Project 3

Chloe Chen & Kangrui Liu

9/19/2023

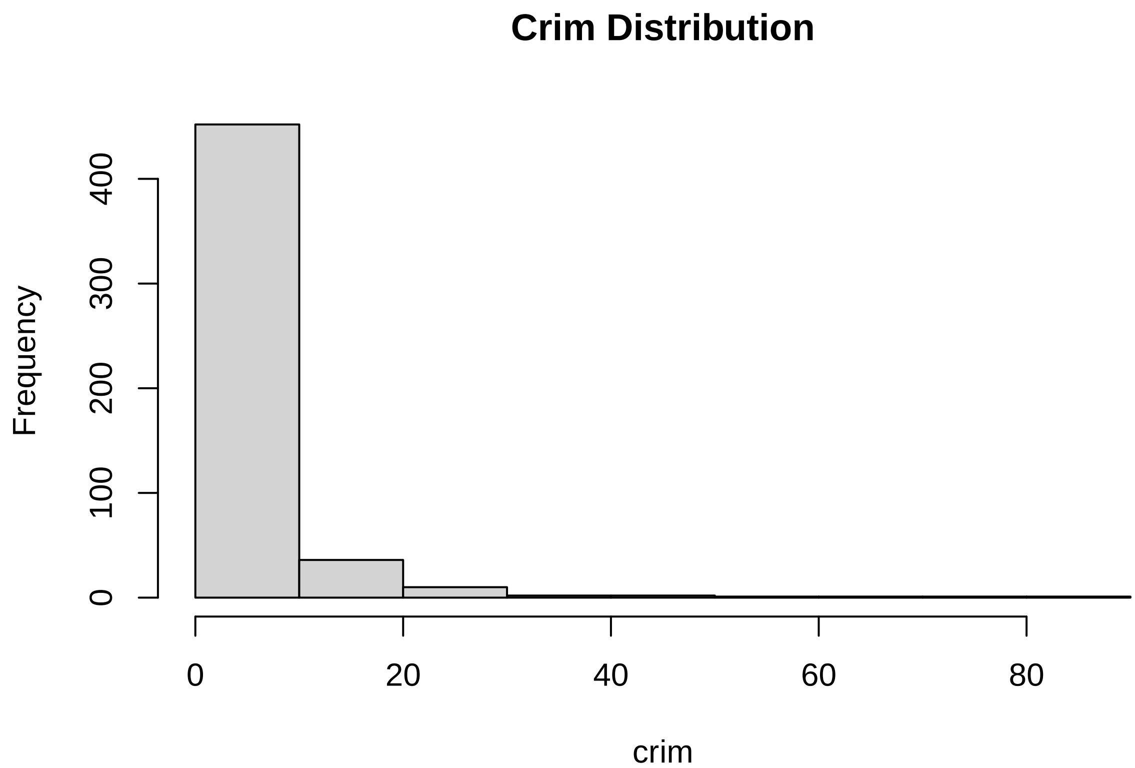
1. JWHT Chapter 2. Modified Exercise 10.

- This exercise involves the Boston housing data set. Assume that we are interested in per capita crime rate, crim. ## A. Examine crim with summary() and in a histogram.

```
library(MASS)
data("Boston")
summary(Boston$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```
hist(Boston$crim, main = "Crim Distribution", xlab = "crim", ylab = "Frequency")
```



B. Focus on suburbs with the crime rate above 25.

How many suburbs fall into this group? What are the pupil-teacher ratios like in those suburbs? How about property tax rates? How about median home values? How do the pupil-teacher ratios, property tax rates and median home values compare between these suburbs and the remaining suburbs?

```
highcrim_sub <- subset(Boston, Boston$crim > 25)
lowcrim_sub <- subset(Boston, Boston$crim <= 25)
n_highcrim_sub <- dim(highcrim_sub)
n_highcrim_sub
```

```
## [1] 11 14
```

```
summary1 <- summary(highcrim_sub[c("ptratio", "tax", "medv")])
summary2 <- summary(lowcrim_sub[c("ptratio", "tax", "medv")])
print("Summary for highcrim_sub")
```

```
## [1] "Summary for highcrim_sub"
```

```r
print(summary1)
```

```
##     ptratio          tax            medv       
##  Min.   :20.2   Min.   :666   Min.   : 5.000  
##  1st Qu.:20.2   1st Qu.:666   1st Qu.: 6.300  
##  Median :20.2   Median :666   Median : 8.800  
##  Mean   :20.2   Mean   :666   Mean   : 9.355  
##  3rd Qu.:20.2   3rd Qu.:666   3rd Qu.:10.650  
##  Max.   :20.2   Max.   :666   Max.   :16.300  
```

```r
print("Summary for lowcrim_sub")
```

```
## [1] "Summary for lowcrim_sub"
```

```r
print(summary2)
```

```
##     ptratio           tax             medv      
##  Min.   :12.60   Min.   :187.0   Min.   : 6.30  
##  1st Qu.:17.00   1st Qu.:278.0   1st Qu.:17.40  
##  Median :18.90   Median :330.0   Median :21.40  
##  Mean   :18.42   Mean   :402.5   Mean   :22.83  
##  3rd Qu.:20.20   3rd Qu.:666.0   3rd Qu.:25.05  
##  Max.   :22.00   Max.   :711.0   Max.   :50.00  
```

```r
# Use t-test to further compare
ttest_ptratio <- t.test(highcrim_sub$ptratio, lowcrim_sub$ptratio)
ttest_ptratio
```

```
## 
##  Welch Two Sample t-test
## 
## data:  highcrim_sub$ptratio and lowcrim_sub$ptratio
## t = 18.258, df = 494, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.591331 1.975134
## sample estimates:
## mean of x mean of y 
##  20.20000  18.41677 
```

```r
ttest_tax <- t.test(highcrim_sub$tax, lowcrim_sub$tax)
ttest_tax
```

```
##
##  Welch Two Sample t-test
##
## data:  highcrim_sub$tax and lowcrim_sub$tax
## t = 35.335, df = 494, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  248.8397 278.1421
## sample estimates:
## mean of x mean of y
##  666.0000  402.5091
```

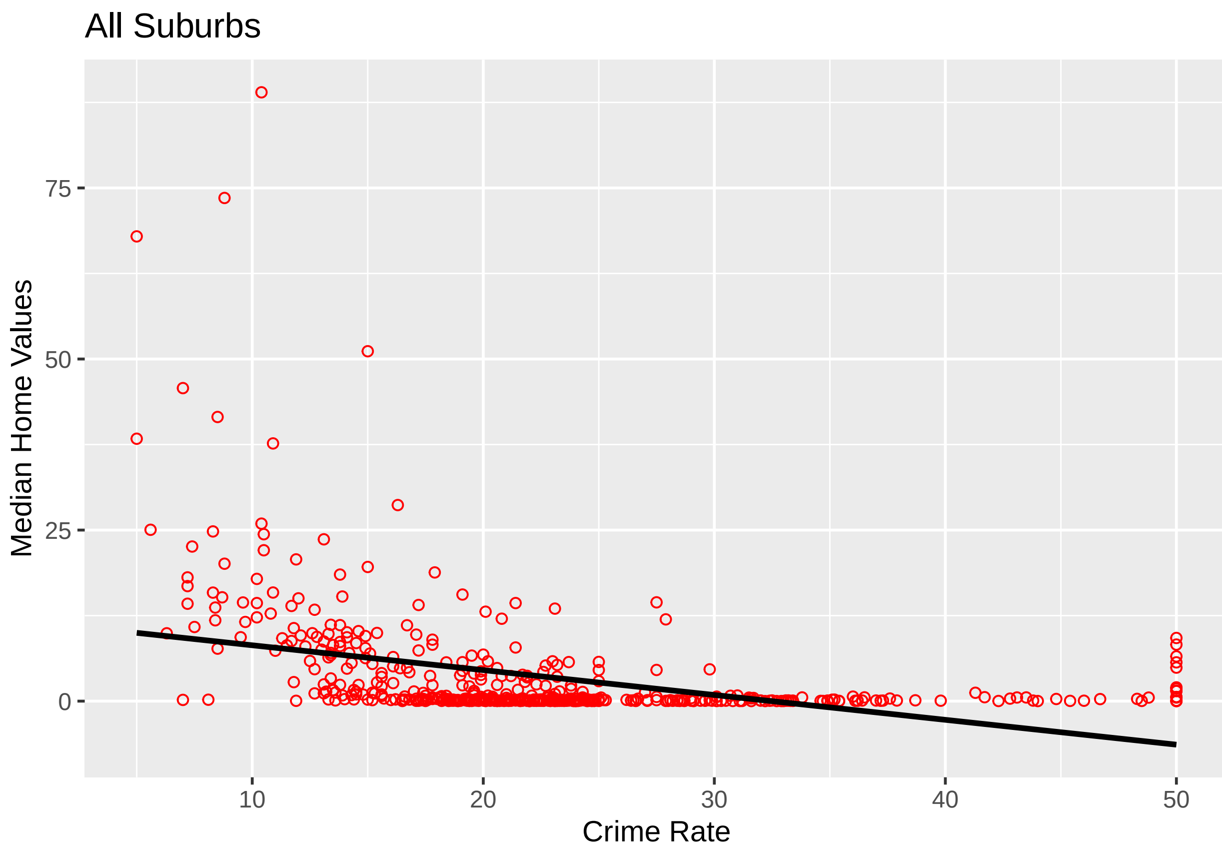
```r
ttest_medv <- t.test(highcrim_sub$medv, lowcrim_sub$medv)
ttest_medv
```

```
##
##  Welch Two Sample t-test
##
## data:  highcrim_sub$medv and lowcrim_sub$medv
## t = -11.116, df = 12.709, p-value = 6.498e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.09539 -10.84683
## sample estimates:
## mean of x mean of y
##  9.354545 22.825657
```

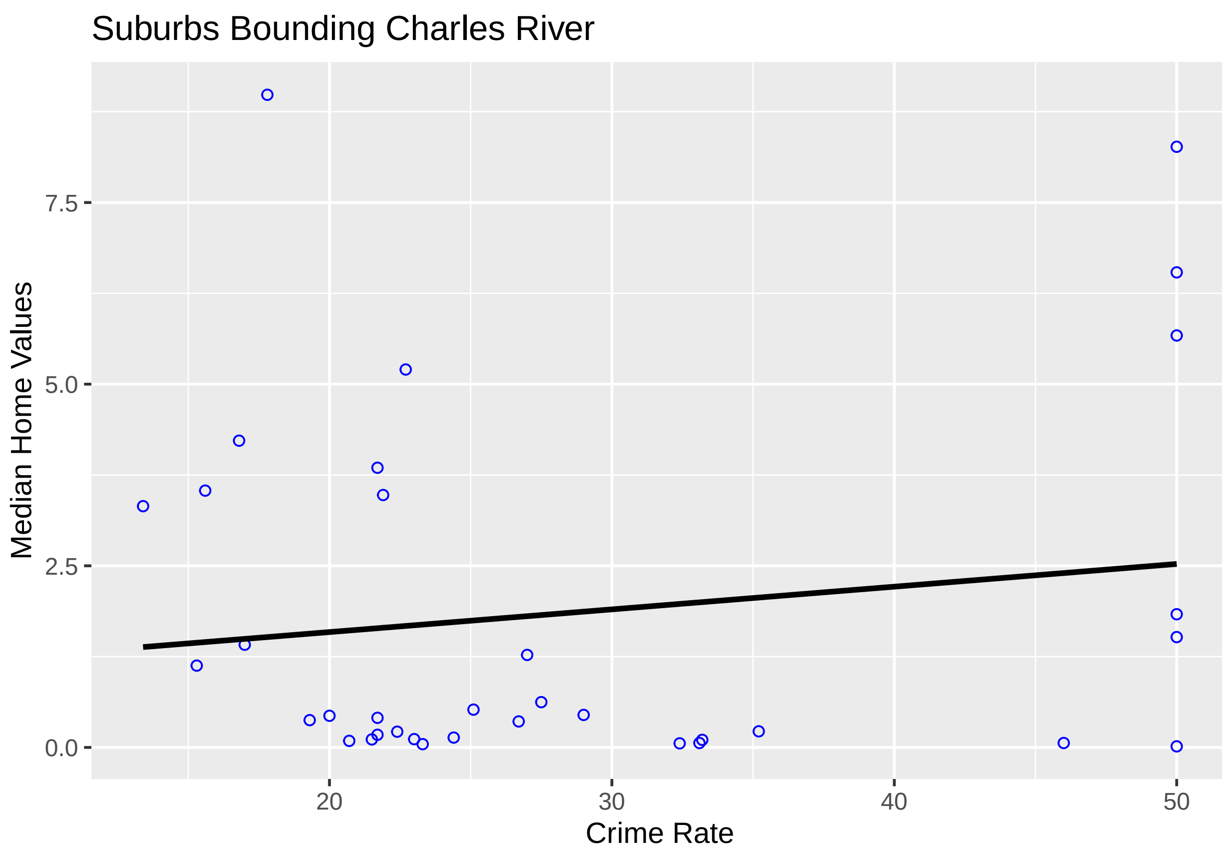
- There are 11 suburbs with the crime rate above 25.
- The ptratio and tax are evenly distributed in high crime suburbs. As we can see from the summary, the mean and median for ptratio are both 20.2, and mean and median for tax are both 666.
- For medv in high crime suburbs, it has s mean of 9.355 and a median of 8.8.
- We compared the three variables between two areas using *Welch Two Sample t-test*. The three variables are all significanntly different from each other at a 95% confidence level.

C. Create a scatter plot of the crime rates and the median home values for 1) all suburbs, 2) suburbs bounding Charles River, and 3) suburbs not bounding Charles River. What do you observe?

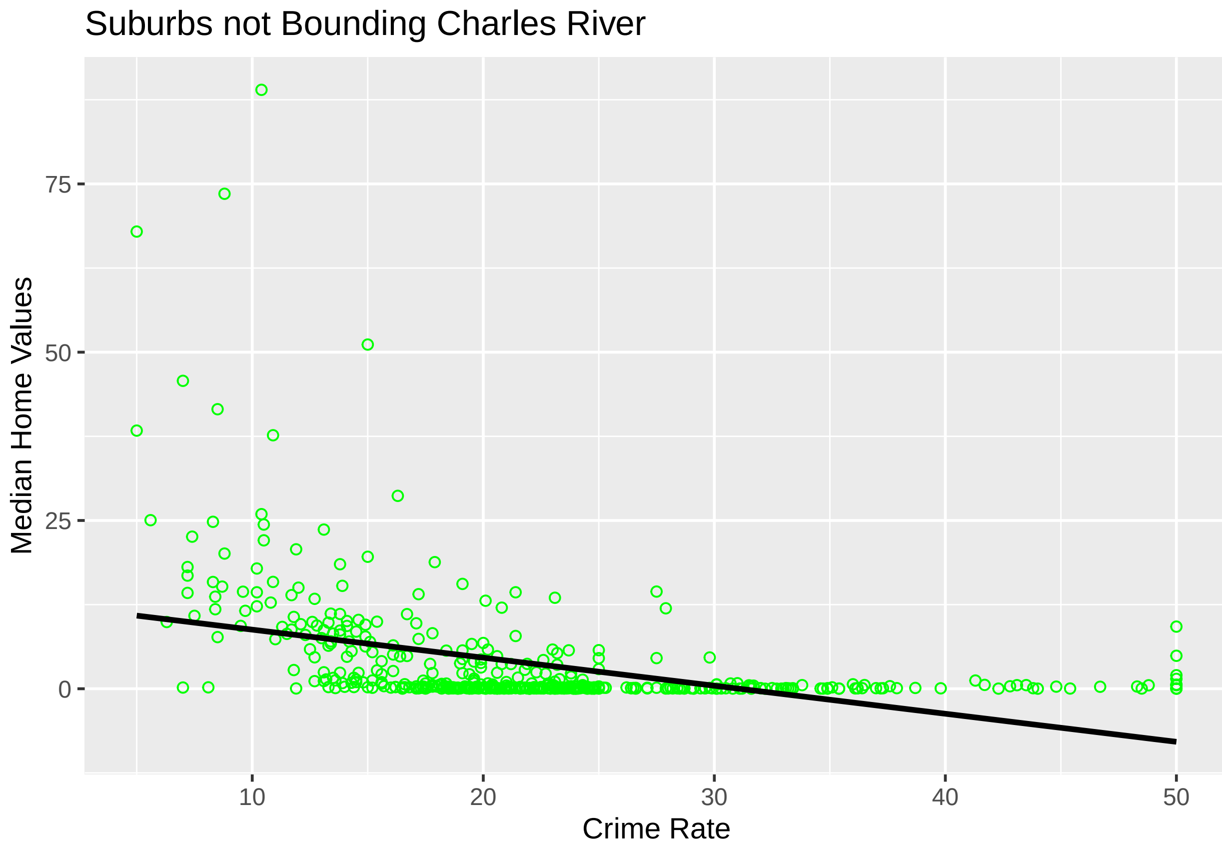
```
library(ggplot2)
ggplot(Boston, aes(y=crim, x=medv))+
  geom_point(shape=1, color="red")+
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title="All Suburbs",
       y ="Median Home Values", x = "Crime Rate")
```




```
ggplot(subset(Boston, chas == 1), aes(y=crim, x=medv))+
  geom_point(shape=1, color="blue")+
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title="Suburbs Bounding Charles River",
       y ="Median Home Values", x = "Crime Rate")
```




```
ggplot(subset(Boston, chas == 0), aes(y=crim, x=medv))+
  geom_point(shape=1, color="green")+
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title="Suburbs not Bounding Charles River",
        y ="Median Home Values", x = "Crime Rate")
```



- For all suburbs and these not bounding Charles River, the crime rates and the median home values seems to be negatively correlated with each other
- The relations between these two variables are hard to be observed in the suburbs bounding Charles River as the plots are more scattered, this is partly because the highest crime rate observed in this area is quite low(at 10).

D. Analyze the crime rates as a function of median home values in a simple linear regression with an intercept.

Report what the regression coefficients mean in lay terms.

```
model_obj_1<-lm(crim~medv,Boston)
summary(model_obj_1)
```

```
##
## Call:
## lm(formula = crim ~ medv, data = Boston)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
coef(model_obj_1)
```

```
## (Intercept)        medv
##  11.7965358  -0.3631599
```

- The intercept 11.79654: Can be interpreted as saying a suburb with a median home value of 0 has a mean expected per capita crime rate of 11.79654.
- The regression coefficient for medv -0.36316: Median home values and per capita crime rates are negatively correlated, on average, with per capita crime rates decreasing by 0.36316 if the median home value increase by $10,000.

E. Calculate the coefficients reported in D as well as their standard errors by hand.

```r
s_XY<-cov(Boston$medv,Boston$crim)
s_XX<-var(Boston$medv)
SS_XY<-sum((Boston$medv-mean(Boston$medv))*(Boston$crim-mean(Boston$crim)))
SS_X<-sum((Boston$medv-mean(Boston$medv))^2)

beta1<-s_XY/s_XX
beta0<-mean(Boston$crim)-beta1*mean(Boston$medv)
beta0;beta1
```

```
## [1] 11.79654
```

```
## [1] -0.3631599
```

```r
Boston$h_crim_medv<-beta0+beta1*Boston$medv
Boston$residual_medv<-Boston$crim-Boston$h_crim_medv #calculate the residual for each p

#standard error for beta1
h_sigma_sq_medv<-sum(Boston$residual_medv^2)/(dim(Boston)[1]-2)
h_sigma_sq_medv
```

```
## [1] 62.95551
```

```r
V_beta1<-h_sigma_sq_medv/SS_X
SE_beta1<-sqrt(V_beta1)

V_beta0<-h_sigma_sq_medv*(1/dim(Boston)[1]+mean(Boston$medv)^2/SS_X)
SE_beta0<-sqrt(V_beta0)
SE_beta0;SE_beta1
```

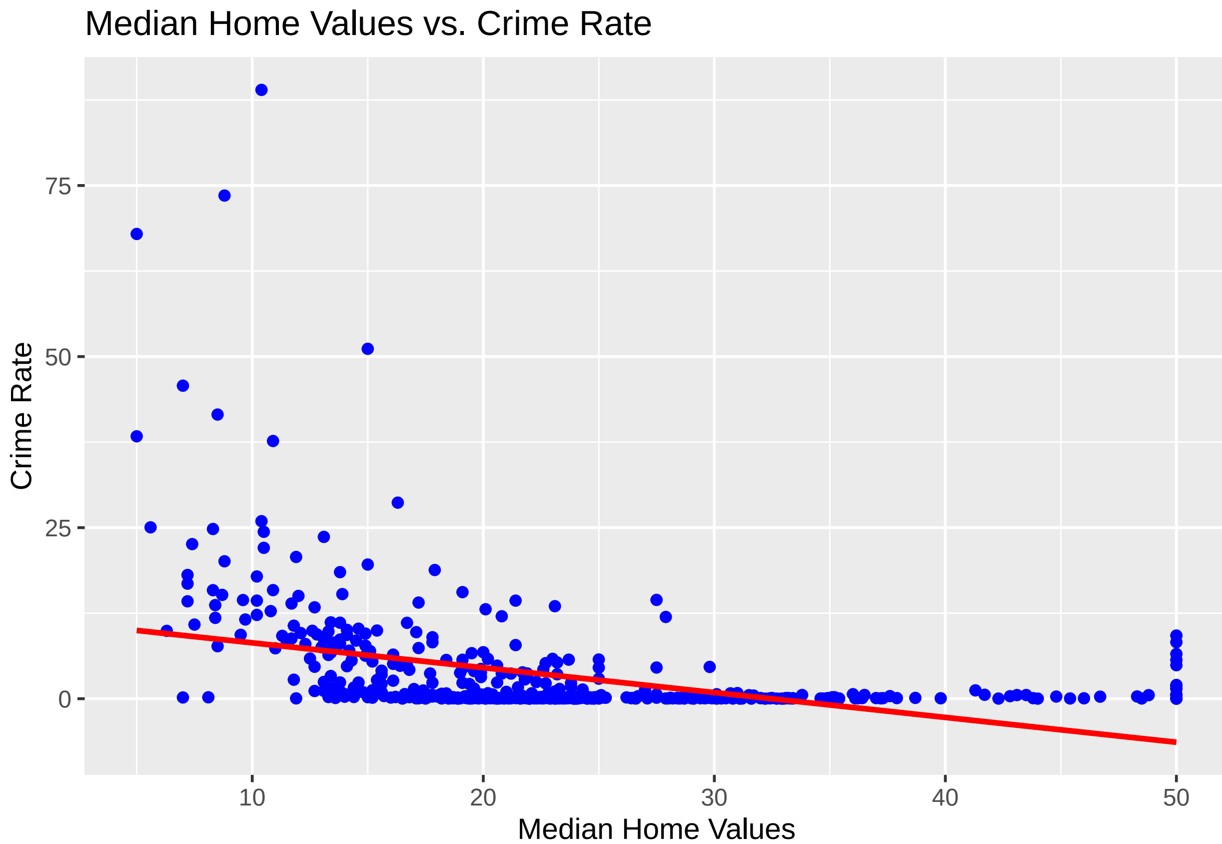
```
## [1] 0.9341892
```

```
## [1] 0.03839017
```

- Coefficients are 11.79654 and -0.3631599.
- Standard error for /beta0 is 0.9341892, standard error for /beta1 is 0.03839017.

F. Create a scatter plot of the crime rates and the median home values with a regression line. Is the regression line a good summary of the crime rates? Examine residuals to assess this.

```r
library(ggplot2)
ggplot(Boston, aes(y=crim, x=medv))+
  geom_point(color="blue")+
  geom_smooth(method='lm', color="red", se=FALSE)+
  labs(title = "Median Home Values vs. Crime Rate", x = "Median Home Values", y = "Crime
```




```r
summary(resid(model_obj_1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -9.071  -4.022  -2.343   0.000   1.298  80.957
```

```r
summary(model_obj_1)$sigma^2
```

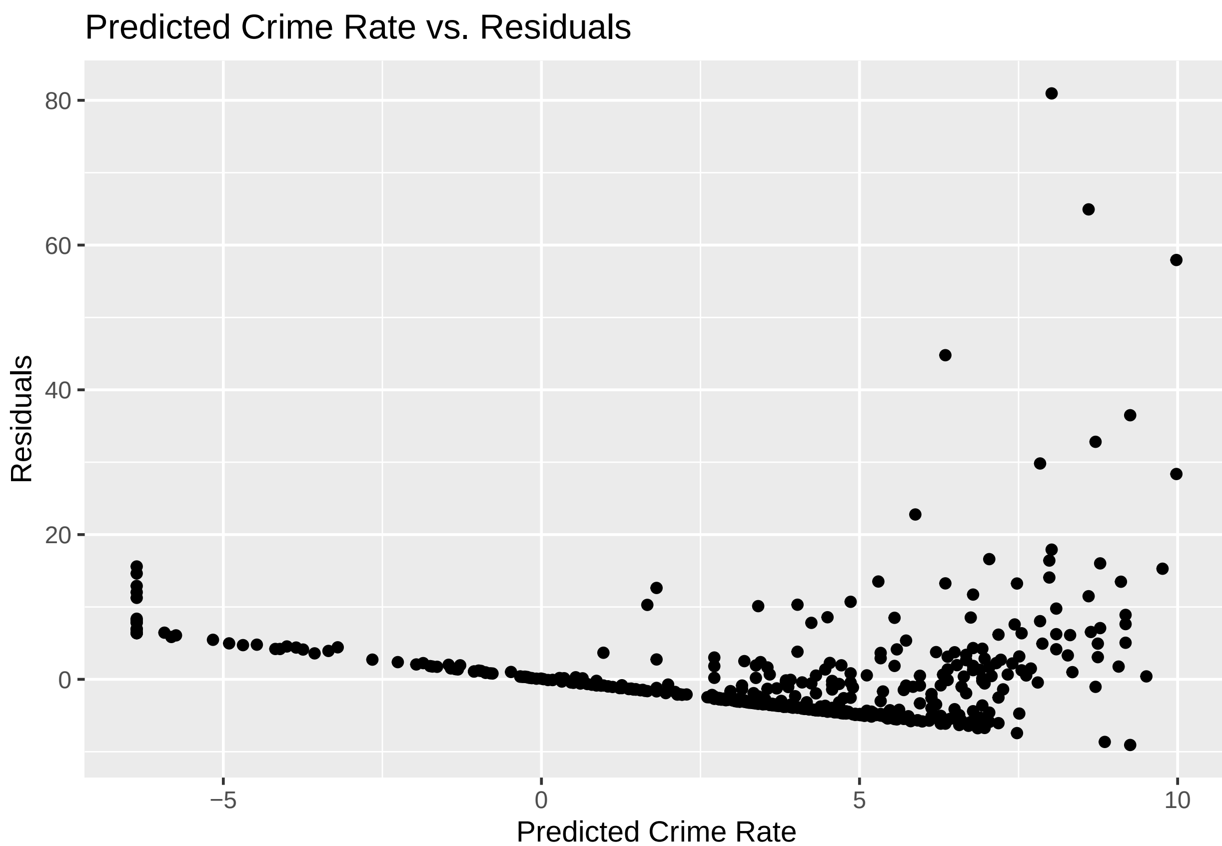
```
## [1] 62.95551
```

- The regression line might not be a good summary of the crime rates, because we expect residuals normally distribute around the regression line, which is not the case for this regression line.

G. Create a scatter plot of predicted crim and residuals. What do you observe?

```r
pred <- predict(model_obj_1)
resid <- resid(model_obj_1)

ggplot(data = Boston, aes(x = pred, y = resid)) +
  geom_point() +
  labs(title = "Predicted Crime Rate vs. Residuals", x = "Predicted Crime Rate",
       y = "Residuals")
```



- We observe a clear pattern of the points, but we expect a random scatter of residuals around the x-axis. So the model is not a good fit for our data.