Project 3

Chloe Chen & Kangrui Liu

9/19/2023

1. JWHT Chapter 2. Modified Exercise 10.

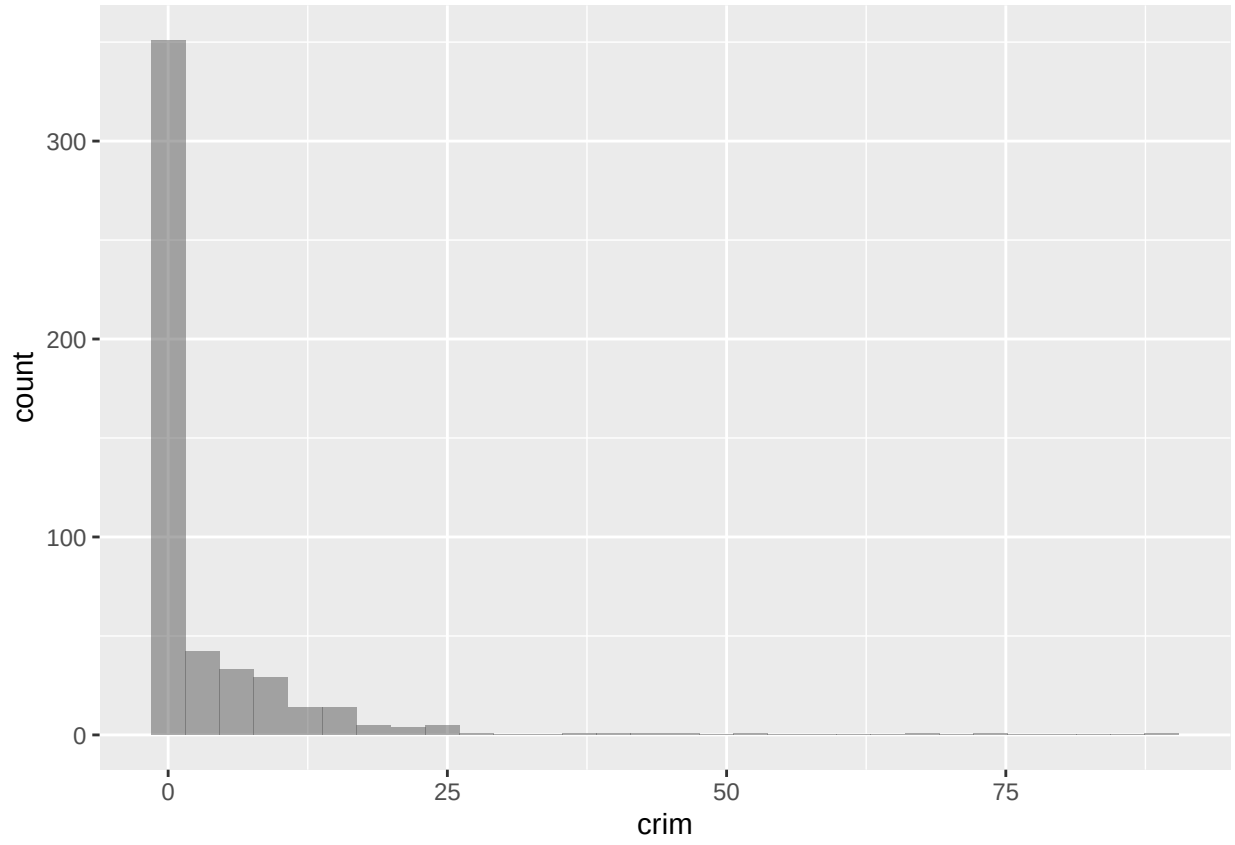
- This exercise involves the Boston housing data set. Assume that we are interested in per capita crime rate, crim. ## A. Examine crim with summary() and in a histogram.

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

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
```

```
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```r
library(ggplot2)
ggplot(Boston, aes(x=crim)) +
   geom_histogram(position="identity", alpha=0.5)
```



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?

```
subset1 <- subset(Boston, Boston$crim > 25)
dim(subset1)[1]
```

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

```
summary(subset1$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.2    20.2    20.2    20.2    20.2    20.2
```

```
summary(subset1$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     666     666     666     666     666     666
```

```
subset2 <- subset(Boston, Boston$crim <= 25)
dim(subset2)[1]
```

```
## [1] 495
```

```
summary(subset2$ptratio)
```

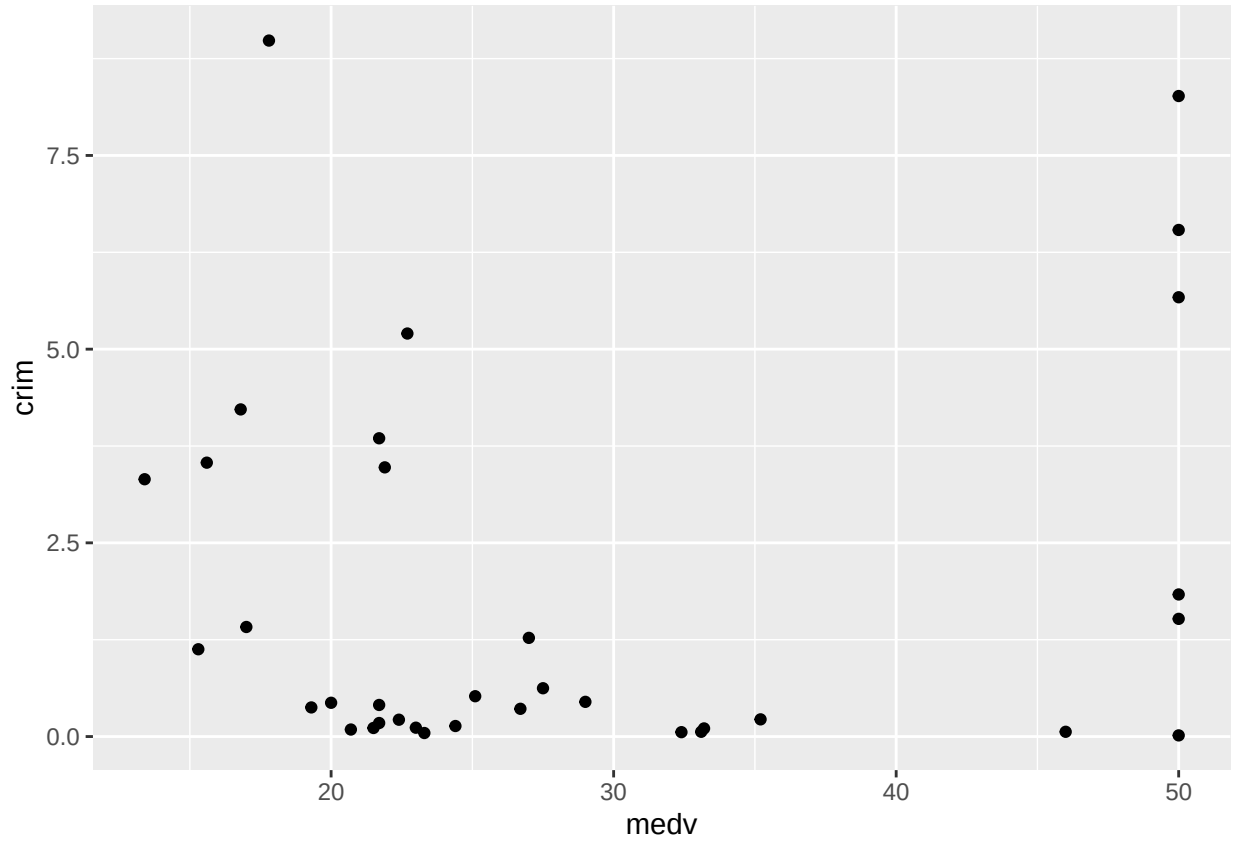
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.00   18.90   18.42   20.20   22.00
```

```
summary(subset2$tax)
```

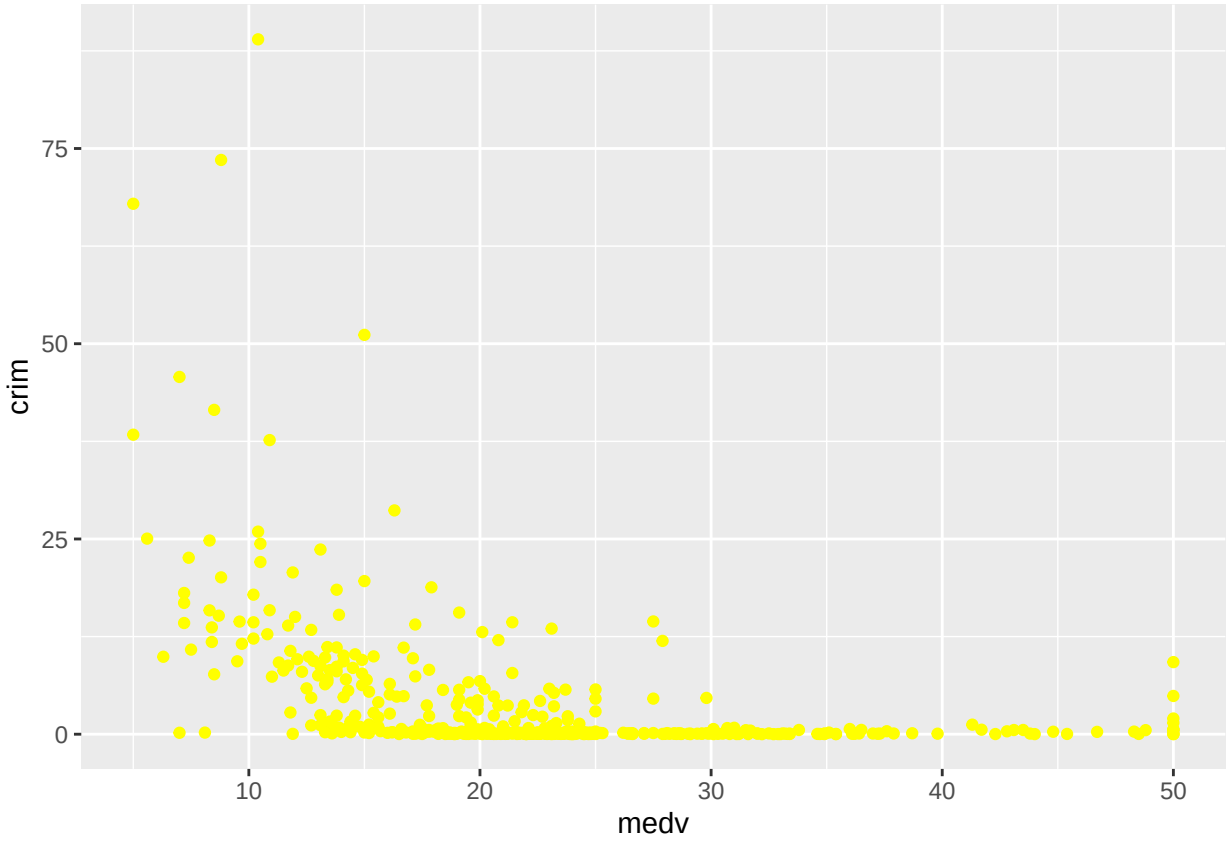
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   278.0   330.0   402.5   666.0   711.0
```

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?

```
subset3 <- subset(Boston, Boston$chas == 1)
subset4 <- subset(Boston, Boston$chas == 0)
library(ggplot2)
ggplot(subset3, aes(y=crim, x=medv))+
  geom_point(color="black")
```




```
ggplot(subset4, aes(y=crim, x=medv))+
  geom_point(color="yellow")
```



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
```

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

- From OLS (lecture note slides 33-36),
  - $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \equiv \frac{SS_{XY}}{SS_X} = \frac{s_{XY}}{s_{XX}}$, where $s_{XY} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ and $s_{XX} = \frac{\sum(x_i - \bar{x})^2}{n-1}$
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

```
SS_XY <- sum((Boston$medv-mean(Boston$medv))*(Boston$crim - mean(Boston$crim)))
SS_X <- sum((Boston$medv-mean(Boston$medv))^2)
beta_1 <- SS_XY / SS_X

S_XY <- SS_XY / (dim(Boston)[1] - 1)
S_XX <- SS_X / (dim(Boston)[1] - 1)
beta_0 <- mean(Boston$crim) - beta_1 * mean(Boston$medv)
```

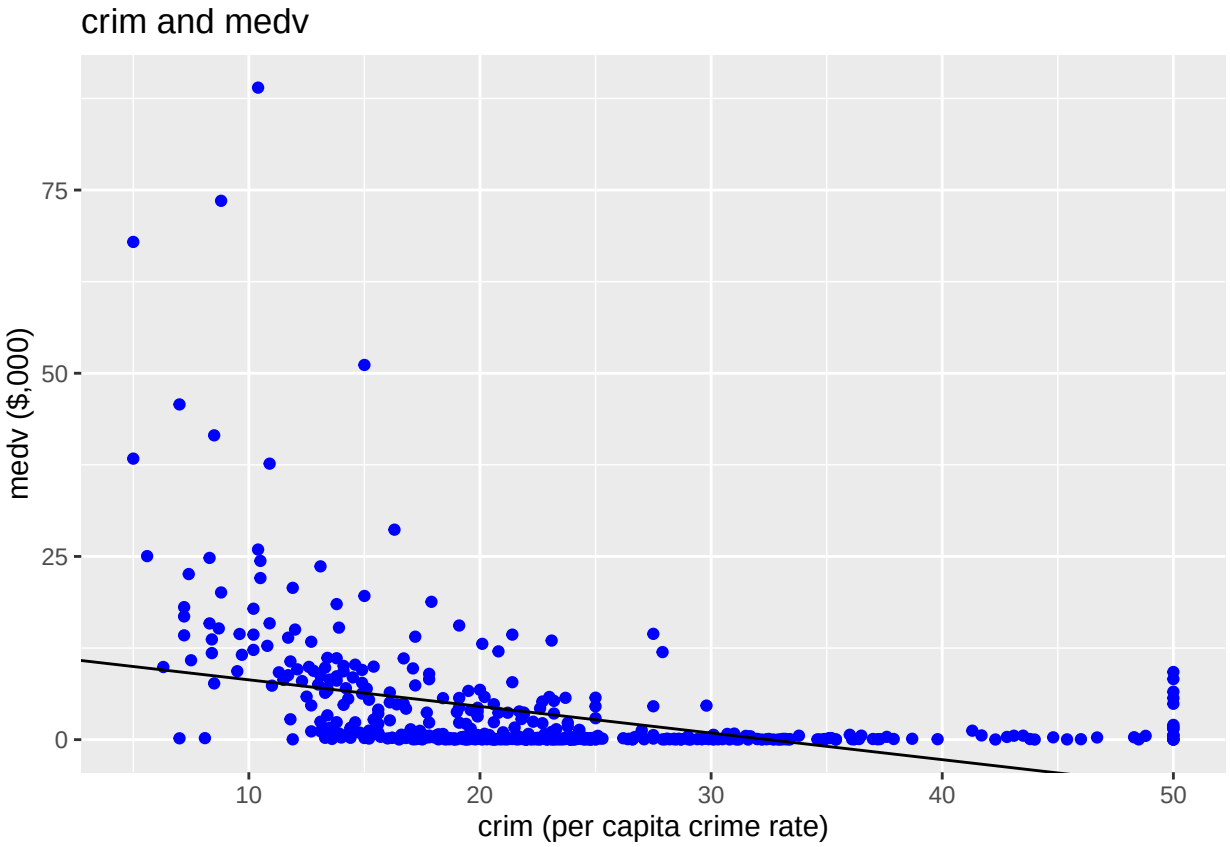
- From OLS (lecture note p.24 and p.25),
  - $\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SS_X}$, where $\hat{\sigma}^2$ is estimated error variance (or residual variance) as $\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2}$
  - $\hat{V}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_X} \right)$

```
coef_crim_medv <- coef(model_obj_1)
Boston$hat_crim <- coef_crim_medv[1] + coef_crim_medv[2]*Boston$medv
Boston$resid_crim_medv <- Boston$crim - Boston$hat_crim
hat_sigma <- sum(Boston$resid_crim_medv^2)/dim(Boston)[1]-2
V_beta_1 <- hat_sigma / SS_X
```

```r
SE_beta_1 <- sqrt(V_beta_1)
SE_beta_0 <- hat_sigma * (1/dim(Boston)[1] + mean(Boston$resid_crim_medv)^2/SS_X)
```

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
ggplot(Boston, aes(y=crim, x=medv))+
  geom_point(color="blue")+
  geom_abline(slope = coef_crim_medv[2],intercept = coef_crim_medv[1])+
  labs(title="crim and medv",
       y ="medv ($,000)", x = "crim (per capita crime rate)")
```



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

```
library(reshape)
library(dplyr)
Boston_sub<-melt(Boston %>% select(crim,hat_crim))
Boston_sub <- Boston_sub%>%
  mutate(type=ifelse(variable=="crim","Observed","Predicted"),crim=value)
ggplot(Boston_sub, aes(x=crim, color=type, fill=type)) +
   geom_histogram(position="identity", alpha=0.5)
```