Q4

Huiyan Li

4/18/2020

Assignment 2: Question 4

```
# Only change this following line, remove 261 and put your student id
student_id=1001145664
# do not change anything below
set.seed(student_id)
x= round(rnorm(15,mean=18,sd=4),2)
y= round(50+1.5*x+rnorm(15, mean=0, sd=5),2)
```

A) Calculate the maximum likelihood estimates of β_1 and β_2 (Let's call them b1 and b2)

- $\overline{x} = 18.26933$
- $\overline{y} = 75.612$
- $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 291.5976$
- $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 160.4735$
- $b_2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{291.5976}{160.4735} = 1.817108$
- $b_1 = \overline{y} - b_2\overline{x} = 75.612 - 1.817108 * 18.26933 = 42.41465$
- $y = 42.41465 + 1.817108x$

```
b2 = sum((x-mean(x))*(y-mean(y))) / sum((x-mean(x))^2)
b1 = mean(y) - b2 * mean(x)

b1
```

```
## [1] 42.41465
```

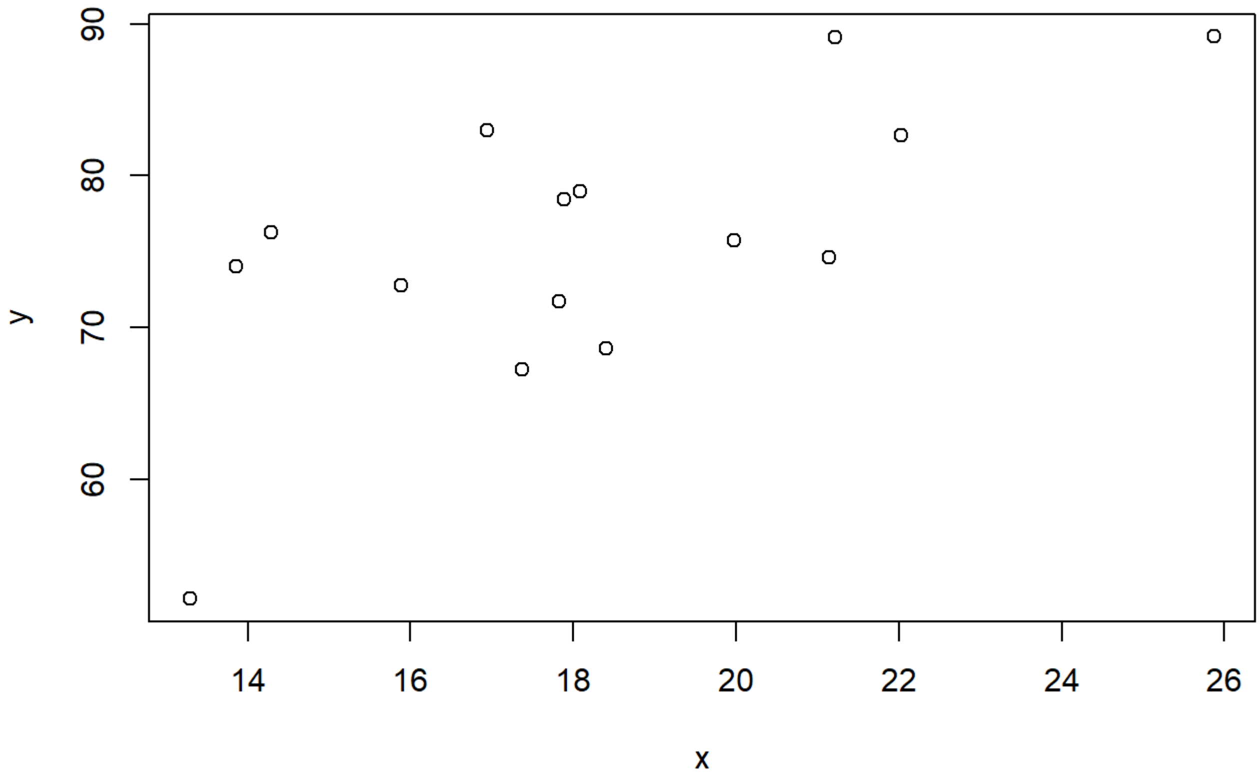
```
b2
```

```
## [1] 1.817108
```

B) Interpret b1 and b2

- The linear regression model equation shows that, when a person doesn't study at all, he/she will has a low score of ≈ 42
- The score increases by a factor of ≈ 1.82 as the number of hour they increases/decreases

```
plot(x, y)
```



C) Construct a 95% confidence interval for β_2 and interpret.

- $SE(B_2) = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 0.5610173$
- Margin of Error $ME = t_{1-\alpha/2, df=n-2} * SE = qt(0.975, df = 13) * 0.5610173 = 1.212004$
- Confidence Interval $CI = \beta_2 \pm ME = 1.817108 \pm 1.212004 = (0.6051036, 3.029112)$
- Which indicates the value of β_2, which is the slope/score fluctuate within the range of (0.6051036, 3.029112) for every unit of x

```
alpha = 0.05
dfSize = length(x) - 2
y_pred = b1 + b2*x
S2 = sum((y-y_pred)^2)/dfSize
SE_B2 = sqrt(S2/sum((x-mean(x))^2))
SE_B2
```

```
## [1] 0.5610173
```

```
ME = qt(1 - alpha / 2, df=dfSize) * SE_B2
ME
```

```
## [1] 1.212004
```

```
CILow = b2 - ME
CIHigh = b2 + ME

CILow
```

```
## [1] 0.6051036
```

```
CIHigh
```

```
## [1] 3.029112
```

D) At 5% level of significance, test $H_0 : \beta_2 = 1.5$ vs $H_a : \beta_2 \neq 1.5$ and write your conclusionin plain English.

- We will reject the Null hypothesis $H_0 : \beta_2 = 1.5$ if and only if $|T| > t_{1-\alpha/2,df=n-2}$, where the T-test statistics value exceeds the critical value
- $CriticalValue = t_{1-\alpha/2,df=n-2} = t_{0.975,13} = qt(0.975, df = 13) = 2.160369$
- $\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - b_1 - b_2x_i)^2 = \frac{656.59741}{15-2} = 50.50749$
- $T = \frac{B_2-\beta_2}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} = \frac{1.817108-1.5}{0.5610173} = 0.5652375$

```
alpha = 0.05
dfSize = length(x) - 2
criticalValue = qt(1 - alpha/2, df=dfSize)
criticalValue
```

```
## [1] 2.160369
```

```
sigma2 = (1/(length(x) - 2)) * sum((y - b1 - b2*x)^2)
sigma2
```

```
## [1] 50.50749
```

```
t = (b2 - 1.5) / (sqrt(sigma2/sum((x - mean(x))^2)))
t
```

```
## [1] 0.565237
```

- $|T| > t_{1-\alpha/2,df=n-2} \implies 0.565237 > 2.160369 \implies false$
- Therefore, we will not reject the null hypothesis at the 0.05 significance level, meaning β_2 could be 1.5

E) Calculate an estimate of σ^2 using an unbiased estimator.

- $\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - b_1 - b_2 x_i)^2 = \frac{656.59741}{15-2} = 50.50749$

```
n = 15
sigma2 = (1/(n - 2)) * sum((y - b1 - b2*x)^2)
sigma2
```

```
## [1] 50.50749
```

F) Compute and interpret the coefficient of determination (R^2)

- Coefficient of determination $R^2 = \frac{RSS}{TSS}$
- Total sum of square $TSS = RSS + ESS$
- Error/Residual sum of square $ESS = \sum_{i=1}^{n}(y_i - b_1 - b_x x_i)^2$
- Regression sum of square $RSS = b_2^2 \sum_{i+1} n(x_i - \overline{x})^2$

```
y_pred = b1 + b2*x
TSS = sum((y-mean(y))^2)
TSS
```

```
## [1] 1186.462
```

```
ESS = sum((y-y_pred)^2)
ESS
```

```
## [1] 656.5974
```

```
RSS = TSS - ESS
R2 = RSS/TSS
R2
```

```
## [1] 0.446592
```

- $R^2 = 0.446592 \implies 44.66\%$ variation in Y in X.
- $r = \sqrt{R^2} = \sqrt{0.446592} = 0.6682754 \implies$ There is a positive relationship between X and Y, but not extremely strong. Which is reasonable, which indicates there is a large σ^2 value due to external factors that affects the performance of a person in the exam and their score