HW1_Jiseon Yang

2023-08-27

1.28. **Crime rate.** A criminologist studying the relationship between level of education and crime rate in medium-sized U.S. counties collected the following data for a random sample of 84 counties; X is the percentage of individuals in the county having at least a high-school diploma, and Y is the crime rate (crimes reported per 100,000 residents) last year. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	82	83	84
X_i:	74	82	81	...	88	83	76
Y_i:	8,487	8,179	8,362	...	8,040	6,981	7,582

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.

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) ε_{10}, (4) σ^2.

a. The simple linear regression model is given by Yi= β0 +β1 Xi.
i: counties
Xi: the percentage of individuals in the county having at least a high-school diploma
Yi: the crime rate (reported per 100K residents) last year
β0: intercept
β1: slope

Using R, I derived the estimated regression function (**refer to a1**). The estimated values for the intercept (β0) and slope (β1) are 20,517.60 and -170.58, respectively.
Thus, the estimated prediction equation is:

Yi^ = 20517.60 -170.58 Xi

The scatterplot below illustrates the relationship between crime rate and the percentage of education. The fitted regression line is highlighted in red (**a2 & a3**). While there is noticeable scatter in the data points, **the linear regression line seems to provide a reasonable fit, suggesting a negative correlation between crime rate and the level of education**.

b.
(1) The estimated coefficient β1 in the linear regression model represents the change in the crime rate (response variable Y) for a one-percentage point change in the high school graduation rate (predictor variable X). Therefore, the difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point is given by the

value of **β1, -170.58** **(b1)**. This means that for every one percentage point increase in the high school graduation rate, **the crime rate decreases by 170.58 per 100,000 residents**.

(2) The mean crime rate last year in counties with a high school graduation percentage of X=80 can be estimated using the previously derived prediction equation Yi^ = 20517.60 -170.58 Xi, with substituting in X = 80. As illustrated in **section b2**, **the estimated mean crime rate with the 80% education to be 6,871.585 per 100,000 residents**.

(3) The point estimate of ε10: the residual for the 10th observation (ei = Y – Yi^, i-10). This is the residual between the crime rate last year (Y) and the crime rate predicted (Y^) by our regression model, for the 10th observation (**refer to b3**). From the results provided, **the point estimate for ε10, e[10], is 1401.566**. This indicates that, for the 10th observation in the dataset, **the actual crime rate was 1401.566 per 100,000 residents higher than what the regression model predicted.**

(4) The point estimate of σ^2: the variance of the residuals (estimated variance of error, s^2).
S^2 = MSE (mean squared error) = SSE / (n-p)
SSE (sum squared error) = Σe^2
n = 84 counties
p = 2
e = Y – Y^ (see the answer for 3)

From the provided R results (**refer b4),** the residual variance (the point estimate for σ^2) is 5552112). This means that on average, **the observed crime rates deviate from the predicted crime values (regression model) by the squared amount of 5552112**.

R code and the output to address the problem 1.28.

Import the dataset

```
CH01PR28 <- read.table("D:/Dropbox (ASU)/#0 Jiseon 2019 -/10 Classes/2023_STP530_Regression/CH01PR28.txt", quote="\"", comment.char="")
```

Rename column

```
colnames(CH01PR28) <- c("crime.rate","education")
head(CH01PR28,3)

##   crime.rate education
## 1       8487        74
## 2       8179        82
## 3       8362        81
```

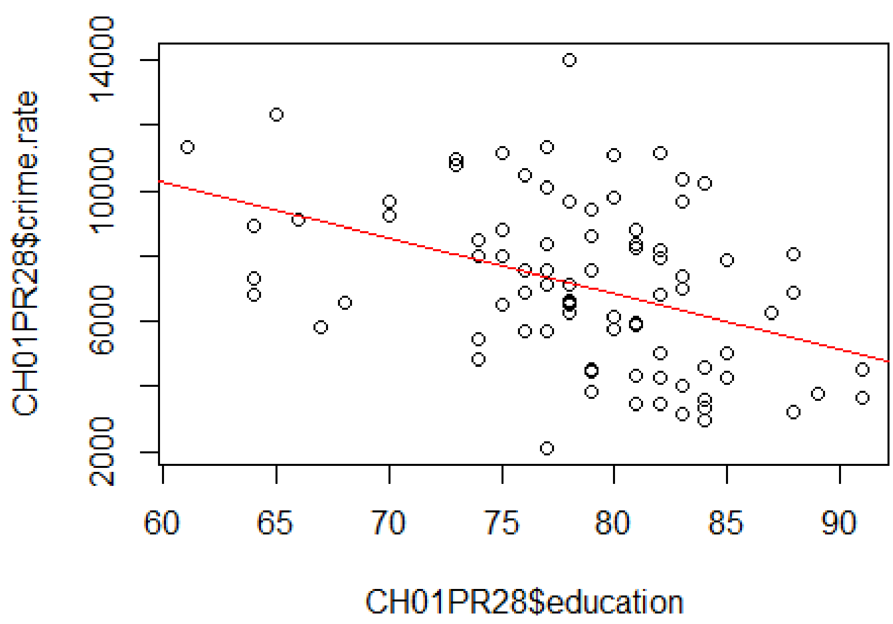
a1. Obtain the estimated regression function

```
Regression <- lm(crime.rate ~ education, data = CH01PR28)
summary(Regression)
```

```
##
## Call:
## lm(formula = crime.rate ~ education, data = CH01PR28)
##
## 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
```

a2 & a3. Scatterplot of the estimated regression function with fitted line

```
plot(CH01PR28$education, CH01PR28$crime.rate)
abline(coef(Regression), col="red")
```



b1. estimated difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point

```
estimate_diff <- coef(Regression)["education"]
estimate_diff

## education
## -170.5752
```

b2. estimated mean crime rate (y.hat) last year in counties with high school graduation percentage X=80

```
last.year <- data.frame(education = 80)
y.hat <- predict(Regression, newdata=last.year)
y.hat

##        1
## 6871.585
```

b3. point estimation of ε10

εi = Yi - E{Yi}, ei = Yi - Yi^

```
y.hat <- predict(Regression)
y.hat

##         1         2         3         4         5         6         7
8
##  7895.036  6530.434  6701.010  6701.010  5677.559  9259.637  8918.487
6701.010
##         9        10        11        12        13        14        15
16
##  7895.036  6530.434  7724.461  6530.434  7212.735  6189.284  6530.434
7042.160
##        17        18        19        20        21        22        23
24
##  7212.735  8065.611  7383.310  9430.213  7383.310  7553.886  7042.160
7042.160
##        25        26        27        28        29        30        31
32
##  7212.735  6189.284  7212.735  6701.010  5336.408  6018.709  7383.310
7895.036
##        33        34        35        36        37        38        39
40
##  6871.585  6189.284  5506.983  7724.461  7383.310  7212.735 10112.513
4995.258
##        41        42        43        44        45        46        47
48
##  6359.859  7383.310  6018.709  8577.337  5506.983  6871.585  6530.434
6530.434
##        49        50        51        52        53        54        55
56
##  6530.434  8577.337  9600.788  7042.160  6359.859  7383.310  7553.886
```

```
6871.585
##        57        58        59        60        61        62        63
64
##  6189.284  6530.434  6701.010  7895.036  6701.010  7553.886  7212.735
7212.735
##        65        66        67        68        69        70        71
72
##  7042.160  6359.859  7042.160  6359.859  6701.010  6189.284  9600.788
9089.062
##        73        74        75        76        77        78        79
80
##  7724.461  8065.611  7383.310  9600.788  7724.461  6871.585  6359.859
6018.709
##        81        82        83        84
##  4995.258  5506.983  6359.859  7553.886

e <- CH01PR28$crime.rate - y.hat
e[10]

##       10
## 1401.566
```

b4. point estimation of σ^2 = s^2 (estimated variation of error)

s^2 = MSE = SSE/(n-p), mean squared error

SSE = sum squared error = Σ(e^2)

s= residual standard error = sqrt(sum(e^2)/(n-p))

s^2 = residual variance = sum(e^2)/(n-p)

```
SSE <- (sum(e^2))
MSE <- SSE/(84-2)
MSE

## [1] 5552112
```