Midterm Exam

Ziyi Bai

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.

```
setwd("/Users/baiziyi/Desktop/MSSP/MA678")
coffee <- read.csv("coffee.info.csv", header =T)
head(coffee)
```

```
##   X    coffee_shop tianyang jiayi tongxu siyuan huanbi chenting kunyu chenghao
## 1 1      Starbucks      120    70      0     70     30        0    60       30
## 2 2          Costa        0     0      0      0      0       35     0        0
## 3 3 Pacific_coffee        0     0      0      0      0        0     0        0
## 4 4 Paris_Baguette        0     0     30     30      0        0     0        0
## 5 5         Luckin        0    36      0     25      0        0    25        0
## 6 6  Self_operated        0     0      0      0     30      210    30       60
##   zhitian ziyi
## 1       0   30
## 2       0    0
## 3       0    0
## 4       0   50
## 5      25   60
## 6      45   70
```

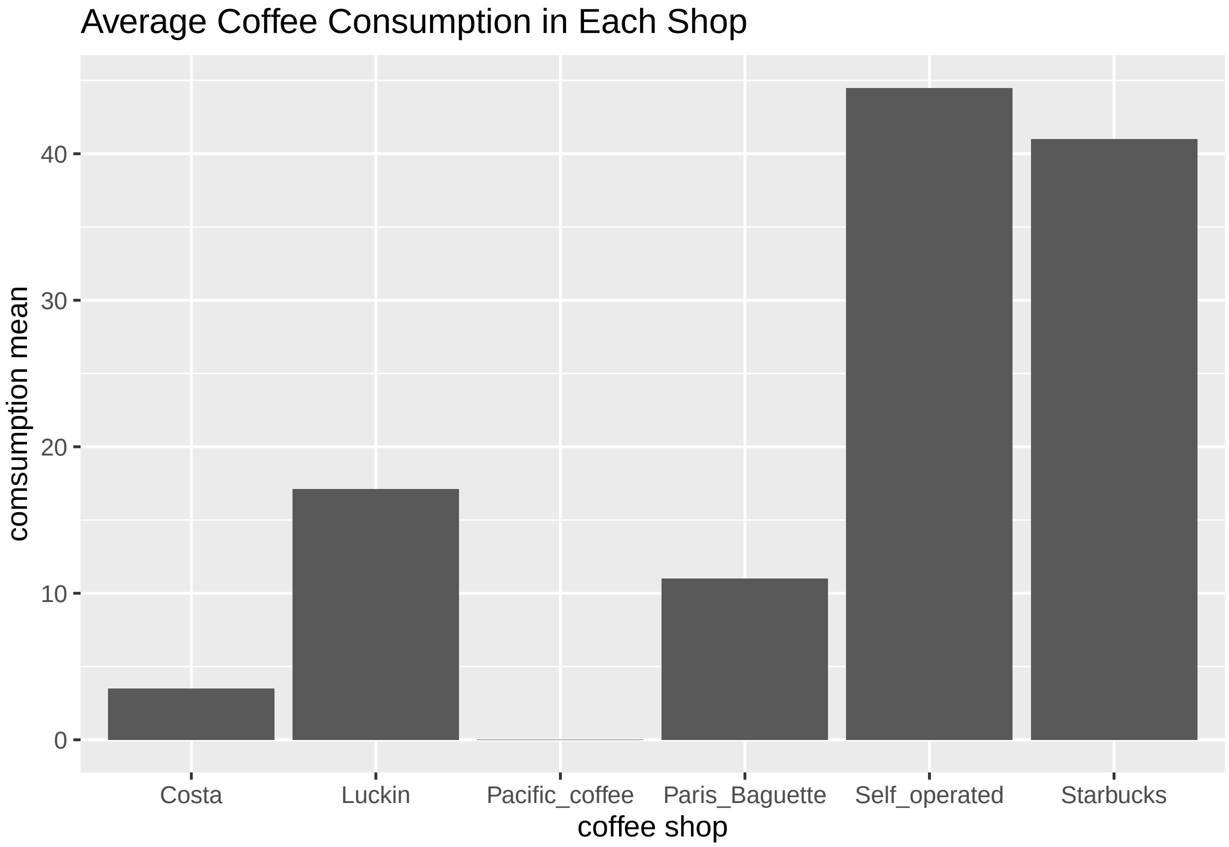
I am a coffee addicter. Every week I spend nearly 100 yuan in my preferred coffee shop, which is a considerable amount of money for a student. I've noticed that most of my friends also drink coffee regularly, so I am wondering how much they spend on coffee each week. In my dataset, I have 10 variables and I list 5 most common coffee shop in China and self-operated column for all other self-operated coffee. Based on how much they spend on each shop, I think I am able to conclude which coffee shop is the most popular one among my friends.

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.

```r
attach(coffee)
coffee$mean <- (tianyang + jiayi + tongxu + siyuan + huanbi + chenting + kunyu + chenghao + zhitian + z

ggplot(coffee, aes(y=mean, x=coffee_shop)) +
  geom_col() +
  labs(title = "Average Coffee Consumption in Each Shop") +
  xlab("coffee shop") +
  ylab("comsumption mean")
```



Based on above graph, it seems that Self-operated coffee shop is the most famous one. I suppose it is because in this kind of coffee shop, customers have more choices and they are able to choose more coffee flavor. Starbucks is the second famous coffee shop among my friends.

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.t2n.test(n1=40,n2=20,d=NULL,sig.level=0.05,power=0.8)
```

```
##
##      t test power calculation
##
##              n1 = 40
##              n2 = 20
##               d = 0.7802503
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

I divide my dataset into 2 groups, the first group is foreign coffee shops, which have 40 observations, the second group is local coffee shops, which have 20 observations. Through power analysis, my d is equal to 0.7802503. Because my d is great than 0.5 and less than 1, so I think my sample size is ok for this 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.

Because my original dataset just includes each person's consumption, which is hard to make any regression. So, I decide to analysis the relationship between gender and their consumption, to figure out which gender tends to consume more coffee.

```
# I create a dataset about each person's total coffee consumption and their gender. I code female into
sum(tianyang)
```

```
## [1] 120
```

```
sum(jiayi)
```

```
## [1] 106
```

```
sum(tongxu)
```

```
## [1] 30
```

```
sum(siyuan)
```

```
## [1] 125
```

```
sum(huanbi)
```

```
## [1] 60
```

```
sum(chenting)
```

```
## [1] 245
```

```
sum(kunyu)
```

```
## [1] 115
```

```r
sum(chenghao)
```

```
## [1] 90
```

```r
sum(zhitian)
```

```
## [1] 70
```

```r
sum(ziyi)
```

```
## [1] 210
```

```r
consum <- c("120", "106","30","125","60","245","115","90","70","210")
gender <- c("0","1","1","1","1","0","1","0","0","1")
coffee_2 <- data.frame(consum, gender)
head(coffee_2)
```

```
##   consum gender
## 1    120      0
## 2    106      1
## 3     30      1
## 4    125      1
## 5     60      1
## 6    245      0
```

```r
coffee_2$consum <- as.numeric(coffee_2$consum)
coffee_2$gender <- as.numeric(coffee_2$gender)
fit_1 <- lm(consum ~ gender, data=coffee_2)
summary(fit_1)
```

```
##
## Call:
## lm(formula = consum ~ gender, data = coffee_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.667 -46.062  -6.458  14.833 113.750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   131.25      34.30   3.826  0.00504 **
## gender        -23.58      44.28  -0.533  0.60881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.6 on 8 degrees of freedom
## Multiple R-squared:  0.03424,    Adjusted R-squared:  -0.08648
## F-statistic: 0.2836 on 1 and 8 DF,  p-value: 0.6088
```

Because I only have two variables, so I use lm model to fit my regression. The p-value of my intercept is 0.00504, which is very significant and the p-value of gender is 0.60881.

Validation (10pts)

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

```r
fit_2 <- stan_glm(consum ~ gender, data= coffee_2, refresh=0 )
print(fit_2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      consum ~ gender
##  observations: 10
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 130.7   35.9
## gender      -22.6   46.7
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 70.4   17.4
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
# Leave-one-out cross validation
loo_1 <- loo(fit_2)
print(loo_1)
```

```
##
## Computed from 4000 by 10 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo    -58.8 2.2
## p_loo         2.7 1.0
## looic       117.5 4.5
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     8     80.0%   1286
##  (0.5, 0.7]   (ok)       2     20.0%   355
##    (0.7, 1]   (bad)      0      0.0%   <NA>
##    (1, Inf)   (very bad) 0      0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

I use leave-one-out cross validation to check the validation of my model. In the first step, I refit my model using stan_glm (the family is gaussian). Based on the result, my Monte Carlo SE of elpd_loo is 0.1, which is the smaller the better, so my model makes sense but can be better if I have larger dataset. All pareto k estimates are less than 0.7, which is ok for this model.

```
# K-fold cross validation
kfold_1 <- kfold(fit_2, k=10)
```

```
## Fitting model 1 out of 10

## Fitting model 2 out of 10

## Fitting model 3 out of 10

## Fitting model 4 out of 10
## Fitting model 5 out of 10

## Fitting model 6 out of 10

## Fitting model 7 out of 10

## Fitting model 8 out of 10

## Fitting model 9 out of 10

## Fitting model 10 out of 10
```

```r
print(kfold_1)
```

```
##
## Based on 10-fold cross-validation
##
##            Estimate  SE
## elpd_kfold    -58.9 2.4
## p_kfold          NA  NA
## kfoldic       117.7 4.7
```

```r
fit_3 <- update(fit_2, prior=hs())
kfold_2 <- kfold(fit_2, k=10)
```

```
## Fitting model 1 out of 10

## Fitting model 2 out of 10

## Fitting model 3 out of 10

## Fitting model 4 out of 10

## Fitting model 5 out of 10

## Fitting model 6 out of 10

## Fitting model 7 out of 10

## Fitting model 8 out of 10

## Fitting model 9 out of 10

## Fitting model 10 out of 10
```

```r
loo_compare(kfold_1, kfold_2)
```

```
##       elpd_diff se_diff
## fit_2  0.0       0.0
## fit_2 -0.1       0.1
```

To double check my model, I use k-fold cross validation as well (which may not necessary in validation check). After I made update of my model, the elpd_diff changes from into -0.1, which negative is favored. se_diff remains at 0.0. So, I do make a little bit enhancement of my model, but overall my original model is ok.

Inference (10pts)

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

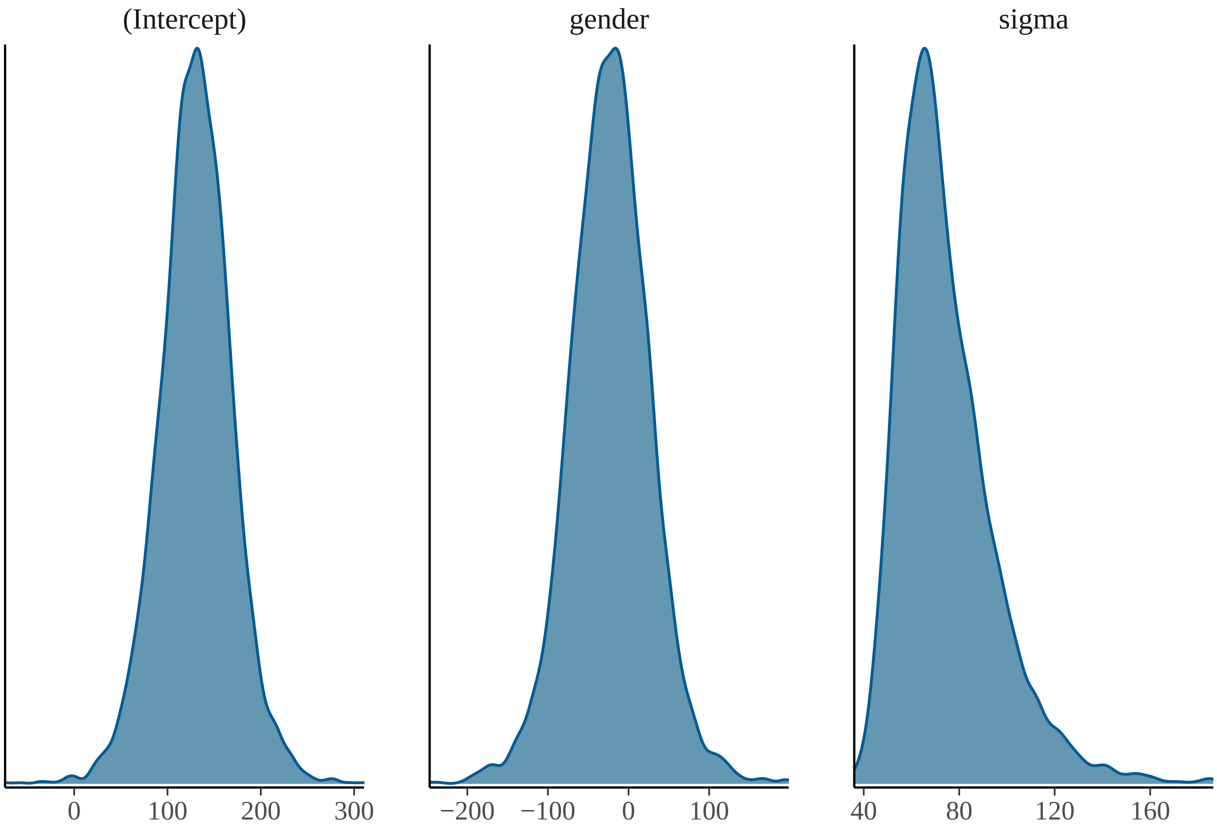
```r
# Calculate CI
confint(fit_1)
```

```
##                  2.5 %    97.5 %
## (Intercept)   52.14943 210.35057
## gender      -125.70173  78.53506
```

```
# Bayesian Inference
sims <- as.matrix(fit_2)
head(sims)
```

```
##           parameters
## iterations (Intercept)     gender    sigma
##       [1,]    127.3629  -46.40336 47.16917
##       [2,]    132.9948  -45.39732 45.76946
##       [3,]    137.9123  -45.67497 65.94604
##       [4,]     99.8508   22.39078 74.24789
##       [5,]    154.6277  -71.06771 57.93413
##       [6,]    167.4617 -101.11678 55.72393
```

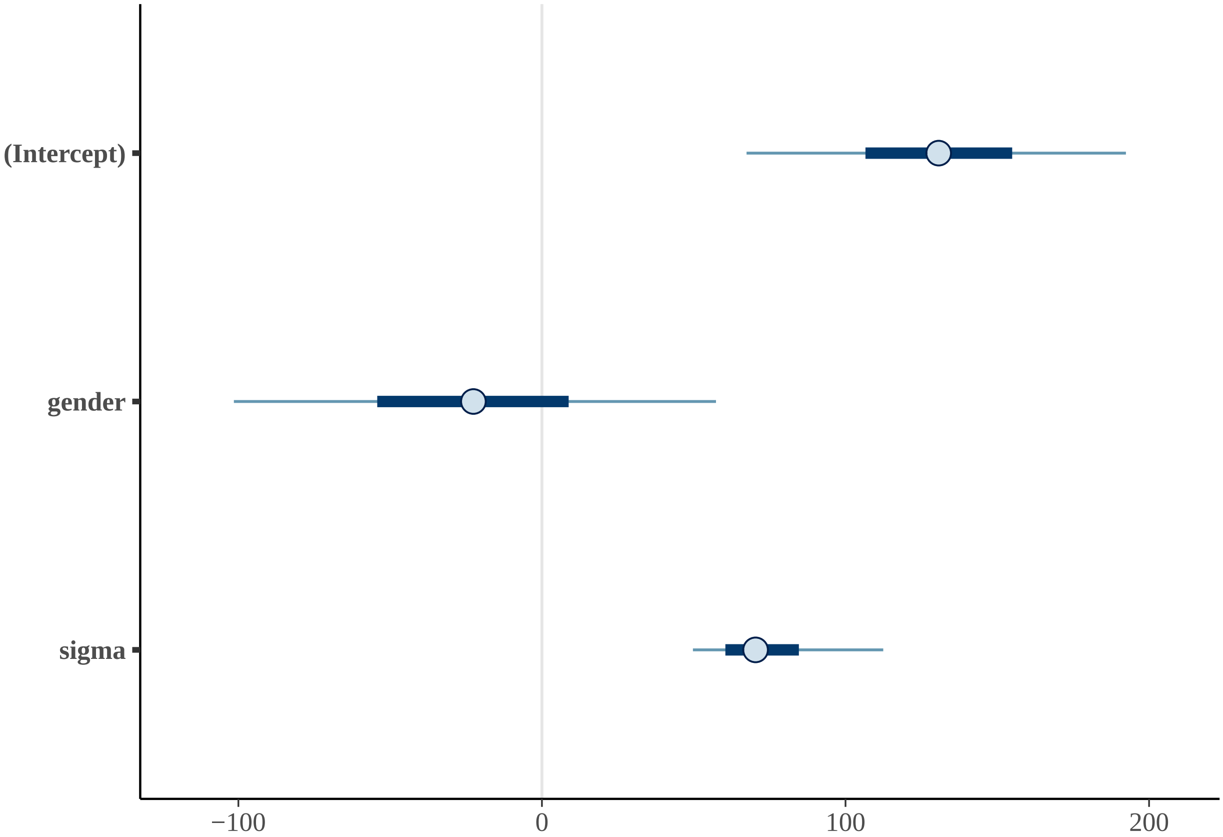
```
mcmc_dens(sims)
```



```
posterior_interval(fit_2)
```

```
##                      5%       95%
## (Intercept)    67.40397 192.37360
## gender       -101.48033  57.34023
## sigma          49.74239 112.43763
```

```
mcmc_intervals(sims)
```



The 95% confidence interval of intercept is from 52.14943 to 210.35057 and the 95% confidence interval of gender is from -125.70173 to 78.53506.

`as.matrix()` allows me to access the matrix of posterior simulations to express uncertainty about a parameter estimate.

The intercept credible interval of my model is from 67.87207 to 189.63582 at 95% confidence and the coefficient credible interval of gender is from -102.37199 to 56.16007.

Discussion (10pts)

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

First of all, because my friends consume most self-operated coffee shop, so my original question is successfully answered: Self-operated coffee shop is the most famous type of coffee shop among my friends.

Secondly, I assume are girls or boys tend to consume more coffee. Based on my model, the intercept of my model is 131.25, so that boys tend to spend 131.25 yuan on coffee per week; the coefficient of gender is -23.58, so girls tend to spend 23.58 yuan less than boys a week when keeping everything else constant.

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.

One of the limitation of my dataset is that it is not large enough. I only did research about 10 of my friends, which is not representative. Also, because my data is estimated coffee consumption in each coffee shop based on the memory of my friends, so there is a bias in my dataset. Then, another thing may influence my data is that if there is only Starbucks near a person's home, he or she may spends more on Starbucks but he or she may not really likes it's coffee.

To fix these problem, I should make my data more reliable. Enlarge my dataset is one thing. Also, I can change the tested people into my neighbors, so that they have same choices of coffee shops and then calculate one month average coffee consumption on each coffee shop.

Comments or questions

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

I think this is a good experience for me to act consulting process, which is not as simple as I imagine. To make a successful consulting, I should take lots of influences into consideration. Maybe it is normal that I cannot get any conclusions when I work on other projects, but the process is fun!