# Assignment 3

Aisha Njie

Today's date here: 10/27/2021

**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

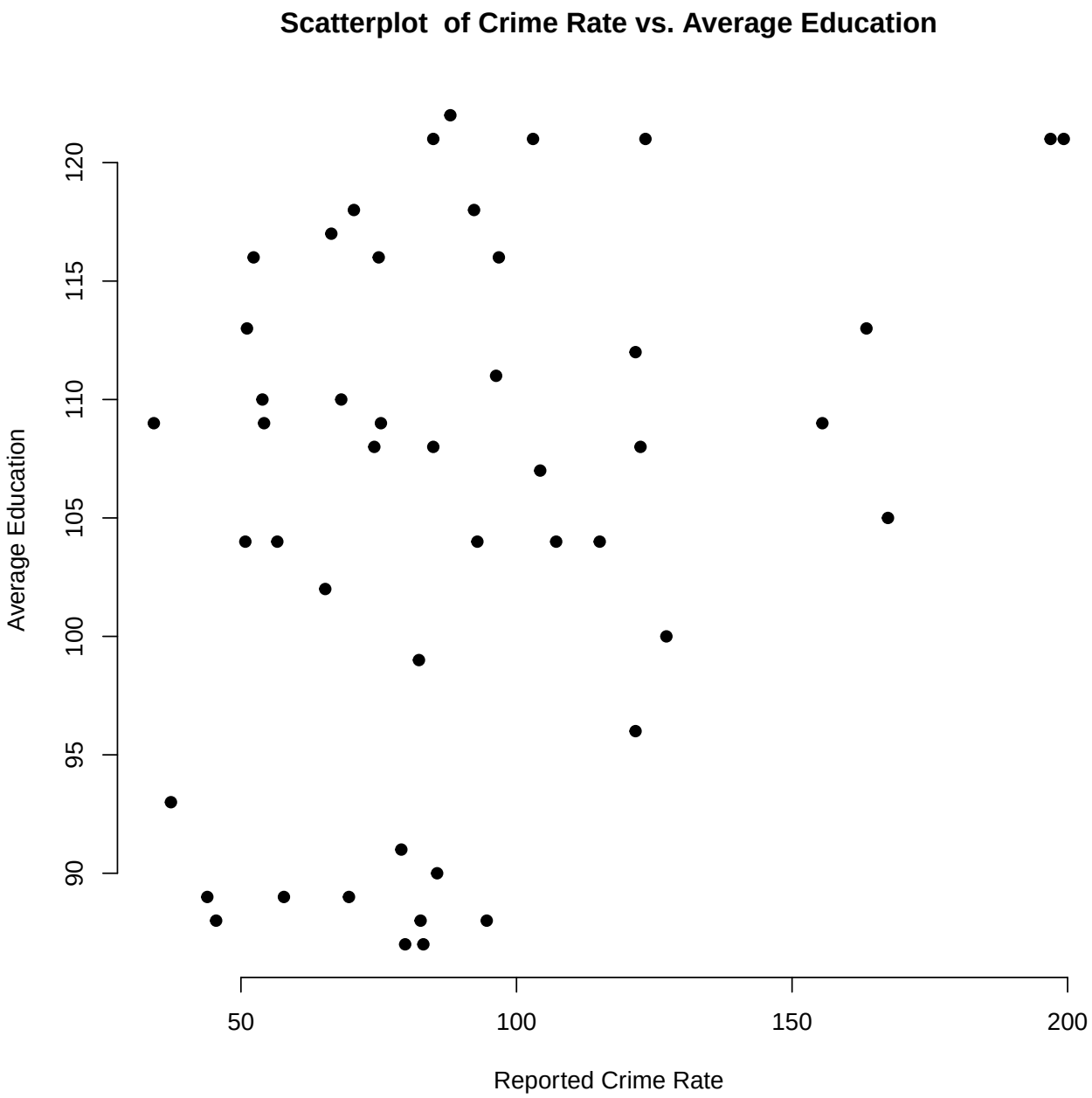
We are interested in checking whether the reported crime rate (# of offenses reported to police per million population) and the average education (mean number of years of schooling for persons of age 25 or older) are related.

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

**There are 47 observations in this dataset. The observations correspond to 47 U.S. states. This information is given in the introduction of the codebook.**

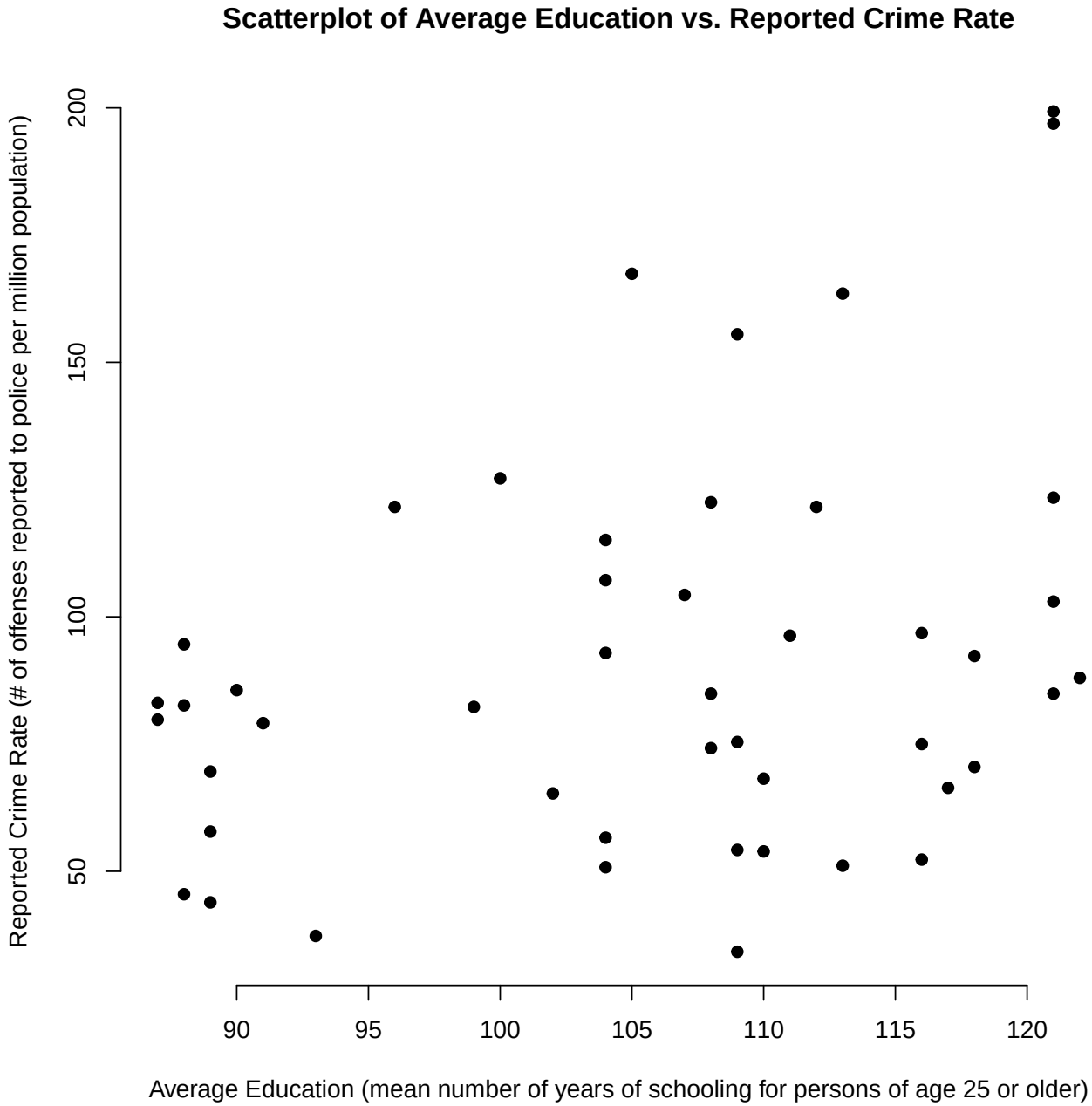
2. Draw a scatterplot of the two variables. Calculate the correlation between the two variables. Can you come up with an explanation for this relationship?

```
x <- dat.crime$R
y <- dat.crime$Ed
plot(x, y, main = "Scatterplot  of Crime Rate vs. Average Education",
     xlab = "Reported Crime Rate", ylab = "Average Education",
     pch = 19, frame = FALSE)
```




**It appears that my initial scatterplot organizes the data in a way that is hard to read so I inverted the variables.**

```
x <- dat.crime$Ed
y <- dat.crime$R
plot(x, y, main = "Scatterplot of Average Education vs. Reported Crime Rate",
     xlab = "Average Education (mean number of years of schooling for persons of age 25 or older)", ylab
     pch = 19, frame = FALSE)
```




```
cor(x, y, method = c("pearson", "kendall", "spearman"))
```

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

**The correlation coeeficient = 0.3228349. This is reflected in the scatter plot because the data certainly does not follow the line of best fit and there is a very weak correlation between the two variables. The only basis someone would have to have to assume there'd be a correlation is that those who are more educated are smart enough to not commit crimes. In this case, they are assuming intellect also signifies morality, which is not always the case. Also, there's no accounting of potentially unreported cases that would be committed by highly intellectual criminals. I think there are too many factors in play, such as access to schooling, racial bias, unreported cases, etc for their to be a strong correlation between the two variables.**

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).
```

```
crime.lm<- lm(R ~ Ed, data = dat.crime)
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.)

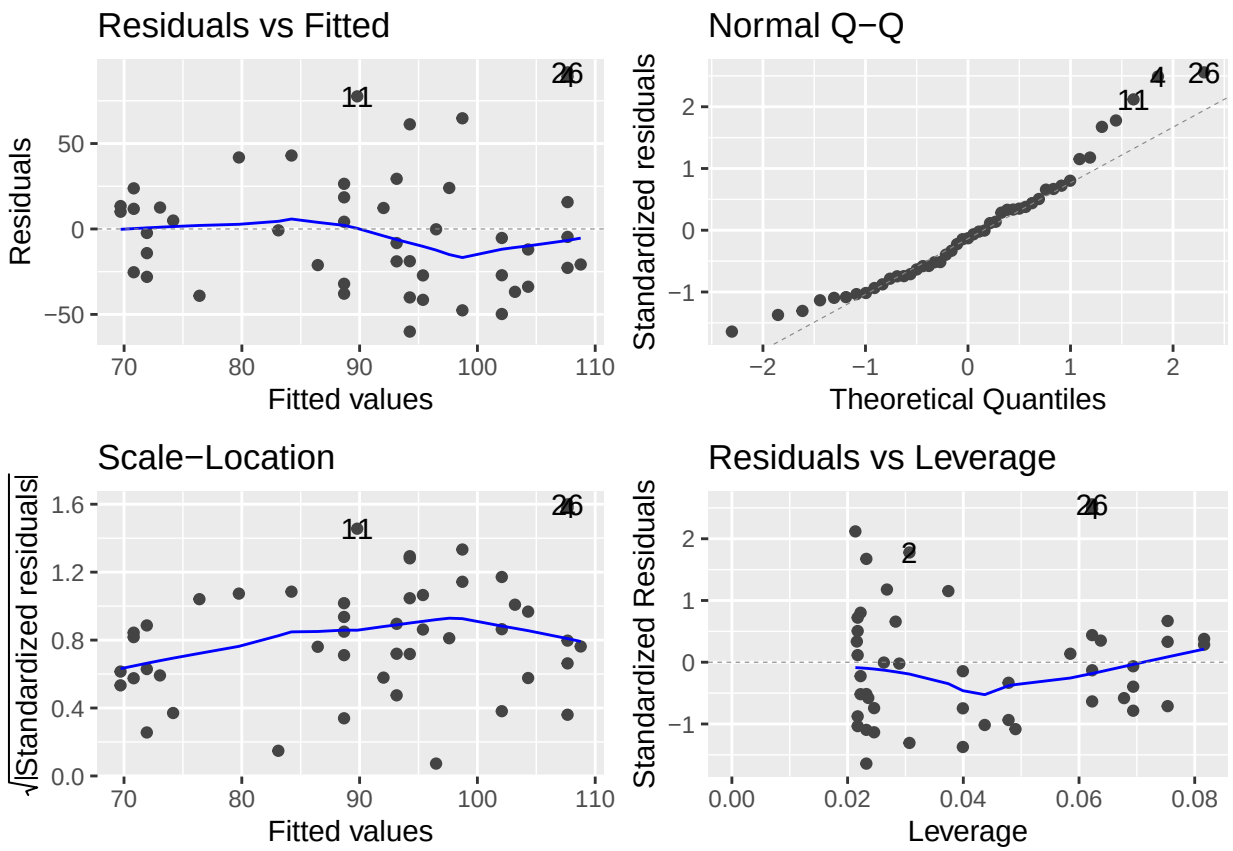
```
install.packages("ggplot2",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/nh/m2s_dxnj099clcww6tpc2vy00000gn/T//Rtmpa8F00e/downloaded_packages
```

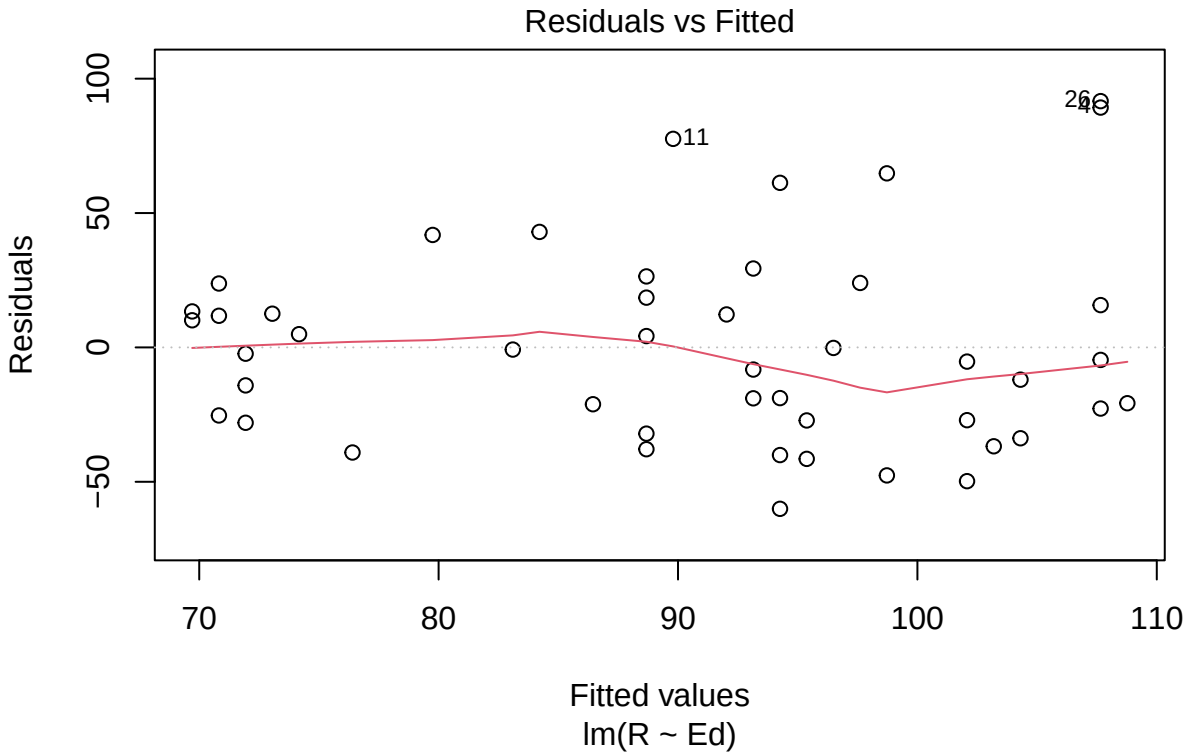
```
install.packages("ggfortify",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/nh/m2s_dxnj099clcww6tpc2vy00000gn/T//Rtmpa8F00e/downloaded_packages
```

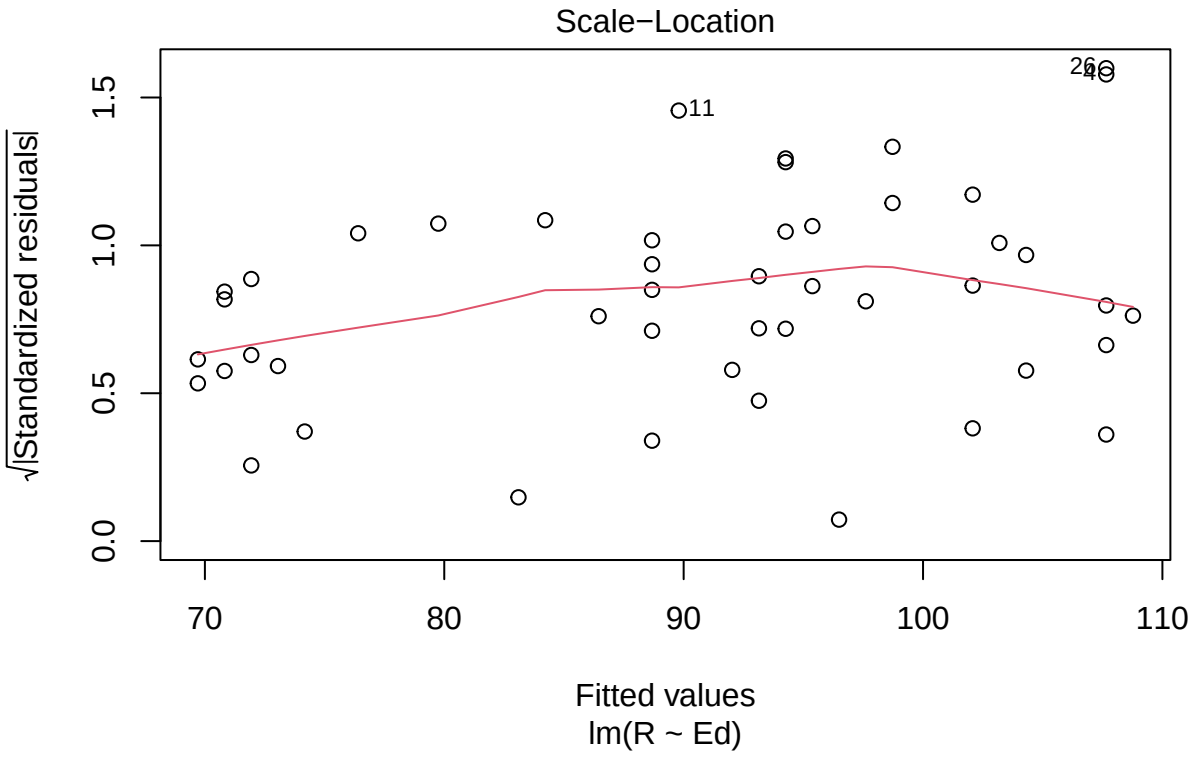
```
library (ggplot2)
library(ggfortify)
autoplot(crime.lm)
```



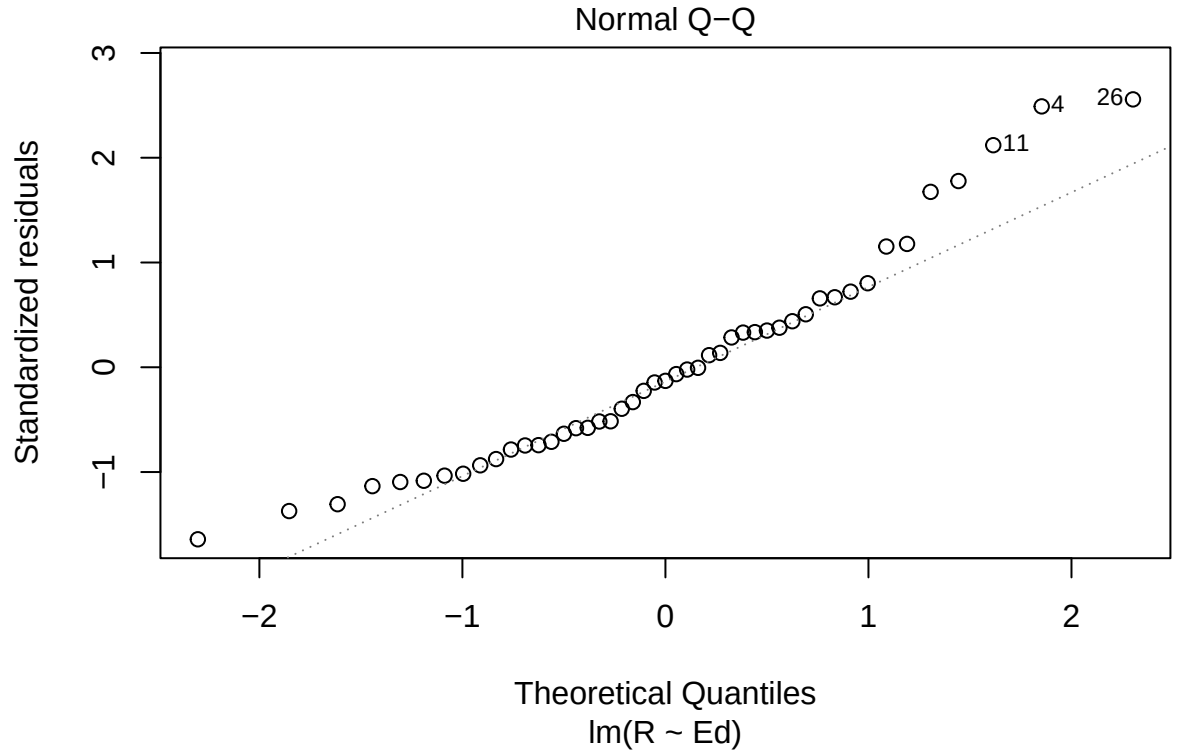

```
plot(crime.lm, 1)
```



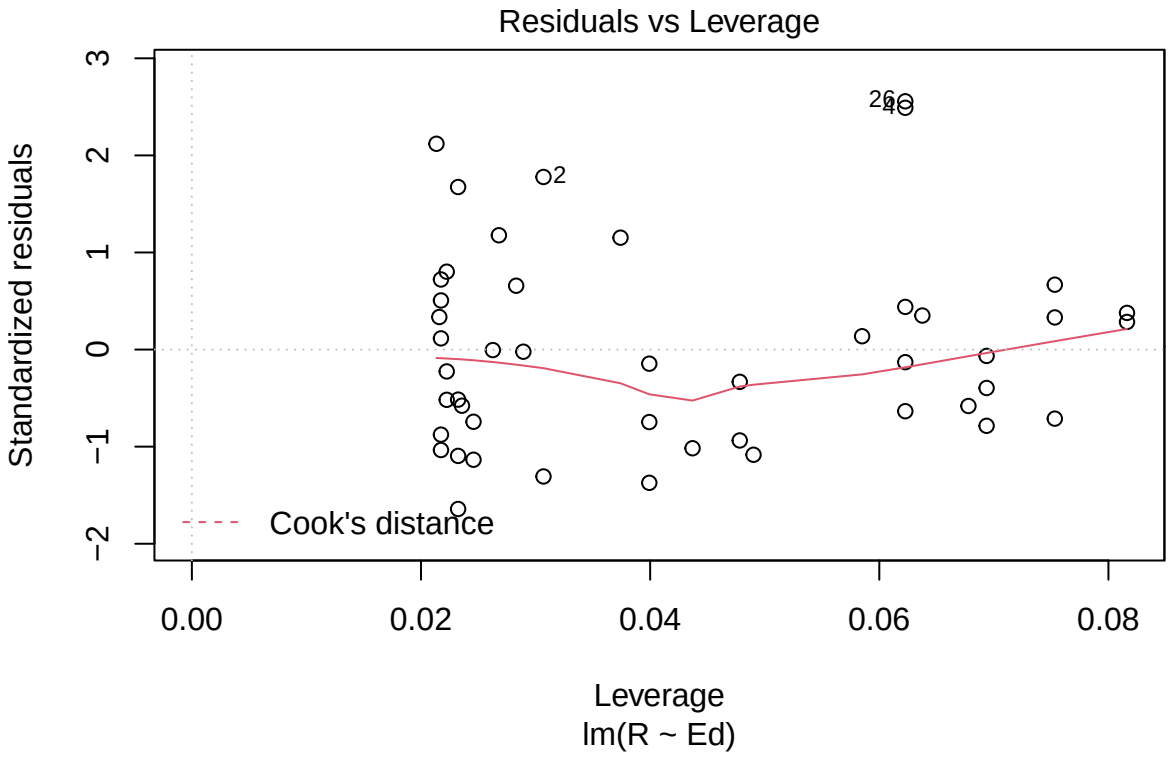
```
plot(crime.lm, 3)
```




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




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




**The assumption of linearity is satisfied because the residual plot shows no fitted pattern, as in the red line is approximately horizontal at zero. The assumption of homoscedasticity is not satisfied because there isn't a horizontal line with equally spread points since the variability of the residual points increases for a bit of time then decreases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or heteroscedasticity).The assumption of normality is not satisfied because all the data points do not fall on the reference line. The assumption of independence is not satisfied because the slight patterns shown indicates a linear relationship between the predictors and the outcome variable...**

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 relationship between reported crime and average education is not statistically significant. In fact, neither the assumption of homoscedasticity, normality, nor independence is satisfied, and so we cannot rely on the linear model to make conclusions about the data. The estimated coefficient of the slope = 1.1161, the standard error=0.4878 (residual standard error =37.01 on 45 degrees of freedom) and the p value =0.02688. For the relationship to be statistically significant, we would have a better chance of being right in finding that a relationship exists between two variables. In other words, the probability of being wrong is small.**

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, reported crime rate increases 1.1161 (per million) per state. The slope which determines this answer was calculated in the problem above.**

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

**The data does not give us any information to answer this question, especially because of the lack of correlation between the variables. However, based on other studies and their statistical findings, I would say for individuals who received more education, their crimes would be**

**reported less. They are also not likely to be committing crimes that would be detected by the everyday, typical officer. They'd also be more likely to commit the crime in a more methodical and rational way so as not to get caught..**