HW 3

Helinda He

2022-11-02

```
class <- read.csv("Class_Data_A-1.csv", header = TRUE, sep = ",")
```

```
handout <- read.csv("Handout 1.csv", header = TRUE, sep = ",")
```

Question 1

a)

The relationship between the age of husbands' and wives' is a linear, negative, strong.

b)

There is a very weak relationship between the height of husbands' and wives'.

c)

The plot of the age has a stronger correlation because the points are close to each other and we estimate that the correlation will be around 0.8.

d)

The conversion will not affect the correlation between husbands' an wives' heights.

Question 2

a)

(2)

b)

(1)

c)

(4)

d)

(3)

Question 3

a)

There is a linear, positive, moderate strong relationship between hip girth and weight.

b)

The relationship will not change based on the the change of one variable.

Question 4

a)

Wage of Man = Wage of Woman + 5000

b)

Wage of Man = 1.25 * Wage of Woman

c)

Wage of Man = 0.85 * Wage of Woman

Question 5

The regression will be under-estimate because the residual is 0.5, which is positive. The equation of the residual is: actual value - predicted value. When the residual is positive that means the actual value is bigger than the predicted value.

Question 6

a)

$$\hat{time} = 50.599 + 0.726 * \text{distance}$$

b) Interpretation of Slope: For every one unite increase in distance, time is expected to increase by 0.726.

Interpretation of Intercept: When distance is 0, time is expected to be 50.599.

c)

$$R^2 = 0.404496$$

Interpretation of

$$R^2$$

: 40.45% of the variability in time that is explained by distance.

d)

0.726 * 103 +50.599 = 125.377

The estimate time of travel will be 125.377 minutes. ### e) residual = actual - predicted = 168 - 125.377 = 42.723 Since the residual is positive which means that the actual value is greater than the predicted value. That tells people the regression expression is under-estimate.

f)

No, because the regression expression is estimate the Amtrak train and if the stop is added then there are more things to consider which make the regression not suitable anymore.

Question 7

a)

$$\hat{Annual Murder} = -29.901 + 2.559 * \text{Poverty}$$

b)

When percent in poverty is 0, annual murders per million is expected to be -29.901.

c)

For every one unite increase in percent in poverty, annual murders per million is expected to increase by 2.559.

d)

70.52% of the variability in annual murders per million that is explained by percent in poverty.

e)

r =

$$\sqrt{0.7052}$$

= 0.8398

Question 8

a)

$$\hat{HeartWeight} = -0.357 + 4.034 * \text{Body Weight}$$

b)

When body weight is 0, heart weight is expected to be -0.357.

c)

For every one unite increase in body weight, heart weight is expected to increase by 4.034.

d)

64.66% of the variability in heart weight that is explained by body weight.

e)

r =

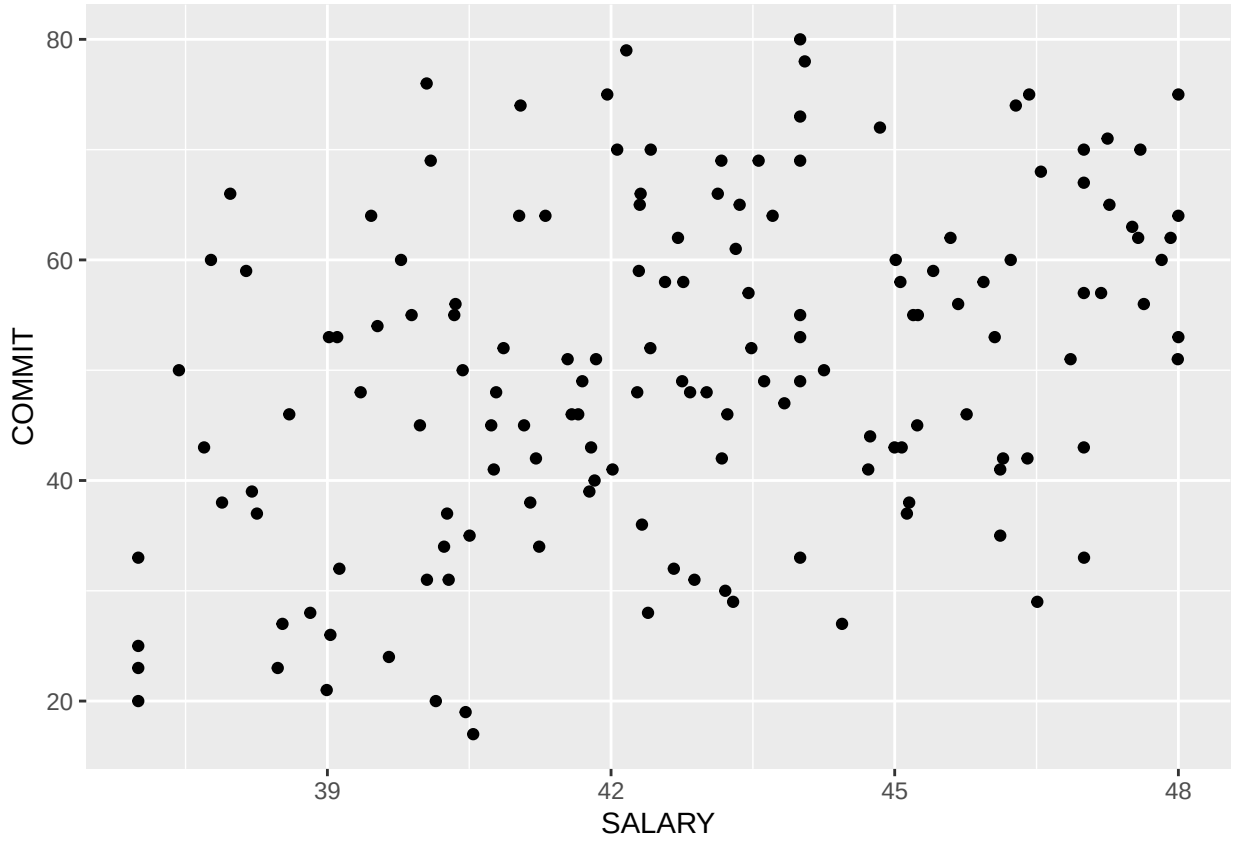
$$\sqrt{0.6466}$$

= 0.8041

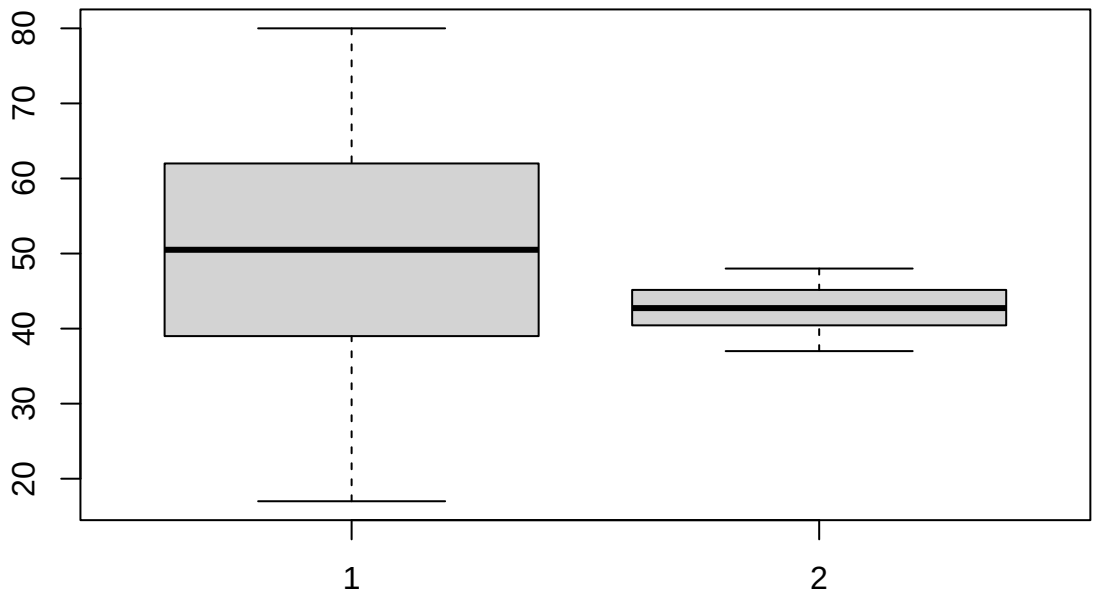
Question 9

a)

```
ggplot(data = handout, aes(x = SALARY, y = COMMIT)) + geom_point()
```




```
boxplot(handout$COMMIT, handout$SALARY)
```



The relationship between COMMIT and SALARY is linear, positive, weak. There is no outlier.

b)

```
lm(COMMIT ~ SALARY, data = handout)
```

```
##
## Call:
## lm(formula = COMMIT ~ SALARY, data = handout)
##
## Coefficients:
## (Intercept)       SALARY
##     -31.866        1.914
```

The regression is:

$$\hat{COMMIT} = -31.866 + 1.914 * \text{SALARY}$$

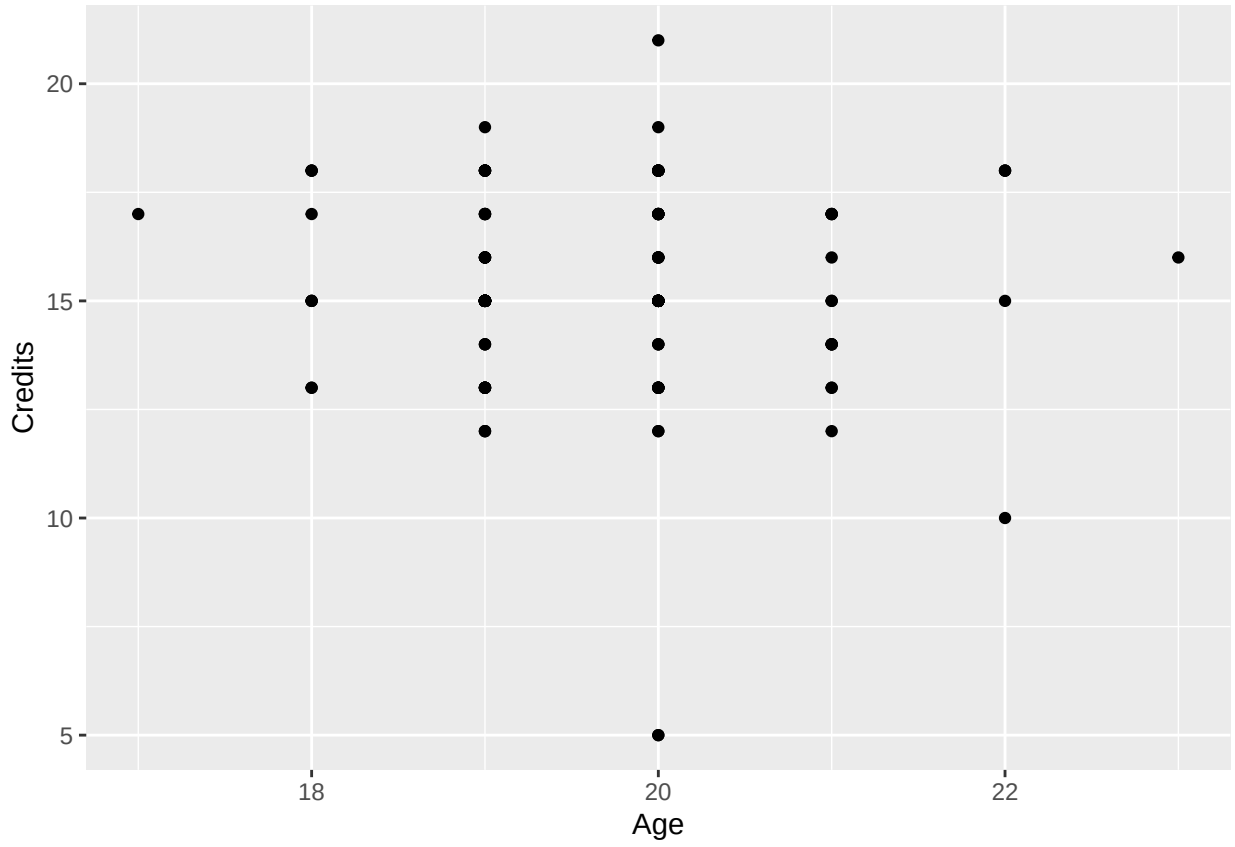
c)

Interpretation of intercept: When SALARY is 0, COMMIT is expected to be -31.866. Interpretation of slope: For every one unite increase in SALARY, COMMIT is expected to increase by 1.914.

Question 10

a)

```
ggplot(data = class, aes(x = Age, y = Credits)) + geom_point()
```




```
m <- lm(Credits ~ Age, data = class)
tidy(m)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)  16.9        3.65      4.63  0.00000778
## 2 Age          -0.0838     0.186    -0.451 0.653
```

The relationship between Age and Credits is linear relationship, negative.

b)

```
glance(m)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma stati~1 p.value    df logLik   AIC   BIC devia~2
##       <dbl>         <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1   0.00131      -0.00513  2.13   0.203   0.653     1  -340.  686.  696.    701.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: statistic, 2: deviance
```

R^2is 0.00131

c)

```
lm(Credits ~ Age, data = class)
```

```
##
## Call:
## lm(formula = Credits ~ Age, data = class)
##
## Coefficients:
## (Intercept)          Age
##    16.90000     -0.08384
```

The regression equation is:

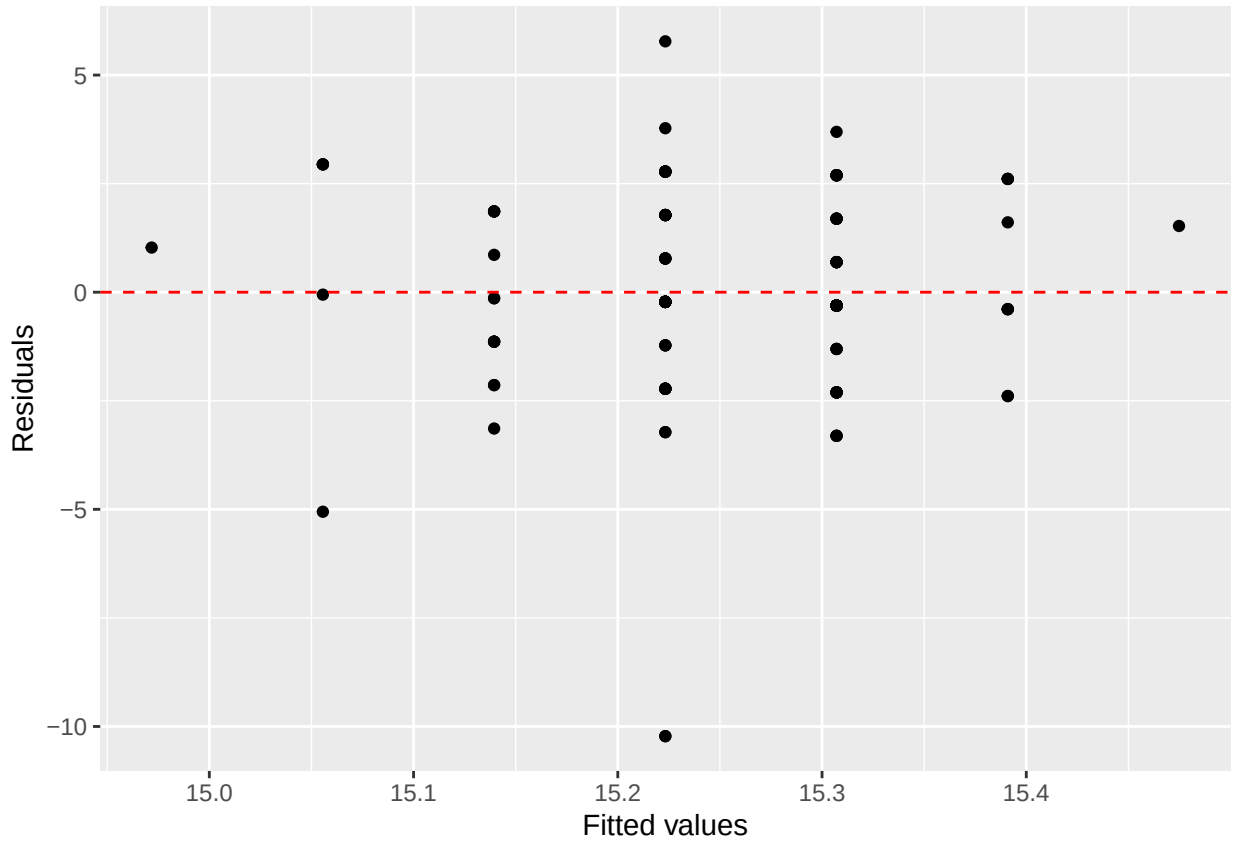
$$\hat{Credits} = 16.9 - 0.08384 * \text{Age}$$

Interpretation: - Slope: For every one unite increase in Age, Credits is expected to decrease by 0.08384. - Intercept: When Age is 0, Credits is expected to be 16.9.

d)

```
m_aug <- augment(m)
```

```
ggplot(data = m_aug, aes(x = .fitted, y = .resid)) + geom_point() + geom_hline(yintercept = 0, linetype
```



People can see from the graph that