Midterm Exam

Yu Du

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.

The data is collected from the survey I made online and seven friends had answered this survey, so there are seven observations with six variables. The data basically collects the information about how many cups of coffee my friends drink in a week, the prices they would pay etc.

The project is to estimate the relationship between the outcome my friends' weekly spending on drinking milktea and two variables: the number of drinks my friends have every week, the cup of size they would buy.

Since my friends and I are all love drinking Milktea, I am interested in knowing about their answers and working on this project to share my result with them later.

```
Mkt<-read.csv("/Users/duyu/Desktop/MA 678/Yu Du Data Collection/Data Collection.csv",header=TRUE)

#Calculate the mean of prices my friends would pay.
Mkt[3,7]<-mean(c(15:30))
Mkt[4,7]<-mean(c(15:30))
Mkt[5,7]<-mean(c(10:20))
Mkt[7,7]<-mean(c(20:30))
Mkt$Price.in.RMB=as.numeric(Mkt$Price.in.RMB)

#Code 0 for male, code 1 for female.
Mkt$Sex=ifelse(Mkt$Sex=="Female",1,0)
```

```r
colnames(Mkt)[2]<-"female"

#Rename the column names make them more easier to read and write in the next questions.
colnames(Mkt)[4]<-"drinks"
colnames(Mkt)[6]<-"cup"
colnames(Mkt)[5]<-"healthy.drinks"

#Calculate the weekly spending for each person.
Mkt%<>%mutate(Total.Spending=paste(Price.in.RMB*drinks))
Mkt$Total.Spending<-as.numeric(Mkt$Total.Spending)
Mkt<-Mkt[,-1]
Mkt
```

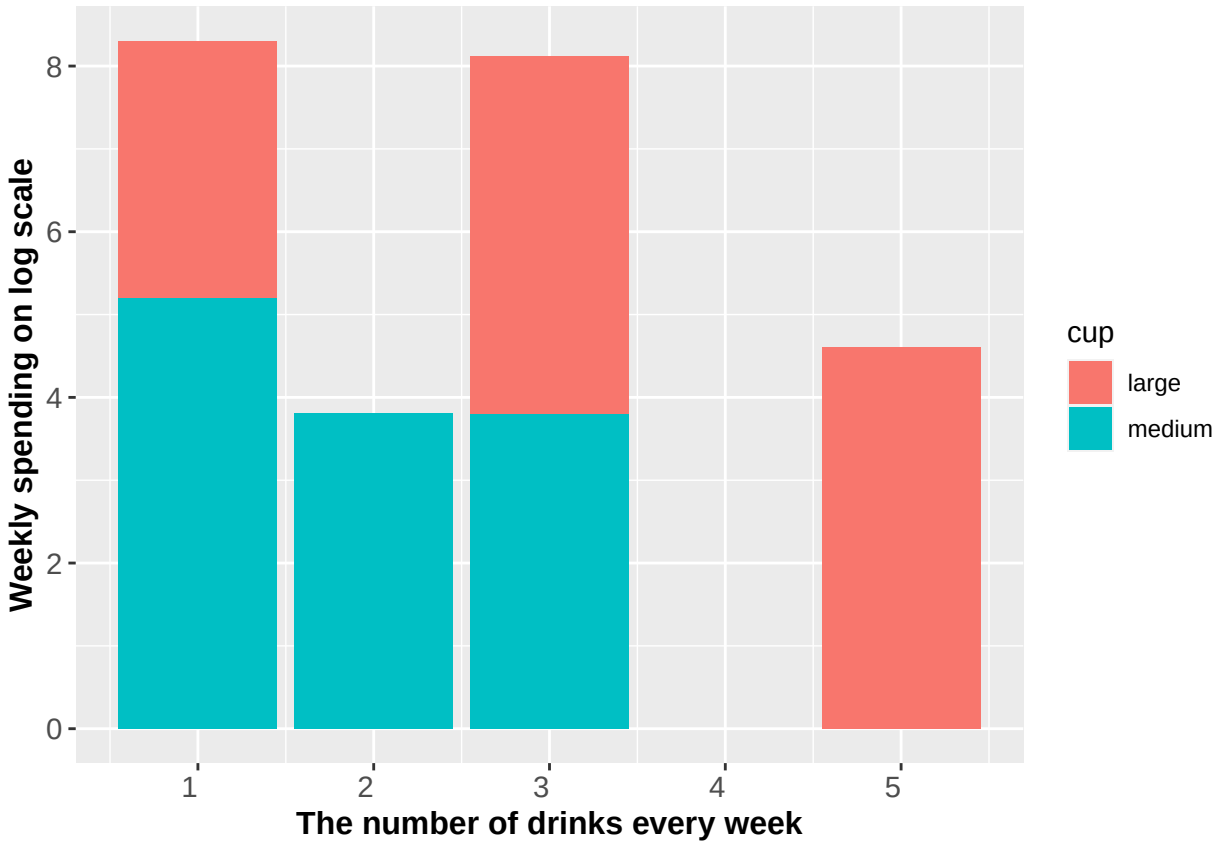
```
##   female Age drinks healthy.drinks    cup Price.in.RMB Total.Spending
## 1      1  22      5              2  large         20.0          100.0
## 2      1  22      3              1 medium         15.0           45.0
## 3      1  22      2              1 medium         22.5           45.0
## 4      1  22      1              1  large         22.5           22.5
## 5      1  22      1              0 medium         15.0           15.0
## 6      1  21      1              1 medium         12.0           12.0
## 7      1  22      3              1  large         25.0           75.0
```

Variables are: female: male=0, female=1 Age:age in numeric value. drinks: the number of drinks every week. healthy.drinks: the number of drinks considered as drinking in a healthy way. cup:the size of cup (medium or large). Price.in.RMB: the price of Milktea they normally would pay in RMB. Total.Spending: Weekly spending on drinking Milktea.

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
ggplot(Mkt,aes(x =drinks,y =log(Total.Spending), fill=cup)) + geom_bar(stat = "identity")+
  theme(axis.text.x = element_text(angle = 0, hjust = 1),
        axis.text = element_text(size = 11),
        axis.title = element_text(size = 12, face = "bold")) +
  labs(x = "The number of drinks every week",y="Weekly spending on log scale")
```



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.

```
library(pwr)
pwr.f2.test(u =2, v =7-2-1 , f2 =NULL , sig.level= 0.05 , power =0.8 )
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 4
##              f2 = 3.20854
##       sig.level = 0.05
##           power = 0.8
```

```
#Calculate the sample size for the effect size is estimated as 0.5.
pwr.f2.test(u =2, v =NULL , f2 =0.5 , sig.level= 0.05 , power =0.8 )
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 19.55271
##              f2 = 0.5
```

```
##       sig.level = 0.05
##           power = 0.8
```

From the first calculation, f2=3.2 as using 80% power and sample size n=7. The effect size is very large.

From the second calculation, we can obtain that n=23 with d=0.5 (medium effect size), so it shows that the sample size n=7 is not enough. Effect size is not the same as the statistical significance.

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.

```
#Code medium=0, large=1 in the column cup
Mkt$cup=ifelse(Mkt$cup=="large",1,0)

#Fit two models and choose one better model later.
Model1= stan_glm(Total.Spending~drinks+cup, data=Mkt)
```

```
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 9.2e-05 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0.92 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 0.087109 seconds (Warm-up)
## Chain 1:                0.060759 seconds (Sampling)
## Chain 1:                0.147868 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 2e-05 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.2 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
```

```
## Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2:
## Chain 2:  Elapsed Time: 0.104147 seconds (Warm-up)
## Chain 2:                0.080014 seconds (Sampling)
## Chain 2:                0.184161 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 2.5e-05 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.25 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3:
## Chain 3:  Elapsed Time: 0.095507 seconds (Warm-up)
## Chain 3:                0.072262 seconds (Sampling)
## Chain 3:                0.167769 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 4.2e-05 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.42 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
```

```
## Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 4:
## Chain 4:  Elapsed Time: 0.091124 seconds (Warm-up)
## Chain 4:                0.054423 seconds (Sampling)
## Chain 4:                0.145547 seconds (Total)
## Chain 4:
```

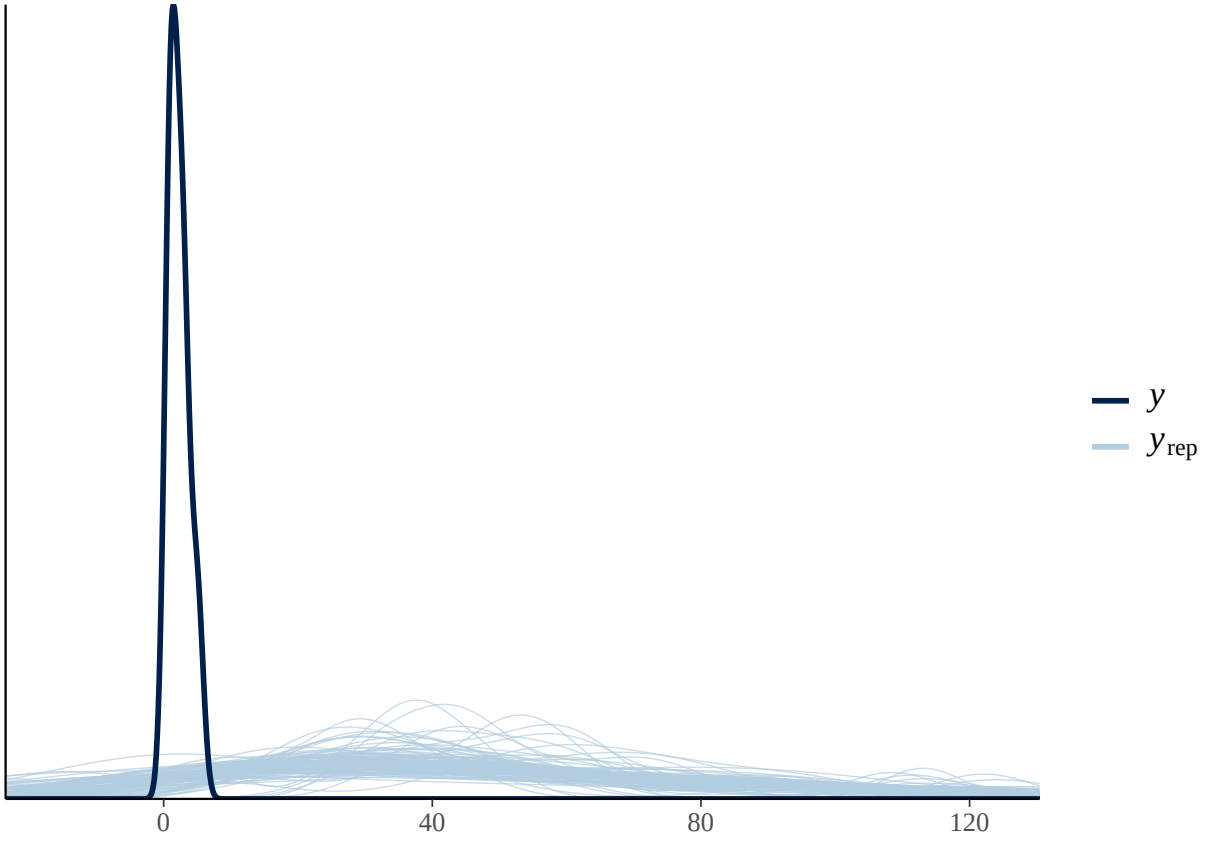
```
print(Model1,2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      Total.Spending ~ drinks + cup
##  observations: 7
##  predictors:   3
## ------
##             Median MAD_SD
## (Intercept) -3.48   7.98
## drinks      18.77   3.37
## cup         12.97   9.39
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 10.93   4.20
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

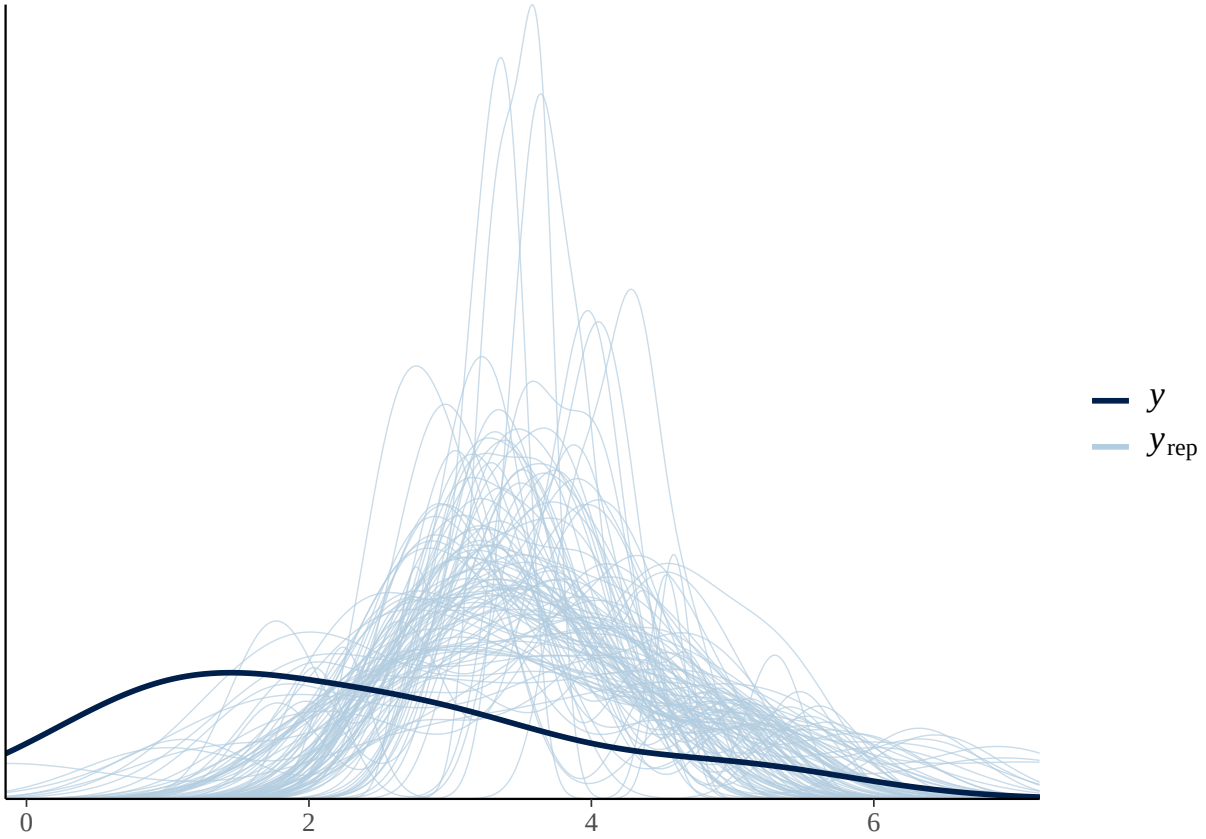
```
Model2= stan_glm(log(Total.Spending)~drinks+cup, data=Mkt, refresh=0)
print(Model2,2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      log(Total.Spending) ~ drinks + cup
##  observations: 7
##  predictors:   3
## ------
##             Median MAD_SD
## (Intercept) 2.43   0.33
## drinks      0.44   0.14
## cup         0.26   0.37
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.45   0.16
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
yrep_1 <- posterior_predict(Model1)
ppc_dens_overlay(Mkt$drinks,yrep_1[1:100,])
```




```
yrep_2 <- posterior_predict(Model2)
ppc_dens_overlay(Mkt$drinks, yrep_2[1:100,])
```



Compare model 1 and model 2: The uncertainties in parameters for model 2 are smaller. Based on the posterior predictive checks, model 2 shows a better fit than model 1 though both models do not perform a good fit. My choice is Model 2 and I will check validations in the next step to further confirm my thought.

Error model: Gaussian Link: identity

Choosing this model: the outcome is a continuous variable with continuous predictor and categorical predictor.

Validation (10pts)

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

```
#Cross Validation:
loo_1=loo(Model1,k_threshold = 0.7)
print(loo_1)
```

```
##
## Computed from 4000 by 7 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo    -28.7 0.9
## p_loo         3.2 0.6
## looic        57.5 1.8
## ------
## Monte Carlo SE of elpd_loo is 0.2.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     2     40.0%   781
##  (0.5, 0.7]   (ok)       3     60.0%   469
```

```
##    (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.
```

```
loo_2=loo(Model2,k_threshold = 0.7)
print(loo_2)
```

```
##
## Computed from 4000 by 7 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo     -6.6 1.3
## p_loo         3.3 1.0
## looic        13.2 2.5
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     3     50.0%   765
##  (0.5, 0.7]   (ok)       3     50.0%   201
##    (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.
```

```
loo_compare(loo_1, loo_2)
```

```
##        elpd_diff se_diff
## Model2   0.0       0.0
## Model1 -22.1       1.4
```

From the loo_compare, model 2 is appropriate since elpd_diff=0.

Inference (10pts)

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

```
sims <- as.matrix(Model2)
#sims
#Confidence Interval:
##for coefficient of drinks:
quantile(sims[,2], c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1103499 0.7565165
```

```
##for coefficient of cup:
quantile(sims[,3], c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -0.5971357  1.1982934
```

Notice that both 95% confidence intervals shown above are too wide, indicating that we have a little knowledge about the effects of variables and further information should be included.

Discussion (10pts)

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

Both the variables drinks and cup have a positive effect on the weekly spending based on the model 2 shown above, From the estimates of coefficients: The interpretations can be: If drinks my friend have every week increased by one cup, their weekly spending will increase by 44%, holding cup at constant. If cup my friend buy change from medium to large, their weekly spending will increase by 26%, holding drinks at constant.

**However, from the posterior predictive check, validation, and inference, the model does not give a best fit due to the small dataset, so we we cannot conclude that both variables have a positive effect on weekly earnings. Both the dataset and model should be improved.

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.

Limitations: 1. The data is small, including only 7 observations. 2. The prices my friends would pay on drinking Milktea are not for one specific brand and prices may change due to different brands, so the total spending here may not equal to their actual weekly spending. 3. Some markets also sell Milktea in a fast food packaging, so they may spend more on buying Milktea every week. 4, There is not much variation in the variables female and Age, so it is hard to include them in a model if I want add more predictor to the model.

Future Opportunities: 1. Asking more friends to increase the number of observations in the data. 2. Choosing one specific brand and asking how many times they visit and how much they pay. 3. Asking if they would buy Milktea in markets. 4. Asking my relatives or friends' relatives, therefore the dataset can have different ages and including both female and male. Adding more predictors is possible.

Comments or questions

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

This is the first time I analyze on my own data and on a very small data, I am very happy to to discover the limitation and future opportunity of my own data through the data analysis.