ASSIGNMENT # 1

Submitted by

Julhas Sujan

Student ID: 2025166681

Course Name: Biostatistics-II

Submitted to

Dr. Ahmed Hossain

Professor, Department of Public Health

North South University

Email: ahmed.hossain@utoronto.ca

ahmed.hossain@northsouth.edu

Date of submission: 16 May 2021

Q1. Suppose a cohort of 231 patients at a cancer Institute with stage I (operable) lung cancer have been followed for a minimum of 5 years, and that 140 of the patients survived at least 5 years. Find the estimate of the 5-year survival probability and its 95% confidence interval. Suppose that Cancer Institute statistics indicate that the 5-year survival probability for stage I lung cancer is 0.60. Do the data support the claim that the population with this disease have a different 5-year survival probability?

Answer:

Here,

Total lung cancer patient = 231

Number of Patients survive in five years = 140

Confidence interval =95%

$$P(il) = \frac{140}{231}$$

Therefore the probability of surviving is 0.6086 = 60.86%

H0: P(survival)=0.61
H1: P(survival)≠0.61

The estimated standard error under the null hypothesis is Standard Error (S.E) = $\sqrt{\frac{p(1-p)}{n}}$

$$= \sqrt{\frac{0.61\ (1-0.61)}{231}}$$

$$= 0.03223$$

To calculate confidence interval, this equation is used: CI = probability of surviving $\pm$ z * S.E

$$= 0.61 \pm 1.96 \times 0.0322$$
$$= 0.61 \pm 0.063$$

Therefore the 95% confidence interval is between **0.**547**, 0.673**

As the confidence interval contains the probability of survival 0.61 (null hypothesis), we can say that we don't have enough evidence to reject the null hypothesis.

Q2. The following is a description of the variables we have selected from the study for the purpose of this assignment:

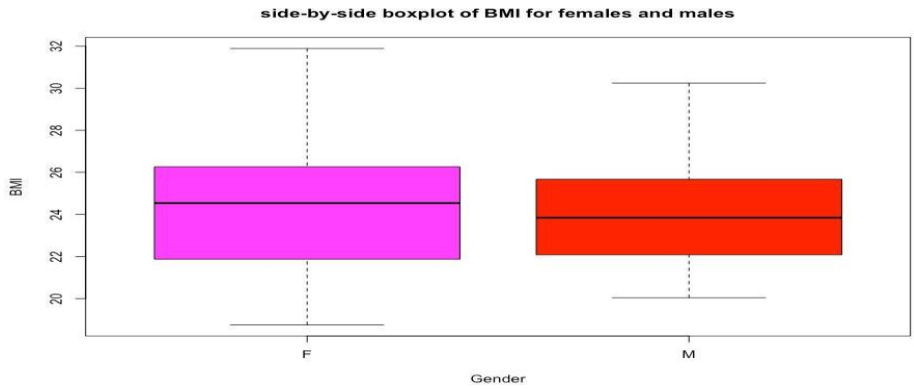
1 Id =Subject Number,2 Gender= Gender (M-Male, F-Female),3 Age= Age,4 Height=Height (inches), Weight= Weight (kg)

The data can be downloaded in excel format from the facebook group (NSUBIO2AH). Answer the following questions using the data:

Q2/1) Obtain a side-by-side boxplot of BMI for females and males. Paste the boxplot into your report.

Answer:

```
1
2
3  # xlsx files
4  my_data <- read_excel("/Users/mosharofct/Downloads/DataQ2_setC.xlsx")
5  my_data
6  bmi <- my_data$WEIGHT/((my_data$HEIGHT * 0.0254) ^ 2)
7  bmi
8  boxplot(bmi~my_data$SEX, col = c(6,2), main = "side-by-side boxplot of BMI for females and
   males", xlab = "Gender", ylab = "BMI")
9
10 factor(my_data$SEX)
11
12 #ID: P35 was marked as 3 (for sex) wrongly by the data entry person. I have edited in excel
   according to the name of the participant and the boxplot is based on the updated excel file.
13
14
15
```



Q2/2) Given the side-by-side boxplot obtained in part (1), what are the appropriate measures of cen-ter and spread to compare the two distributions? Compare the centers, spreads, and shapes(symmetric, skewed) of the two distributions

Answer:

```
my_data %>% mutate(bmi) %>% group_by(SEX) %>% summarise(mean =
mean(bmi), median = median(bmi), mode = mode(bmi), q1 = quantile
(bmi,.25), q2 = quantile(bmi,.50), q3 = quantile(bmi,.75), q4 =
quantile(bmi,1),sk = skewness(bmi, type =3))


options(dplyr.width = Inf)
```

Shape :

Female: Q2 - Q1 > Q3 - Q2 , so, we can say, the distribution of females negatively skewed.

Male: Q2 - Q1 < Q3 - Q2, so, we can say, the distribution of males positively skewed.

The width of the box is more for females than for males. This indicates more variability. Besides, the maximum and minimum values and the range is larger for females than for males.

The appropriate measures of centre and spread to compare the two distributions are median and interquartile range as both distributions are skewed.

Median BMI for F: 24.5 and for M: 23.8. Median BMI is higher in females than males. Mean

for F and for M: 24.5 and 24.1 respectively.

InterQuartile Range (IQR):

For Female

Q3= 26.3 Q1= 21.9, Now IQR= 26.3-21.9= 4.4

For Male

Q3=25.70 Q1=22.1, Now IQR= 25.7-22.1= 3.60

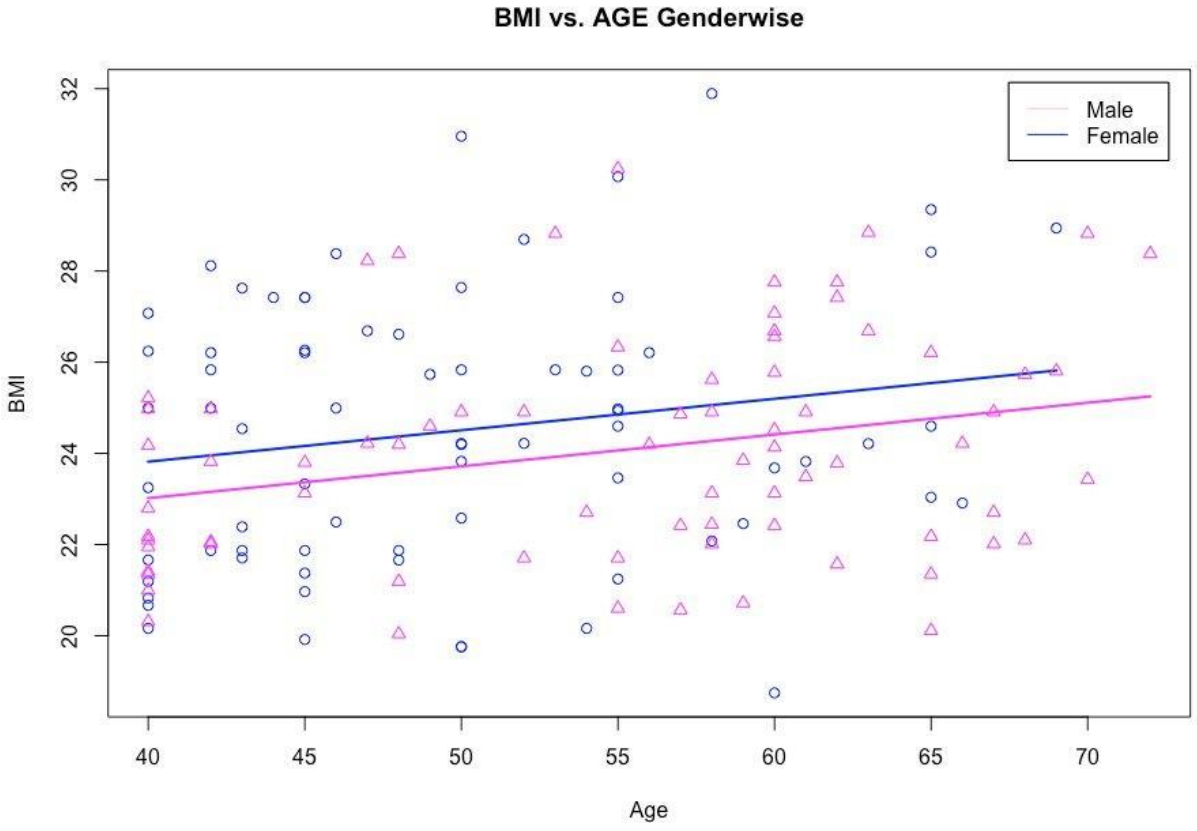
There is more variability of BMI in females than males.

Q2/ 3. Obtain a scatterplot of BMI vs. age with different marking symbols for each gender. Paste The scatterplot into your report. (title, names of the axes, and the legend for the two gender groups)

Answer:

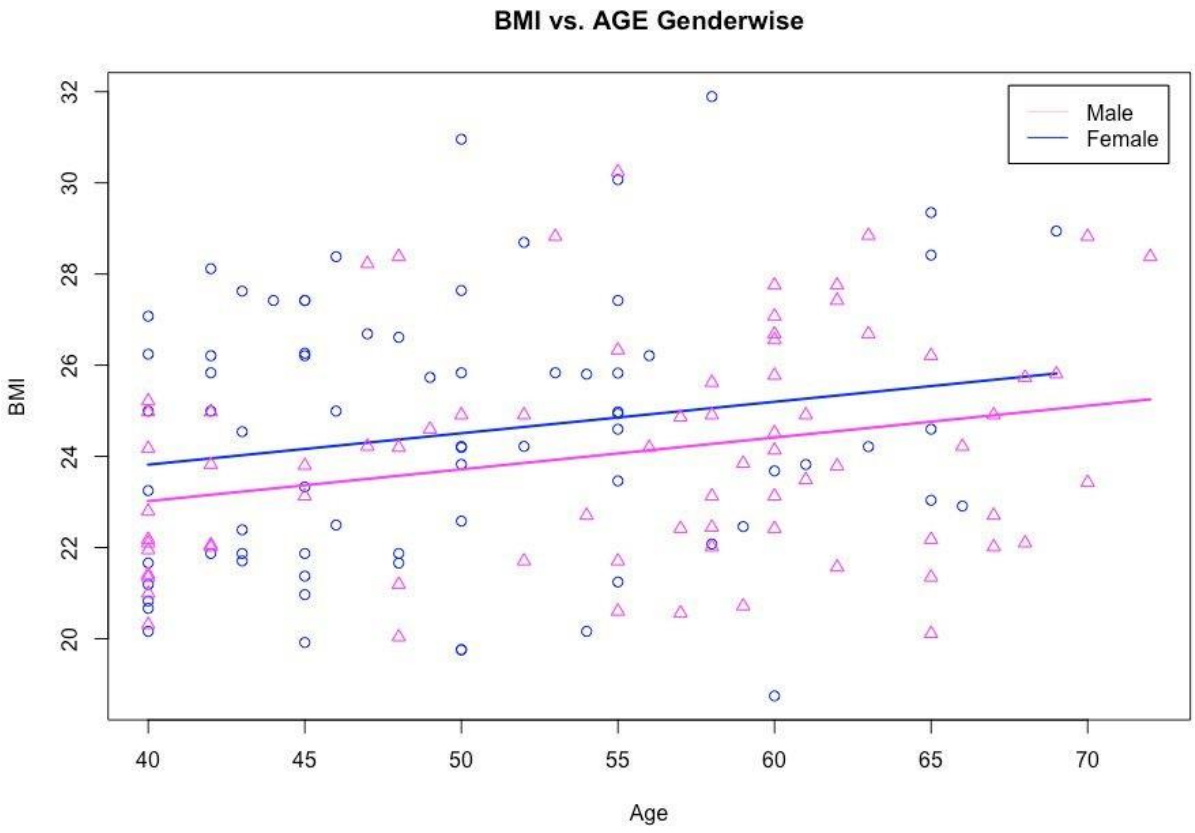
```
my_data <- my_data %>% mutate(bmi)

scatterplot(my_data$bmi ~ my_data$AGE  | my_data$SEX,smooth = FALSE, grid = FALSE, regLine =
TRUE, main = "BMI vs. AGE Genderwise", xlab = "Age", ylab = "BMI",legend("topright", lty=1
,col=c("pink", "blue"), legend = c ("Male", "Female"), inset=0.02))
```



Q2/4. Use the scatter plot obtained in part (3) to describe the relationship between BMI and age for each gender. In particular, comment on the overall form (line, curve), direction (positive or negative) and strength (size of the scatter) of the relationship for males and females.

Answer:



Direction: As the Age increases for Male and Female, BMI also increases. Therefore the relationship between age and bmi positive.

```
spI?t_/ry_data        spI   I    ny_data,    ny_da ta SEX

sp I? I: my da ta
```

```
Restdua Is:
        Min.      10    Wedd an       30        Nax
—6 4S31 —2 3196       —8 2894     2 2S34    6  8329


                    EsUmate  SId.       Errow   I va I ue Pr(> I | )
(Zn te rcep I)      21.05892        2.23685     9.415   4.04e—14 +++
fvualo_nart5AC5                                 1.555       8.134

S qn f.  codes:   0   +++   0.001   ++   0.01 '+   0.05   .' 0.1        1

Rest dna I  stun dv rd e rro r: 2. 988 on 71 deg rees o L  I reedoo

Nu I t1  p I e R—s quo red:   8. 83292,      Ad just ed  R—squa red:   8. 8193
```

```
> summary(male_lm)

Call:
lm(formula = male_part$bmi ~ male_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6520 -1.8098 -0.2233  1.4025  6.1800

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.22227    1.62882  12.415   <2e-16 ***
male_part$AGE  0.06988    0.02916   2.396   0.0191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.386 on 73 degrees of freedom
Multiple R-squared:  0.07293,   Adjusted R-squared:  0.06023
F-statistic: 5.743 on 1 and 73 DF,  p-value: 0.01912
```

Strength: The multiple R-squared values from above 2 screenshot imply that Age and BMI has weak positive relationship (for both males and females)

Q2/5. How does the relationship between BMI and age for the males differ from the one for the females?

Answer:

```
cor(male_part$bmi, male_part$AGE)
cor(female_part$bmi, female_part$AGE)
```

```
> cor(male_part$bmi, male_part$AGE)
[1] 0.2700631
> cor(female_part$bmi, female_part$AGE)
[1] 0.1814411
>
```

The correlation between age and BMI (for male) is 0.27 and The correlation between age and BMI (for Female) is 0.18 . This implies a weak positive relationship between age and BMI for both genders but in the case of male, the relationship between age and BMI is slightly better compared with females.

Q2/6) Obtain the correlation coefficients between BMI and age for males and females. Paste the related output into your report.

Answer:

```
> summary(female_lm)

Call:
lm(formula = female_part$bmi ~ female_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4531 -2.3196 -0.2894  2.2534  6.8329

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     21.05892    2.23685   9.415 4.04e-14 ***
female_part$AGE  0.06899    0.04437   1.555    0.124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.908 on 71 degrees of freedom
Multiple R-squared:  0.03292,   Adjusted R-squared:  0.0193
F-statistic: 2.417 on 1 and 71 DF,  p-value: 0.1245
```

```
> summary(male_lm)

Call:
lm(formula = male_part$bmi ~ male_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6520 -1.8098 -0.2233  1.4025  6.1800

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.22227    1.62882  12.415   <2e-16 ***
male_part$AGE  0.06988    0.02916   2.396   0.0191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.386 on 73 degrees of freedom
Multiple R-squared:  0.07293,   Adjusted R-squared:  0.06023
F-statistic: 5.743 on 1 and 73 DF,  p-value: 0.01912
```

Simply stated: the **R-squared** value is simply the square of the **correlation coefficient R** .

So, the correlation coefficient in between BMI and age for males: sqrt(0.07293) = 0.27

And the correlation coefficient in between BMI and age for females: sqrt(0.0329) = 0.18

Q2/7) Do the signs and magnitudes of the coefficients confirm your conclusions you have reached in Question 5? Explain briefly

Answer:

Yes, it confirms.

Correlation coefficient for male: **R= 0.27** (Positive) it appears that there is positive weak correlation between age and BMI of male.

Correlation coefficient for male: **R= 0.181438** (Positive) it appears that there is also weak positive correlation between age and BMI of male.

Correlation coefficient for male: R= 0.27 greater than Correlation coefficient for female: R = 0.18. So, we can say that in case of male, the dependent variable BMI can be explained a bit better using independent variable Age (compared to female)

Q2/8) Find the equations of two least-squares regression lines to predict BMI from age for each gender group. Compare the slopes of the least-squares regression lines. Which BMI increases faster with age, the one for men or the one for women?

Answer:

```
> summary(female_lm)

Call:
lm(formula = female_part$bmi ~ female_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4531 -2.3196 -0.2894  2.2534  6.8329

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     21.05892    2.23685   9.415 4.04e-14 ***
female_part$AGE  0.06899    0.04437   1.555    0.124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.908 on 71 degrees of freedom
Multiple R-squared:  0.03292,   Adjusted R-squared:  0.0193
F-statistic: 2.417 on 1 and 71 DF,  p-value: 0.1245
```

Least square regression line equation for female:

female_part$bmi =21.05892 + 0.06899 * female_part$AGE

```
> summary(male_lm)

Call:
lm(formula = male_part$bmi ~ male_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6520 -1.8098 -0.2233  1.4025  6.1800

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.22227    1.62882  12.415   <2e-16 ***
male_part$AGE  0.06988    0.02916   2.396   0.0191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.386 on 73 degrees of freedom
Multiple R-squared:  0.07293,   Adjusted R-squared:  0.06023
F-statistic: 5.743 on 1 and 73 DF,  p-value: 0.01912
```

Least square regression line equation for male::

male_part$bmi =20.22227 + 0.06988 * male_part$AGE Comparison

between two slopes:

For male: If age increases by 1 year, the bmi will be increased by 0.06988 kg/m^2

For female: If age increases by 1 year, the bmi will be increased by 0.06899 kg/m^2

So we can say, the male BMI increases slightly faster with age [compared to female]

Q2/9) What percent of the variation in BMI for males and females is explained by their age?

Answer:

```
> summary(female_lm)

Call:
lm(formula = female_part$bmi ~ female_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4531 -2.3196 -0.2894  2.2534  6.8329

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     21.05892    2.23685   9.415 4.04e-14 ***
female_part$AGE  0.06899    0.04437   1.555    0.124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.908 on 71 degrees of freedom
Multiple R-squared:  0.03292,   Adjusted R-squared:  0.0193
F-statistic: 2.417 on 1 and 71 DF,  p-value: 0.1245
```

As multiple R-squared value for female model: 0.03292, so (0.03292 * 100)% = 3.292% of the variation in BMI for females (dependent variable) is explained by their age (independent variable).

```
> summary(male_lm)

Call:
lm(formula = male_part$bmi ~ male_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6520 -1.8098 -0.2233  1.4025  6.1800

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.22227    1.62882  12.415   <2e-16 ***
male_part$AGE  0.06988    0.02916   2.396   0.0191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.386 on 73 degrees of freedom
Multiple R-squared:  0.07293,   Adjusted R-squared:  0.06023
F-statistic: 5.743 on 1 and 73 DF,  p-value: 0.01912
```

As multiple R-squared value for male model: 0.07293, so (0.07293 * 100)% = 7.293% of the variation in BMI for male (dependent variable) is explained by their age (independent variable).

Q2/10) Predict the BMI of a man and a woman of 40 years old. Would you be able to predict the BMI of a man and a woman of 55 years old? Explain briefly.

Answer:

Least square regression line equation for female:

female_part$bmi =21.05892 + 0.06899 * female_part$AGE So, for

a 40 years woman,

female_part$bmi =21.05892 + 0.06899 * 40 = 23.81852 kg/m^2 And

for a 55 years woman,

female_part$bmi =21.05892 + 0.06899 * 55 = 24.85337 kg/m^2

Least square regression line equation for male::

male_part$bmi =20.22227 + 0.06988 * male_part$AGE So, for a 40 years male,

male_part$bmi =20.22227 + 0.06988 * 40 = 23.01747 kg/m^2 And

for a 55 years male,

male_part$bmi =20.22227 + 0.06988 * 55 = 24.06567 kg/m^2

Q2/11) Use the Summary Statistics (Column) feature in the Stat menu to calculate the mean, standard deviation, median, quartiles, minimum and maximum of the residuals for each gender. Paste The outputs into your report

Answer:

Residuals summary of male model:

```
> summary(male_lm)

Call:
lm(formula = male_part$bmi ~ male_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6520 -1.8098 -0.2233  1.4025  6.1800
```

Residual standard error: 2.386 on 73 degrees of freedom The

mean of residuals in linear regression is always zero

Residuals summary of female model:

```
> summary(female_lm)

Call:
lm(formula = female_part$bmi ~ female_part$AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4531 -2.3196 -0.2894  2.2534  6.8329
```

Residual standard error: 2.908 on 71 degrees of freedom The

mean of residuals in linear regression is always zero

Q2/12) What is the mean, and standard deviation of the residuals for each gender group? Identify the female and male subjects with largest residual.

Answer:

The mean of residuals in linear regression is always zero (True for both male and female models)

The estimated standard deviation of the residuals called "residual standard error" So,

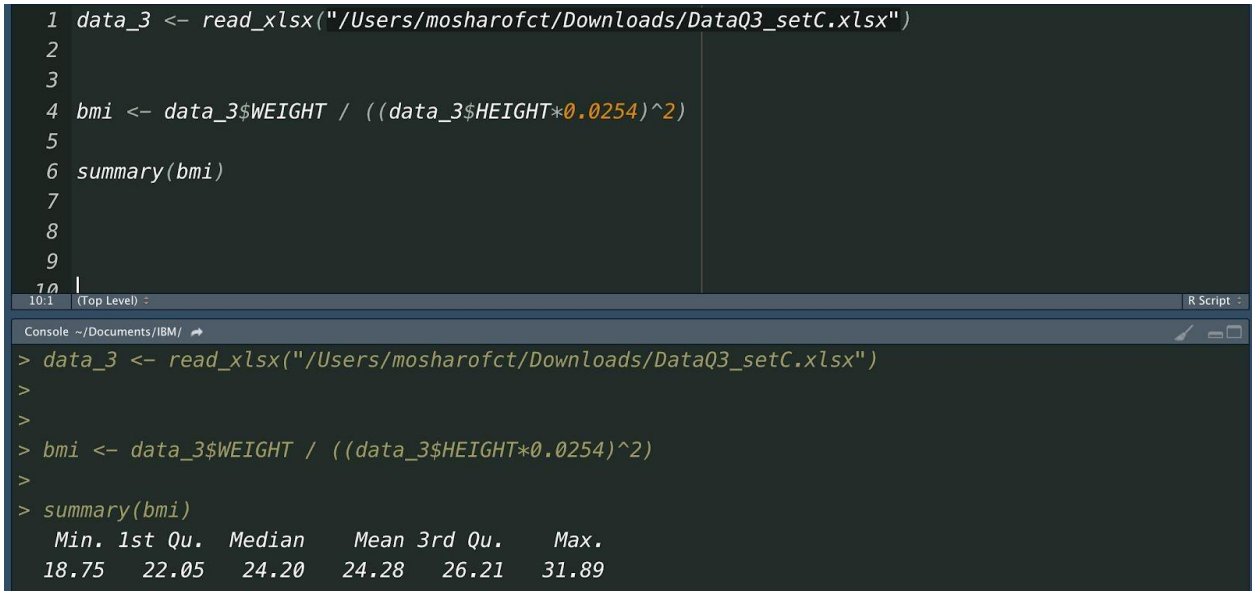
Residual standard error: 2.386 on 73 degrees of freedom (for male)

And, Residual standard error: 2.908 on 71 degrees of freedom (for female)

Q3) This dataset has a binary response (outcome, dependent) variable called disease (1= yes and 0= no). There are three predictor variables: age, sex and BMI. We have the variables BMI as continuous. The data is given in excel format. Answer the following questions using the data

Q3/1) Summarize the data for BMI and categorize it for normal, overweight and obese group.

Answer:



```
data_3 <- read_xlsx("/Users/mosharofct/Downloads/DataQ3_setC.xlsx")


bmi <- data_3$WEIGHT / ((data_3$HEIGHT*0.0254)^2)

summary(bmi)
```

10:1 (Top Level) R Script

Console ~/Documents/IBM/

```
> data_3 <- read_xlsx("/Users/mosharofct/Downloads/DataQ3_setC.xlsx")
>
>
> bmi <- data_3$WEIGHT / ((data_3$HEIGHT*0.0254)^2)
>
> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.75   22.05   24.20   24.28   26.21   31.89
```

```
 8  data_3 <- data_3 %>% mutate(bmi)
 9
10  bmi_cat <- (cut(bmi,c(0,25,30,100),labels=c("normal","overweight","obese")))
11
12
13  data_3 <- data_3 %>% mutate(bmi_cat)
14
15  table(data_3$bmi_cat)
16
17
```

16:1 (Top Level) R Script

Console ~/Documents/IBM/

```
> data_3 <- data_3 %>% mutate(bmi)
>
> bmi_cat <- (cut(bmi,c(0,25,30,100),labels=c("normal","overweight","obese")))
>
>
> data_3 <- data_3 %>% mutate(bmi_cat)
>
> table(data_3$bmi_cat)

   normal overweight      obese
       98         46          4
```

Q3/2) Find two-way contingency table of categorical outcome and BMI. Is there any relationship between them?

Answer:

```
20  table(data_3$Disease, data_3$bmi_cat)
21
22  chisq.test(table(data_3$Disease, data_3$bmi_cat))
23
```

20:1 (Top Level)

Console ~/Documents/IBM/

```
> table(data_3$Disease, data_3$bmi_cat)

   normal overweight obese
0      27         10     1
1      71         36     3
>
```

```
22  chisq.test(table(data_3$Disease, data_3$bmi_cat))
23
```

22:1 (Top Level)

Console ~/Documents/IBM/

```
> chisq.test(table(data_3$Disease, data_3$bmi_cat))

        Pearson's Chi-squared test

data:  table(data_3$Disease, data_3$bmi_cat)
X-squared = 0.5551, df = 2, p-value = 0.7576

Warning message:
In chisq.test(table(data_3$Disease, data_3$bmi_cat)) :
  Chi-squared approximation may be incorrect
>
```

As the p-value is > 0.05%, we don't have enough evidence to reject the null that there has no relationship between disease outcome and bmi.

Q2/3) Fit a logistic regression model with the outcome disease.

```
26  lrm <- glm(data_3$Disease ~ data_3$AGE + factor(data_3$SEX)+ factor(data_3$bmi_cat), family =
    "binomial")
27
28  summary(lrm)
```

26:53 (Top Level) R Script

Console ~/Documents/IBM/

```
Call:
glm(formula = data_3$Disease ~ data_3$AGE + factor(data_3$SEX) +
    factor(data_3$bmi_cat), family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2832  -1.0920   0.5828   0.7677   1.2653

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -3.30616    1.23147  -2.685 0.007259 **
data_3$AGE                       0.09114    0.02574   3.541 0.000399 ***
factor(data_3$SEX)M             -0.54354    0.42279  -1.286 0.198580
factor(data_3$bmi_cat)overweight  0.04900    0.45373   0.108 0.913996
factor(data_3$bmi_cat)obese      -0.39831    1.20735  -0.330 0.741470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 168.61  on 147  degrees of freedom
Residual deviance: 153.51  on 143  degrees of freedom
AIC: 163.51

Number of Fisher Scoring iterations: 4
```

log(odds ratio) =

-3.30616+0.09114 * data_3$AGE-0.54354 * data_3$SEX+0.04900 * overweight -.39831 * obese

Q3/4) Interpret the logistic regression coefficients.

Answer:

BMI and Sex don't have significant effect on disease outcome at 5% level of significance (as p-values > 0.05) but Age has significant effect on disease outcome (p-value < 0.05)

AGE: e^ (0.09114) =1.0954. Odds of developing disease are 1.095 times with every 1 year increase of age while considering other factors constant.

SEX: e^ (-0.54354) = 0.5806. Odds of developing disease are 0.42 times or 42% less for females than for males.

Overweight: e^ (0.04900) = 1.05. Odds of developing disease are 5% more for overweight compared to normal.

Q3/5) Find overall effect of BMI is statistically significant. Obese: e^ (-0.39831) = 0.6714. Odds of developing disease are approximately 33% less for obese compared to normal.

Answer:

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -3.30616    1.23147  -2.685 0.007259 **
data_3$AGE                       0.09114    0.02574   3.541 0.000399 ***
factor(data_3$SEX)M             -0.54354    0.42279  -1.286 0.198580
factor(data_3$bmi_cat)overweight  0.04900    0.45373   0.108 0.913996
factor(data_3$bmi_cat)obese     -0.39831    1.20735  -0.330 0.741470
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 168.61  on 147  degrees of freedom
Residual deviance: 153.51  on 143  degrees of freedom
AIC: 163.51
```

Which implies that the effect of BMI on outcome is not statistically significant at 5% level of significance..

Q3/6) Test whether the difference between normal and obesity are statistically significant.

Answer:

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -3.30616    1.23147  -2.685 0.007259 **
data_3$AGE                      0.09114    0.02574   3.541 0.000399 ***
factor(data_3$SEX)M            -0.54354    0.42279  -1.286 0.198580
factor(data_3$bmi_cat)overweight  0.04900    0.45373   0.108 0.913996
factor(data_3$bmi_cat)obese    -0.39831    1.20735  -0.330 0.741470
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 168.61  on 147  degrees of freedom
Residual deviance: 153.51  on 143  degrees of freedom
AIC: 163.51
```

Here, normal BMI is reference category and obese used as dummy variable. As the influence of obese on outcome compare to influence of normal is not statistically significant, so, we can say that we don’t have enough evidence to reject the null hypothesis.