FS21 STT 180 Homework 3

Kaitlyn Watson

November 11-27, 2021

Setting up:

Load `tidyverse` (which includes `dplyr`, `ggplot2`, `tidyr`, and other packages), `infer`, `knitr` and `broom` packages.

```
library(tidyverse)
library(infer)
library(knitr)
library(broom)
```

Homework 3 has two sections. In Section 1, you will deal with inference. In Section 2, you will work with linear models.

General Instructions:

- This is an individual assignment. You may consult with others as you work on the assignment, but each student should write up a separate set of solutions.
- Rather than creating a new Rmd file, just add your solutions to the supplied Rmd file. **Zip and submit both the Rmd file and the resulting HTML file to D2L.**
- Please note if you are **compiling to a pdf, change the option in the YAML and also delete the .css lines in the file.**
- Except for questions, or parts of questions, that ask for your commentary, use R in a code chunk to answer the questions.
- The code chunk option `echo = TRUE` is specified in the setup code chunk at the beginning of the document. Please do not override this in your code chunks.
- A solution will **lose points if the Rmd file does not compile**. If one of your code chunks is causing your Rmd file to not compile, you can use the `eval = FALSE` option. Another possibility is to use the `error = TRUE` option in the code chunk.
- This Homework is due on **Saturday, November 27th, 2021 on or before 11 pm.**

Section 1

For the first section of this homework will use the `Breast_Cancer.csv` file. There are 10 quantitative variables and a binary dependent variable indicating the presence or absence of breast cancer. The predictors are anthropometric data which can be gathered in routine blood analysis.

Read in the data and convert the data frame to a tibble.

```
breast_cancer <- read.csv("Breast_Cancer.csv", header = TRUE)
breast_cancer <- as_tibble(breast_cancer)
```

A glimpse of the data:

```
glimpse(breast_cancer)
```

```
Rows: 116
Columns: 10
$ Age            <int> 48, 83, 82, 68, 86, 49, 89, 76, 73, 75, 34, 29, 25, 24,~
$ BMI            <dbl> 23.50000, 20.69049, 23.12467, 21.36752, 21.11111, 22.85~
$ Glucose        <int> 70, 92, 91, 77, 92, 92, 77, 118, 97, 83, 78, 82, 82, 88~
$ Insulin        <dbl> 2.707, 3.115, 4.498, 3.226, 3.549, 3.226, 4.690, 6.470,~
$ HOMA           <dbl> 0.4674087, 0.7068973, 1.0096511, 0.6127249, 0.8053864, ~
$ Leptin         <dbl> 8.8071, 8.8438, 17.9393, 9.8827, 6.6994, 6.8317, 6.9640~
$ Adiponectin    <dbl> 9.702400, 5.429285, 22.432040, 7.169560, 4.819240, 13.6~
$ Resistin       <dbl> 7.99585, 4.06405, 9.27715, 12.76600, 10.57635, 10.31760~
$ MCP.1          <dbl> 417.114, 468.786, 554.697, 928.220, 773.920, 530.410, 1~
$ Classification <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

The variables in the data set are:

Variable	Description
`Age`	age in years.
`BMI`	the body mass index.
`Glucose (mg/dL)`	the fasting glucose level (mg/dL).
`Insulin (µg/mL)`	amount of insulin.
`HOMA`	Homeostasis Model Assessment.
`Leptin (ng/mL)`	type of adipocytokines
`Adiponectin (µg/mL)`	a protein hormone.
`Resistin (ng/mL)`	cysteine-rich peptide hormone.
`MCP-1(pg/dL)`	Monocyte chemoattractant protein-1 (MCP-1)
`Classification`	1= Healty control, 2= Breast Cancer Patients.

Make sure to familiarize yourself with the data by reading about the variables on the website. Note that the data comes the study https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1.

According to the CDC (https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html), a BMI between 18-25 is considered normal.

Using the `breast_cancer` data, let's investigate whether breast cancer patients have a normal BMI on average (considering 25 as normal).

Question 1

What is the parameter for the above investigation?

The parameter is the mean

Calculate sample statistic.

```
 avgBMI_BC<-breast_cancer%>%
 filter(Classification==1) %>%
 summarise(avgBMI_BC=mean(BMI))
avgBMI_BC
```

```
# A tibble: 1 x 1
  avgBMI_BC
      <dbl>
1      28.3
```

```
avgBMI_Healthy<-breast_cancer%>%
 filter(Classification==2) %>%
 summarise(avgBMI_healthy=mean(BMI))
avgBMI_Healthy
```

```
# A tibble: 1 x 1
  avgBMI_healthy
           <dbl>
1           27.0
```

Is it a continuous or categorical sample statistic?

This is a continuous sample statistic.

Question 2

a. Set up and test the hypotheses to determine whether breast cancer patients have higher than normal BMI (25). (Hint: Follow the hypothesis process step-wise as you have done in your Module 5 group assignments and think about the direction of the alternative hypothesis.)

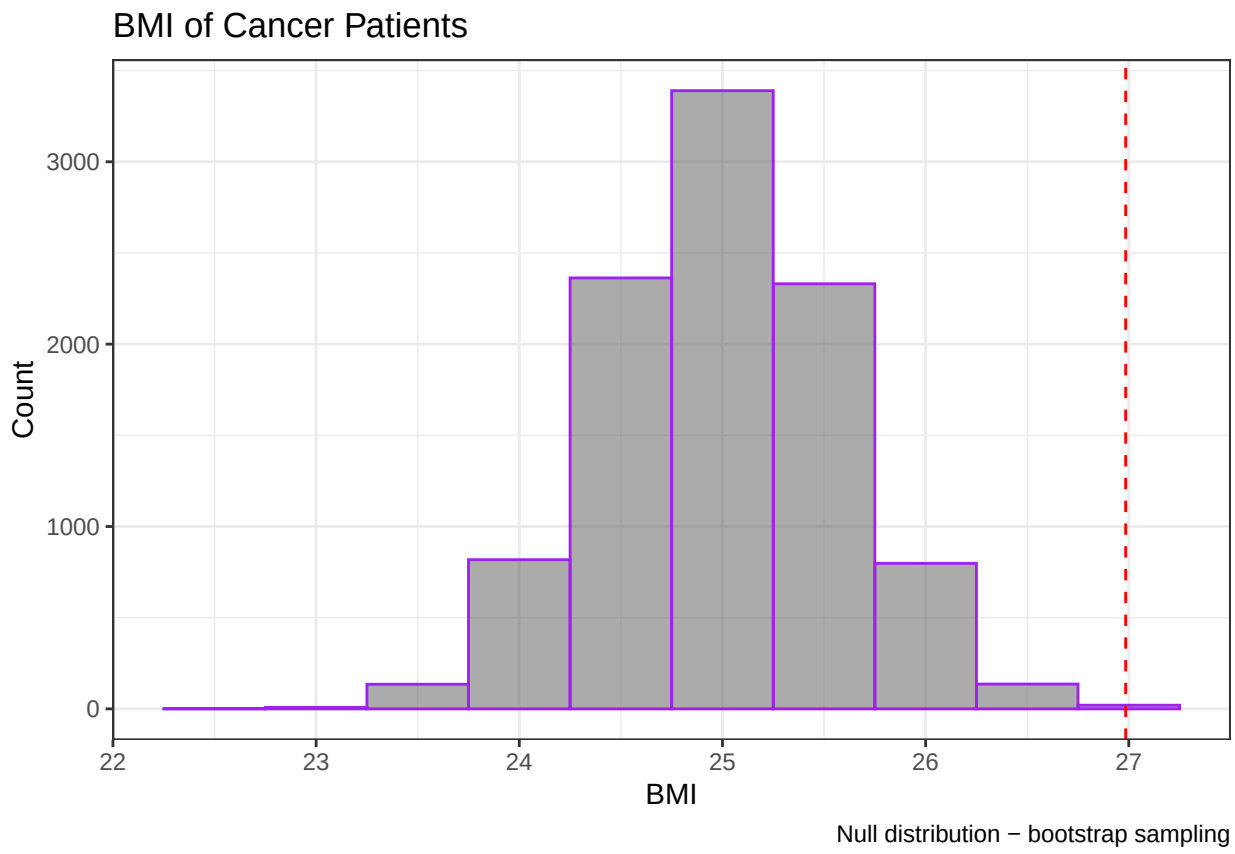
- State the null and alternative hypotheses.

The null hypotheses: Cancer Patients have a normal mean BMI. The alternative Hypothesis: Cancer patients have a mean BMI >25.

- Generate the null distribution and plot the distribution.

```
set.seed(54321)
null.dist <- breast_cancer %>%
  filter(Classification==2) %>%
  specify(response = BMI) %>%
  hypothesize(null = "point", mu = 25) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

null.dist %>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .5, color = "purple", alpha = .5) +
  labs(x = "BMI", y = "Count",
       title = "BMI of Cancer Patients",
       caption = "Null distribution - bootstrap sampling") +
  theme_bw()+
  geom_vline(xintercept = 26.98474, color="red", linetype="dashed")
```



b. Determine the p-value and compare it to $\alpha = 0.05$.

```
null.dist %>%
  filter(stat > 25) %>%
  summarise(p_value = 2 * n() / nrow(null.dist))
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.998
```

c. Conclude and interpret the results.

Because the calculated p-value is 1.0092 and the data was tested at the 0.05 significance level, there is not sufficient evidence to reject the null hypothesis. Therefore, we cannot confidently say that the cancer patients have a higher than normal BMI (25).

Question 3

a. Estimate 95% confidence interval for average BMI of breast cancer patients.

```r
set.seed(4321)
# bootstrap samples
boot.means <- breast_cancer %>%
  filter(Classification==2) %>%
  specify(response = BMI) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
# cutoff bounds
boot.means %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975),
            lower99 = quantile(stat, probs = 0.005),
            upper99 = quantile(stat, probs = .995))
```

```
# A tibble: 1 x 4
  lower95 upper95 lower99 upper99
    <dbl>   <dbl>   <dbl>   <dbl>
1    25.8    28.1    25.5    28.4
```

```r
# save as vector
results<-c(boot.means)
```

b. Interpret the 95% confidence interval.

Using the confidence interval above, we can conclude with 95% confidence that the population mean for the BMI of breast cancer patients falls between 25.8446 and 28.10133.

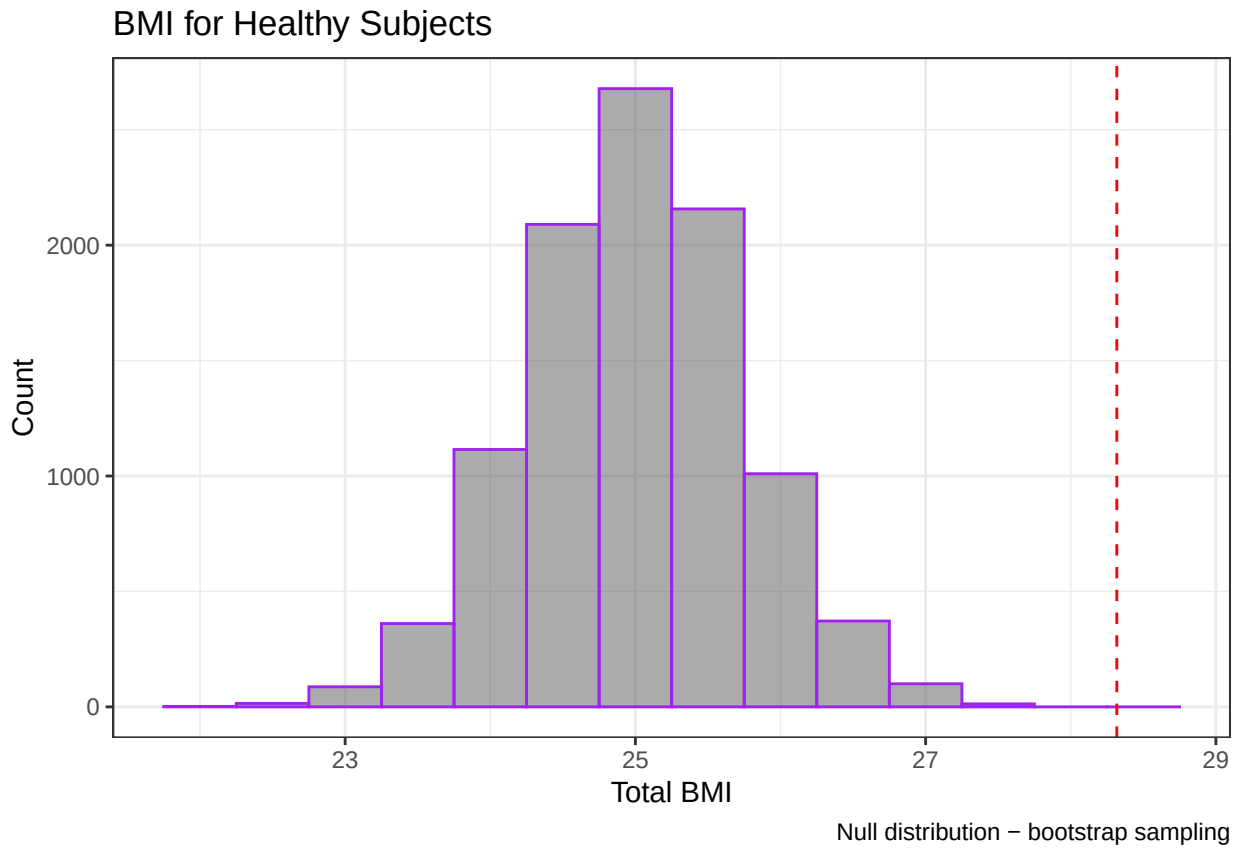
Question 4

Is having a higher than normal BMI an indicator of increased risk of breast cancer given your results in 2 and 3? (Hint: Consider the BMI of people that don't have breast cancer (healthy control). Run the hypothesis test and estimate the 95% confidence interval to check the your conclusion.)

```r
#null hypothesis: Healthy People have a normal risk of breast cancer with any BMI.
#Alternative Hypothesis: Healthy people have a higher risk of breast cancer with a BMI>25.

null.dist2 <-breast_cancer %>%
  filter(Classification==1) %>%
  specify(response = BMI) %>%
  hypothesize(null = "point", mu = 25) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

  null.dist2 %>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .5, color = "purple", alpha = .5) +
  labs(x = "Total BMI", y = "Count",
       title = "BMI for Healthy Subjects",
       caption = "Null distribution - bootstrap sampling") +
  theme_bw()+
  geom_vline(xintercept = 28.31734, color="red", linetype="dashed")
```



Null distribution − bootstrap sampling

```
  null.dist2 %>%
  filter(stat > 25) %>%
  summarise(p_value = 2 * n() / nrow(null.dist))
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1    1.01
```

```
  boot.means2 <- breast_cancer %>%
  filter(Classification==1) %>%
  specify(response = BMI) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

  boot.means2 %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975),
            lower99 = quantile(stat, probs = 0.005),
            upper99 = quantile(stat, probs = .995))
```

```
# A tibble: 1 x 4
  lower95 upper95 lower99 upper99
    <dbl>   <dbl>   <dbl>   <dbl>
1    26.8    29.8    26.4    30.2
```

```
  results2<-c(boot.means2)
```

Justify your answer in 3-4 sentences.

After running the hypothesis test and calculating the p-value 0f 0.9898, we fail to reject the null hypothesis at the 0.05 significance level. Therefore, there is not sufficient evidence to reject the null hypothesis which states that healthy people have a normal risk of heart disease at any BMI. Using a 95% confidence interval, we can confidently say the population mean is between the values 26.89 and 29.76.

Section 2

For this section of this homework will use the`abalone.csv` file from UCI repository (https://archive.ics.uci.edu/ml/datasets/Abalone).

The number of rings in the shell of an abalone is indicative of its age. This is done by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. In this section, we will analyze the relationship between age (measured by the number of rings) and a few different variables present in the data.

```
ab <- read.csv("abalone.csv")
glimpse(ab)
```

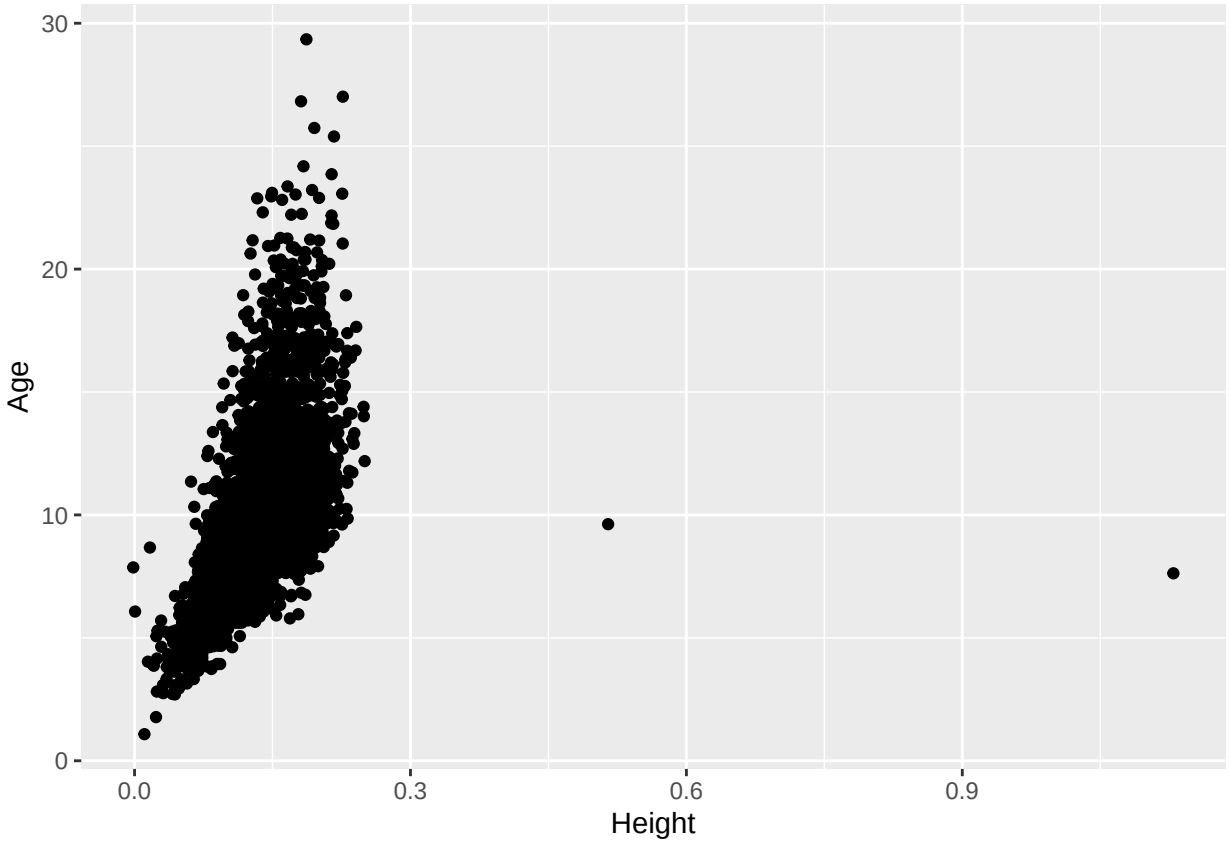
```
Rows: 4,177
Columns: 11
$ Sex      <chr> "M", "M", "F", "M", "I", "I", "F", "F", "M", "F", "F", "M", "~
$ Length   <dbl> 0.455, 0.350, 0.530, 0.440, 0.330, 0.425, 0.530, 0.545, 0.475~
$ Diameter <dbl> 0.365, 0.265, 0.420, 0.365, 0.255, 0.300, 0.415, 0.425, 0.370~
$ Height   <dbl> 0.095, 0.090, 0.135, 0.125, 0.080, 0.095, 0.150, 0.125, 0.125~
$ Whole    <dbl> 0.5140, 0.2255, 0.6770, 0.5160, 0.2050, 0.3515, 0.7775, 0.768~
$ Shucked  <dbl> 0.2245, 0.0995, 0.2565, 0.2155, 0.0895, 0.1410, 0.2370, 0.294~
$ Viscera  <dbl> 0.1010, 0.0485, 0.1415, 0.1140, 0.0395, 0.0775, 0.1415, 0.149~
$ Shell    <dbl> 0.150, 0.070, 0.210, 0.155, 0.055, 0.120, 0.330, 0.260, 0.165~
$ Rings    <int> 15, 7, 9, 10, 7, 8, 20, 16, 9, 19, 14, 10, 11, 10, 10, 12, 7,~
$ X        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ X.1      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

Question 1

We will start by analyzing a simple bivariate relationship between age and height.

(a) Plot a scatter plot to get an idea about the relationship between height and age.

```
ab %>%
  ggplot(mapping=aes(x=Height, y=Rings))+
  geom_jitter()+
  labs(x="Height", y="Age")
```



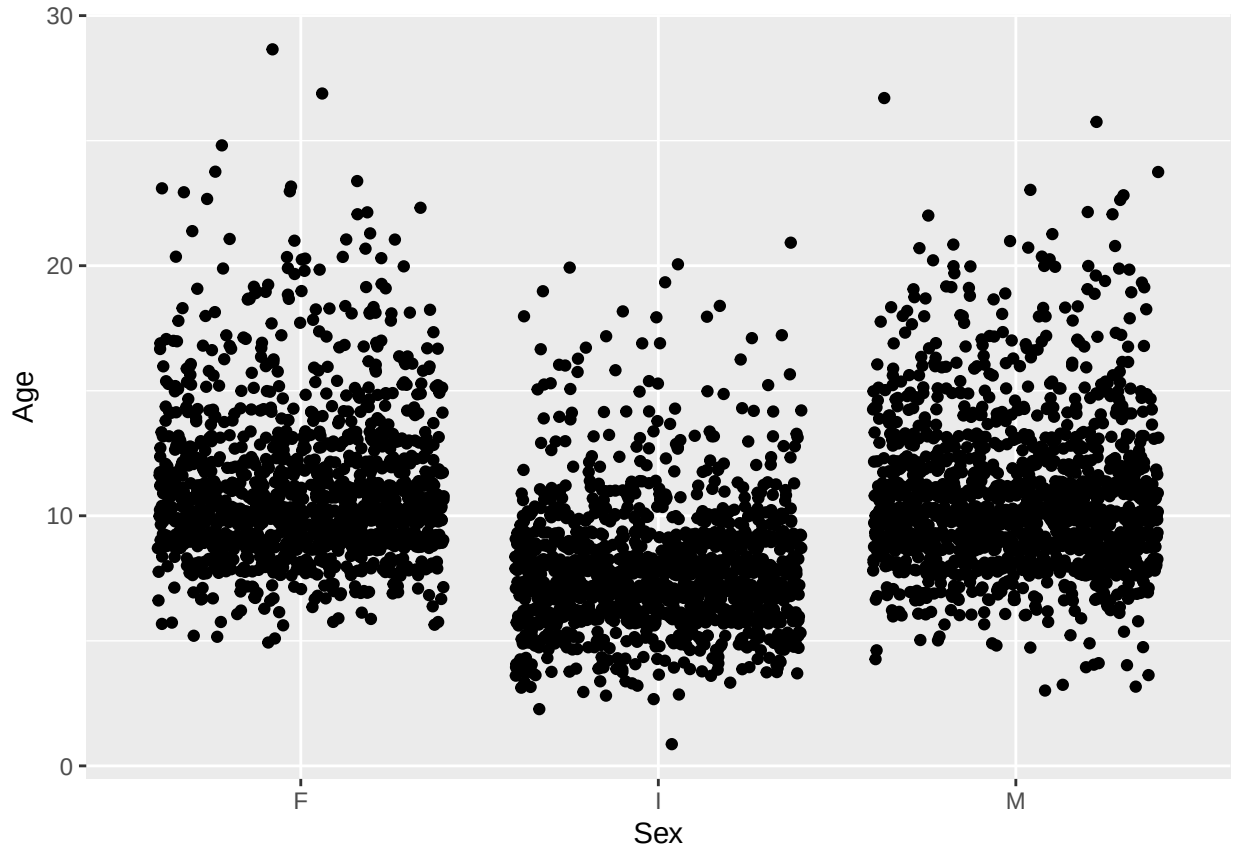
(b) Comment (1-3 sentences) on the plot.

The scatter plot clearly shows an increasing trend. There seems to be two clear outliers, while the rest of the plot is clustered together in majority of the same area on the plot.

Question 2

(a) Modify the plot in Question 1 to reflect the effect of the variable `Sex` in the plot.

```
ab %>%
  ggplot(mapping=aes(x=factor(Sex), y=Rings))+
  geom_jitter()+
  labs(x="Sex", y="Age")
```



(b) Is there any significant pattern or distribution based on the `Sex` of the abalone?

All of the plots seems to look relatively similar. However, the infant category clearly has a much shorter age range, clustering majority around the age of 10. There clearly seems to be a decrease in abalones as the age increases in all three categories.

Question 3

One of the goal is to study if there is significant difference in the age of the abalone based on shell weight, height, and diameter.

(a) Fit a multiple regression model to test the effect of the three variables on the age of the abalone.

```
m.ab<- lm(Rings~Shell + Height+ Diameter, data=ab)
m.ab %>%
tidy()
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     5.88     0.249     23.6  6.37e-116
2 Shell          12.6      0.678     18.6  3.58e- 74
3 Height         11.2      1.74       6.45 1.26e- 10
4 Diameter       -1.28     0.993     -1.29 1.97e-  1
```

```
m.ab %>%
  glance() %>%
  select(r.squared, adj.r.squared)
```

```
# A tibble: 1 x 2
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.400         0.400
```

(b) Comment on the effect of the three variables based on their direction, magnitude, and significance.

Rings = 5.881870+12.612867 x Shell + 11.198283 x Height - 1.281921 x Diameter

The shell and the height clearly produce an increasing effect with a high magnitude on the Age, while the Diameter produces a decreasing effect and smaller magnitude. The p-value for shell and height also have very small p-values meaning that they will have a large significance on the calculated R squared which therefore indicates a large impact on the the dependent variable of age. On the other hand, the diameter has a relatively high p-value which means that it will produce a lower R squared value and does not contribute much to the variability present.

Question 4

Can the model in Question 3 be improved to make it more parsimonious? Does it significantly change the model fit parameters?

Yes, we can remove the Diameter as a predictor.

```
m2.ab<- lm(Rings~Shell + Height, data=ab)
m2.ab %>%
tidy()
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     5.63     0.151     37.3  2.40e-263
2 Shell          12.0      0.482     24.9  1.38e-127
3 Height         10.3      1.60       6.44 1.29e- 10
```

```
m2.ab %>%
  glance() %>%
  select(r.squared, adj.r.squared)
```

```
# A tibble: 1 x 2
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.400         0.400
```

After removing the Diameter, the Shell and Height parameters decreased only slightly in the equation. Additionally, the p-values changed slightly as well and resulted in actually a lower R squared value which is not what we would want for a best fit model. ### Question 5

(a) How about the variable `Sex`? Does it have any significant impact on predicting the `Age` if included in the model from Question 4?

```
m3.ab<- lm(Rings~Shell + Height+factor(Sex), data=ab)
m3.ab %>%
tidy()
```

```
# A tibble: 5 x 5
  term         estimate std.error statistic   p.value
  <chr>           <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     6.53     0.188      34.8  3.21e-233
2 Shell          10.9      0.494      22.2  6.84e-103
3 Height          8.23     1.61        5.12 3.27e-  7
4 factor(Sex)I   -0.926    0.113      -8.18 3.89e- 16
5 factor(Sex)M   -0.150    0.0936     -1.60 1.09e-  1
```

```
m3.ab %>%
  glance() %>%
  select(r.squared, adj.r.squared)
```

```
# A tibble: 1 x 2
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.410         0.410
```

Yes, using sex as a factor created a higher R squared and adjusted R squared variable. Therefore, this would be the ideal model to use as a higher R squared value indicates that these predictors have a greater impact on age and will provide explanation for any variability in the results.

(b) Write out the best fit model based on your results.

Age = 6.525 + 10.937 x Shell + 8.233 x Height - 0.926 x Sex(I) - 0.150 x Sex (M)

Question 6

Interpret the results from the model in Question 5. Interpret each coefficient in the model. What does each coefficient signify?

The y-intercept is the age of the abalone if Shell, Height, and Sex were all set to zero. The coefficient of shell signifies that for each value put in for shell, the change in mean would increase by 10.937 if all other predictors were held constant. This is similar for height as well except the the change in mean would increase by 8.233 while all other values are held constant. Sex is a factor represented by male, female, or infant. In this model, female is set to zero. Therefore the sex infant has a coefficient of -0.926 signifying the different from the female sex. Furthermore, the sex male has a coefficient of -0.150 also indicating the difference from the mean of the female sex, Overall, all of these predictors contribute to the age of the abalone and produce the best fit model.

Question 7:

(a) What is the predicted value and the prediction interval of the `Age` of a female abalone whose `Shell` weight is 0.768 gram and `Height` is 0.95mm?

```
p1=data.frame(Sex ="F",Shell = 0.768,Height=0.95)
predict(m3.ab, p1, interval="predict")
```

```
      fit     lwr      upr
1 22.7459 17.4192 28.07259
```

(b) What is the predicted value and the prediction interval of the `Age` of a infant abalone whose `Shell` weight is 0.0010 gram and `Height` is 1.25mm? (Note the values of the predictor variables.)

```
p2=data.frame(Sex ="I",Shell = 0.0010,Height=1.25)
predict(m3.ab, p2, interval="predict")
```

```
       fit      lwr      upr
1 15.90111 9.796825 22.00539
```

The values of the predictor variables are much smaller indicating that we would also result in both a smaller prediction and overall interval. This is due to the model we used above.

Reference for both prediction equations: R Tutorial Accessed On November 27,2021. http://www.r-tutor.com/elementary-statistics/simple-linear-regression/prediction-interval-linear-regression

Essential details

Deadline and submission

- The deadline to submit Homework 3 is **11:00pm on Saturday, 27th November, 2021.**
- This is a individual assignment. Save your file with your name and write your name in the YAML header. Points will be deducted if these are not done.
- Submit your work by uploading your RMD and HTML/PDF files through D2L. Kindly **double check your submission to note whether the everything is displayed in the uploaded version of the output in D2L or not.** If submitting HTML outputs, please zip the .rmd and html files for submission.
- Please do not disable the `echo=TRUE` option in your global code chunk.
- Please **name all the code chunks**.
- **Late work will not be accepted except under certain extraordinary circumstances.**

Help

- Post general questions in the Teams HW 3 channel. If you are trying to get help on a code error, explain your error in detail
- Feel free to visit us in during our virtual office hours or make an appointment.
- Communicate with your classmates, but do not share snippets of code.
- The instructional team **will not answer any questions within the first 24 hours of this homework being assigned, and we will not answer any questions after 6 P.M of the due date**.

Academic integrity

This is an individual assignment.You may discuss ideas, how to debug code, and how to approach a problem with your classmates in the discussion board forum. You may not copy-and-paste another's code from this class. As a reminder, below is the policy on sharing and using other's code.

> Similar reproducible examples (reprex) exist online that will help you answer many of the questions posed on group assignments, and homework assignments. Use of these resources is allowed unless it is written explicitly on the assignment. You must always cite any code you copy or use as inspiration. Copied code without citation is plagiarism and will result in a 0 for the assignment.

Grading

Use the R Markdown blank file that is provided. If you want you can use your own formatting. Self-formatting is at your discretion but is graded. Use the in-class assignments and resources available online for inspiration. Another useful resource for R Markdown formatting is available at: https://holtzy.github.io/Pimp-my-rmd/

Topic	**Points**
Questions(total 11) and communication of results	85
R Markdown formatting and knitting	7
Code style	8
Total	**100**

Please note: Code style includes code efficiency, naming code chunks, etc.

Reference

https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

https://archive.ics.uci.edu/ml/datasets/Abalone