STP530 HW1 Solution

1.28

a.

After reading the data into R, use the *lm* function to fit a regression model to the data. A sample code is shown below. You can also use other methods to read in the data.

```
# Set the working directory to where you saved the data file

setwd("~/Documents/ASU/STP530-YiZheng/HW-HaozhenXu/HW1")

# Read data from txt file into R and assign data to an object called "HW1.data"

HW1.data <- read.table("CH01PR28.txt")

# Examine the data, especially pay attention to the column names

head(HW1.data)

# Because the dataset doesn't contain column names in its first row, the
# column names were set to V1 and V2 by default. Let's change them to X
# and Y to align with the problem description in the book

colnames(HW1.data) <- c("Y", "X")
head(HW1.data)

# Fit the linear regression model on HW1.data

my.mod <- lm(Y ~ X, data = HW1.data)
summary(my.mod)
```

```
> summary(my.mod)

Call:
lm(formula = Y ~ X, data = HW1.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5278.3 -1757.5  -210.5  1575.3  6803.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
X            -170.58      41.57  -4.103 9.57e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2356 on 82 degrees of freedom
Multiple R-squared:  0.1703,    Adjusted R-squared:  0.1602 
F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

We can get the estimation of the regression function in two ways:

1) From the R output directly.

   The *summary* function gives us the result, which is $\hat{Y} = 20517.60 - 170.58X$

2) Using formulas in the equation (1.10) on page 17 in the textbook.

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

   We can use the following R code to do the calculation,

```
# Attach the dataset so you can directly call the objects

attach(HW1.data)

b1 <- sum((X - mean(X)) * (Y - mean(Y)))/sum((X-mean(X))^2)
b0 <- mean(Y) - b1 * mean(X)
```

```
# View results

b0; b1
```

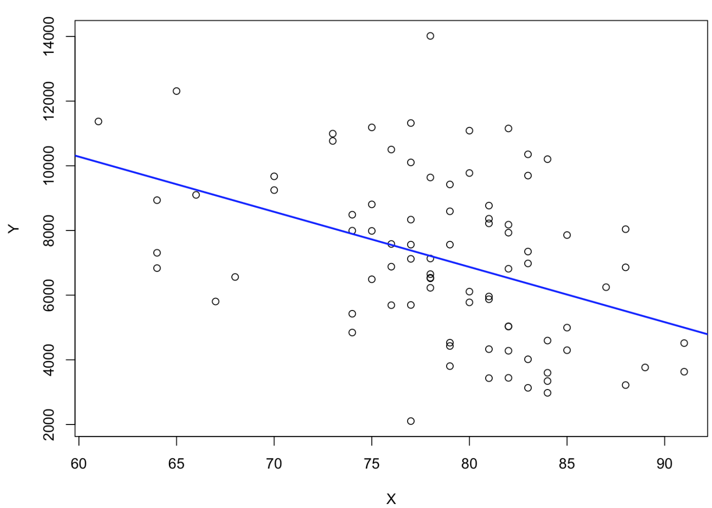
```
> b1 <- sum((X - mean(X)) * (Y - mean(Y)))/sum((X-mean(X))^2)
>
> b0 <- mean(Y) - b1 * mean(X)
>
> b0; b1 # View results
[1] 20517.6
[1] -170.5752
```

So we get the same result $\hat{Y} = 20517.60 - 170.58X$ as using the first method.

For plotting the estimated regression function and the data:

```
# Plot the estimated regression function and the data

plot(X,Y)
abline(my.mod, col="blue", lwd=2)
```



From the plot, we can see lots of points far from the regression line. It is not a great fit. (The multiple R-squared is 0.1703, which means there are only 17.03% information of data can be explained by the fitted model.)

b.

(1)

The slope coefficient in the fitted regression equation gives the change in the expected Y value for every 1-unit increase in X. From the fitted regression function above, we have $b1 = -170.58$. Therefore, we can conclude that when the high school graduate rate (X) for a county increases by 1 percentage point (i.e., 1-unit), the mean (i.e., expected) crime rate (Y) of that county decreases by a count of 170.58 per 100,000 residents.

(2)

The mean crime rate last year in counties with high school graduation percentage $X = 80$ can be calculated as the following:

$$\hat{Y} = b_0 + b_1 X = 20517.60 - 170.58 \times 80 = 6871 \text{ (count per 100,000 residents)}$$

(3)

$$\hat{\varepsilon}_{10} = Y_{10} - \hat{Y}_{10} = 7932 - (20517.60 - 170.58 \times 82) = 7932 - 6530.04 \approx 1402$$

Using R code below is quicker, giving a same result (1401.566).

```
my.mod$residuals[10]
```

(4)

In Section 1.7, the equation (1.22) gives a point estimate of σ^2, which is

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Calculation can be done in R:

```
sum(my.mod$residuals^2)/(dim(HW1.data)[1]-2)
```

```
> sum(my.mod$residuals^2)/(dim(HW1.data)[1]-2)
[1] 5552112
```