Midterm Exam

Simu Huang

11/2/2020

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.

```
#import the data
data_sum <- read.csv("D:\\Documents\\R Project\\678\\homework\\Data Collection\\datamidterm1.csv",heade
data_1 <- read.csv("D:\\Documents\\R Project\\678\\homework\\Data Collection\\datamidterm2.csv",header=

#rename the variable
names(data_sum) <- c("name", "type", "total_number", "price")
rename(data_1, respondent = "X.U.FEFF.")
```

```
##    respondent YSL.416 YSL.12 Dior.999 Armani400 Armani405 Mac.chili
## 1           A       0      1        1         1         0         1
## 2           B       1      0        1         0         1         1
## 3           C       0      0        0         0         0         0
## 4           D       0      0        0         0         1         0
## 5           E       0      1        1         0         1         1
## 6           F       0      0        0         0         0         1
## 7           G       1      0        1         1         1         1
## 8           H       0      0        1         0         1         0
## 9           I       0      0        1         0         0         0
## 10          J       1      0        1         0         0         0
## 11      count       3      2        7         2         5         5
```

```
##    Mac.marrakash Estee.lauder.333 Estee.Lauder.360 TF.16
## 1              0                0                0     0
## 2              0                1                0     0
## 3              1                1                0     0
## 4              1                0                0     0
## 5              0                0                0     1
## 6              0                0                0     0
## 7              1                1                1     1
## 8              0                0                0     1
## 9              0                0                0     0
## 10             0                0                0     1
## 11             3                3                1     4
```

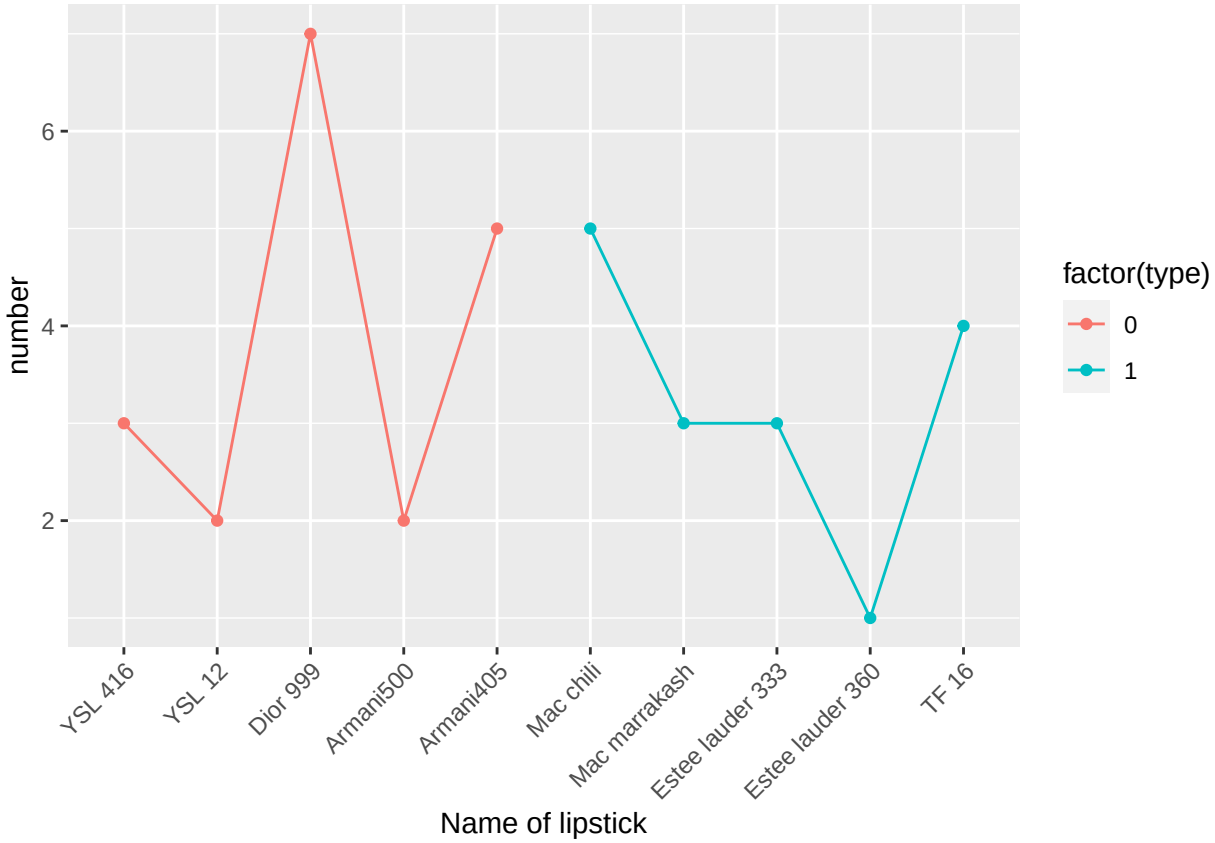
This data included responses from 10 of my friends. They were asked to answer that in the top 10 hottest lipsticks/lip gross in 2020, how many did they buy. In the "data_sum", the first column "name" concludes the names of the top 10 hottest lipsticks/lip gross that I collected from some fashion websites. The second column "type" indicates that they are lipstick or lip gross. In this column, value "0" means it is a lip gross and value "1" means it is a lipstick. The third column "number" is how many people out of ten own this lipstick.(The scale is 0 to 10.) The last column "price" is the price of each lipstick, which I collected from their web shops. In the "data_1", the first column concludes the answer of my ten friends. Value "0" means that she/he own this lipstick and value "1" means that she/he do not own it.

Depending on this dataset, I plan to find out that which lipstick people like most in this year. However, just looking at which lipstick sells the most does not reflect the most popular lipstick. Price also affects whether people buy it. It is possible that a person likes one lipstick very much, but she thinks it is too expensive, so she buys another. Therefore, the price of each lipstick should be taken into consideration.

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.

```
ggplot(data = data_sum, mapping = aes(x = factor(name, levels = unique(name)), y = total_number, color =
```



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.

```
pwr.f2.test(u=1, v=8, f2= NULL, sig.level=0.05, power=0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 8
##              f2 = 1.024617
##       sig.level = 0.05
##           power = 0.8
```

The effect size here is 1.02, which represents a large effect size. It means that these two variable in this model are substantially different. Therefore the sample size in this dataset is too small.

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.

```r
fit_1 <- stan_glm( total_number ~ price + factor(type), data = data_sum, family = gaussian(), refresh =
summary(fit_1)
```

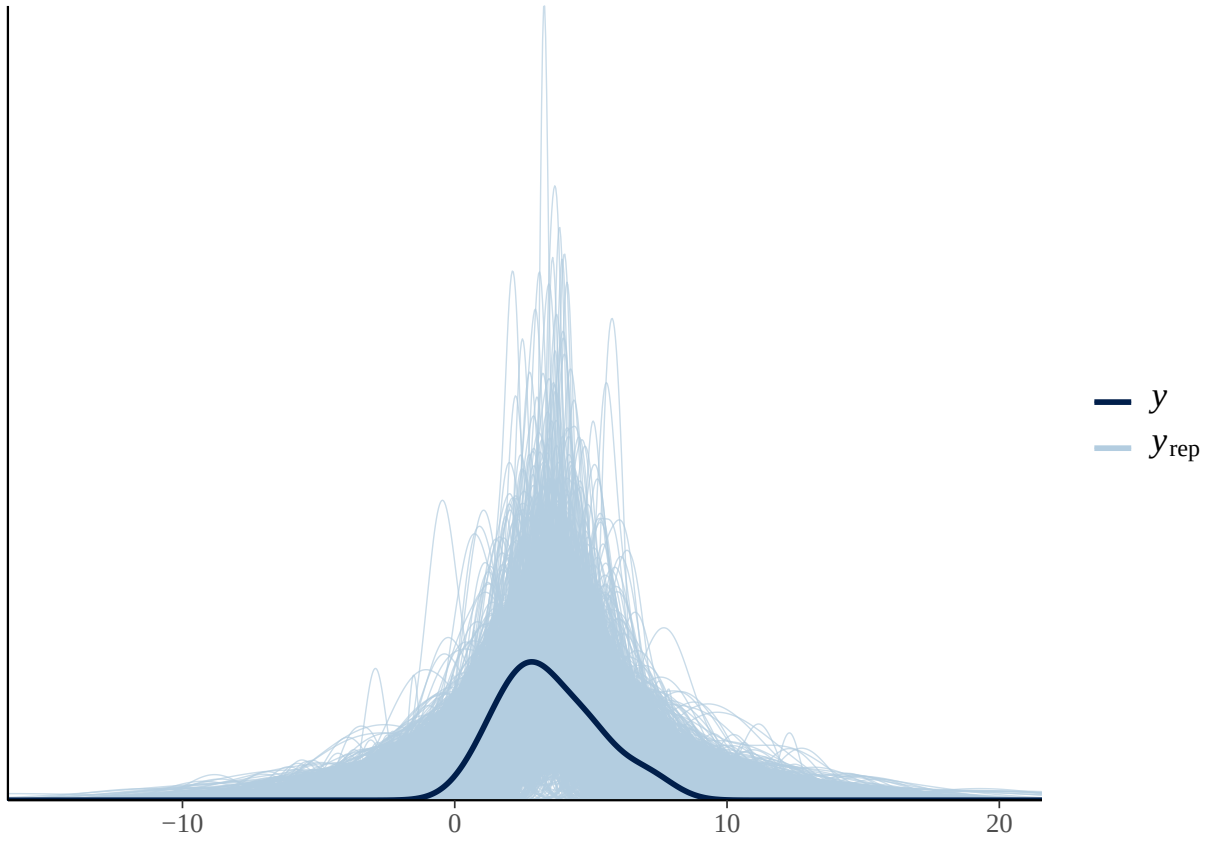
```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      total_number ~ price + factor(type)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 10
##  predictors:   3
##
## Estimates:
##                  mean   sd   10%   50%   90%
## (Intercept)     4.5    3.2  0.6   4.5   8.3
## price           0.0    0.0  0.0   0.0   0.0
## factor(type)1 -0.6    1.5 -2.4  -0.6   1.2
## sigma           2.2    0.6  1.5   2.1   3.0
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 3.5    1.0  2.3   3.5   4.7
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.1  1.0  2806
## price         0.0  1.0  3067
## factor(type)1 0.0  1.0  2748
## sigma         0.0  1.0  1799
## mean_PPD      0.0  1.0  2895
## log-posterior 0.1  1.0  1221
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

In this data, the outcome is the numerical value. And in this case, I want to explore that whether there is a linear relationship between the predictors and the outcome.

Validation (10pts)

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

```r
predict <- posterior_predict(fit_1)
ppc_dens_overlay(data_sum$total_number, predict)
```



From the ppc distribution plot below, we can find that the predicted value already cover the actual value, which means that the model is appropriate.

Inference (10pts)

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

```
data_gross    <- data_sum %>% filter( type == "0")
data_lipstick <- data_sum %>% filter( type == "1")

fit_2 <- lm(total_number ~ price, data = data_gross)
fit_3 <- lm(total_number ~ price, data = data_lipstick)
print(fit_2)
```

```
##
## Call:
## lm(formula = total_number ~ price, data = data_gross)
##
## Coefficients:
## (Intercept)        price
##      38.593       -0.113
```

```
print(fit_3)
```

```
##
## Call:
## lm(formula = total_number ~ price, data = data_lipstick)
##
```

```
## Coefficients:
## (Intercept)        price
##   3.2425144   -0.0001586
```

In the first model, for the lipstick, the higher the price of the lipstick, the fewer people would buy this lipstick. We can estimate that for YSL 416(the price is 320 CNY), every 10 people should have 2 unit.

In the second model, for the lip gross, it also shows that the higher the price, the fewer people would buy this lip gross. In this model, we can predict that for Estee lauder 333 (The price is 270 yuan.), every 10 people will own 3 unit of it.

Discussion (10pts)

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

Comparing these two model, we can roughly say that people are more influenced by the price when buying lipstick than when buying lipstick. According to the data, we can conclude that Dior 999 is the most popular one among my ten friends among the five most popular lip glazes. And the MAC chili is the most popular lipstick among my friend in these 5 hottest lip gross. Just from these ten limited observations, there is no obvious preference for lip stick or lip gloss for my friends.

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.

From the simple analysis result above, we can find that the observations there is too little. There is only ten responses, which means that there must be some bias. For example, most of my friends are in the same age-group, they may all like a particular color code or brand. This can lead to bias in data collection. Meanwhile, the samples here are not representative enough. The top 10 hottest lipstick/lip gross I chose there might not conclude all the choices of people. Therefore, in the future study, I should widen the scope of my investigation to include more factors that may influence the result, such as the different age_group, the makeup style they like and the package of the lipstick.

Comments or questions

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