Project 1, Part C: Analysis and Conclusions
Due Sunday, July 5 (Aim to be done sooner)

For each part, neatly present the indicated results and discussions in a Word document. Copy and paste the required plots and R output into the Word document next to your explanations. Save to PDF for upload to Canvas. There will be an upload spot for each part.

(If you don't already use it, you may want to learn about the "Snipping Tool" or "Snip and Sketch" that likely came with your PC. Free versions are available for the Mac. This will help you easily insert plots, etc, into your documents.)

You will also need to upload the PDF from the Jupyter notebook for each part. Use markdown cells to **clearly label the steps, particularly for Part 4.** This will make it easier for the TAs to comment on any problems with code or with your decisions.

Make sure you include your team number at the top of Part 1.

If your team chose to use a response variable constructed from before and after measurements: If your response variable (post – pre) has both negative and positive values and you need to transform your response variable, you'll find that certain transformations will not be possible. In that case, try using the ratio (post/pre) as your response variable.

Do not upload this page

Team #: 54

Part 1. Identify the following for your Project 1

Your TA needs to see the most current information on your project.

1. **Research questions**

   (a) (1 marks) Research Question 1:
   Does the sex of an adult Islander influence the level of blood cholesterol?

   (b) (1 marks) Research Question 2:
   Does the weight of an adult Islander influence the level of blood cholesterol?

2. **Variables**

   (a) (1 mark) What is your quantitative response variable?
   Blood cholesterol. Measurement unit: mg/dL.

   (b) (1 mark) What is your categorical explanatory variable?
   Sex of a person.

   (c) (1 mark) What is your quantitative explanatory variable?
   Weight of a person. Measurement unit: kg.

3. **Study details**

   (a) (0.5 mark) What is the target population?
   All adult Islanders.

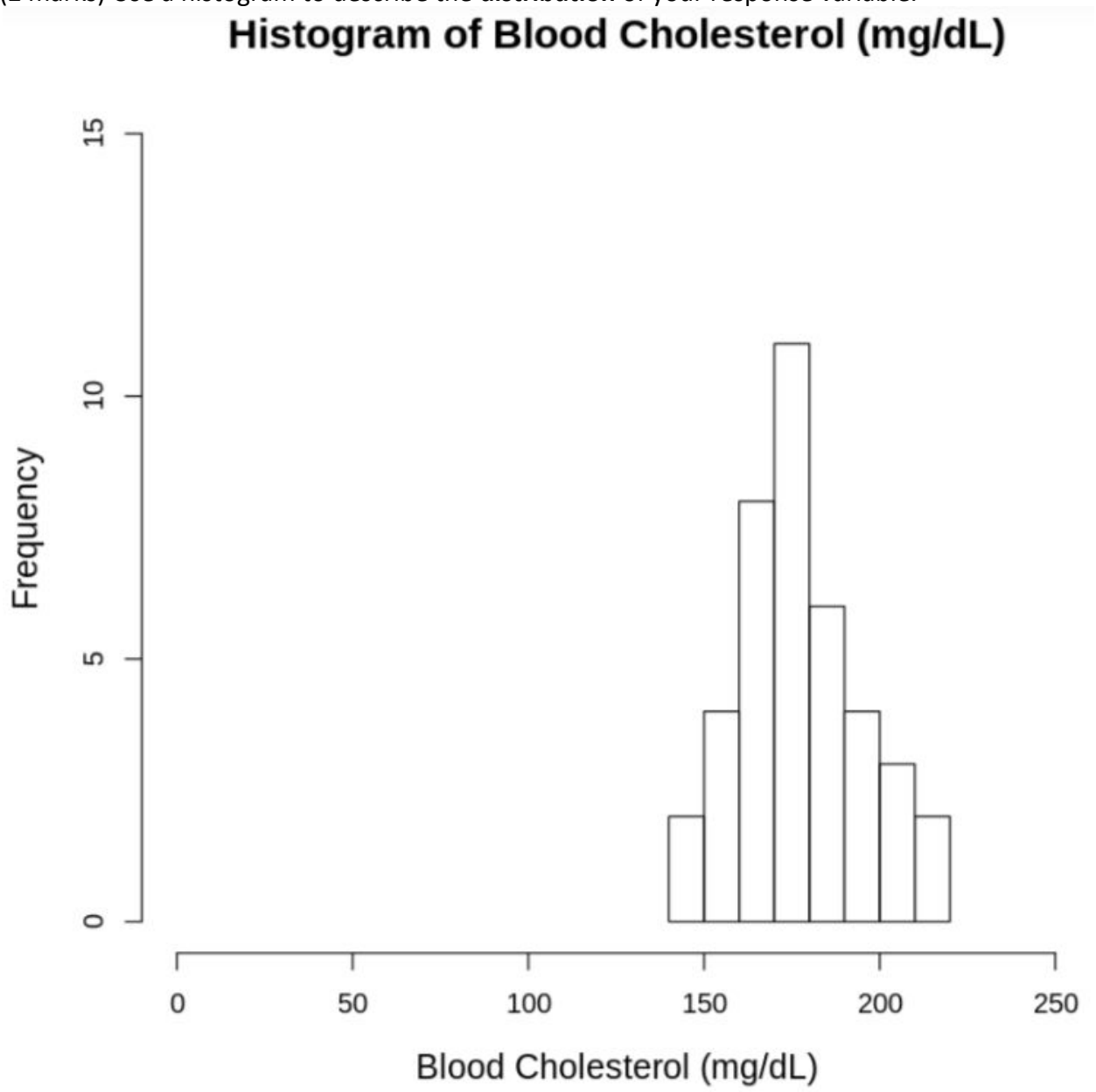
   (b) (0.5 mark) What is the study population?
   All adult Islanders that gave consent to the observational study.

   (c) (1 mark) How does your study population differ from your target population?
   There is a small difference between the target population and the study population because Islanders we select randomly may give or not give consent. So our study population is smaller than the target population.

Part 2. Data Verification (Med)

2.1 Continuous Response variable

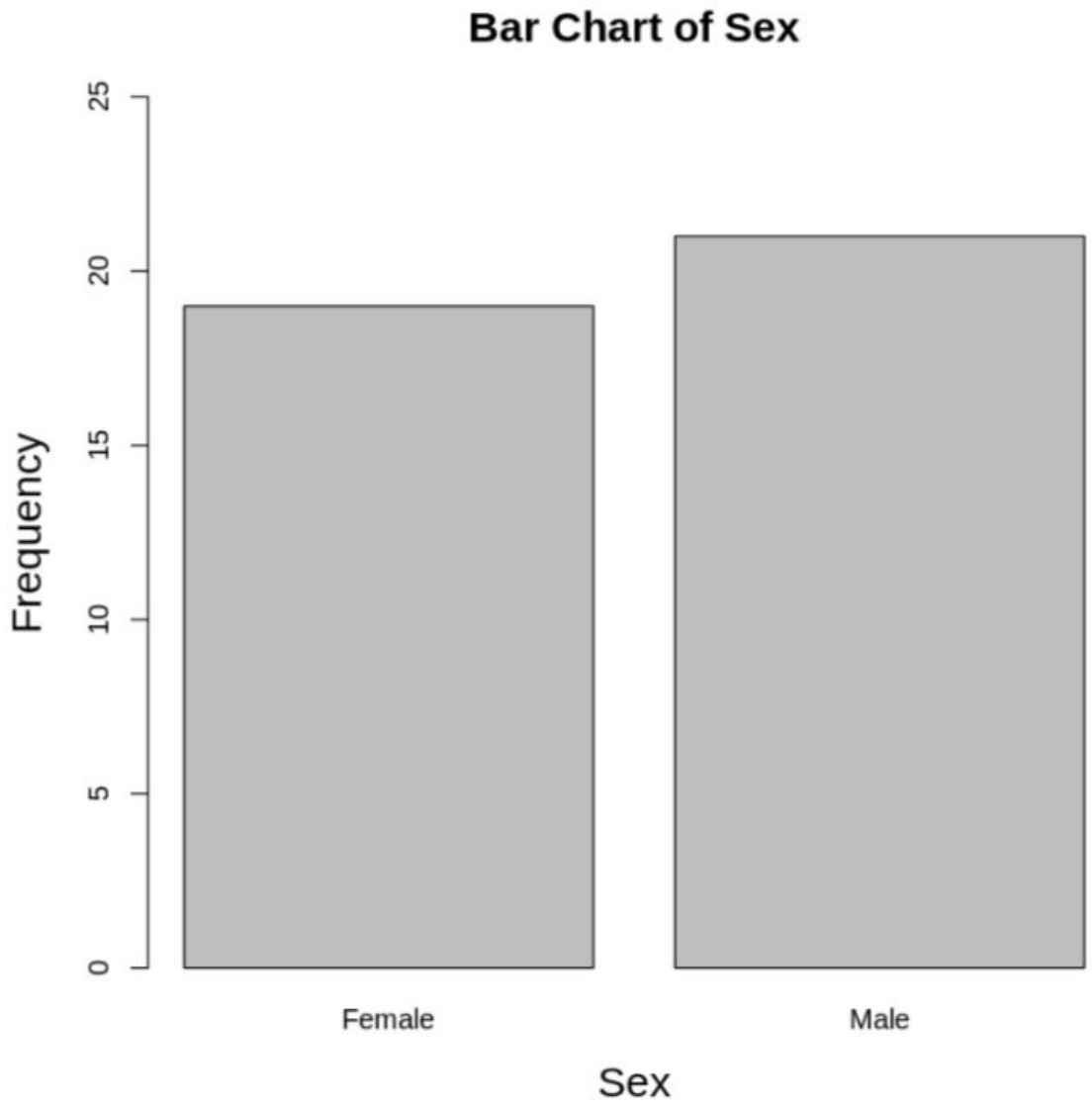
(a) (2 marks) Use a histogram to describe the **distribution** of your response variable.



(b) Make sure any outliers are not data entry errors (i.e. typos).
All observations fall within [Q1 - 1.5IQR] and [Q3+ 1.5IQR], so there are no outliers.

2.3 Categorical Explanatory variable

(a) (2 mark) Use a bar chart (bar plot) to describe the distribution of the data among the categories. (E.g. compare the categories.)

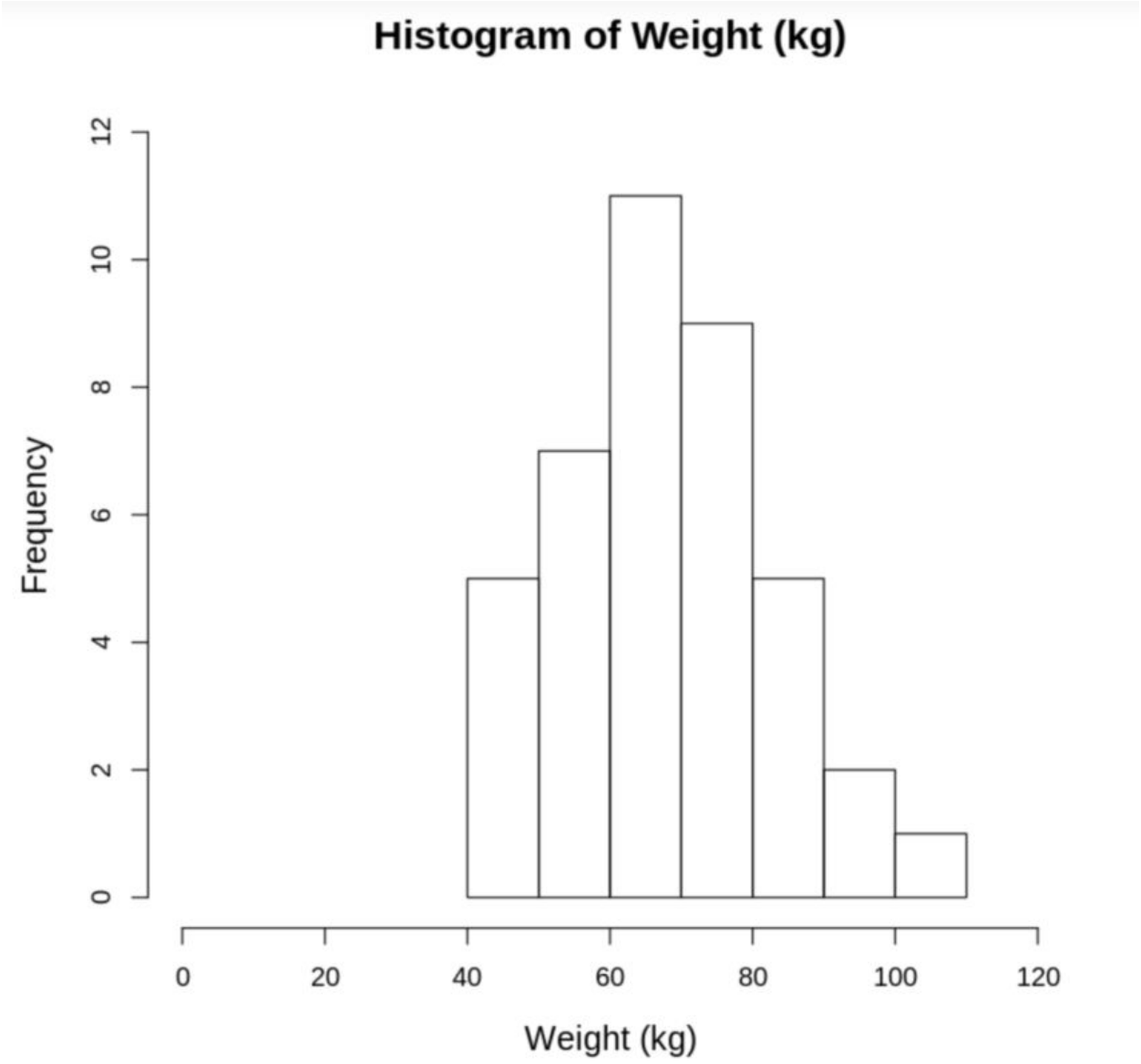


The distribution of sex looks correct. There are slightly more males than females. It is approximately half and half. The difference might have been smaller, if the sample size was larger.

(b) If any unexpected bars show up in your chart, this may be because of data entry error.

2.2 Quantitative Explanatory variable

(a) (2 marks) Use a histogram to describe the distribution of your quantitative explanatory variable.



(c) Make sure any outliers are not data entry errors (i.e. typos).
All observations fall within [Q1 - 1.5IQR] and [Q3+ 1.5IQR], so there are no outliers.

Part 3. Analysis for Response ~ Categorical Explanatory Variable

Perform a two-sample t-test to determine whether there is a difference in mean response between the categories of your categorical explanatory variable. Lecture 10 walks through an example. Example code for the pig study is provided in Canvas for Stat 302 under "Computing Resources." Clearly label all the steps in your write-up.

1. (3 marks) Hypotheses (Define your parameters. Also, include your choice of α-level.)

The null hypothesis is that there is no difference in mean response between female and male in our data, that is H_0: u1-u2 = 0, and our alternative hypothesis is that there is a difference, that is H_a: u1-u2 <0. The type-I error rate, α = 0.05.

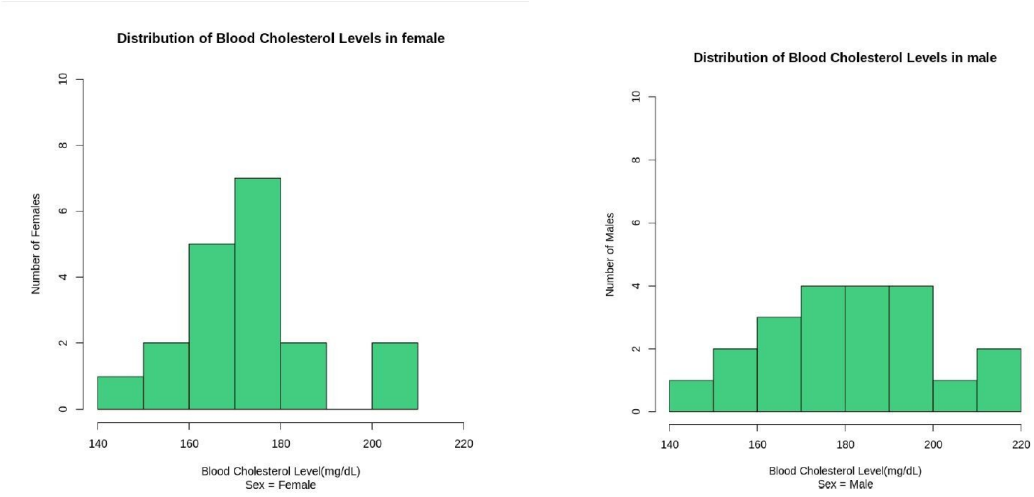
2. (2 marks) Assess whether the conditions for the test have been met.

40 samples have been randomly selected from the islands and there are four conditions that must be met to apply the two sample t-test.

i) The categorical explanatory variable is sex of adults. Because of random sampling we have two categories, which are male and female and can be separated into two simple random samples.

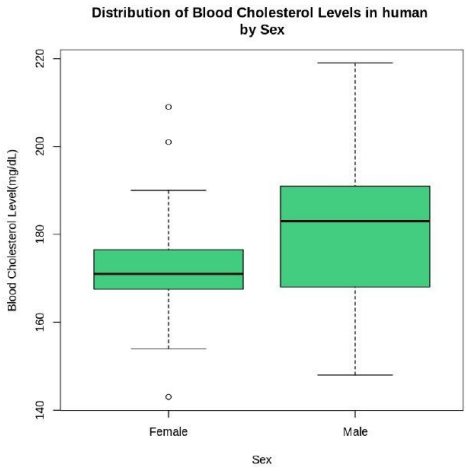
ii) As there is no way to prove that the samples are independent, it is assumed that they are.

iii)



Both the female and male blood cholesterol levels have a bi-modal shape, but we can still use the two sample t-tests that are robust to departures in normality.

iv)



From the box plot, it can be seen that the two sex have difference centres and interquartile range. The female group has few outliers as well. While the female plot looks more left skewed, the male plot looks more right skewed.

As the boxplots are not enough to decide on the fourth condition, there is a test for equality of variances.

```
#SUMMARY

#FEMALE
summary(Female.islanders$BC)
sd(Female.islanders$BC)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  143.0   167.5   171.0   173.3   176.5   209.0
```

```
15.4813602315779
```

```
#MALE
summary(Male.islanders$BC)
```



```
sd(Male.islanders$BC)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  148.0   168.0   183.0   181.1   191.0   219.0
```

```
19.2049100864485
```

```
#Difference between means = 181.1 - 173.3 = 7.8 mg/dL

#EQUALITY OF VARIANCE
  #RULE OF THUMB
  #Ratio of standard deviations = 15.48/19.20 = 0.81
```

Using the rule of thumb, the ratio of the standard deviations is 0.81, which is between 0.5 to 2, and therefore we can use the pooled version of the t-test.

3. (2 marks) Compute the test statistic and obtain the p-value **using R.**

```
#create a data frame for each group
# var.equal argument indicates whether the variances are equal (TRUE or FALSE)
Female.islanders = subset(data.islanders, Sex=="Female")
Male.islanders = subset(data.islanders, Sex=="Male")
t.test(x=Female.islanders$BC, y=Male.islanders$BC, alternative="less", var.
 ↪equal=TRUE)
```

```
Two Sample t-test

data:  Female.islanders$BC and Male.islanders$BC
t = -1.4094, df = 38, p-value = 0.08343
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.535984
sample estimates:
mean of x mean of y
 173.3158  181.1429
```

The p-value is 0.08343.

4. (2 marks) Write your conclusion, both in terms of the p-value and hypothesis, and in terms of your research question.

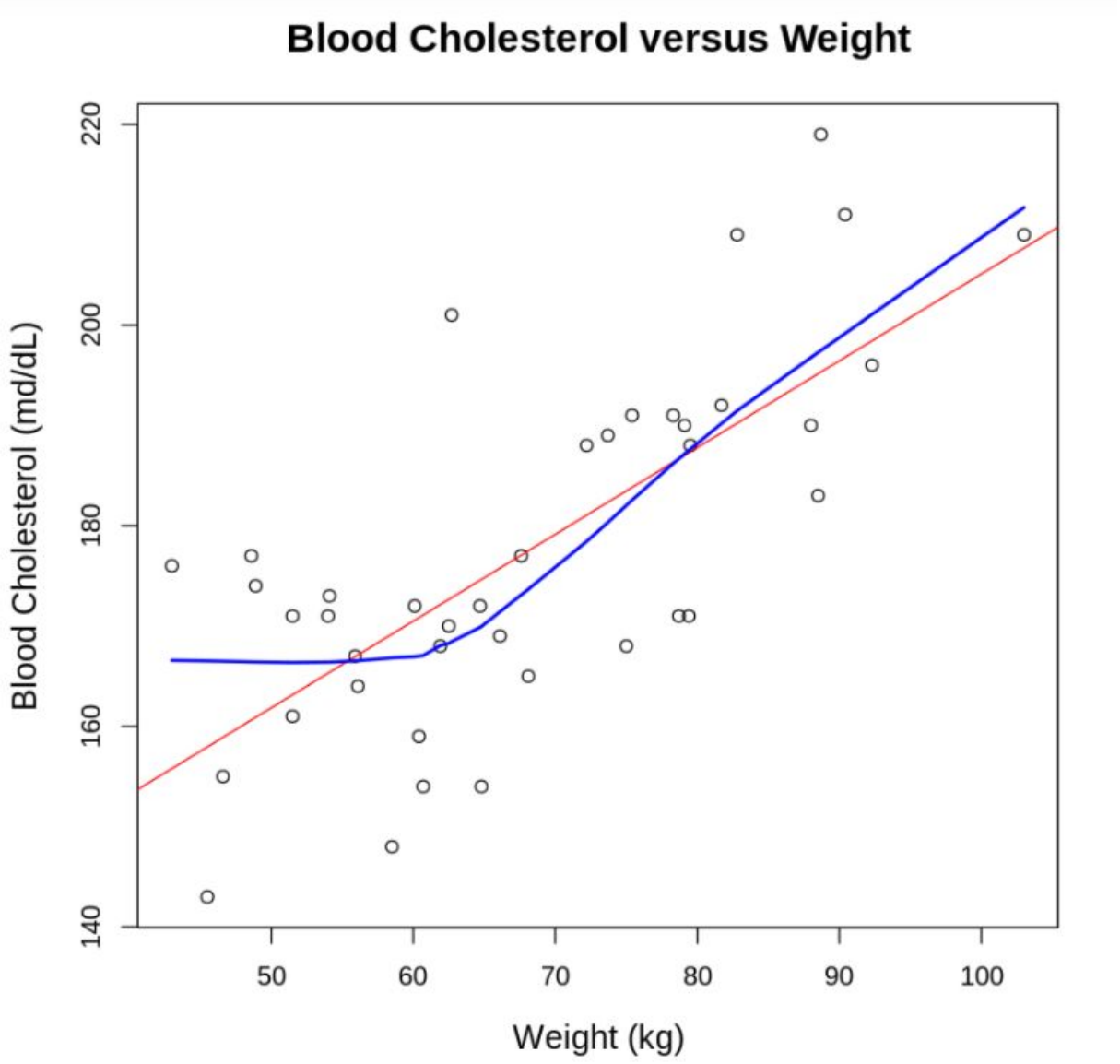
The p-value is 0.08343 and greater than the choice of type-I error rate which is 0.05. Therefore, there is no strong evidence to reject the null hypothesis. We cannot conclude that there is a significant difference in mean response of blood cholesterol levels between the two sex.

Part 4. Simple Linear Regression Analysis for Response ~ Quantitative Explanatory Variable

Example code for the Syringe scenario is provided in Canvas for Stat 302 under "Computing Resources."

Step 1: (Med)

(a) (2 marks) Create an appropriate scatterplot of your variables. Fit the scatterplot with a regression line. Add a lowess curve, if it helps you. Use good labelling.



(b) (2marks) Describe the scatterplot using the criteria provided in lecture. Do you anticipate needing to transform one or more variable?

Direction: Increasing, positive

Strength: r = 0.7261, so it is (moderately) strength

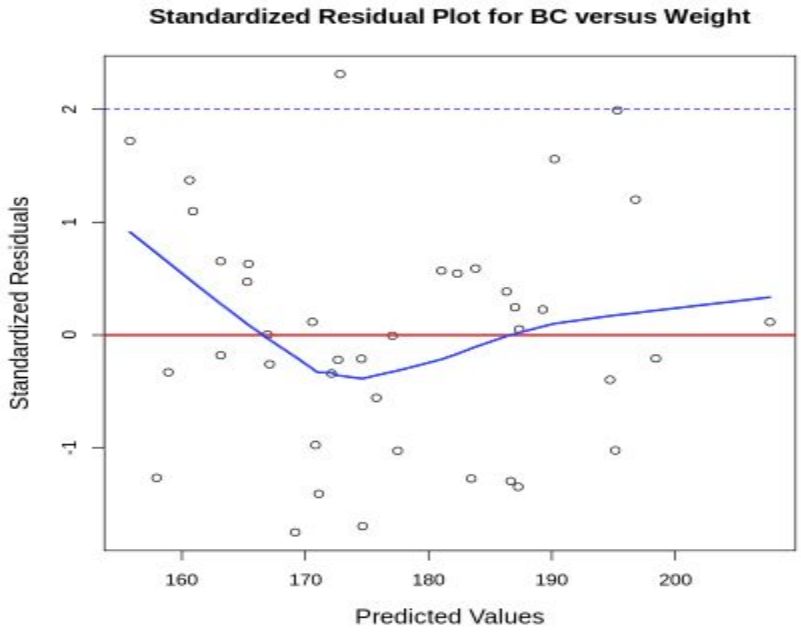
Form: Linear
Deviations: Nothing serious

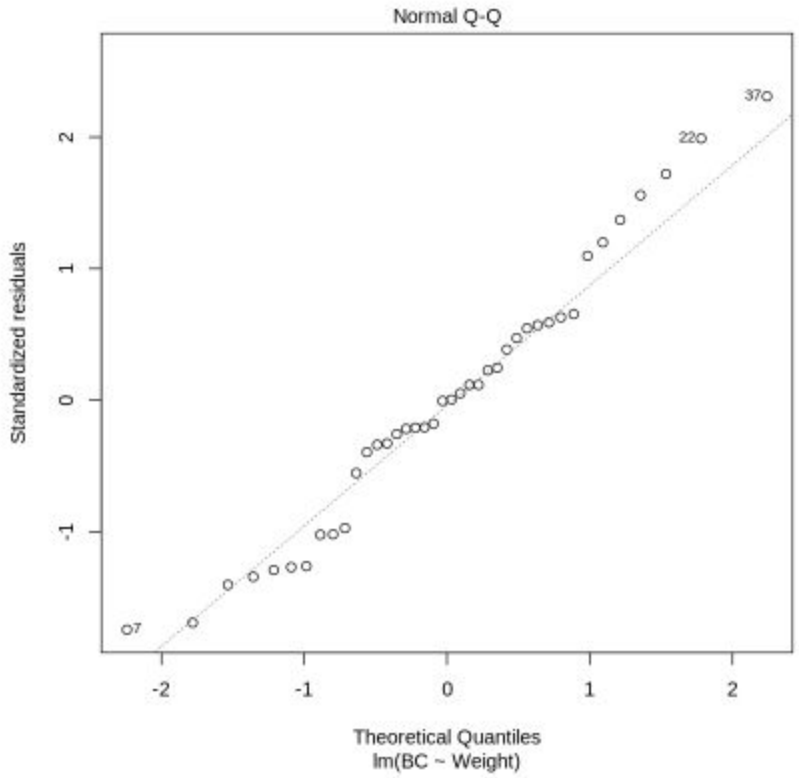
Step 2:(Jin)

(a) (1 mark) Fit a model using your original, **untransformed** response and explanatory variables. Present the output from the summary() function.

```
Call:
lm(formula = BC ~ Weight, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.1967  -8.0381  -0.0079   7.0108  28.1703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.5942     9.2459   12.83 2.21e-15 ***
Weight        0.8650     0.1329    6.51 1.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.37 on 38 degrees of freedom
Multiple R-squared:  0.5273,    Adjusted R-squared:  0.5148
F-statistic: 42.38 on 1 and 38 DF,  p-value: 1.139e-07
```

(b) (2 marks) Present a **standardized residual plot** and a **Normal Q-Q plot** (i.e. Normal Quantile plot) based on the model in Step2(a).





(c) (3 marks) Use the plots from Step2(b) and your knowledge about your sampling process to indicate whether each of the conditions have been adequately met. Also, describe any possible outliers.

- **Linearity** – this has not been met. There is an apparent curvature in the lowess line in Standardized Residual plot.
- **Constant variance** – this has been met. The vertical spread of residuals is sufficiently constant.
- **Independence** – this has been met. There is nothing showing that any other subject might influence the value.
- **Randomness** – this has been met. The data is randomly selected.
- **Normality** – this has been met. The Normal Q-Q plot shows some departure from normality for the highest and lowest residuals, but normality has been adequately met.
- **Outlier** – From the Standardized Residual plot, there is one point outside the second quartile, which is an outlier with low leverage. And there are two other points on the border of the second quartile with higher leverage.

(d) (1 mark) Can you use the model from Step2(a) as your "final" model? If not, what should your next step be? Explain.

No, because the condition of linearity has not been met, it is necessary to transform X to correct the issue of non-linearity.

If transformations are required, go to Step 3. If transformations are not needed, skip Step 3 and go on to Step 4.

Step 3: (Only if needed.) If you need this step and get stuck, contact your TA or visit Marie's office hours.

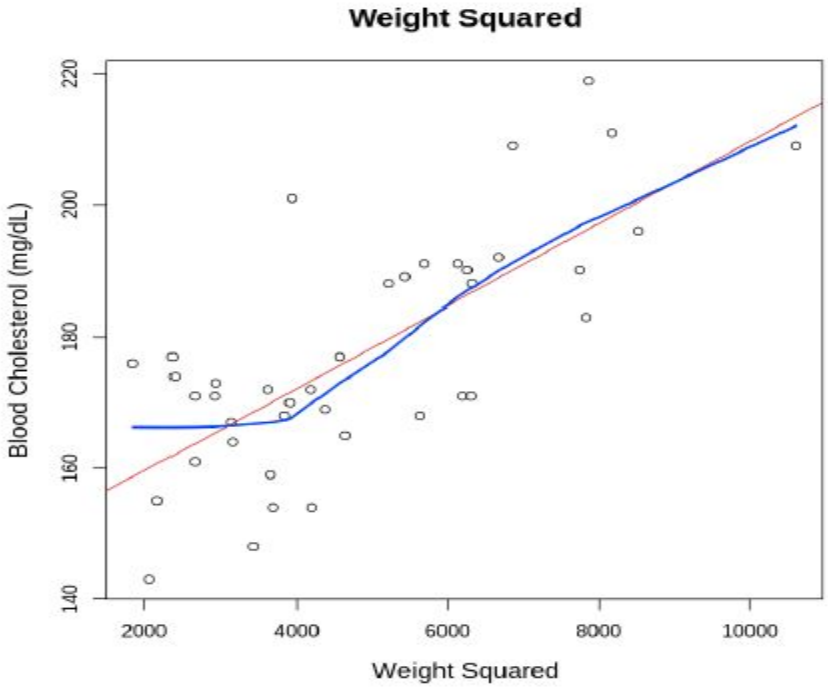
(a) Select 3 possible transformation functions and present a **standardized** residual plot for each. (No need to list the regression output at this stage.)

- Example code *(syringes.ipynb*) under Computing Resources on Canvas for Stat 302 generates the residual plots for several transformation choices. You can adapt this code to your needs. (The advanced code (*example with For loops.ipynb*) generates the plots more efficiently, but might be hard for a beginning coder to adapt.) Delete unneeded code (e.g. you don't need studentized residual plots.)
- If you transformed Y (response) to reduce non-constant variance, but still have a lack of linearity, remember that you can try adding a transformation of X.

(b) Select the model that results in the best improvement in the conditions. You can use the residual plots and r^2 to make this judgement.

Chosen model with X = (Weight)^2.

The variability looks fairly constant and the relationship looks efficiently linear.



Step 4: (For everyone)

4. (2 marks) Present the summary() output for your final model. Write out your final fitted model.

```
Call:
lm(formula = BC ~ Wt_sq, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-20.5462  -7.0349  -0.5042   7.6361  29.2713

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.472e+02  4.857e+00  30.298  < 2e-16 ***
Wt_sq       6.252e-03  9.218e-04   6.782 4.86e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.1 on 38 degrees of freedom
Multiple R-squared:  0.5476,    Adjusted R-squared:  0.5357
F-statistic:    46 on 1 and 38 DF,  p-value: 4.858e-08
```

The final fitted model:

BC = 147.2 + 0.006252Wt_sq

5. (3 marks) Report the **t-statistic and p-value for your slope.** Provide a conclusion for this test. Also provide a 95% confidence interval for the slope.

t = 6.782 **p- value = $4.858*10^{-8}$**

Conclusion:

For the p-value < 0.0001, we can reject H_0 and conclude that the slope is different from zero. So, there is a linear association between Blood Cholesterol and (Weight)^2.

95% confidence interval for the slope:

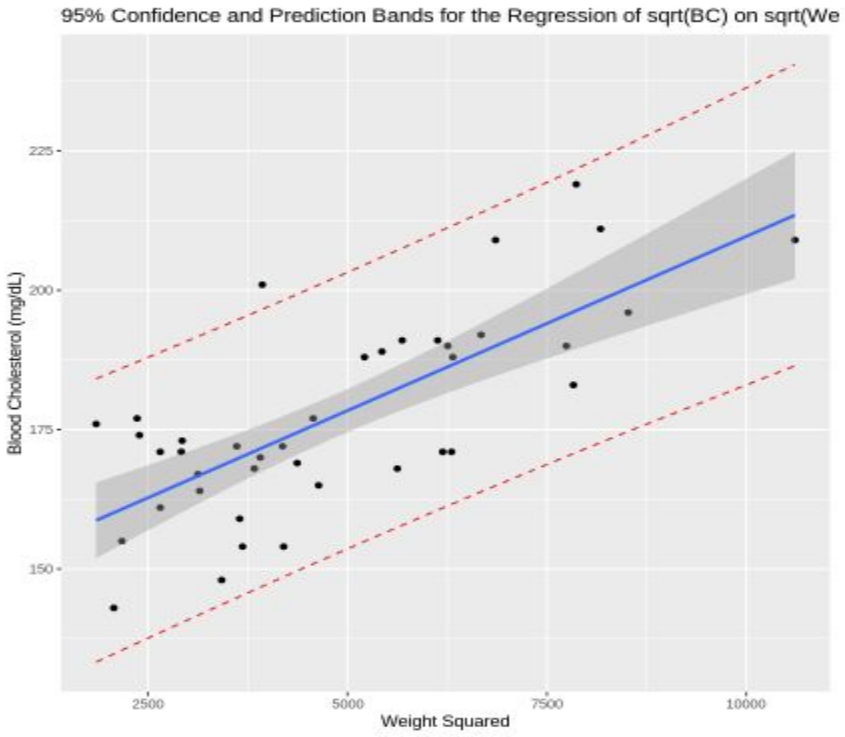
	2.5 %	97.5 %
(Intercept)	1.373187e+02	1.569827e+02
Wt_sq	4.385732e-03	8.118061e-03

6. (1 mark) Use the coefficient of determination (r^2) from your final model to discuss how much of the variation in your response variable is explained by your final model.

r^2 = 0.5476 **r = $\sqrt{0.5476}$ = 0.7400**

This indicates that 74.00% of the variation in Blood Cholesterol is explained by the regression on Weight.

7. (2 marks) Create a scatterplot with the variables from your **final** model. Add the confidence band and prediction band to the scatterplot. Present the resulting plot. (Example code is on Canvas.)



- No need to comment. Just compare the two bands and notice their shape. Where do you get the best precision (narrowest interval)? Where do you get the worst precision?

Best precision: $(\text{weight})^2$ = around 5000
Worst precision: at the end, around 10000

8. Choose a value of the explanatory variable within the range of your actual data.
choose Weight = 70kg, Wt_sq = 4900

   i. (1 mark) Compute the predicted value at that value of your explanatory variable.
from fitted model: BC = 147.2 + 0.006252*4900 = 177.8348 mg/dL

compute predicted BC = 177.785 mg/dL

   ii. (2 marks) Compute the confidence interval for that predicted value. Interpret the confidence interval.

	fit	lwr	upr
1	177.785	173.9098	181.6601

CI: [173.9098, 181.6601]

iii. (2 marks) Compute the prediction interval for the same predicted value. Interpret the prediction interval.

	fit	lwr	upr
1	177.785	152.9812	202.5887

PI: [152.9812, 202.5887]

Note:

- If you transformed your **explanatory** variable, don't forget to transform your selected value for X before using the model to make predictions/estimations.
- If you transformed your **response** variable, then you will need to back-transform your point estimates and interval endpoints to get your final intervals.