lab0

9/1/2020

- Loading USArrests data

```
data("USArrests")
```

- Loading coords

```
fileName <- 'data/stateCoord.txt'
coords_txt <- readChar(fileName, file.info(fileName)$size)

.extract_city <- function(s){
  city <- strsplit(s, ' ')[[1]][1]
  city <- gsub('-', ' ', city)
  return(city)
}

.extract_long <- function(s){
  str_split <- strsplit(s, ' ')[[1]]

  return(as.numeric(str_split[length(str_split)-1]))
}

.extract_lan <- function(s){
  str_split <- strsplit(s, ' ')[[1]]

  return(as.numeric(str_split[length(str_split)]))
}


lines <- strsplit(coords_txt, '\n')[[1]]


cities <- unlist(lapply(lines[2:length(lines)], FUN = .extract_city))
long <- unlist(lapply(lines[2:length(lines)], FUN = .extract_long))
lan <- unlist(lapply(lines[2:length(lines)], FUN = .extract_lan))

coords = data.frame(long=long, lan=lan)
rownames(coords) = cities
```

Manipulating the data

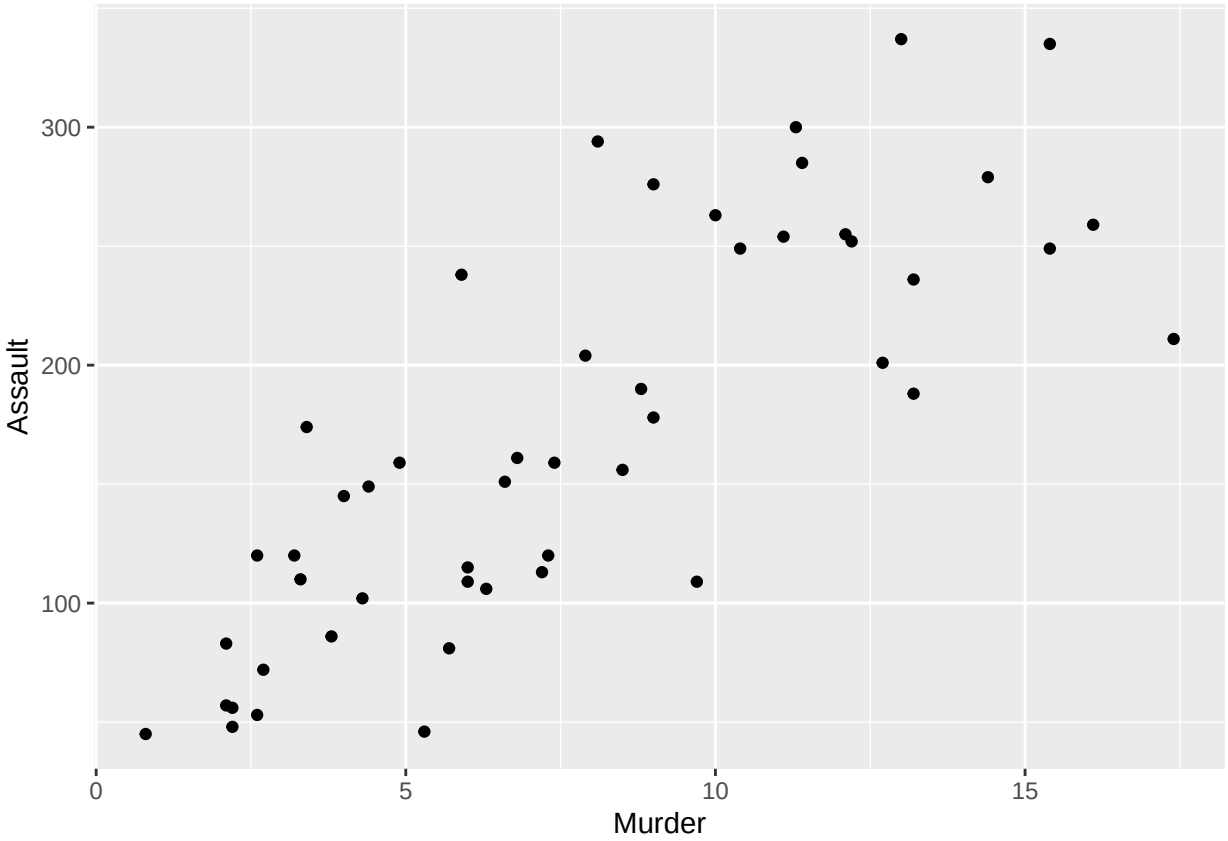
- Merging datasets

```
arr <- tibble::rownames_to_column(USArrests, "region")
coo <-  tibble::rownames_to_column(coords, "region")


arrests = dplyr::full_join(arr, coo,by="region")
```

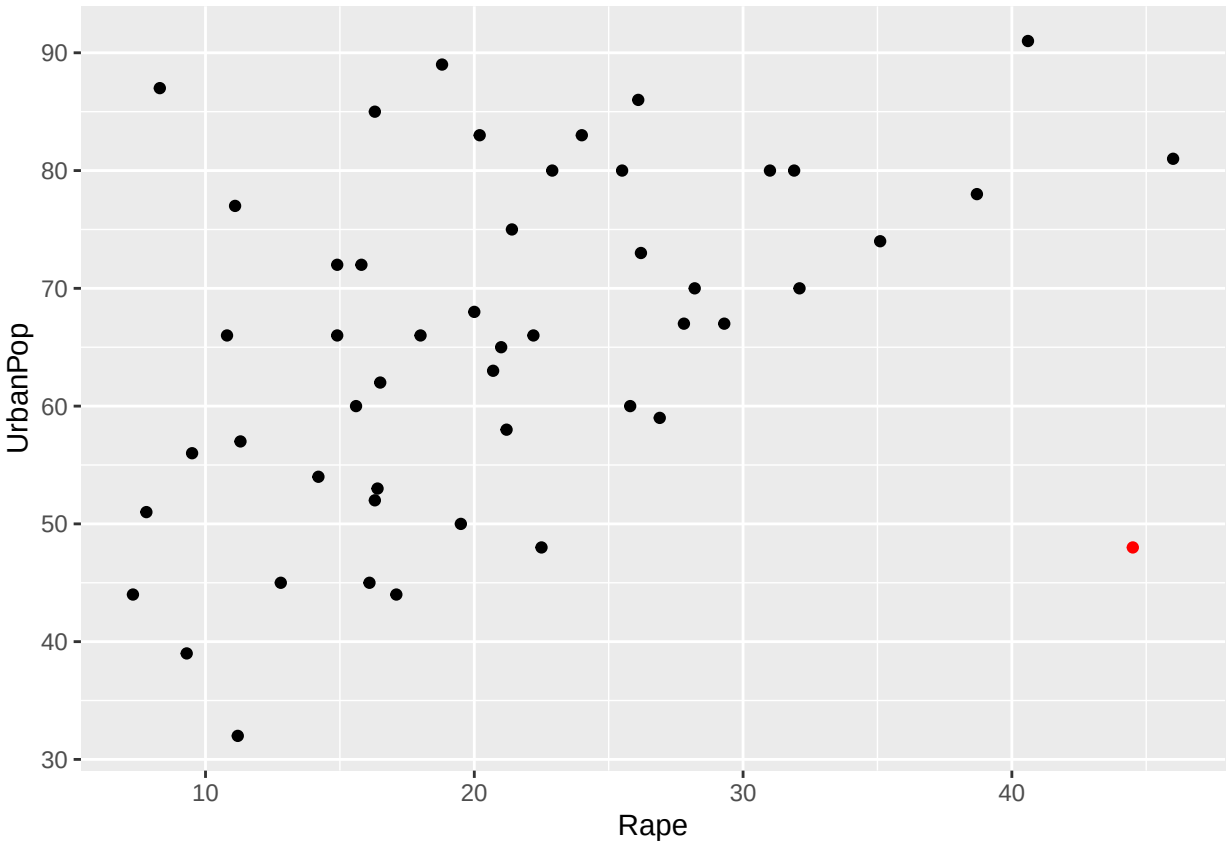
Visualizing the data

- Murder vs Assault

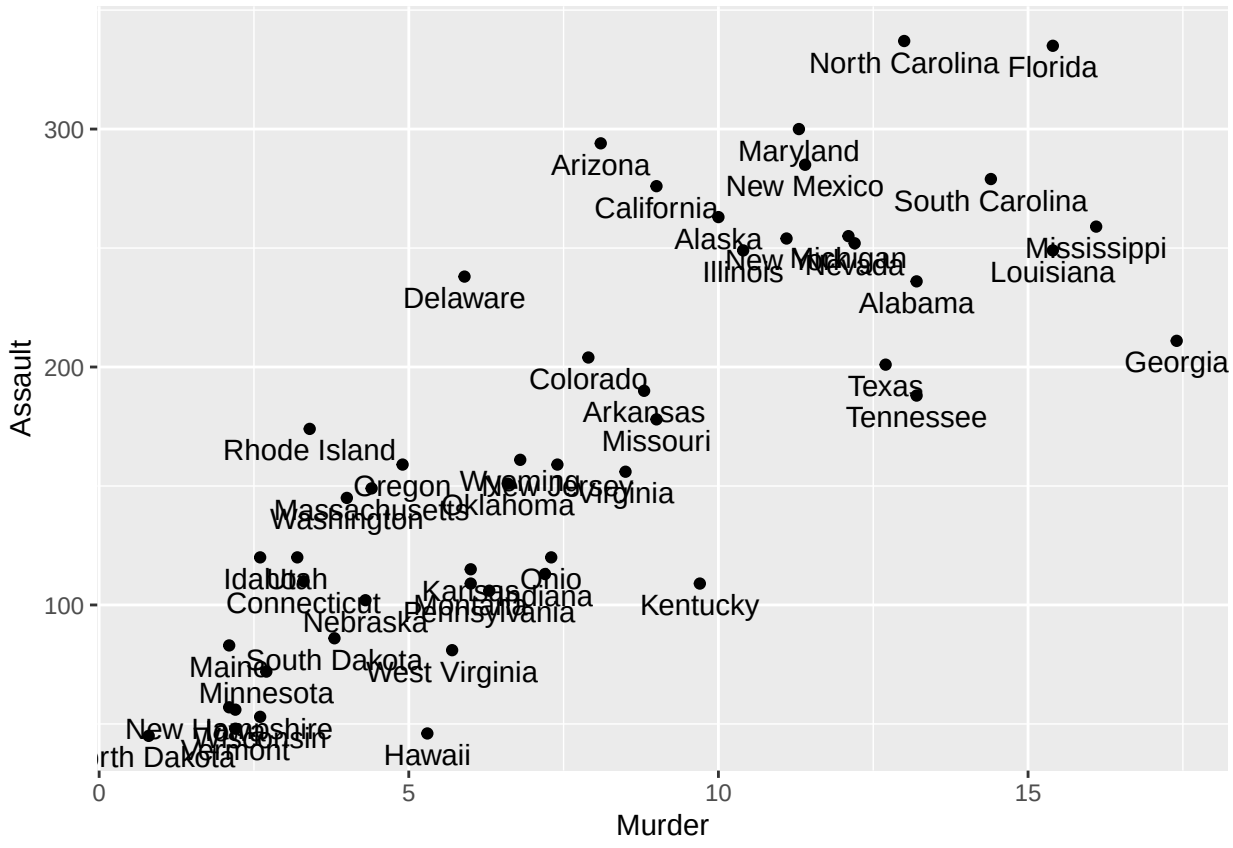


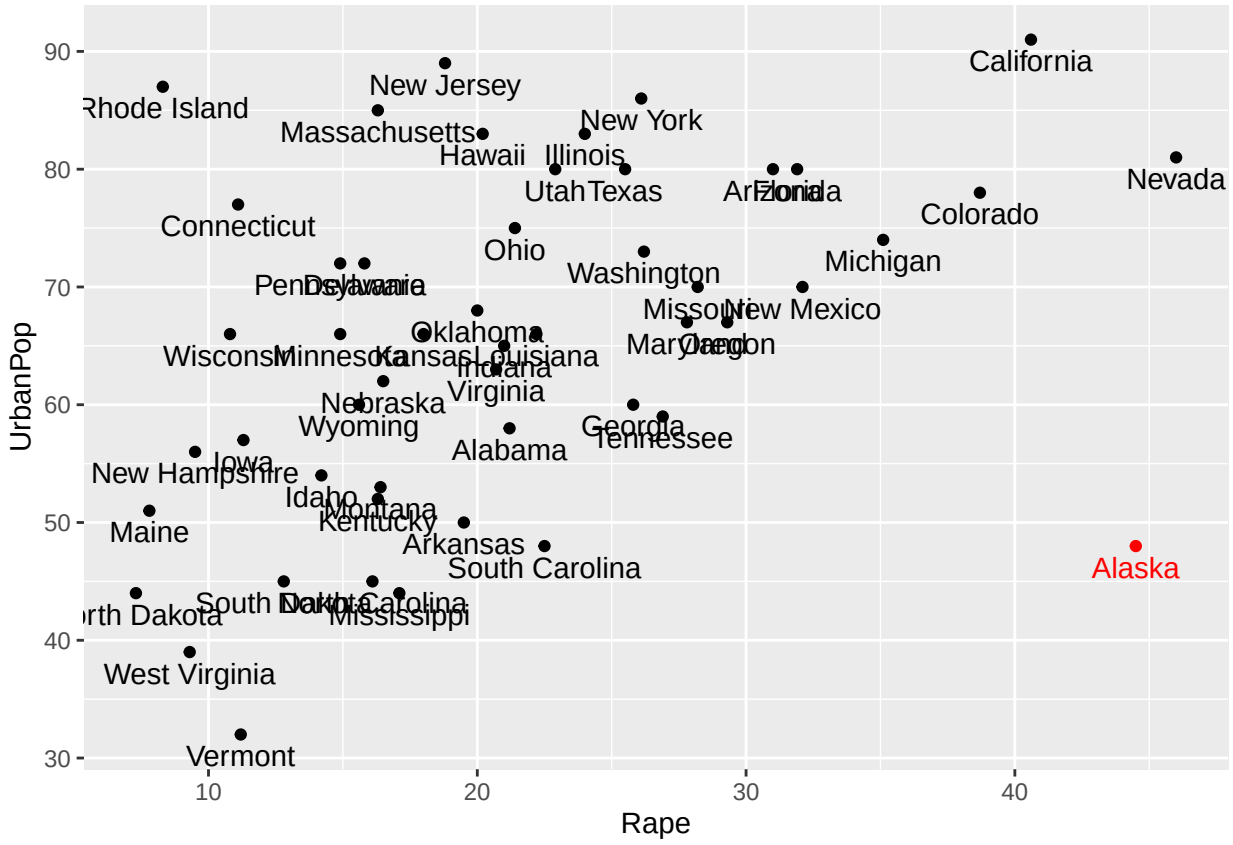
It seems like there is a linear connection between murder and assault, across states.

- Rape vs urban population



- Now with names



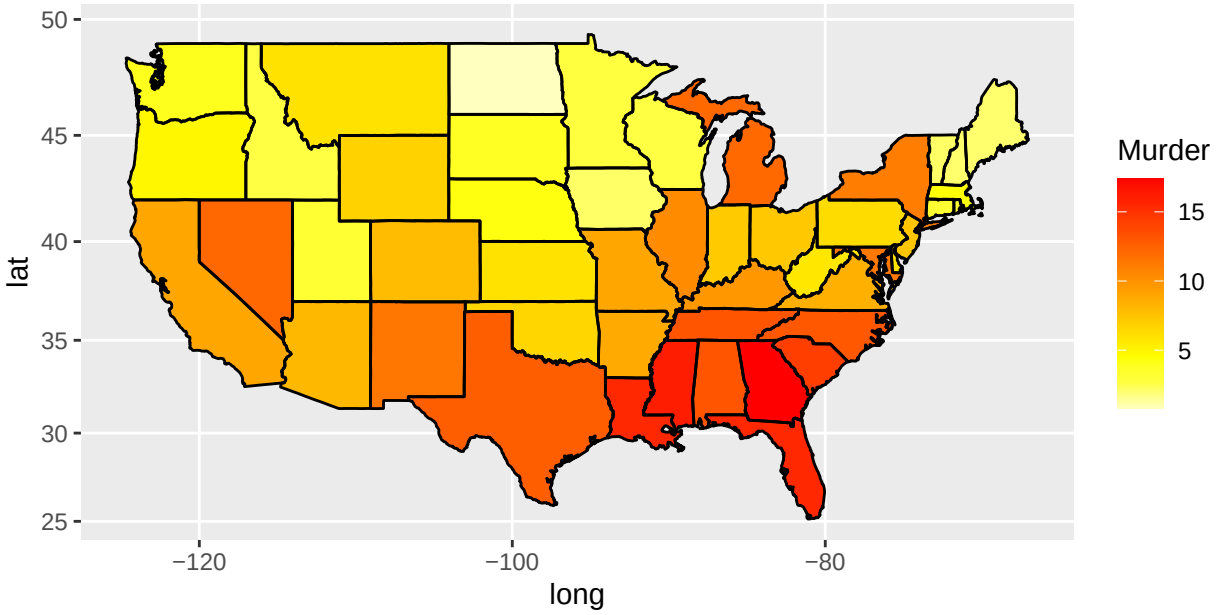


Alaska is the outlier, the rape rate in alaska in extremly high, at first I thought this was a mistake in the

data but looking online (https://edition.cnn.com/interactive/2014/02/opinion/sutter-change-alaska-rape/) it seems like a real problem.

- Challenge

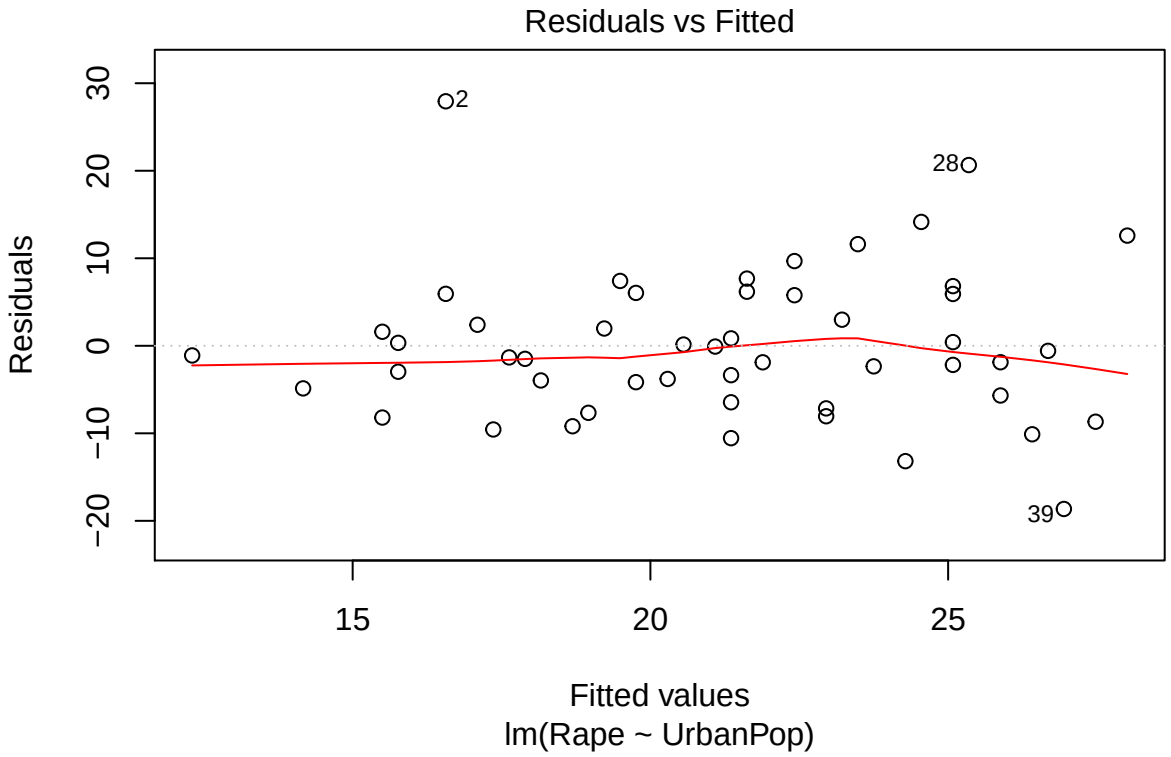
```
library(maps)
library(mapproj)
states <- map_data("state")
arsts <- arrests[c('region', 'Murder')]
arsts$region <- tolower(arsts$region)
map.df <- merge(states,arsts, by="region", all.x=T)
map.df <- map.df[order(map.df$order),]
ggplot(map.df, aes(x=long,y=lat,group=group))+
  geom_polygon(aes(fill=Murder))+
  geom_path()+
  scale_fill_gradientn(colours=rev(heat.colors(10)),na.value="grey90")+
  coord_map()
```



Regression

1-3.

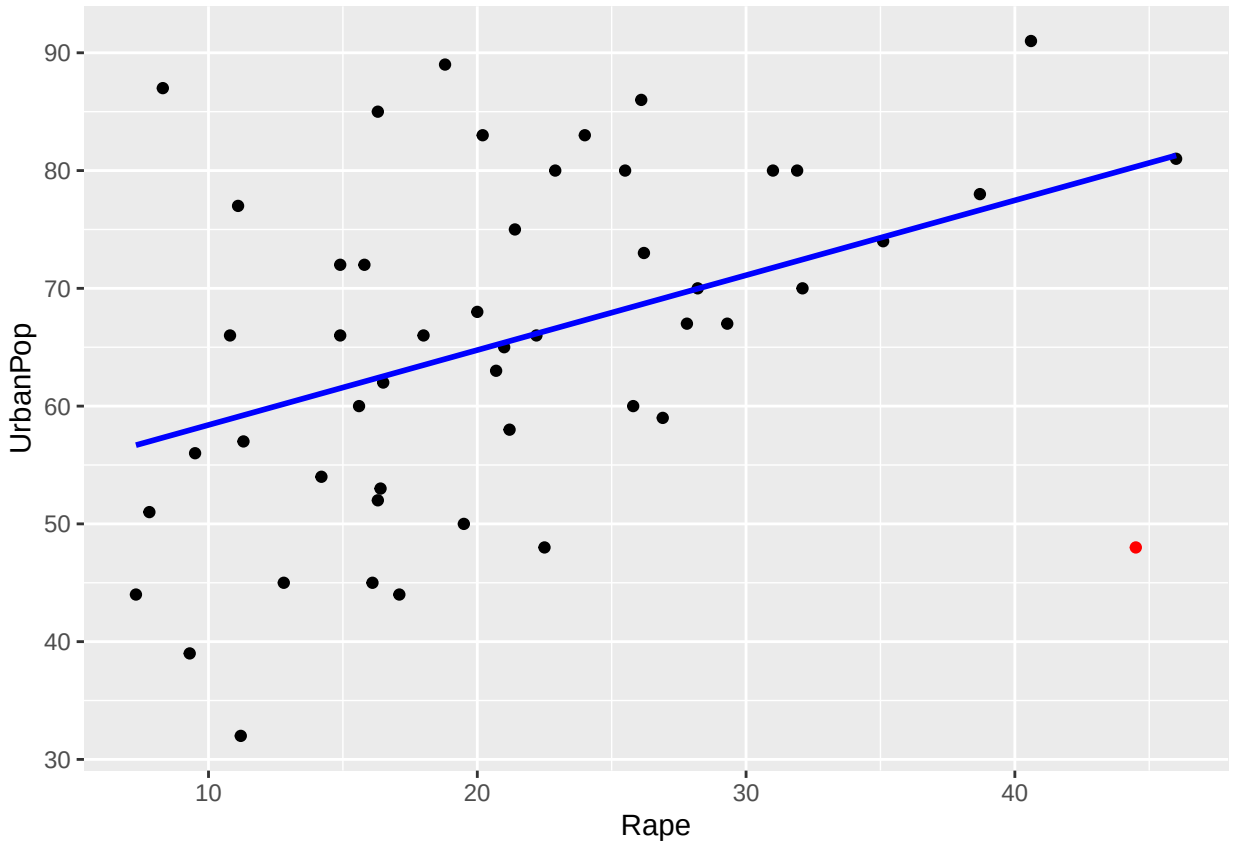
```
arrests_f <- arrests %>% dplyr::select(-Murder, -Assault)
linear_fit <-lm(Rape~UrbanPop, data = arrests_f)
plot(linear_fit,1)
```



Looking at the residuals helps us to get a sense of how our model assumptions do in reality, the first assumption is that the expected value of the residuals is 0, looking at this plot there's no clear evidence to say that this assumption is violated. Another thing to look at is homoscedasticity of the residuals, which means that the variance is the same across all population sizes. From the plot I can say that the variance of the residuals increases with the populations size (the coefficient is positive, hence we can conclude that an increase in fitted values comes from an increase population).

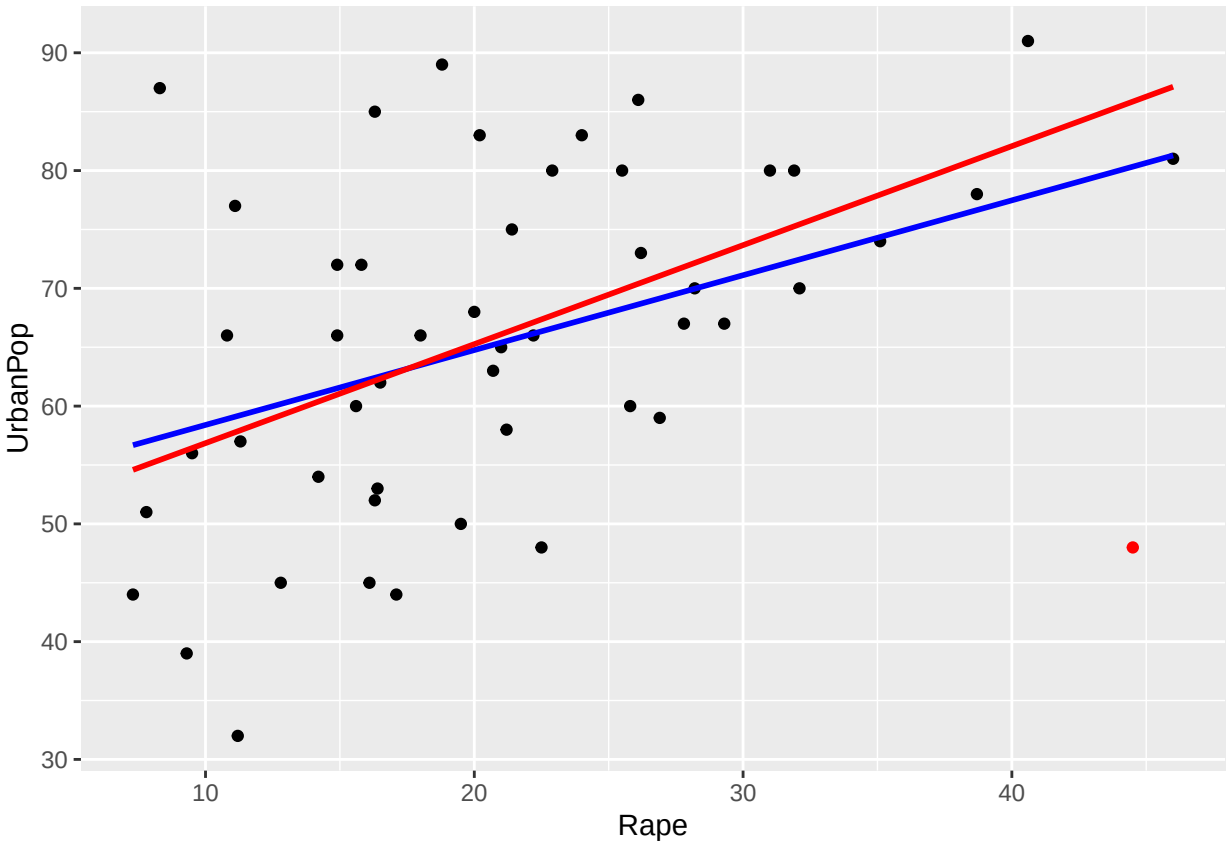
4.

```
arrests %>% ggplot(aes(Rape, UrbanPop, label=region)) +
  geom_point(color = case_when(arrests$Rape>40 &
                                 arrests$UrbanPop < 50~ "red"
                               , TRUE ~ "black")) +
 geom_smooth(method='lm', se = FALSE, colour='blue')
```

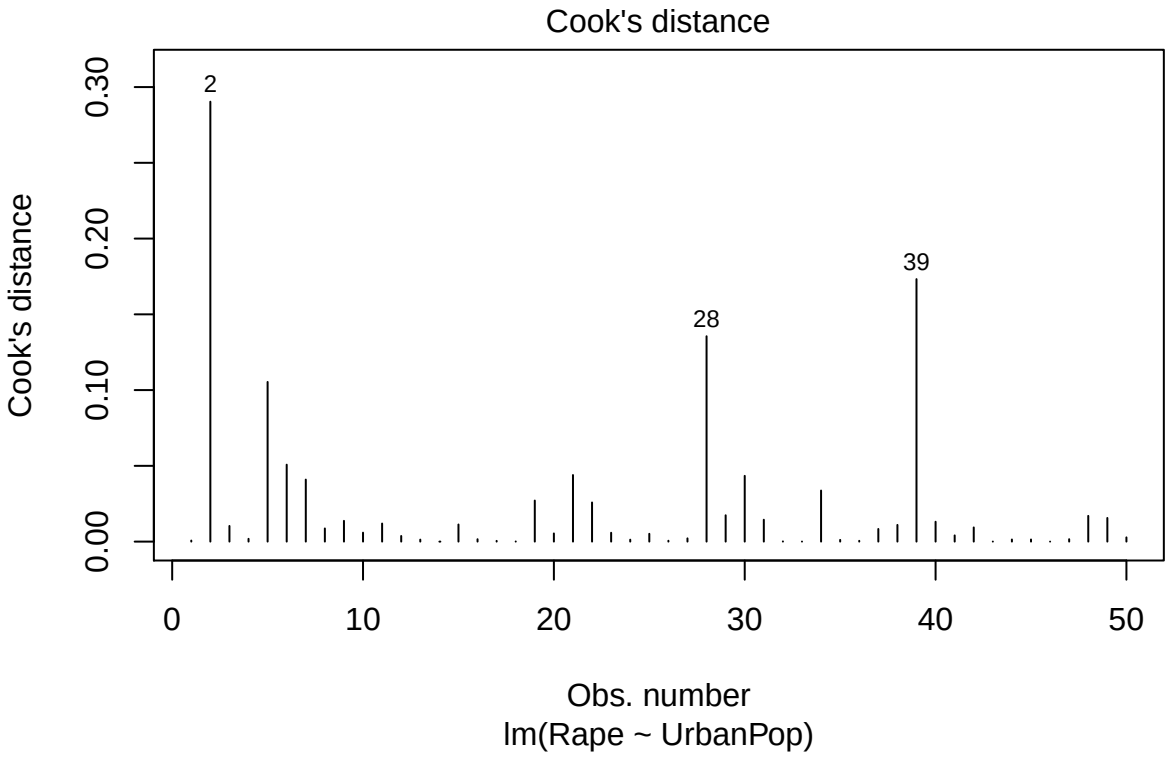


5.

```
arrests %>% ggplot(aes(Rape, UrbanPop, label=region)) +
  geom_point(color = case_when(arrests$Rape>40 &
                               arrests$UrbanPop < 50~ "red", TRUE ~ "black")) +
 geom_smooth(method='lm', se = FALSE, colour='blue') +
 geom_smooth(data=arrests[arrests$region!='Alaska',],method='lm',
             se = FALSE, colour='red')
```

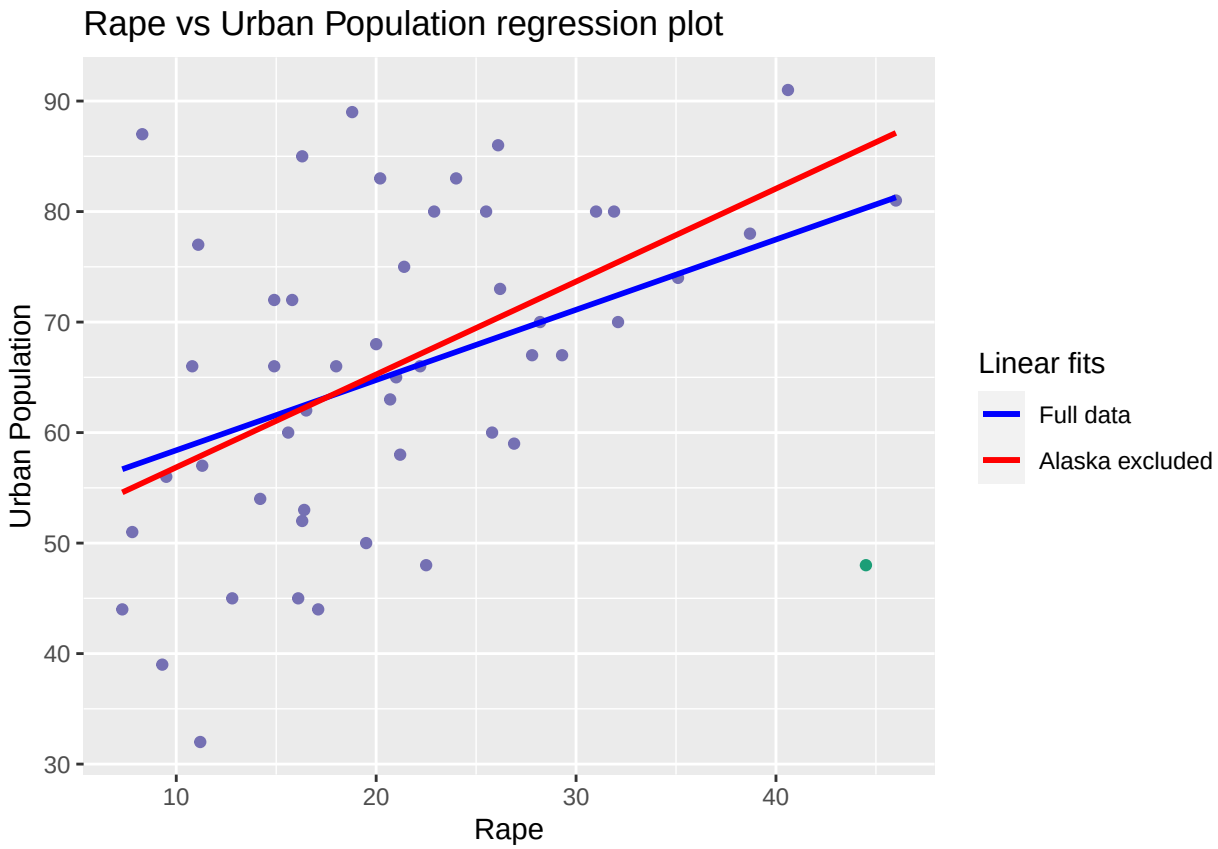


- I think that most of the variation in the data isn't explained by any of the linear models. What I get looking at the two models is that both are saying that there is a positive correlation between population size and rape.
- With regard to Alaska (data point 2) we can observe the effect of observation with high cook's distance have on our regression line, shifting it significantly down.



7.

```
arrests %>% ggplot(aes(Rape, UrbanPop, label=region)) +
  geom_point(color = case_when(arrests$Rape>40 &
                                 arrests$UrbanPop < 50~ "#1b9e77", TRUE ~ "#7570b3")) +
 geom_smooth(method='lm', se = FALSE, aes(color = "Full data")) +
 geom_smooth(data=arrests[arrests$region!='Alaska',],method='lm',
             se = FALSE, aes(color = "Alaska excluded")) +
  ggtitle('Rape vs Urban Population regression plot')+
  xlab('Rape')+
  ylab('Urban Population')+
     scale_color_manual(name = "Linear fits",
                        breaks = c("Full data", "Alaska excluded"),
                        values = c("Full data" = "blue", "Alaska excluded" = "red") )
```


Rape vs Urban Population regression plot
Urban Population
Rape
90
80
70
60
50
40
30
10
20
30
40
Linear fits
Full data
Alaska excluded