SML 201 Mini Project 2

Bill Haarlow & Weston Carpenter

2019-11-10

Project 2 is due by 11:59pm on Thursday November 14. Please submit both a .Rmd and a .pdf file on Blackboard by the deadline **and** drop off a hard copy of the pdf file at 26 Prospect Avenue by 5pm of the **next weekday** of the due date. To look for the drop-off cabinet, after you enter the building turn to the left to enter the lounge area and the file cabinet is to your right with an open slot with the label "SML 201 Homework"; note that the building might be locked after 6pm and on the weekends. You are also welcome to bring your PDF copy to any lecture **before** the deadline and I will drop off the copy for you.

Late **projects** will be penalized at intervals rounded up to multiples of 24 hours. For example, if you are 3 hours late, 10% off or if you are 30 hours late, 20% off.

This project can be completed in groups of up to 3 students. It is okay to work by yourself, if this is preferable. You are welcome to get help (you can either ask questions on Piazza or talk to instructors in person during office hours) from instructors; however, please do not post code/solutions on Piazza on a public post.

You are encouraged to get help from the instructors (either through Piazza or in person) if you need help to understand the definitions of the variables of the dataset or the procedure of the experiment.

When working in a group it is your responsibility to make sure that you are satisfied with all parts of the report and the submission is on time (e.g., we will not entertain arguments that deficiencies are the responsibility of other group members). We expect that you each work independently first and then compare your answers with each other once you all finish, or you all work together on your own laptops. Failing to make contributions and then putting your name on a report will be considered a violation of the honor code. **Please do not divide work among group mates.** Everyone in your group is responsible for being able to explain the solutions in the report.

For all parts of this problem set, you **MUST** use R commands to print the output as part of your R Markdown file. You are not permitted to find the answer in the R console and then copy and paste the answer into this document.

If you are completing this project in a group, please have only **one** person in your group turn in the .Rmd and .pdf files; the other person in your group should turn in the list of the people in your group in the *Text Submission* field on the submission page. This means that **everyone should make a submission**–either a file-upload or a text submission–regardless of whether you are working in a group or not.

The physical pdf report that you drop off and the .pdf file that you submit on Blackboard should be identical. Modifying your report after the deadline could result in a penalty as much as getting a zero score for the assignment.

Please type your name(s) after "Digitally signed:" below the honor pledge to serve as digital signature(s). Put the pledge and your signature(s) at the beginning of each document that you turn in.

> I pledge my honor that I have not violated the honor code when completing this assignment.

Digitally signed: Bill Haarlow & Weston Carpenter

In order to receive full credits, please have sensible titles and axis labels for all your graphs and adjust values for all the relevant graphical parameters so that your plots are informative. Also, all answers must be written in complete sentences.

(-3 pts each if any of these are not satisfied: code runs, code has annotations, answers are in complete sentences)

Before you start: loops are not allowed for this project. Please report all numerical answers to 2 digits after the decimal. Remember not to round intermediate calculations and please avoid hard-coding.

We will use the `bdims` dataset from the `openintro` package for questions 1-3. The dataset consists of body girth and skeletal diameter measurements, as well as the age, weight, height and gender on 507 (247 men and 260 women) physically active individuals.

(Hypothetical) Clair works for the Red Cross in an underdeveloped country. Each day she travels to the villages in the area to see adult patients. For some of the medicines in order to prescribe the correct dosages, Clair will need to know the approximate weight of the patient. However, it is not always feasible for her to carry a scale with her for her visits. In this project we will develop a formula for Clair to use so that she could get a rough estimate on a patient's weight (in kilograms) by using the the sum of the patient's knee diameters (in centimeters). It is much more convenient to carry a ruler and a calculator than to carry a scale.

You can assume that the dataset `bdims` contains the information on all the adult residents of the villages that Clair visits.

As usual please read the info on the help manual about the dataset. The variables that we are going to use for this project are: `kne.di`, `wgt`, and `sex`. Below are the variable definitions from the help manual.

- **kne.di**: A numerical vector, respondent's knee diameter in centimeters, measured as sum of two knees.
- **wgt**: A numerical vector, respondent's weight in kilograms
- **sex**: A categorical vector, 1 if the respondent is male, 0 if female.

For simplicity, we will use the variable name `Knee Diameter` to refer to 'kne.di' which corresponds to "the sum of knee diameters".

Question 1 (15 pts) Investigate the relationships between the variables

Part a (2 pts)

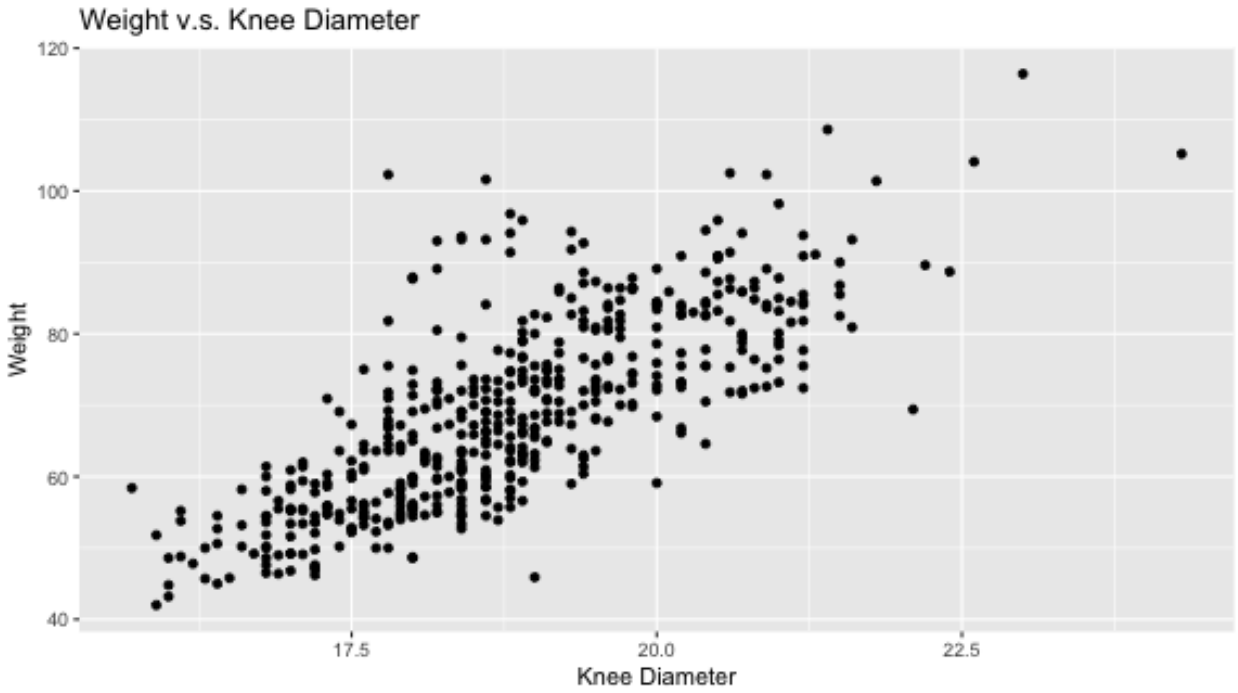
Extract out the variables needed for this project and named the resulting data frame `body.dims`. Make sure that the variables are in the correct data types.

```
body.dims = bdims[c("kne.di", "wgt", "sex")]
# Creates new data frame with needed variables
```

Part b (5 pts)

Use the ggplot() package to make a scatterplot of Weight v.s. Knee Diameter. Do the two variables have a linear relationship?

```
# Creates scatterplot of Weight v.s. Knee Diameter using ggplot
ggplot(body.dims) + geom_point(aes(x = kne.di, y = wgt)) + labs(x = "Knee Diameter",
    y = "Weight", title = "Weight v.s. Knee Diameter")
```



Part c (5 pts: 2 for multiple choice question; 3 for the explanation)

Based on your scatterplot in Part b and without doing any calculations choose the correct answer for this question: The correlation between the two variables Weight and Knee Diameter should be

(A.) Negative
(B.) Positive
(C.) Cannot be determined without any calculations

Answer: B

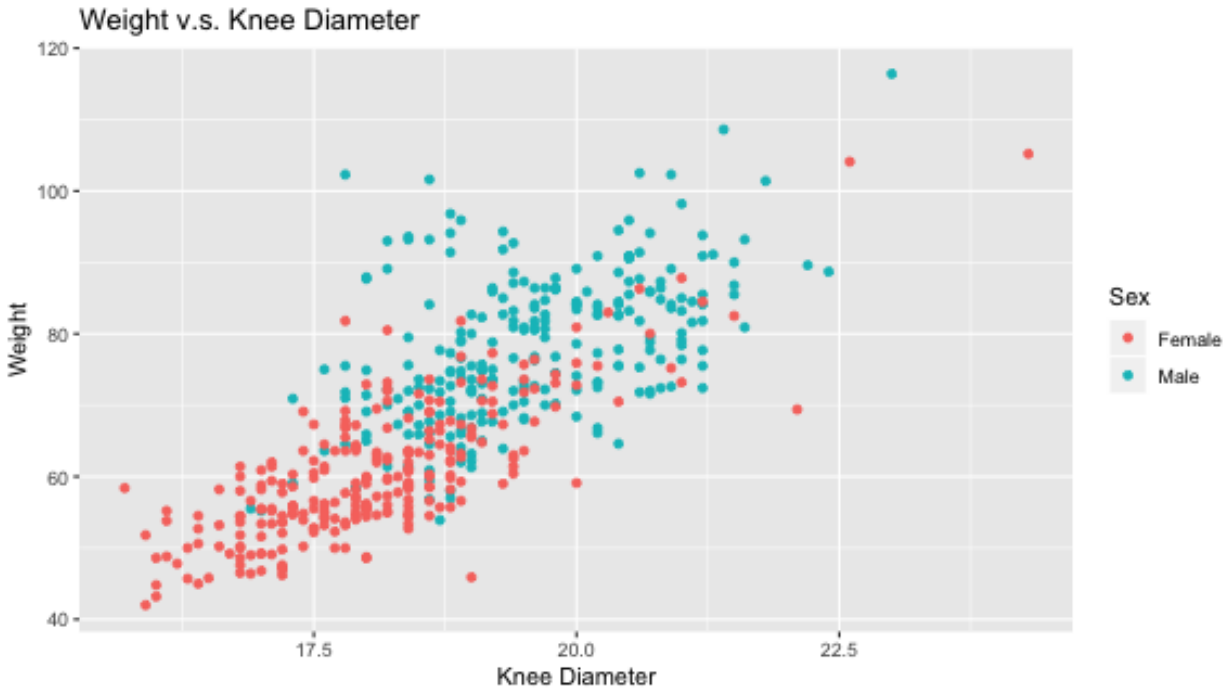
Explain your reasoning.

Explanation: The scatterplot generated in Part B shows that the points for weight vs. knee diameter are generally clustered in a group that slopes upwards and to the right, indicating a positive correlation.

Part d (3 pts)

Modify the code for your scatterplot in part b and make a new scatterplot for the two variables by coloring the data points with different colors for different Sex categories. Does the relationship between the two variables look linear within each Sex category?

```
# Creates scatterplot of Weight v.s. Knee Diameter using ggplot
# but differentiates males and females by color
ggplot(body.dims) + geom_point(aes(x = kne.di, y = wgt, color = as.factor(sex))) +
    labs(x = "Knee Diameter", y = "Weight", title = "Weight v.s. Knee Diameter",
        colour = "Sex") + scale_colour_discrete(labels = c("Female",
    "Male"))
```



Question 2 (46 pts) Model fitting

For each of the following models, report the fitted model (i.e., report the formula that you will use to predict the weight of a chosen person) and answer the additional question listed in each part.

Part a

Model 1

For each person in the dataset we will model the person's weight as

$$Weight = \beta_0 + \beta_{KneeDiameter} KneeDiameter + Error$$

Part (i) (5 pts)

The fitted model is (complete the following)

$$\widehat{Weight} = \beta_0 + \beta_{KneeDiameter} KneeDiamater$$

Part (ii) (5 pts)

Interpret your fitted model: on average if person A's knee diameter sum is 1 cm longer than person B's knee diameter sum, how much heavier or lighter (in kilograms) do you expect person A will be compared to person B?

Answer: Person A will be about $\beta_{KneeDiameter}$ kilograms heavier than person B.

Part b

Model 2

For each person in the dataset we will model the person's weight as

$$Weight = \beta_0 + \beta_{1_{male}} 1_{male} + \beta_{KneeDiameter} KneeDiameter + Error$$

Part (i) (8 pts)

(Complete the following.) According to our fitted model, for a female resident we will use the following formula to predict her weight in kg:

$$\widehat{Weight} = \beta_0 + \beta_{KneeDiameter} KneeDiameter$$

and for a male resident we will use the following formula:

$$\widehat{Weight} = \beta_0 + \beta_{1_{male}} 1_{male} + \beta_{KneeDiameter} KneeDiameter$$

Part (ii) (5 pts)

Interpret the meaning for the coefficient estimate $\hat{\beta}_{1_{male}}$.

Answer: This coefficient estimate is a dummy variable which helps distinguish between the separate estimates between males and females. Without this, the model would not account for discrepancies in the data between the two sexes.

Part c

Model 3

For each person in the dataset we will model the person's weight as

$$Weight = \beta_0 + \beta_{1_{male}} 1_{male} + \beta_{KneeDiameter} KneeDiameter + \beta_{(KneeDiameter\ 1_{male})} (KneeDiameter\ 1_{male}) + Error$$

Part (i) (8 pts)

(Complete the following.) According to our fitted model, for a female resident we will use the following formula to predict her weight in kg:

$$\widehat{Weight} = \beta_0 + \beta_{KneeDiameter} KneeDiameter$$

and for a male resident we will use the following formula:

$$\widehat{Weight} = \beta_0 + \beta_{1_{male}} 1_{male} + \beta_{KneeDiameter} KneeDiameter + \beta_{(KneeDiameter\ 1_{male})} (KneeDiameter\ 1_{male})$$

Part (ii) (5 pts)

According to our fitted model, on average one centimeter increment in the knee diameter sum will result in a bigger estimated weight increment for a male or a female? Use numbers to support your answer.

```
# Creates linear model corresponding to fitted model
Weight = lm(wgt ~ kne.di * sex, data = body.dims)
summary(Weight)  # Returns summary of linear model

Call:
lm(formula = wgt ~ kne.di * sex, data = body.dims)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-20.393  -5.092  -0.827   4.343  32.996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.4793     7.2773  -7.349 8.14e-13 ***
kne.di        6.3038     0.4013  15.710  < 2e-16 ***
sex          33.4741    11.5228   2.905  0.00383 **
kne.di:sex   -1.2864     0.6074  -2.118  0.03468 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.663 on 503 degrees of freedom
Multiple R-squared:  0.6723,    Adjusted R-squared:  0.6703
F-statistic: 343.9 on 3 and 503 DF,  p-value: < 2.2e-16
```

Answer: According to our fitted model, on average a one centimeter incremental increase in the knee diameter sum will result in a 6.3038 kg increase in the estimated weight increment for females. For males, a one centimeter incremental increase in the knee diameter sum will result in a 5.0174 kg increase in the estimated weight increment, which is lower than for females. However, once we adjust for the `sex` dummy variable, which adds 33.4741 kg to the male weight increment overall, the average one centimeter incremental increase in the knee diameter sum will result in a bigger estimated weight for males.

Part d (12 pts: 5 for the True/False question; 7 for the explanation.)

Compare Model 3 in Part c with Model 1 in Part a. Without doing any computation, select the correct answer about the statement: The MSE produced by Model 3 cannot be bigger than the MSE produced by Model 1.

(A.) True
(B.) False

Answer: A

Justify your choice by using the key concept behind linear regression. (Hint 1: How is the best line chosen when fitting a regression model? Hint 2: How much freedom (in terms of the choices of the slopes and the intercepts) does Model 3 have? Compare this to the freedom that Model 1 has.)

Justification: The best line chosen when fitting a regression model is the line that minimizes the mean squared error. It is possible to reduce the mean square error by increasing the freedom the model has through increasing the choices of slopes and intercepts. Model 3 has many more choices and intercepts than Model 1.

Question 3 (12 pts)

The 3rd quartile of the weights is about 79 kg. Is the event that a randomly selected village resident weighs at least 79 kg independent of the event that this resident is male? Support your answer with numbers. (Hints: What is the chance that a randomly selected resident weighs at least 79 kg? Does this chance change if we knew that the resident chosen is a male?)

```
# Total number of weights greater than or equal to 79 over total
# number of weights
nrow(body.dims[body.dims$wgt >= 79, ])/nrow(body.dims)
[1] 0.2485207
# Creates new data frame that only has the data for males
```

```
body.dims.m = body.dims[body.dims$sex == 1, ]
# Total number of male-only weights greater than or equal to 79
# over total number of male-only weights
nrow(body.dims.m[body.dims.m$wgt >= 79, ])/nrow(body.dims.m)
[1] 0.4615385
```

Answer: The event that a randomly selected village resident weighs at least 79 kg is not independent of the event that this resident is male. The chance that a randomly selected resident weighs at least 79 kg is about 0.2485207, while the chance that a randomly selected resident, known to be male, weighs at least 79 kg is about 0.4615385. If the events were independent, the chances in both circumstances would be the same.

Question 4 (25 pts)

We are interested in finding out the proportion of adults in the United States who cannot cover a $400 unexpected expense without borrowing money or going into debt. In a simple random sample of 765 adults in the United States, 322 say they could not cover a $400 unexpected expense without borrowing money or going into debt.[1]

Part a (5 pts)

What population is under consideration in the data set?

Answer: The population in the data set is the population of adults in the United States.

Part b (5 pts)

What parameter is being estimated?

Answer: The parameter being estimated is the proportion of adults in the United States who cannot cover a $400 unexpected expense without borrowing money or going into debt.

Part c (5 pts)

What is the point estimate for the parameter?

Answer: The point estimate for this parameter is 322/765 adults in the United States, or 0.4209.

Part d (10 pts)

An economist claims that the true proportion (of adults in the United States who cannot cover a $400 unexpected expense without borrowing money or going into debt) is 0.35. You suspect that this proportion is too low. To look for evidence against his claim you decide to run some simulations. You will draw 100,000 samples, each of size 765, from a population consisting of thirty-five 1's and sixty-five 0's. The draws should be done with replacement so that the chance of getting a 1 remains to be .35 for all draws.

```
set.seed(500)  # Sets the seed at 500 for consistency
# Creates matrix `mat1` with 100000 columns of size 765 where 0
# has a .65 chance and 1 has a .35 chance of being drawn
mat1 = matrix(sample(x = (0:1), size = 765 * 1e+05, replace = T,
    prob = c(0.65, 0.35)), ncol = 1e+05)
```

Calculate the sample mean for each sample. Now, what proportion of your sample means is greater than 322/765? Report this number. Note that if this number is very small then this gives strong evidence against the economist's claim. For example, suppose this number is close to zero. Then, this means that if the economist were right, the chance that you would observe a sample with such high sample proportion should

be almost zero. However, you did observe such sample; thus, the economist claim might be false. This problem demonstrates the key concept in hypothesis testing. We will discuss hypothesis testing in week 8.

```
# Calculates the sample mean for each sample
sample.mean = apply(mat1, MAR = 2, FUN = mean)
# Confirms that there are 100000 sample means
length(sample.mean)
[1] 100000
# Outputs number of sample means greater than 322/765
table(sample.mean > (322/765))

FALSE  TRUE
99998     2
```

Answer: The proportion of our sample means greater than 322/765 is 2/100000.

[1]Question was taken and modified based on a question from Diez, D. M., Barr, C. D., and Çetinkaya-Rundel, M. (2019). *OpenIntro Statistics* third edition.