Midterm Exam

Jenna Moscaritolo

11/7/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

Note: This data description portion is taken from the Rmd file from the data collection exercise. It serves to help the reader understand how the data was collected and what the original question is.

Introduction

People are diverse in what type of clothing they own. For instance, some people do not care for fanny packs while others strive to incorporate fanny packs into their fashion or collect them. This is the same for hats. People wear hats for eye protection, face shade, to cover up a bad hair day, or for fashion.

Goal

This project's goal is to look at the correlation (if any) of the number of hats students own with binary factors of gender and exercise levels.

Data Collection

I have recorded data from 14 Boston University (BU) undergraduate students who participate in the athletic bands at BU who are also in the honorary band service fraternity of Kappa Kappa Psi or sorority of Tau

Table 1: Complete Dataset

student	gender	exercise	no_hats
1	male	less	14
2	male	less	7
3	female	less	14
4	female	more	4
5	male	less	4
6	female	more	4
7	female	less	6
8	female	less	1
9	female	less	3
10	male	less	2
11	female	less	3
12	female	less	12
13	female	less	25
14	female	less	5

Beta Sigma. They were asked for their gender (binary), if they exercise three/more than three times a week or less than three times a week (binary), and the number of hats they each own.

This information was collected by Google Forms in three questions:

1. What is your gender?
   - Female
   - Male
   - Other
2. Do you exercise more or less than 3 times a week?
   - More
   - Less
3. How many hats do you own?
   - (Participants typed their numeric answer.)

Here is a table of the collected data:

```
hat_data <- read.csv("hat_data.csv", header = TRUE)
data_table <- kable(hat_data, booktabs = TRUE, align=rep('c', 5), caption = "Complete Dataset") %>% kabl
data_table
```

EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

What is known about the data is that "gender" and "exercise" are binary, and that "no_hats" is count.

Gender: * 1: female * 0: male Exercise: * 1: less than 3 times a week * 0: more than 3 times a week

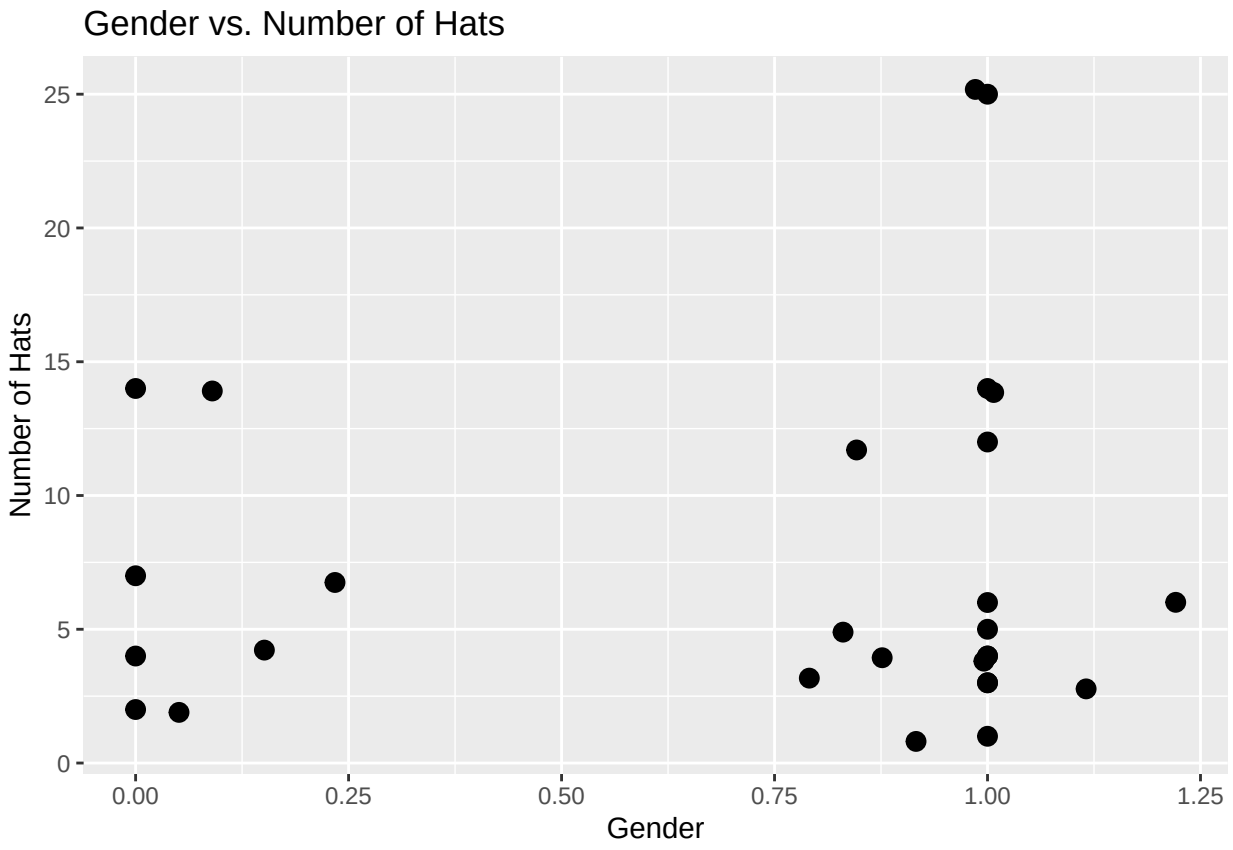
```
# Pulling the column "no_hats"
hats <- hat_data$no_hats

# Changing the binary outputs from strings to integers:
gender <- as.numeric(hat_data$gender == "female")
```

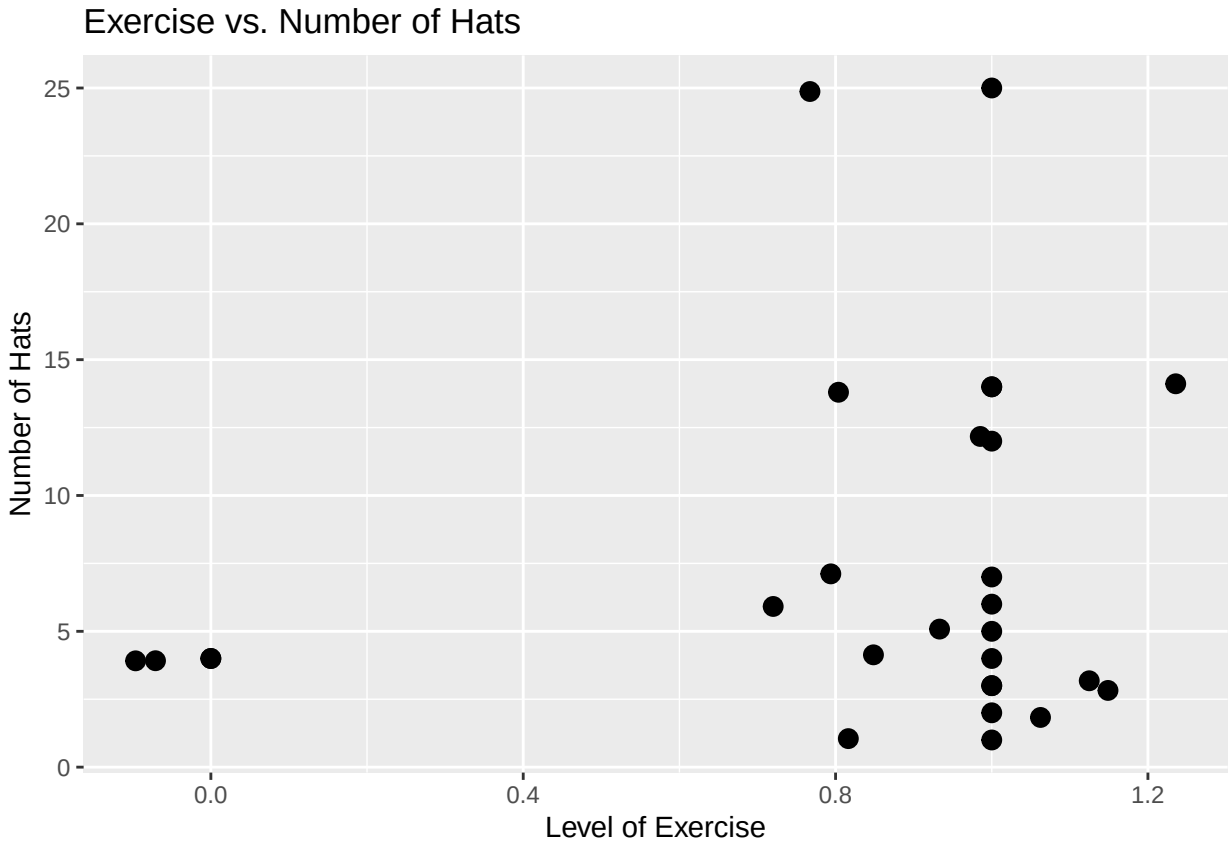
```r
exercise <- as.numeric(hat_data$exercise == "less")

# Creating a data frame with integer data only
df_all <- data.frame(hats, gender, exercise)

# Plotting the data
ggplot(df_all, aes(gender, hats)) +
    geom_point(size = 3) +
    geom_jitter(size = 3, width = 0.3, height = 0.3) +
    labs(title = "Gender vs. Number of Hats", y = "Number of Hats", x = "Gender")
```




```r
ggplot(df_all, aes(exercise, hats)) +
    geom_point(size = 3) +
    geom_jitter(size = 3, width = 0.3, height = 0.2) +
    labs(title = "Exercise vs. Number of Hats", y = "Number of Hats", x = "Level of Exercise")
```



Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used, and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

To begin, let's calculate the numerator and denominator of the degrees of freedom from each linear regression. This is calculated in the R chunk below.

```
# Number of groups = k
k <- length(unique(df_all$hats))
# Total sample size = n
n <- nrow(df_all)


# Degrees of freedom (numerator)
u <- k - 1
# Degrees of freedom (denominator)
v <- n - k


# Power analysis for linear models:
pwr.f2.test(u, v, f2 = NULL, power = 0.80)
```

```
##
##      Multiple regression power calculation
```

```
##
##               u = 9
##               v = 4
##              f2 = 5.397114
##       sig.level = 0.05
##           power = 0.8
```

For this power analysis, I needed to know the numerator and denominator for the degrees of freedom, which is calculated above (u = numerator, v = denominator). Then, I plugged these values and the power of 80% into the pwr.f2.test() function to find f2 which is the effect size measure. The sig.level is, at default, 0.05.

According to the webpage (cited after this paragraph), they suggest that f2 values of more than 0.35 are large effect sizes. Here, f2 = 5.397114 which is well over the 0.35 mark. Interpreting this, it is clear that more samples are needed as the larger the effect size, the larger the difference between the variables are. The closer the effect size is to zero, the more accurate the model is.

I used Quick-R by Datacamp to help me with my code for this power analysis.

Modeling (10pts)

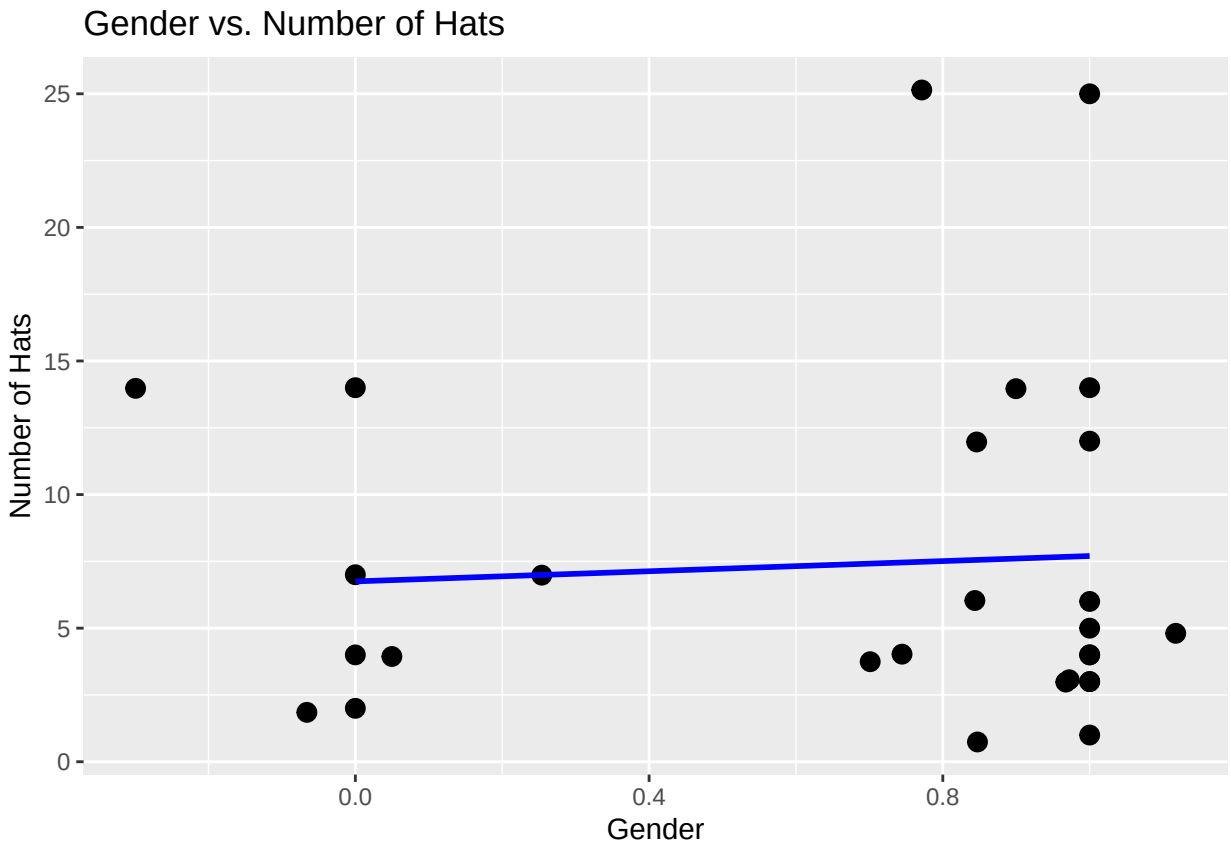
Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

Below is the model I chose. I chose the simple linear regression because, to answer the initial question of looking at the correlation (if any) of the number of hats students own with binary factors of gender and exercise levels, for both comparisons (gender and exercise level), the outcome (number of hats) is count data while the comparisons are binary. So, to see if there is any trend, a linear model was the best choice here.

The next section, "Validation," I will explain why other models would not work for this data. Here are the models:

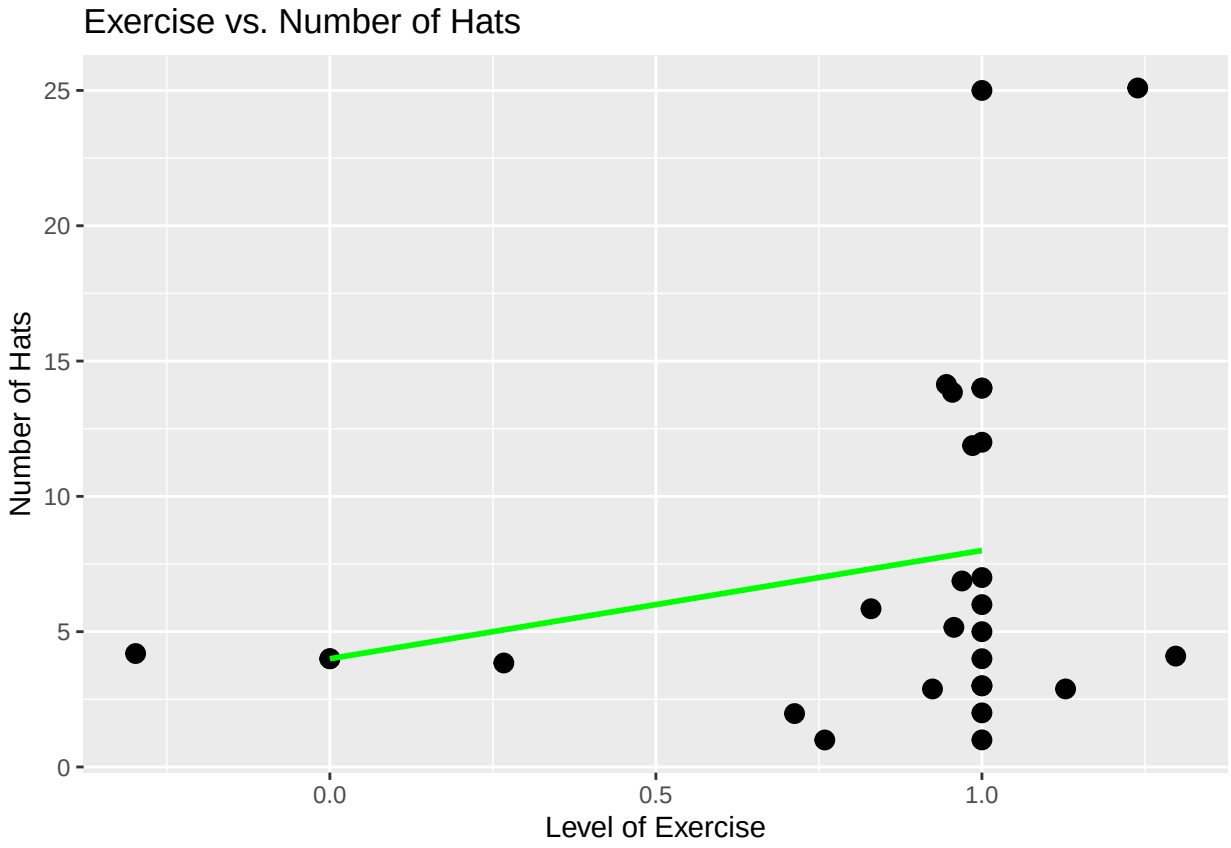
```r
# Linear Model 1: gender vs. hats
fit_gender <- lm(hats ~ gender, data = df_all)

# Plotted data w/ regression line & jitter
ggplot(df_all, aes(gender, hats)) +
    geom_point(size = 3) +
    geom_jitter(size = 3, width = 0.3, height = 0.3) +
    geom_line(data = fortify(fit_gender), aes(x = gender, y = .fitted), size = 1, color = "blue") +
    labs(title = "Gender vs. Number of Hats", y = "Number of Hats", x = "Gender")
```




```
# Linear Model 2: exercise vs. hats
fit_exercise <- lm(hats ~ exercise, data = df_all)

# Plotted data w/ regression line & jitter
ggplot(df_all, aes(exercise, hats)) +
    geom_point(size = 3) +
    geom_jitter(size = 3, width = 0.3, height = 0.2) +
    geom_line(data = fortify(fit_exercise), aes(x = exercise, y = .fitted), size = 1, color = "green") +
    labs(title = "Exercise vs. Number of Hats", y = "Number of Hats", x = "Level of Exercise")
```



Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

Before choosing to use linear models, I went through most of the regressions and found reasons to not use it. I did not go near MLLM because this data is too simple for that level of regression. Here are the reasons why I did not choose other models:

- Logistic regression does not work here because it requires the dependent variable to be binary to be influenced by the independent variable. This dataset only has independent variables, so this regression would not work.
- Poisson regression does not work because, although there is a count data outcome, there are binary outcomes as well. Poisson does well with count data, but works best when the count data happens over a time frame or as a rate.
- Binomial logistic and probit regressions will not work because it predicts a binary outcome rather than predicting the count variable (hats) based on gender and based on exercise level.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

Now that I have two linear models, I will compare the slopes of the two fits and decide if the variables correlate with one another. The coefficients are below, and the interpretation is in the next section called "Discussion."

```
fit_gender[1]
```

```
## $coefficients
## (Intercept)      gender
##        6.75        0.95
```

```
fit_exercise[1]
```

```
## $coefficients
## (Intercept)    exercise
##           4           4
```

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

Looking at the slopes of the lines from the outputs above, where gender is 0.95 and exercise is 4.0, there is little to no correlation between the two fits of hats vs. gender and hats vs. exercise. With the difference of the slopes being high (4.0 - 0.95 = 3.05), it is clear that they are all independent of each other. As a note, from knowing about the data collection variables, we already knew that these variables were all independent – this was just to prove it with numbers. Thus, to answer the question, there is little to no correlation of the number of hats students own and the two variables of gender and athleticism.

In a further analysis, the graphs easily show that, on average, female students have more hats than male students and that students with less than three workouts per week own more hats than students who workout more than three times per week.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

This was a pretty simple dataset and question that was provided from the data collection exercise. Due to the simplicity, I feel like I could have asked better survey questions in my data collection exercise and to have at least one dependent variable. It is tough to go more in depth on things when everything is independent. For the future, I would ask the data collection scientist (myself here) to provide more samples and more survey questions. Perhaps, "how many hours a week do you spend outside" to possibly have the hours depend on how many hats the students owned (or vice versa).

Comments or questions

If you have any comments or questions, please write them here.

I am having trouble understanding power analysis as it was one of the last things we learned. Over time, I will get more familiar with it. I hope I was correct with this project.

Other than that problem, I really enjoyed this exam. Being able to use our own data that we collected (without knowing we were going to use it for this) made the exam super fun and enjoyable.