MS 204 Exam II

Your Name Here

General instructions for Midterms:

- Create a new Markdown file
- Change the heading to include your author name
- Save the R Markdown file (named as: [MikeID]-[MidtermII].Rmd – e.g. "mlopez-MidtermII") to somewhere where you'll be able to access it later (zip drive, My Documents, Dropbox, etc)
- Your file should contain the code/commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. **Plots and answers must show up as code output to receive full credit for any relevant questions**
- Each answer must be supported by written statements (unless otherwise specified) as well as any code used
- A printed HTML or PDF copy of this midterm is due Tuesday at 11:59 PM. I will not answer any questions on this exam after 12:00 noon on Tuesday... I am free over email through that point, and also in my office prior. Note that if you submit after 5:00 on Tuesday, Harder Hall may be locked. If that is the case, submit via email by the deadline, and submit a hard copy by Wednesday at 12:00 noon which includes the written honor code.
- Each student must abide by Skidmore's honor code. You may not use the internet to solicitating answers or communicate with other students, or to check for any help on any question. Your book, class notes, activities, and labs are sufficient for finishing this exam.

Part I: okcupiddata (50 points total, all questions worth 4 points unless indicated otherwise)

Return to the `okcupiddata`, which we first looked at during Chapter 2. From the course description, this data contain cleaned profile data of 59,946 OkCupid users who were living within 25 miles of San Francisco, had active profiles on June 26, 2012, were online in the previous year, and had at least one picture in their profile. The original data and codebook can be found at https://github.com/rudeboybert/JSE_OkCupid.

```
library(oilabs)
library(mosaic)
library(tidyverse)
library(okcupiddata)
head(profiles)
```

1. To start, we need to do a bit of further data cleaning. Please execute the following steps, and make a new data set called `profiles.new`. (3 points)

- Filter to only include subjects between 60 and 80 inches tall inclusive
- Create a new variable, `log.income`, defined as `log.income = log(income)`. Note that R uses the natural log as a default.

The rest of this question will use the `profiles.new` data set.

2. Make plot to compare the association of `log.income` by `height`
3. Edit your plot above to include a smoothed line of best fit using `+ geom_smooth(method = "lm")`. (2 points)
4. In two or fewer non-technical sentences, explain to someone why it may be important to account for gender when trying to learn about the association between height and income.
5. Recreate the plot in (3), done separately within each gender. Compare the two lines. Is there any difference in the association between height and income between the two genders?
6. Using only the males in the data set (`profiles.males <- profiles.new %>% filter(sex == "m")`), estimate the regression line of `log.income` on `height`. (3 points)
7. Pick *one* of the following explanations and justify it (5 points)

- There is a link between height and log income among males on dating websites, but it is likely to be accounted for by chance
- There is a link between height and log income among males on dating websites, which is unlikely to be accounted for by chance. It is most feasible that males who are taller will get paid more
- There is a link between height and log income among males on dating websites, which is unlikely to be accounted for by chance. It is likely accounted for by a confounding variable.
- There is a link between height and log income among males on dating websites, which is unlikely to be accounted for by chance. It is likely accounted for by reverse causation.

8. Consider the sample of Okcupid uses in this data. How does that impact the generalizability of your findings above?

Questions 9-13 will use multiple regression. Please return to the `profiles.new` that was created above

9. Find a multiple regression model of `log.income` on sex, height, age and smoking status. Write the estimated regression line (you can round to the nearest hundreth).
10. Using the regression line in (9), a fellow student comments "So profiles who smoke when drinking boast roughly the same income as those who don't smoke at all." Explain to your fellow student why that statement is not quite correct.
11. Interpret the coeffient on height in your model in Question 9.

12. Zark Muckerberg is a 33-year old male who does not smoke and is 67 inches tall, and has a reported income of 1,000,000. Estimate a predicted value of `Zark`'s income, as well as his residual, which you should report on the dollar scale (6 points)

13. Check the assumption of normally distributed residuals regarding your model above.

Part II bechdel data (50 points)

During April of 2014, Walt Hickey of FiveThirtyEight wrote an article titled "The Dollar-and-cents case against Hollywood's exclusion of women," which can be accessed at this link: https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

```
#install.packages("fivethirtyeight")
library(fivethirtyeight)
head(bechdel)
```

You can learn about the specific variables using `?bechdel`, and for all questions, assume that the population of interest is all movies from 1970 through 2013 that could have been made.

1. Read the article, and confirm that the two metrics in the first sentence of "our findings" are accurate – that there are exactly 1,794 films, of which 53 percent passed the Bechdel test. (6 points)
2. Estimate a 95% confidence interval for the true rate of all movies passing the Bechdel test (using the `binary` variable). Interpret your interval. You do not need to consider assumptions for inference, and you can calculate this interval by hand or using R (8 points)

For the following two questions, state appopriate hypotheses (if applicable), relevant assumptions (which you should check), test-statistic (such as z or t), degrees of freedom (if relevant), p-value, and both technical and non-technical conclusions. You should do the calculations of this question both by hand (using summary statistics) and via the `inference` command. (18 points each)

3. Test whether or not the average film budget (using the variable `budget_2013`, the budget of movies adjusted for inflation in 2013) is the same between movies that passed and failed the Bechdel test.
4. Calculate a 95 percent confidence interval for the difference in rates of movies between `decade_code = 1` (2010 onwards) and `decade_code = 3` (1990 to 1999) that have passed the Bechdel test.

Part III

Write the Skidmore honor code (handwritten): While taking this examination, I have not witnessed any wrongdoing, nor have I personally violated any conditions of the Skidmore College Honor Code.