SS21 STT 180 Homework 3

Derien W

April 1-17, 2021

Contents

Setting up:

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

```
library(tidyverse)
library(infer)
library(knitr)
library(broom)
library(na.tools)
```

This Homework is due on **Saturday, April 17, 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 and parameters which can be gathered in routine blood analysis.

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

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

A glimpse of the data:

```
glimpse(birth_data)
```

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

Let us investigate using the Breast_Cancer data, whether breast cancer patients have normal BMI on average (considering 25 as normal)?

Question 1

Calculate sample statistic. Is it a continuous or categorical sample statistic?

```
summary(birth_data$BMI[birth_data$Classification == 1])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.67   23.10   27.69   28.32   32.33   38.58 
```

```
summary(birth_data$BMI[birth_data$Classification == 2])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.37   22.79   27.41   26.98   30.81   37.11 
```

```
#This is a continuous sample statistic since BMI is a range of values.
```

.

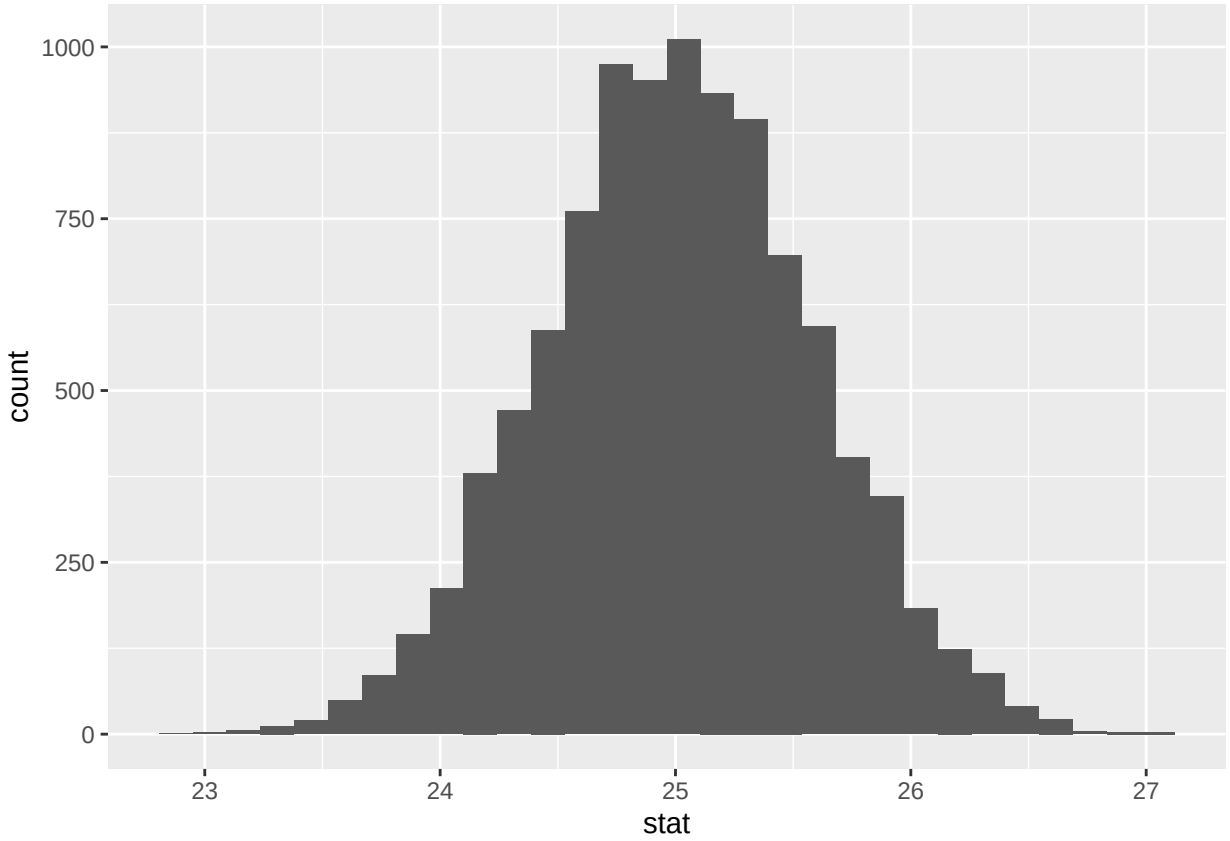
Question 2

a. Set up and test the hypotheses to determine whether breast cancer patients have higher than normal BMI (25) or not (follow the hypothesis process stepwise as you have done in your Module 5 group assignments) .

- State the null and alternative hypotheses. # Null Hypothesis: mu = 25 #Alternative Hypothesis: mu =/= 25
- Generate the null distribution and plot the distribution.

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

ggplot(null.dist, aes(stat)) + geom_histogram()
```



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

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

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.490
```

c. Conclude and interpret the results. #Given the p-value of 0.5057, this is greater than the significance level of alpha. Thus, we fail to reject the null hypothesis. We have enough evidence to suggest that a BMI of 25 is normal. ### Question 3

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

```
# bootstrap samples
boot.means <- birth_data %>%
   filter(Classification == 2) %>%
  specify(response = BMI) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
boot.means %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975))
```

```
# A tibble: 1 x 2
  lower95 upper95
    <dbl>   <dbl>
1    25.8    28.1
```

```
# cutoff bounds
# save as vector
```

b. Interpret the 95% confidence interval. #We are 95% confident that the interval (25.8,28.1) captures the average BMI in the population ### Question 4

Is having a higher than normal BMI, an indicator of increase risk of breast cancer given your results in 2 and 3? (Hint: consider the BMI of people don't have breast cancer (healthy control). Run the hypothesis test and estimate the 95% confidence interval to check the your conclusion). Justify your answer in 3-4 sentences.

```
null.dist1 <- birth_data %>%
  filter(Classification == 1) %>%
  specify(response = BMI) %>%
  hypothesise(null = "point", mu = 25) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

null.dist1 %>%
  filter(stat !=25) %>%
  summarise(p_value=n()/nrow(null.dist))
```

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

```
boot.means1 <- birth_data %>%
   filter(Classification == 1) %>%
  specify(response = BMI) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
boot.means1 %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975))
```

```
# A tibble: 1 x 2
  lower95 upper95
    <dbl>   <dbl>
1    26.8    29.8
```

```
#Given the p-value of 1, this is greater than the significance level of alpha. Thus, we fail to reject
```

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 a 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)
```

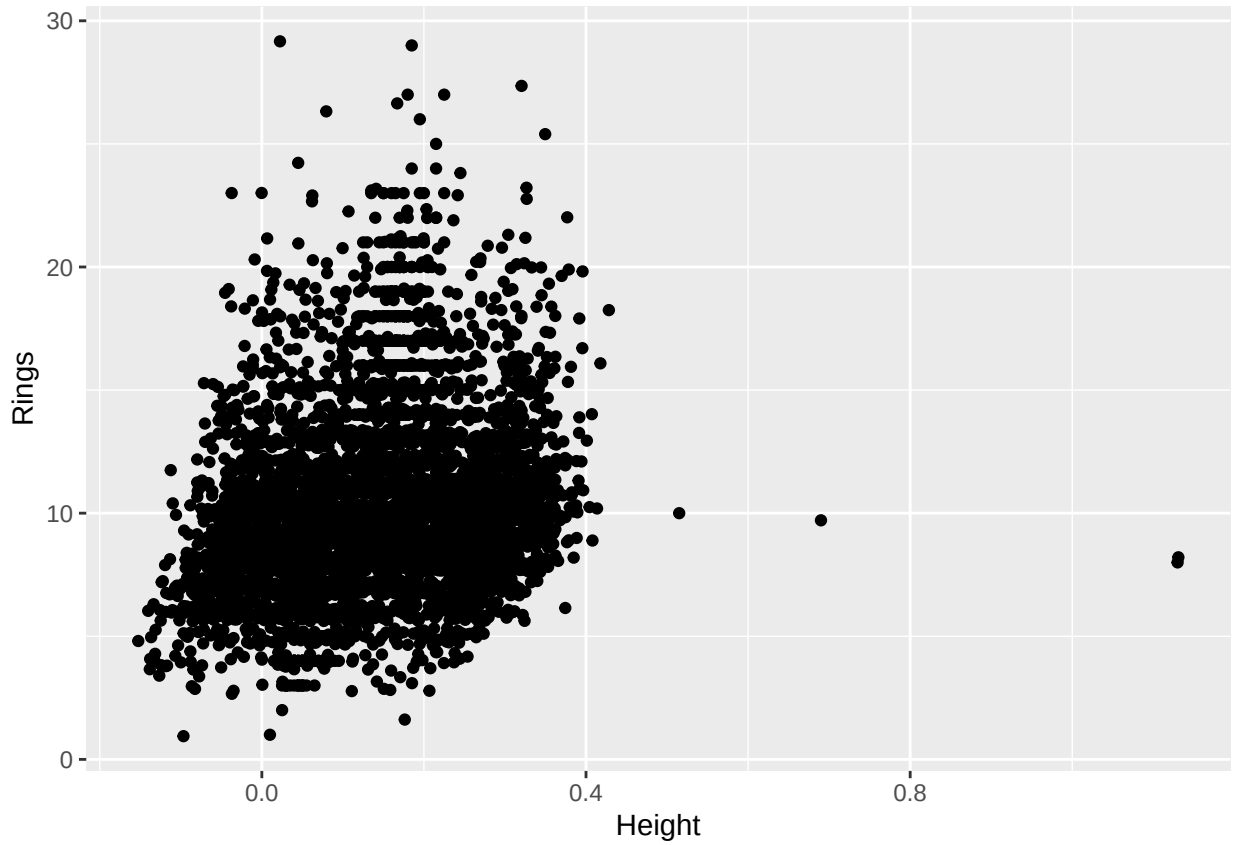
```
Rows: 4,177
Columns: 11
$ Sex      <chr> "M", "M", "F", "M", "I", "I", "F", "F", "M", "F", "F", "M", "~
$ Length   <dbl> 0.455, 0.350, 0.530, 0.440, 0.330, 0.425, 0.530, 0.545, 0.475~
$ Diameter <dbl> 0.365, 0.265, 0.420, 0.365, 0.255, 0.300, 0.415, 0.425, 0.370~
$ Height   <dbl> 0.095, 0.090, 0.135, 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.

Plot a scatter plot to get an idea about the relationship between height and age. Comment (1-3 sentences) on the plot.

```
ggplot(ab, aes(x = Height, y =Rings)) + geom_jitter(position = position_jitter(width = 0.2)) +
    geom_point()
```

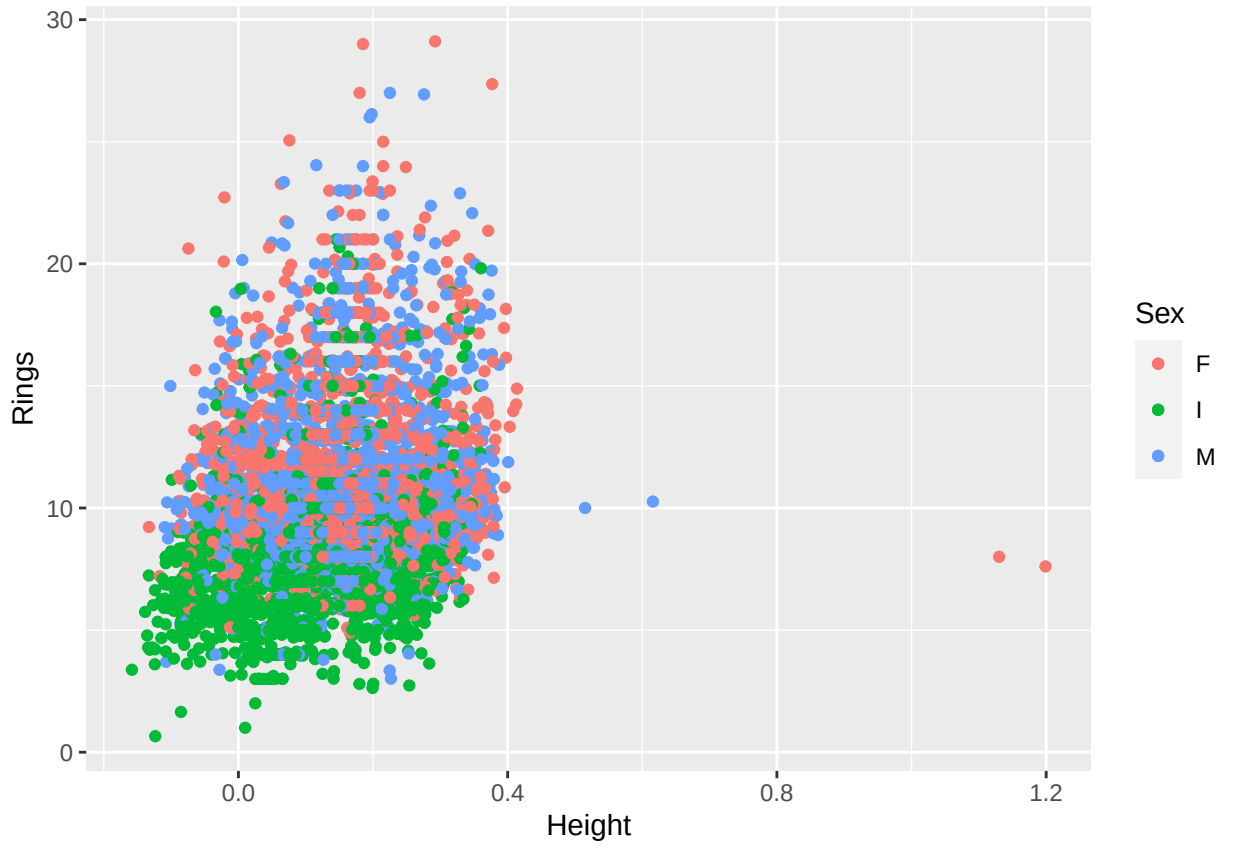



```
# The scatter shows a trend of as the age reaches in between 5-15 range, the height begins to round out,
```

Question 2

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

```
ggplot(ab, aes(x = Height, y =Rings, colour =Sex)) + geom_jitter(position = position_jitter(width = 0.2)
    geom_point()
```



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? Fit a multiple regression model to test the effect of the three variables on the age of the abalone.Interpret the model fit.

```
mod.fit <-  lm(Rings ~ Whole + Height + Diameter, data = ab)
mod.fit
```

```
Call:
lm(formula = Rings ~ Whole + Height + Diameter, data = ab)

Coefficients:
(Intercept)        Whole       Height     Diameter
     2.2105      -0.1781      20.1382      12.4084
```

```
require(MASS)
require(dplyr)

mod.fit %>%
  glance() %>%
  dplyr::select(r.squared, adj.r.squared)
```

```
# A tibble: 1 x 2
```

```
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.350         0.350
```

```
#The Rings coefficient describes the expected differences between score in that specific Weight, Height
```

Question 4

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

```
#Start with a full model
full.model1 <-  lm(Rings ~ Whole + Height + Diameter + Length + Shucked + Viscera + Shell, data = ab)
step(object = full.model1, direction = "backward", trace = TRUE)
```

```
Start:  AIC=6662.1
Rings ~ Whole + Height + Diameter + Length + Shucked + Viscera +
    Shell

           Df Sum of Sq   RSS    AIC
- Length    1      3.65 20510 6660.8
<none>                  20506 6662.1
- Diameter  1    175.45 20681 6695.7
- Viscera   1    279.49 20785 6716.6
- Shell     1    279.98 20786 6716.7
- Height    1    287.02 20793 6718.2
- Whole     1    783.61 21290 6816.7
- Shucked   1   2964.97 23471 7224.2

Step:  AIC=6660.85
Rings ~ Whole + Height + Diameter + Shucked + Viscera + Shell

           Df Sum of Sq   RSS    AIC
<none>                  20510 6660.8
- Shell     1    282.20 20792 6715.9
- Height    1    285.48 20795 6716.6
- Viscera   1    287.66 20797 6717.0
- Diameter  1    676.52 21186 6794.4
- Whole     1    785.25 21295 6815.8
- Shucked   1   3001.27 23511 7229.3

Call:
lm(formula = Rings ~ Whole + Height + Diameter + Shucked + Viscera +
    Shell, data = ab)

Coefficients:
(Intercept)        Whole       Height     Diameter      Shucked      Viscera
      2.896        9.256       11.790       11.634      -20.271       -9.931
      Shell
      8.606
```

```r
#Remove one variable at a time until R^2 does not change. Taking away Shell here. (Other variations of
full.model2 <-  lm(Rings ~ Whole + Height + Diameter + Length + Shucked + Viscera, data = ab)
step(object = full.model2, direction = "backward", trace = TRUE)
```

```
Start:  AIC=6716.75
Rings ~ Whole + Height + Diameter + Length + Shucked + Viscera

           Df Sum of Sq   RSS    AIC
- Length    1       5.9 20792 6715.9
<none>                  20786 6716.7
- Diameter  1     230.6 21016 6760.8
- Height    1     353.3 21139 6785.1
- Viscera   1     557.3 21343 6825.3
- Whole     1    4637.6 25424 7556.0
- Shucked   1    6912.5 27698 7914.0

Step:  AIC=6715.93
Rings ~ Whole + Height + Diameter + Shucked + Viscera

           Df Sum of Sq   RSS    AIC
<none>                  20792 6715.9
- Height    1     351.3 21143 6783.9
- Viscera   1     576.2 21368 6828.1
- Diameter  1     880.9 21673 6887.3
- Whole     1    4677.0 25469 7561.4
- Shucked   1    7079.7 27871 7938.0

Call:
lm(formula = Rings ~ Whole + Height + Diameter + Shucked + Viscera,
    data = ab)

Coefficients:
(Intercept)        Whole       Height     Diameter      Shucked      Viscera
       2.55        13.67        13.01        13.04       -24.19       -13.24
```

```r
#Final full.model shown here where adjusted R^2 has not increased from the previous. Select model with
full.model <-  lm(Rings ~ Whole + Height + Length + Shell + Shucked, data = ab)
step(object = full.model, direction = "backward", trace = TRUE)
```

```
Start:  AIC=6753.04
Rings ~ Whole + Height + Length + Shell + Shucked

          Df Sum of Sq   RSS    AIC
<none>                  20977 6753.0
- Height   1    318.39 21296 6814.0
- Length   1    415.22 21392 6832.9
- Whole    1    506.81 21484 6850.8
- Shell    1    656.96 21634 6879.8
- Shucked  1   2701.78 23679 7257.1

Call:
lm(formula = Rings ~ Whole + Height + Length + Shell + Shucked,
```

```
    data = ab)

Coefficients:
(Intercept)        Whole       Height       Length        Shell      Shucked
      3.251        5.840       12.359        7.370       12.288      -18.582
```

```
#Finding R^2 and adjusted R^2
require(MASS)
require(dplyr)

full.model %>%
  glance() %>%
  dplyr::select(r.squared, adj.r.squared)
```

```
# A tibble: 1 x 2
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.517         0.516
```

```
## The model fit parameters does not change significantly.
```

Question 5

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

```
mod.fit3 <-  lm(Rings ~ Sex + Height + Diameter, data = ab)
mod.fit3
```

```
Call:
lm(formula = Rings ~ Sex + Height + Diameter, data = ab)

Coefficients:
(Intercept)         SexI         SexM       Height     Diameter
     3.9292      -1.1105      -0.1566      17.9455       9.5979
```

```
require(MASS)
require(dplyr)

mod.fit3 %>%
  glance() %>%
  dplyr::select(r.squared, adj.r.squared)
```

```
# A tibble: 1 x 2
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.365         0.365
```

```
#Including Sex does not have a much significant impact on predicting age. It is suggested that the diff
```

Question 6

Interpret the results from the model in Question 5. WHat does each coefficeint signify?

SexI, SeXM, Height, and Diameter signifies the significance it has as a factor when predicting ring age of abalones. Height and Diameter have the most significance of the variables when predicting ring age.

Essential details

Deadline and submission

The deadline to submit Homework 3 is **11:00pm on Saturday, 17 April, 2021.**

- Submit your work by uploading your **Rmd and html file through D2L.Kindly check after submission. If graphs are not displayed in the html after uploading it to d2l, upload the pdf output.**
- Kindly ensure that **the `echo=TRUE` is set in the every chunk option.**
- Please **name all the code chunks**
- Late work **will not be accepted** except under certain extraordinary circumstances.

Help

- Post general questions in the HW3 Teams channel. If you are trying to get help on a code error, explain your error in detail
- Feel free to visit us during our virtual zoom 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, or how to approach a problem with your classmates.You may also post your general questions in the HW3 channel in Teams.But you may not copy-and-paste another individual'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 in-class assignments, pre-class assignments, homework assignments, and midterm exams. 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 10)	80
R Markdown formatting and knitting	5
Communication of results	10
Code style	5
Total	**100**

Please note: Code style includes code efficiency.

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