Midterm Exam

Fan Feng

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.

Introduction of the dataset: The data is collected to explore the differences in the habits of downloading the mobile applications between boys and girls. We collect the data using electronic questionnaire and the people surveyed are mainly my good friends who are graduate students or just start working.

Specifically, we have 18 samples in total. For each sample, we collected gender information and the numbers of 4 different types of applications in their mobile phones, including video, music, shopping, and study.

Firstly, let's import the data and check the summary information:

```
app <- read_excel("D:/appdataset.xlsx")
setnames(app,"sex","gender")
app$gender = as.factor(ifelse(app$gender == 'female', 0, 1))
summary(app)
```

```
##        ID        gender   video_app     music_app      shopping_app
##  Min.   : 1.00   0:9    Min.   :1.0   Min.   :1.000   Min.   : 1.000
##  1st Qu.: 5.25   1:9    1st Qu.:2.0   1st Qu.:2.000   1st Qu.: 4.000
##  Median : 9.50          Median :3.5   Median :2.000   Median : 6.000
##  Mean   : 9.50          Mean   :3.5   Mean   :2.167   Mean   : 6.111
##  3rd Qu.:13.75          3rd Qu.:5.0   3rd Qu.:2.750   3rd Qu.: 6.000
##  Max.   :18.00          Max.   :7.0   Max.   :4.000   Max.   :16.000
##    study_app
##  Min.   :0.00
##  1st Qu.:2.00
##  Median :2.50
##  Mean   :3.50
##  3rd Qu.:5.75
##  Max.   :8.00
```

Based on the output, we can find that we have a balanced data set, each group has 9 samples. And we have no NAs in our data set.

The Comparison of interest: We want explore the differences in the habits of female and male downloading apps by analyzing this data set.

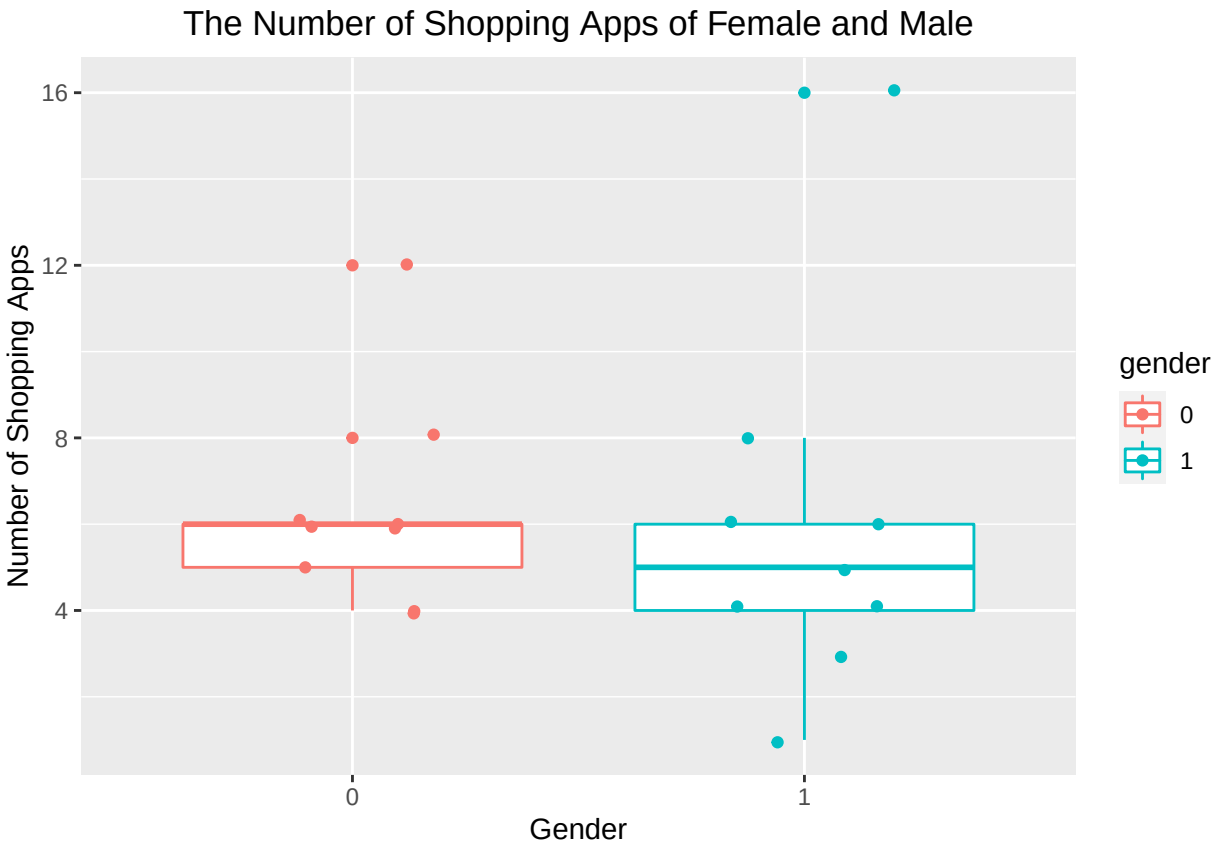
More specifically, we can ask following questions: 1. Do women tend to download more shopping apps than men do? 2. Do women and man have different preferences for downloading video apps? 3. Are there any interactions between different types of apps?

In this analysis, we mainly focus on the first question.

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.

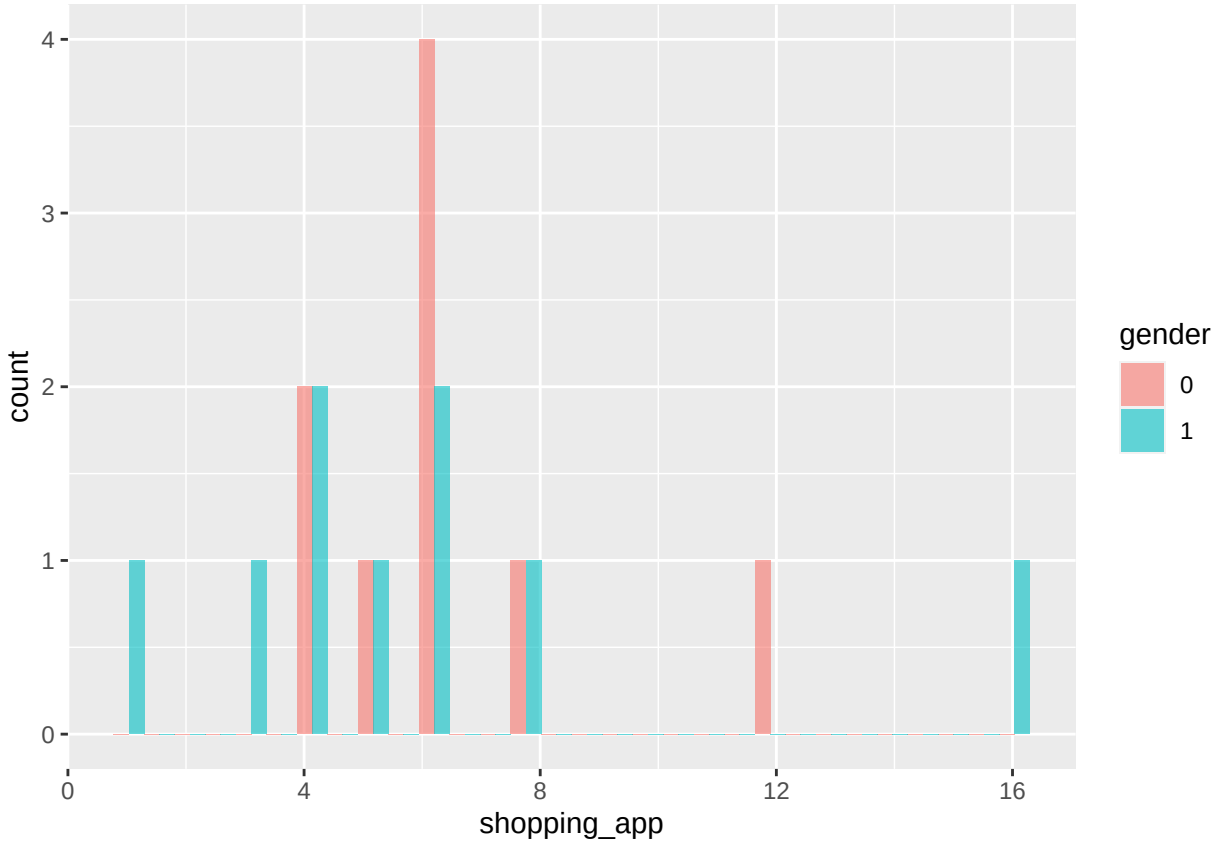
```
ggplot(app, aes(x = gender, y =shopping_app,color = gender)) +
  geom_boxplot() + geom_jitter(width = 0.2, height = 0.1) +
  labs(title="The Number of Shopping Apps of Female and Male",
       x="Gender", y = "Number of Shopping Apps") +
  theme(plot.title = element_text(hjust = 0.5))
```



According to box plot above, it seems that the distributions of the number of shopping apps are different between the male and female. But we can not draw any conclusion hastily without detailed analysis.

```
ggplot(app, aes(x = shopping_app, fill = gender)) +
  geom_histogram(alpha = 0.6, position ='dodge')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

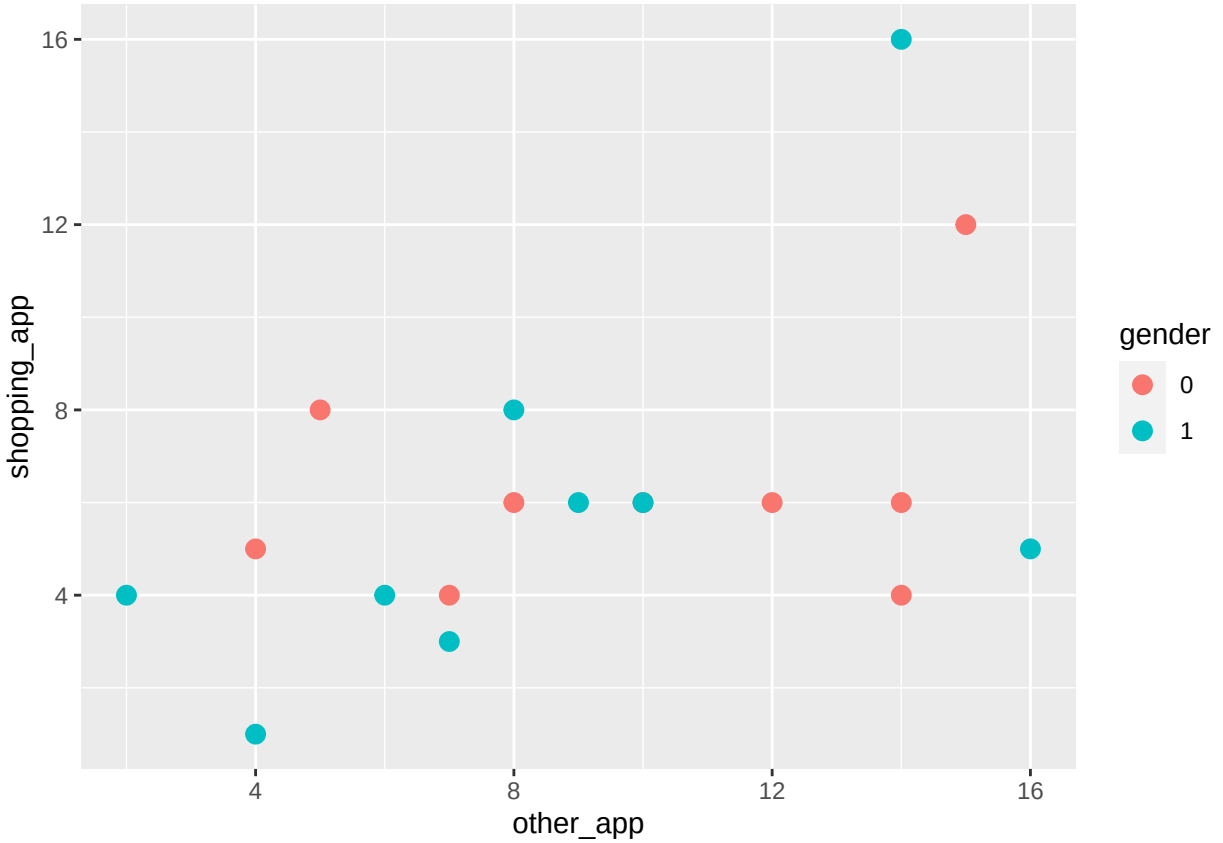


In addition, we can notice that, for the count of shopping apps, there are so called outliers for both genders, which are 12 and 16 respectively. To be honest, it is very rare that one person installs so many apps with similar function. However, after the fact check with my interviewees, I know that those two outliers are real and not a product of errors so we will treat them as 'novelties' in the data.

Next, let's creat a new variable 'other_app' which is the sum of other 3 types of apps and make a plot to show its relationship with 'shopping_app'

```
#creat variable 'total'
app$other_app <- apply(app[,c(3,4,6)],1,sum)
```

```
ggplot(app, aes(x = other_app, y = shopping_app, color = gender)) +
  geom_point(size = 3)
```



From this plot, we can find that two persons who downloaded 12 and 16 shopping apps also downloaded many other apps. We can guess that the reason why they downloaded so many shopping apps is not only because they love shopping, but they also have the habit of downloading a lot of various apps.

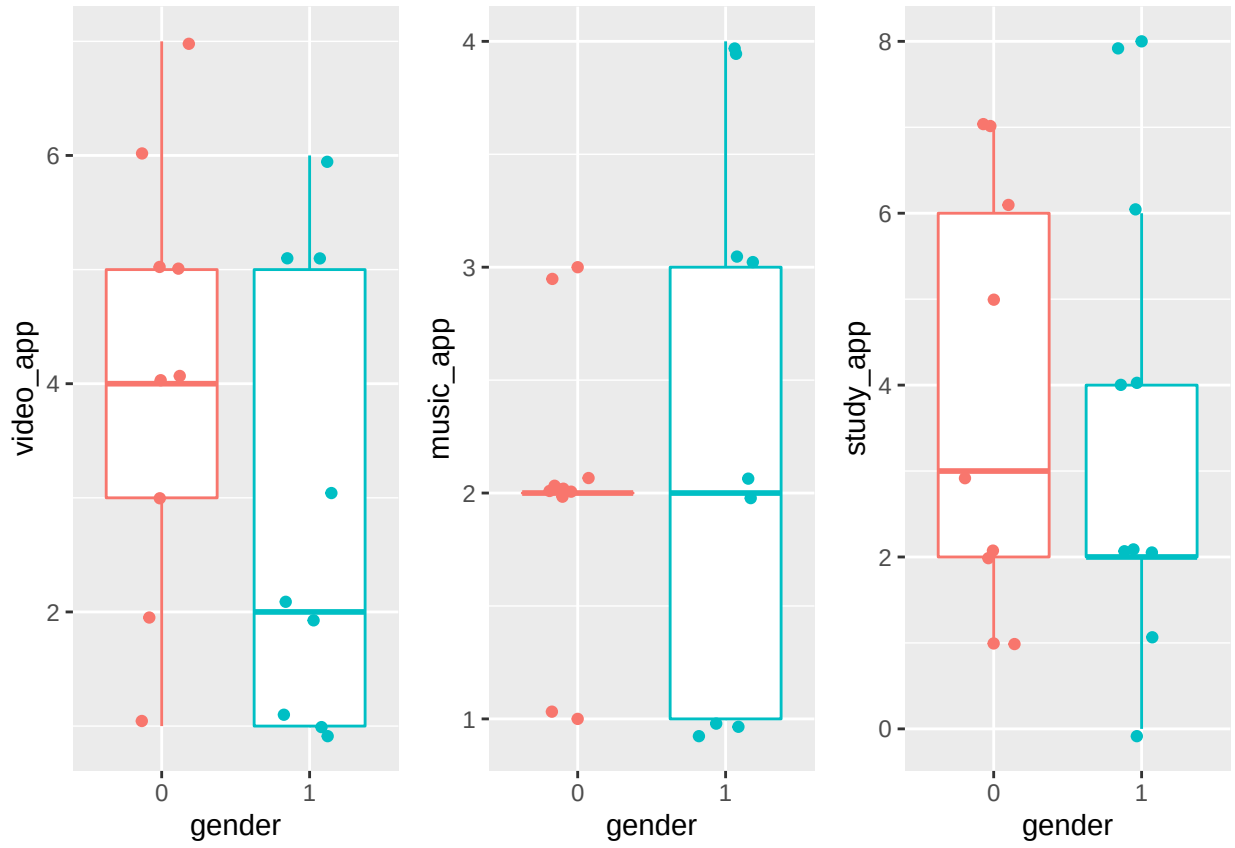
In addition, we also make some plots to check the relationship between the gender and other types of apps.

```
p1 <- ggplot(app, aes(x= gender, y = video_app, color = gender)) +
  geom_boxplot() + geom_jitter(width = 0.2, height = 0.1) +
  theme(legend.position = "none")

p2 <- ggplot(app, aes(x = gender, y = music_app, color = gender)) +
  geom_boxplot() + geom_jitter(width = 0.2, height = 0.1) +
  theme(legend.position = "none")

p3 <- ggplot(app, aes(x = gender, y = study_app, color = gender)) +
  geom_boxplot() + geom_jitter(width = 0.2, height = 0.1) +
  theme(legend.position = "none")

grid.arrange(p1,p2,p3, ncol = 3,nrow = 1)
```



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.t.test(n=8,d=NULL,sig.level=0.05,power = 0.8,type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 8
##               d = 1.50665
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

d = 1.50665

According to the output, the effect size that I will be able to detect is 1.5. In my opinion, this effect size is relatively too large and it means that the distributions of two groups need to be quitely different if I want to detect the differences.

More ideally, I think that it is better that the value of d is between 0.5 and 1. So let's calculate the sample size if d = 0.6

```
pwr.t.test(n = NULL, d = 0.6, sig.level=0.05,power = 0.8,type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 44.58577
##               d = 0.6
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

It shows that if the effect size is 0.6, I need a much larger sample size which means I need 45 samples for each group. It is clear that the sample size of my dataset is far away from enough in such condition.

The effect size from the fitted model should not be used since it is overestimated and we need to detect more tiny differences.

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.

```
fit_1 <- glm(shopping_app ~ gender, data = app)
summary(fit_1)
```

```
##
## Call:
## glm(formula = shopping_app ~ gender, data = app)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8889  -1.8889  -0.3333   0.1111  10.1111
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3333     1.1634   5.444 5.41e-05 ***
## gender1      -0.4444     1.6452  -0.270    0.791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12.18056)
##
##     Null deviance: 195.78  on 17  degrees of freedom
## Residual deviance: 194.89  on 16  degrees of freedom
## AIC: 99.959
##
## Number of Fisher Scoring iterations: 2
```

```r
fit_2 <- glm(shopping_app ~ video_app + music_app + study_app +gender,
           data = app)
summary(fit_2)
```

```
##
## Call:
## glm(formula = shopping_app ~ video_app + music_app + study_app +
##     gender, data = app)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3330  -2.1806  -0.0722   1.0905   6.8475
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.238581   2.420234   0.925   0.3718
## video_app    1.037331   0.524901   1.976   0.0697 .
## music_app   -0.089388   0.871686  -0.103   0.9199
## study_app    0.002368   0.399507   0.006   0.9954
## gender1      0.854516   1.637942   0.522   0.6107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 10.10741)
##
##     Null deviance: 195.78  on 17  degrees of freedom
## Residual deviance: 131.40  on 13  degrees of freedom
## AIC: 98.863
##
## Number of Fisher Scoring iterations: 2
```

```r
fit_3 <- step(fit_2, trace = FALSE)
summary(fit_3)
```

```
##
## Call:
## glm(formula = shopping_app ~ video_app, data = app)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5547  -2.3329   0.2949   1.0825   7.4829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7427     1.4180   1.934   0.0710 .
## video_app     0.9624     0.3551   2.710   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.386487)
##
##     Null deviance: 195.78  on 17  degrees of freedom
```

```
## Residual deviance: 134.18  on 16  degrees of freedom
## AIC: 93.241
##
## Number of Fisher Scoring iterations: 2
```

I decide to use the linear regression because the outcome variable is continuous and linear regression is very interpretable. The link function I choose is the default function which is used for linear regression.

In the fit_1 model, we only use one variable 'gender', the Residual deviance is 194.89 and the p value of gender1 is 0.791 and the p value of intercept is 5.41e-05.

In the fit_2 model, we use all the varibles we have, including gender. The Residual deviance is 131.40 and AIC is 98.863. The p value of gender is 0.6107 and none of the variables has a p value smaller than 0.05.

Based on the model selection, the fit_3 model is relatively the best regression model compared with fit_1 and fit_2. But, unfortunately, gender is not one of the variables being selected. The only variable used in fit_3 is 'video_app'. The Residual standard error of fit_3 model is 2.896 and the p value of intercept is 0.071 and the p value of video_app is 0.0155.

Validation (10pts)

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

Leave-one-out cross validation

cv1<-cv.glm(app, fit_1)$delta[1] print(cv1)

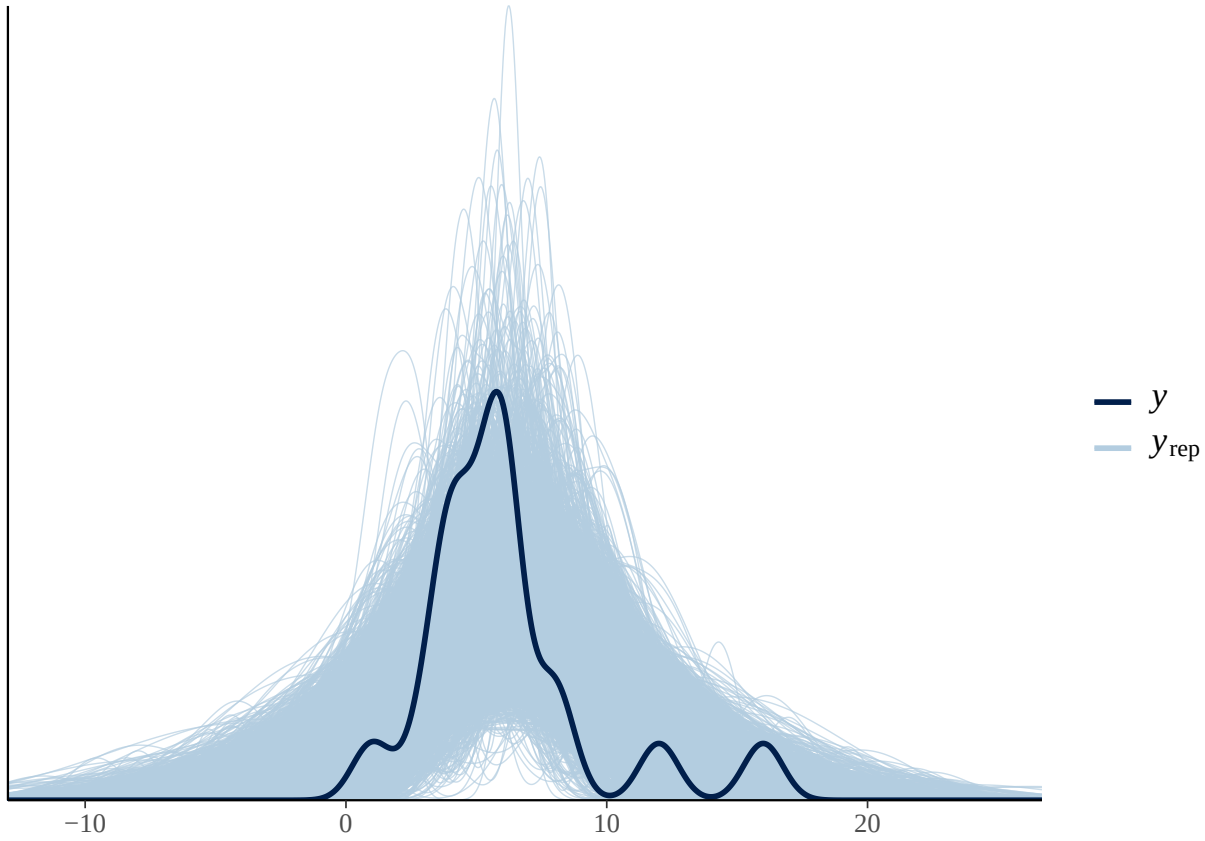
[1] 13.70312

cv2<-cv.glm(app, fit_2)$delta[1] print(cv2)

[1] 14.08581

The fit_1 has smaller value in the loocv so I choose the fit_1 as the more appropriate model.

```
fit_1_stan<- stan_glm(shopping_app ~ gender, data = app, refresh = 0)
ppc_dens_overlay(app$shopping_app, posterior_predict(fit_1_stan))
```




```
posterior_interval(fit_1_stan)
```

```
##                    5%      95%
## (Intercept)  4.277548 8.355611
## gender1     -3.357825 2.449568
## sigma        2.736417 4.851456
```

Inference (10pts)

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

H0: no difference between male and female in the numbers of shopping app
H1: females tend to download more apps than males

```
confint(fit_1)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %   97.5 %
## (Intercept)  4.053199 8.613467
## gender1     -3.669041 2.780152
```

The confidence interval all include 0 and its pvalue is larger than 0.05, so I accept the null hypothesis that there is no difference between female and male in the number of shopping apps.

Discussion (10pts)

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

Based on the analysis above, because my sample size is too small, I can not conclude that there are statistcally significant difference between female and male in regards of the number of shopping apps in the phone.

And based on the result fit_1 model: The intercept means that the predicted number of shopping apps in the phone is for a female is 6.3; The coefficient of gender meas that predicted number of shopping apps in the phone for a male is 0.4 fewer than a female.

Actually, the results of this model did show us that in this data set, the difference of the number of shopping apps is very tiny between male and female. But we still need more samples to verify this result.

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.

Concerns about the analysis:

1. This sample size is too small to represent the whole population. What is more, the samples are all at the similar ages and cities. So, if we really want to explore the differences between two genders, we need collect the data extensively or we just narrow the scope of the questions we are interested in.
2. I am not sure whether or not that two so called outliers cause some effects to our analysis, because it increase the variance of our data greatly. It is better if we consider this factor.

Future study plan

1. Improve the models by increasing the sample sizes and considering the effect of variance.
2. Perform similar analysis on video_app, study_app and music_app to explore the relationship with gender.
3. Consider more genders and collect more useful information such the time spent on the certain kind of app.

Comments or questions

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

I think this question is interesting because many people have such a stereotype that girls always have more shopping desires than boys and spend more time shopping around. However, with the development of e-commerce, is it true today? May be we need much more aspects of data, such as the time spending on shopping apps, to try to answer such kind of question. I am willing to explore this kind of question in depth in the future.