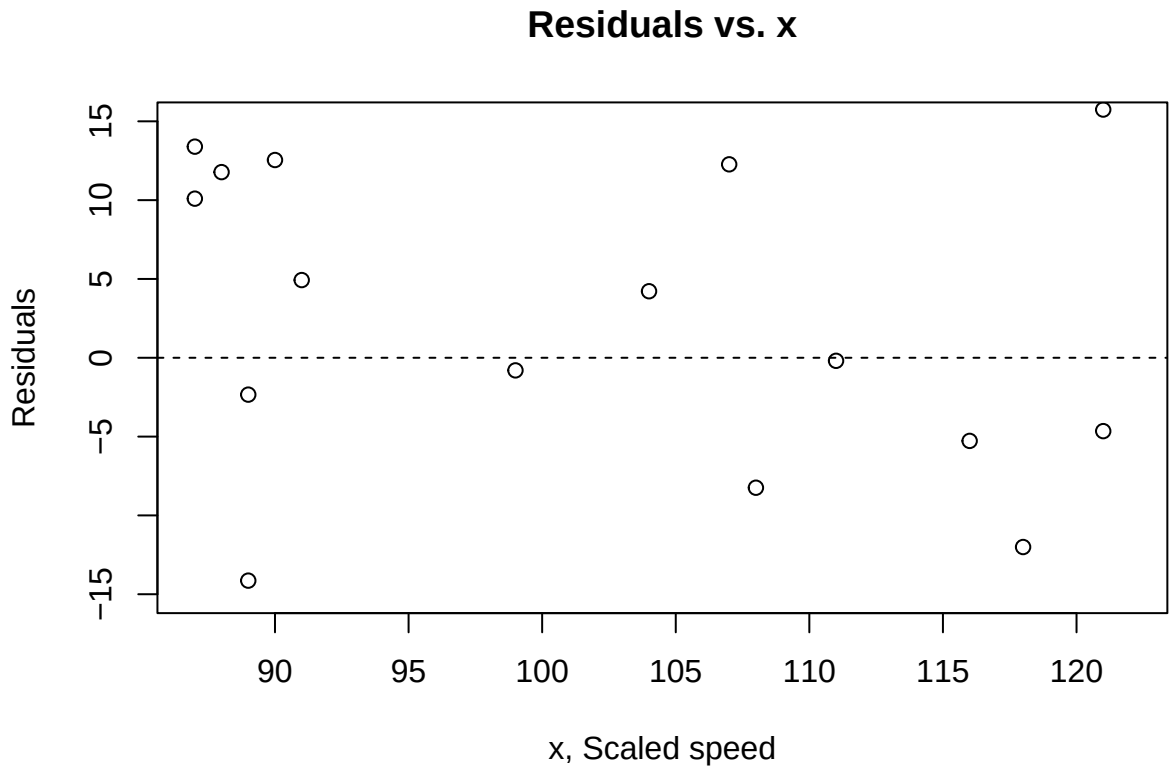
```
# Remember to remove eval=FALSE above!

crime.lm <- lm(formula = dat.crime$R ~ dat.crime$Ed, data = dat.crime)
kable(summary(crime.lm)$coef, digits = 2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.40	51.81	-0.53	0.60
dat.crime$Ed	1.12	0.49	2.29	0.03

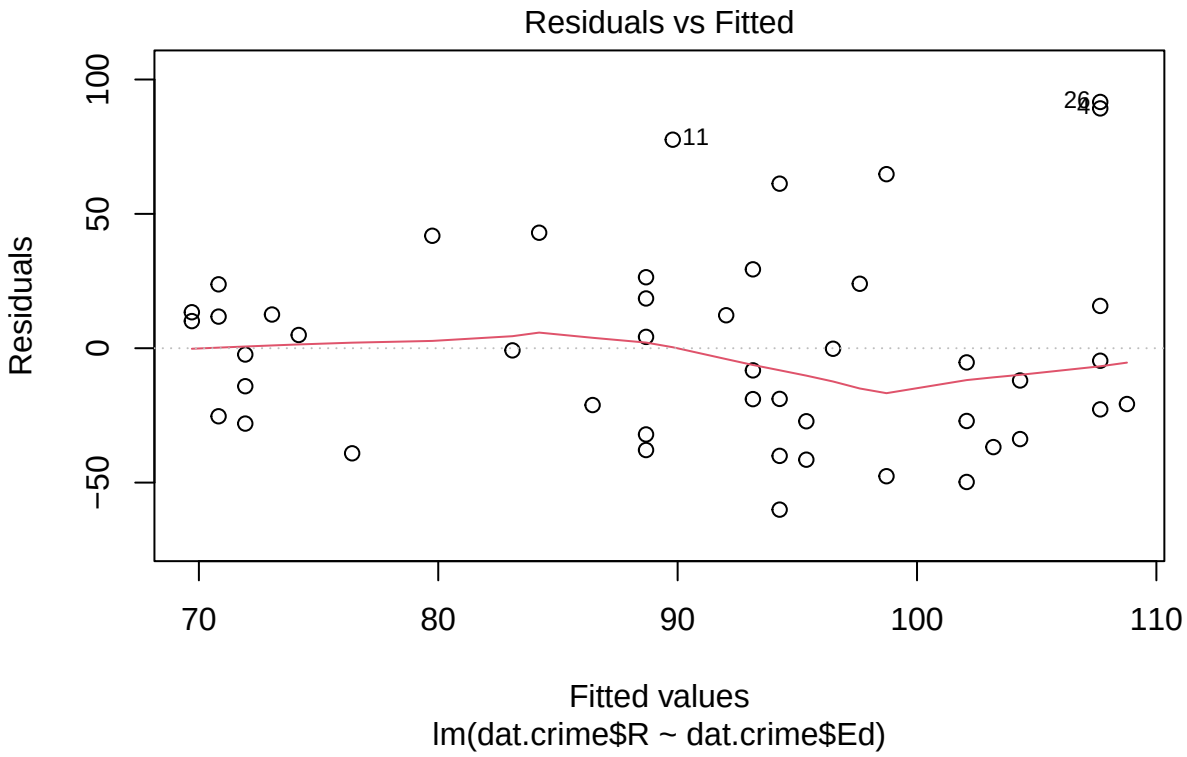
This regression has a slope of 1.12 which means for every unit of education that increases the reported crime increases by 1.12. Additionally, we see that the error for this is 0.49 (which is how much is varys from the average). Additionally we have a p value of about 2% (which means that this is significant) and our regression is 2.29 standard deviations away from the 0. Another thing to take note of is the fact that our intercept estimate is quite different in comparison to our slope. The estimate is larger as well as the error. Because of this our P value is not a significant one for the intercept.

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

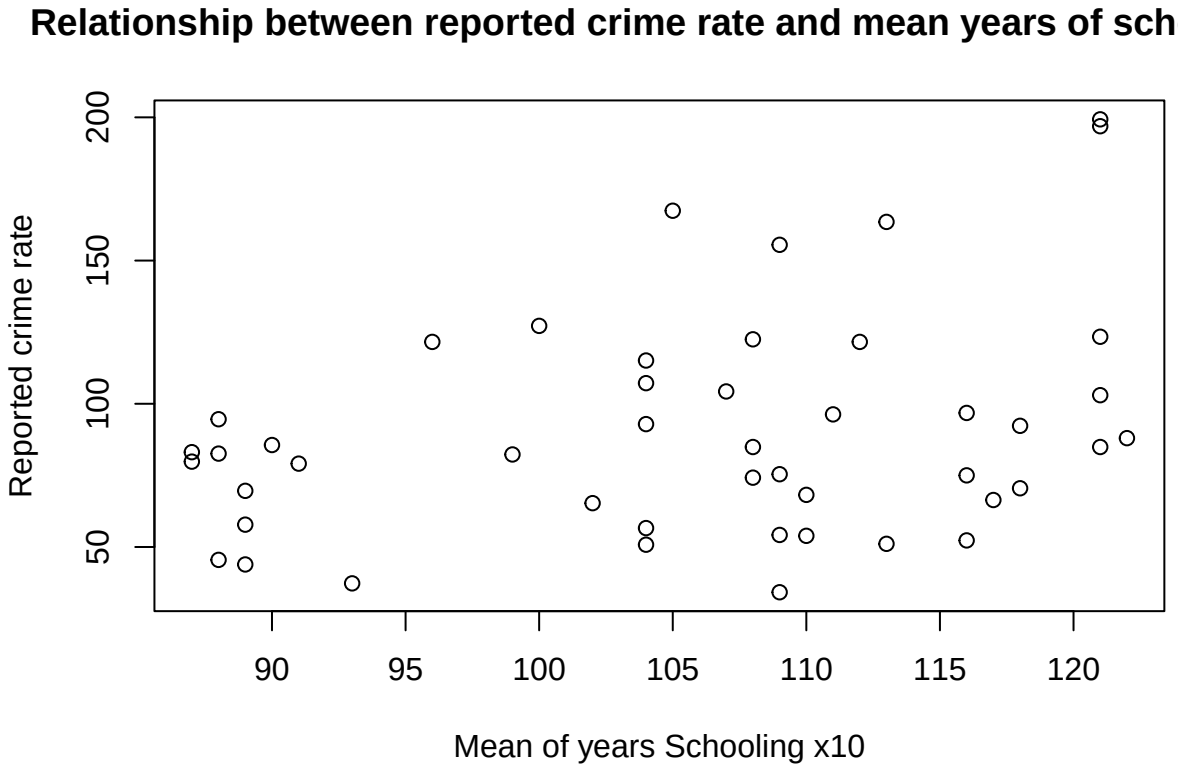
```
plot(dat.crime$Ed, crime.lm$residuals, ylim=c(-15,15), main="Residuals vs. x", xlab="x, Scaled speed", y
abline(h = 0, lty="dashed")
```



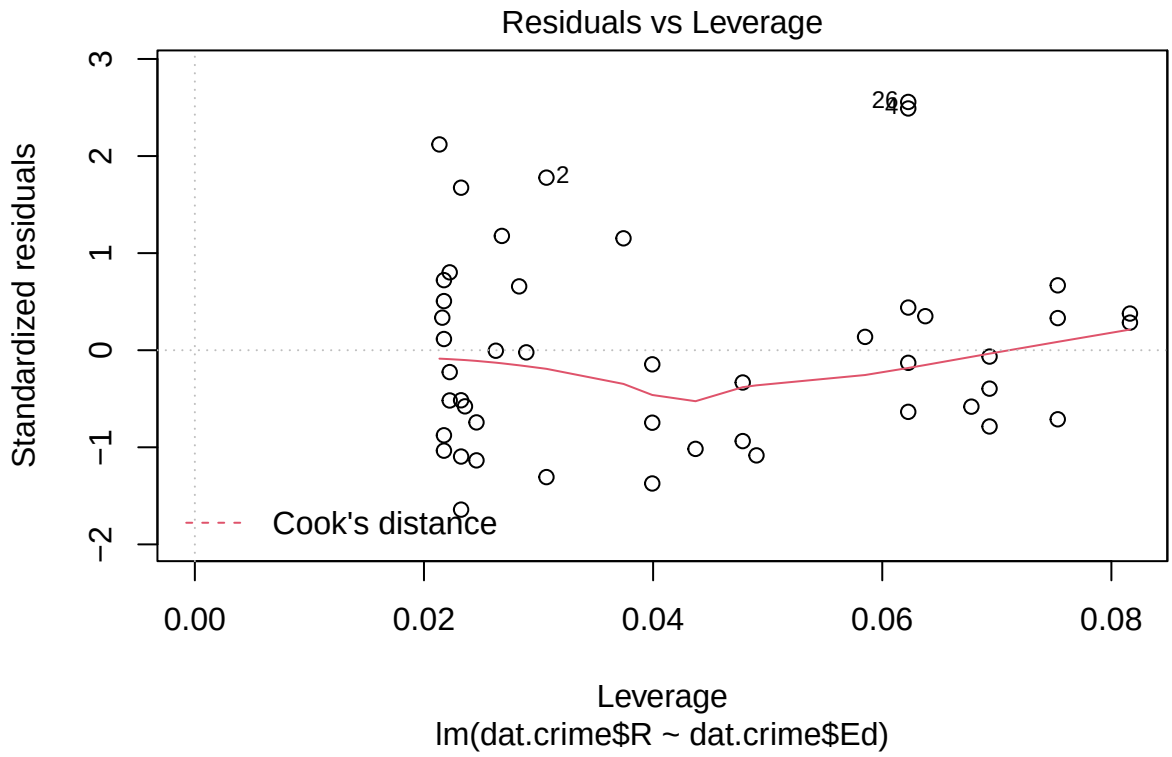

```
plot(crime.lm, which=1)
```



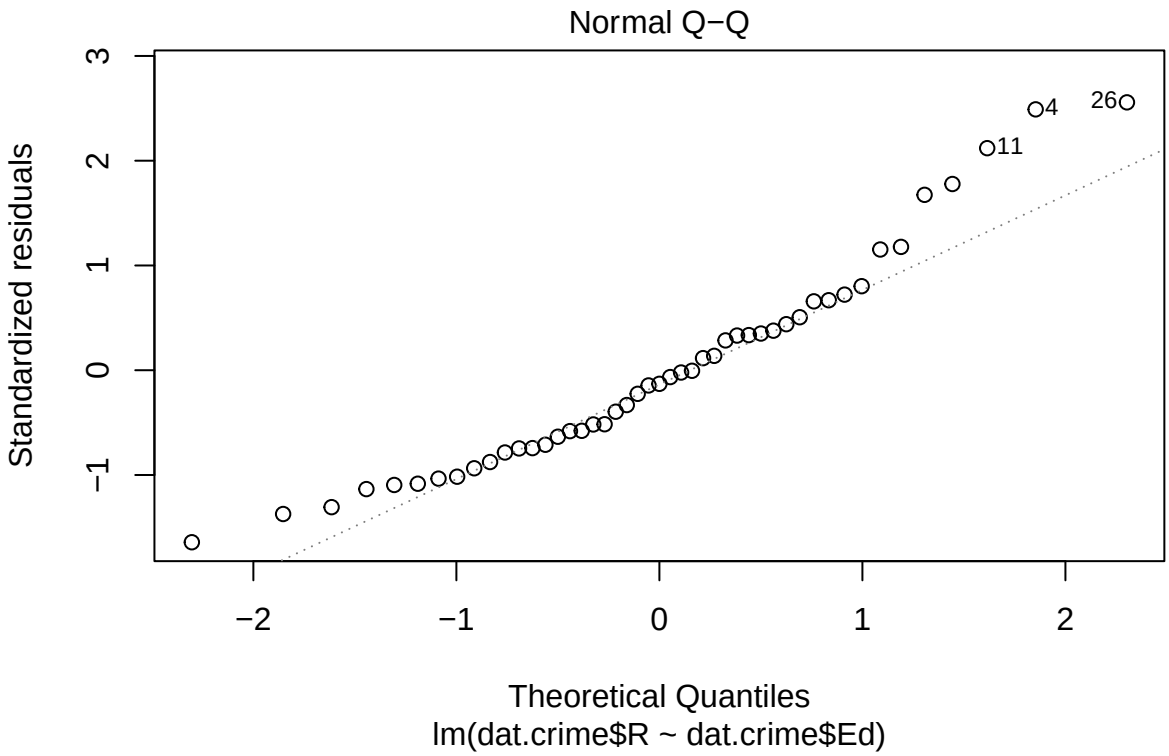

```
plot(dat.crime$Ed, dat.crime$R, main="Relationship between reported crime rate and mean years of schooli
```




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




```
plot(crime.lm, which=2)
```



1.Linearly Assumption: Looking at the first two plots for Residuals vs X and Residuals vs Fitted we see that these have no real pattern for the nodes as the red line is almost completely flat.

2. Independence assumption : Looking at Residuals vs X, there seems to be no pattern and the nodes seem random.

3.Equal variance assumption : There is a slight unequal amount of variable between points of x = 92 to x = 100, therefore this assumption doesn't stand.

4. Normal Population Assumption Our qqplot shows us that our residuals tend to be larger in magnitude and overestimate (on the right side), and our right tail is lighter than the rest for a normal distribution.

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?

```
summary(crime.lm)
```

```
##
## Call:
## lm(formula = dat.crime$R ~ dat.crime$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
## dat.crime$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
```

Our estimated coefficient of slope is 1.1161. Our standard residual error was 37.01 on 45 degrees of freedom. We have a high p value for our intercept and a p value lower for our slope. For the the relationship to be statistically significant means there exists a relationship between reported crime and average education.

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?

Because we have a significant slope value, we can say that we reject the null. For every unit increase of average education, the reported crime rate changes by 1.1161.

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

Based of our linear regression, we can conclude that individuals who receive more education will report crime more often since we obtained a pvalue that was statistically significant. This tells us that there exists some relationship (from the slope), which we can see is increasing as a unit of education increases. However, we need to keep in mind that this test doesn't have enough data points because it is hard to tell if the equal variance and the normal population assumption pass.