STP 530: Applied Regression Analysis

Name : **Sai Swaroop Reddy Vennapusa**

Homework 1

Instructor : **Yi Zheng**

Due Date : 29th August 2023, 10:30AM

Question (a) : Obtain the estimated regression function. Plot the estimated regression function and the data. Does the linear regression function appear to give a good fit here? Discuss.

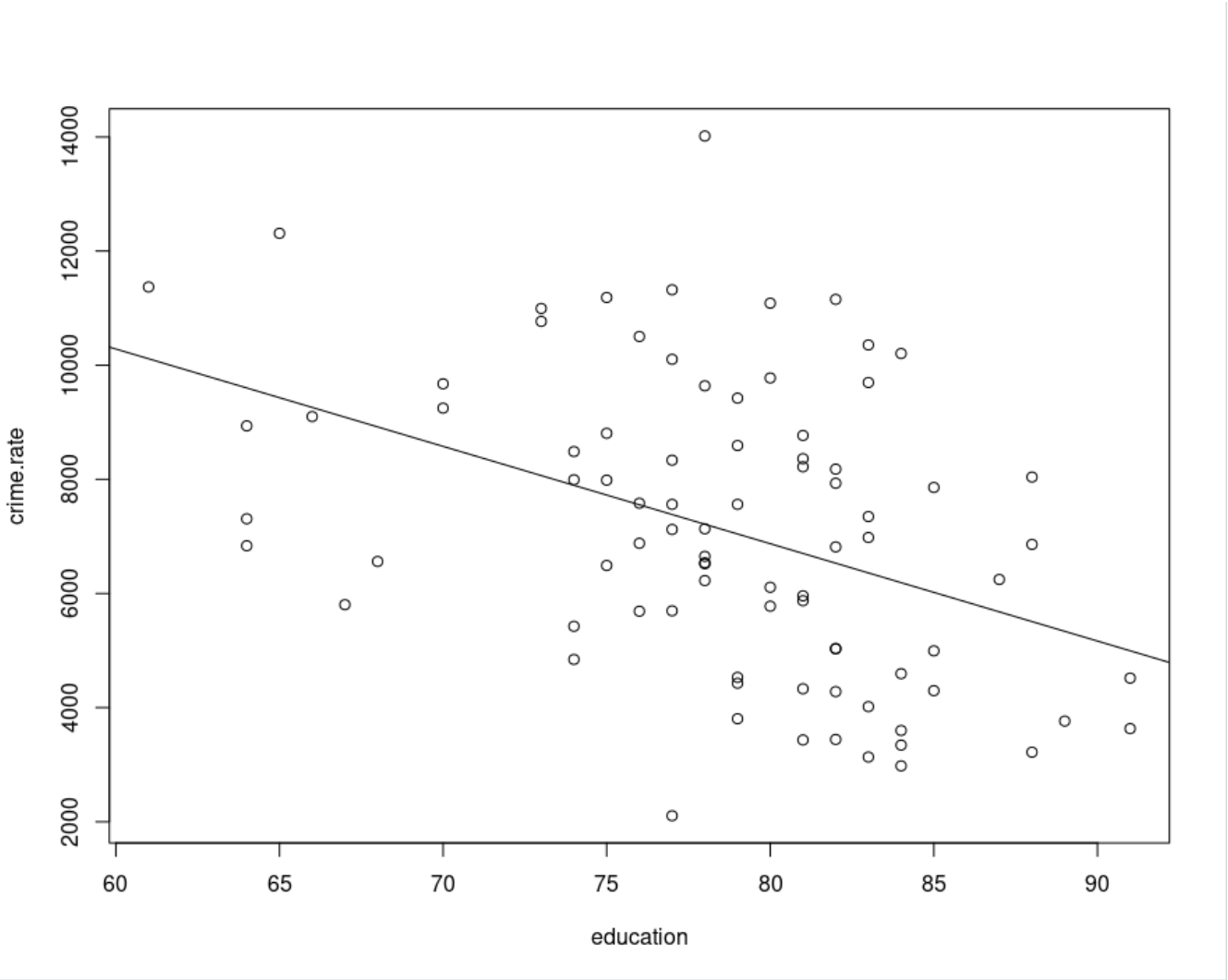
R Code:

```
# Import the dataset
mydata <- read.table("~/Downloads/Assignments/STP530/HW1/CH01PR28.txt", quote="\"", comment.char="")

# Rename the columns
colnames(mydata) <- c("crime.rate", "education")


# Question a
# Run the regression model
m <- lm(crime.rate ~ education, data = mydata)
summary(m)

# Plot the points and the line
plot(crime.rate ~ education, data = mydata) # first option is formula which is y~x. second option is data
abline(coef(m))
```

R Output:

```
Console  Terminal  Background Jobs
R 4.3.1 · ~/
> # Import the dataset
> mydata <- read.table("~/Downloads/Assignments/STP530/HW1/CH01PR28.txt", quote="\"", comment.char="")
> # Rename the columns
> colnames(mydata) <- c("crime.rate", "education")
> # Question a
> # Run the regression model
> m <- lm(crime.rate ~ education, data = mydata)
> summary(m)

Call:
lm(formula = crime.rate ~ education, data = mydata)

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 ***
education    -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

> # Plot the points and the line
> plot(crime.rate ~ education, data = mydata) # first option is formula which is y~x. second option is data
> abline(coef(m))
>
```

Plot:



Answer:

The estimated regression function based on the output from R is: $\hat{Y} = 20517.60 - 170.58 \times \text{education}$.

By visually inspecting the plot, we can gauge how closely the points align with the regression line. Additionally, based on the R-squared value from our `summary(m)` output, approximately 17.03% of the variability in the crime rate is explained by the education level. This is not a particularly high percentage, suggesting that the regression may not be capturing all the factors affecting the crime rate or that the relationship might not be purely linear.

Question (b) : Obtain point estimates of the following: (1) the difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point, (2) the mean crime rate last year in counties with high school graduation percentage X = 80, (3) ephsilon(10)

(4) sigma^2

b.1:

R code:

```
# Question b.1

# The difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point. This
is the coefficient of the variable "education" which is -170.58. This means for every one-percentage-point increase in the
percentage of individuals having at least a high-school diploma, the crime rate decreases by 170.58 crimes per 100,000 residents,
on average.
coef(m)["education"]
```

R output:

```
> coef(m)["education"]
education
-170.5752
```

Answer:

The coefficient of the "education" variable is -170.58. This indicates that for every one-percentage-point increase in the percentage of individuals having at least a high-school diploma, the crime rate decreases by 170.58 crimes per 100,000 residents, on average.

b.2:

R code :

```
# Question b.2

# The mean crime rate last year in counties with high school graduation percentage X = 80.
# Using the regression equation:
# Y^=20517.60−170.58×80=20517.60−13646.4=6871.2
# So, the predicted crime rate in counties with a graduation percentage of 80% is 6,871.2 crimes per 100,000 residents.
predict(m, newdata = data.frame(education = 80))
```

R output:

```
> predict(m, newdata = data.frame(education = 80))
       1
6871.585
```

Answer:
Using the regression equation: Y^=20517.60−170.58×80=20517.60−13646.4=6871.2 So, the predicted crime rate in counties with a graduation percentage of 80% is 6,871.2 crimes per 100,000 residents.

b.3:

R code:

```
# Question b.3

#This refers to the residual for the 10th observation. From the output, epsilon(10) is 1401.566. This means that the actual crime
rate for the 10th county is 1,401.566 crimes per 100,000 residents higher than what the regression line predicts.
# Predict
y.hat <- predict(m)
e <- mydata$crime.rate - y.hat
e[10]

# Get residuals directly
# residuals(m)[10]
# m$residuals[10]
```

R output:

```
> # Predict
> y.hat <- predict(m)
> e <- mydata$crime.rate - y.hat
> e[10]
      10
1401.566
```

Answer:

The residual for the 10th observation is 1401.566, which means that the actual crime rate for the 10th county is 1,401.566 crimes per 100,000 residents higher than what the regression line predicts.

b.4:

R code:

```
# Question b.4

# n is the number of observations (84 counties in this case).
# p is the number of parameters in the model. Since we have an intercept and one predictor (education), p=2
# ei is the residual for the ith observation (difference between observed and predicted values).
# sigma^2 = sum(ei ^ 2) / (n-p)
# Residual variance
sqrt(sum(e ^ 2) / (84 - 2)) # residual standard error
sum(e ^ 2) / (84 - 2) # the square of the above
```

R output:

```
> # Residual variance
> sqrt(sum(e ^ 2) / (84 - 2)) # residual standard error
[1] 2356.292
> sum(e ^ 2) / (84 - 2) # the square of the above
[1] 5552112
```

Answer:

From the output, the variance of the residuals (or the residual variance) is 5,552,112. This is a measure of the spread or variability of the residuals around the regression line. The square root of this value gives the residual standard error, which is 2,356.292 as shown in the output.

Conclusion:

In conclusion, while the regression line provides a way to predict crime rate based on education level, it explains only about 17% of the variability in crime rate. This suggests that there might be other significant factors affecting the crime rate that aren't accounted for in this model.