Midterm Exam

Haoqi Wang

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.

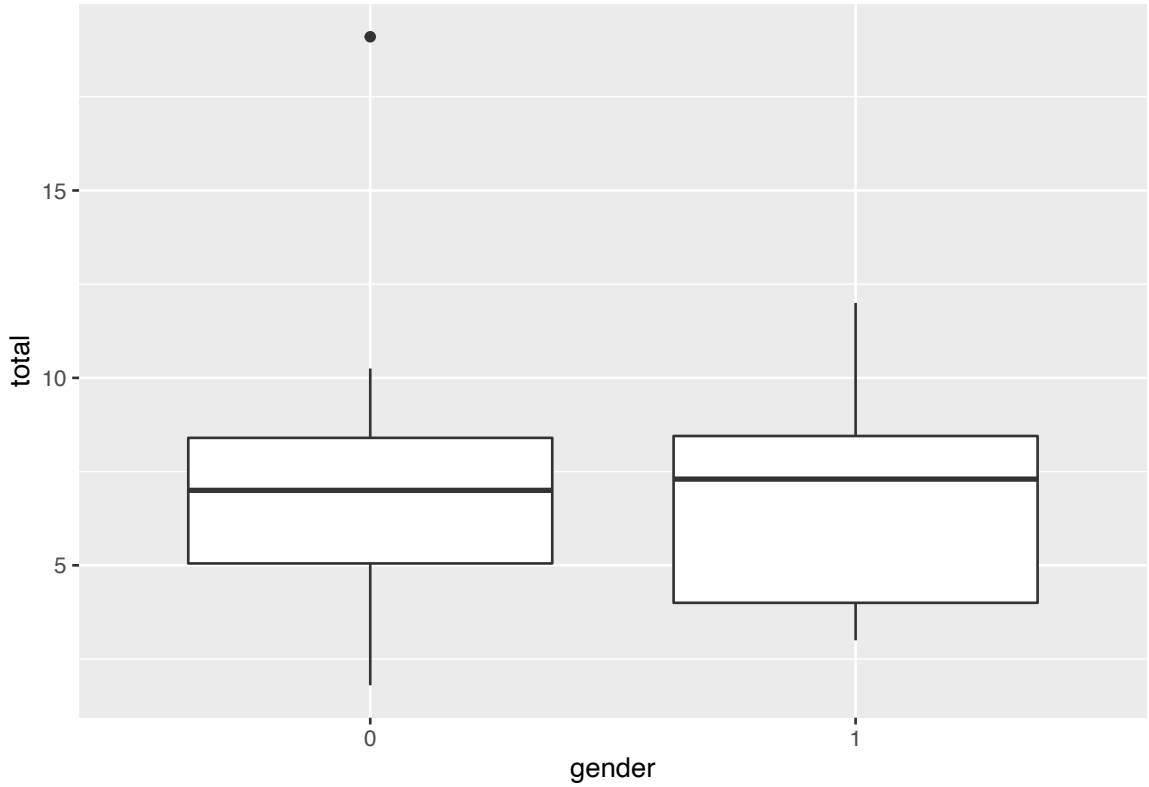
I collected this via online survey (https://www.wjx.cn/jq/95221038.aspx).The data shows the number of hours male and female spent on different phone apps. There are 30 classmates who did this survey totally, and the number of female who did this survey and the number of male are same. So my question is how gender and age play the role in time spent on smartphone. The dataset contains following variables: `user`: the person who did this survey `age`: the age of the person who did this survey `gender`: the gender of the person who did this survey: 0-female, 1-male `social.sharing.apps`: the number of hours that the person who did this survey spent on social sharing apps `entertainment.apps`: the number of hours that the person who did this survey spent on entertainment apps `shopping.apps`: the number of hours that the person who did this survey spent on shopping apps `game.apps`: the number of hours that the person who did this survey spent on game apps `reading.apps`: the number of hours that the person who did this survey spent on reading apps

```
#import data
phone<-read.csv("/Users/wanghaoqi/Desktop/2020 Fall/MA678/data collection.csv")
phone<-transform(phone,total=social.sharing.apps+entertainment.apps+shopping.apps+game.apps+reading.apps
#change `gender` to binary variable
phone$gender[phone$gender==2]<-0
phone$gender<-as.factor(phone$gender)
```

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.

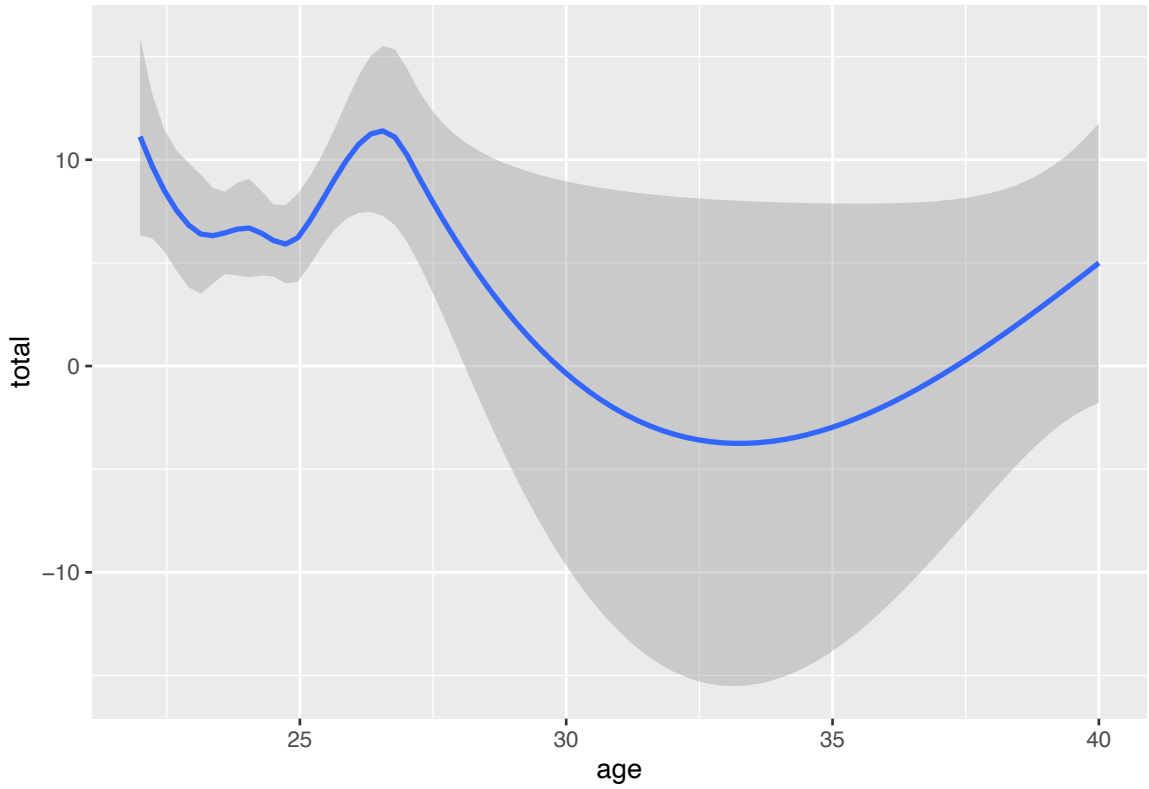
```
ggplot(data=phone,aes(x=gender,y=total))+
  geom_boxplot()
```




```
ggplot(data=phone,aes(x=age,y=total))+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 24

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at 24

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius 1
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0
```



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.

```
#calculate the effect size in my model
sd_male<-sd(phone$total[phone$gender==1])
mean_male<-mean(phone$total[phone$gender==1])
sd_female<-sd(phone$total[phone$gender==0])
mean_female<-mean(phone$total[phone$gender==0])
effect_size<-abs(mean_male-mean_female)/sqrt((sd_male^2+sd_female^2)/2)
effect_size
```

```
## [1] 0.157221
```

```
#detect the sample size
pwr.t.test(d=effect_size,sig.level = 0.05, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 636.0228
```

```
##               d = 0.157221
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
#detect the effect size
pwr.t.test(n =15,sig.level = 0.05, power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 15
##               d = 1.059797
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

When I choose two groups in my dataset, the power analysis shows that I need to have 636 observations per group when the effective size is 0.16. However, my dataset only have 15 observations per groups, which is not enough. When the power is 80% and the significant level is 0.05, the effective size is 1.06, which means the difference of two groups is big. That's the reason why I should not use effected size from the fitted model.

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<-stan_glm(total~gender+age,data = phone,refresh=0)
summary(fit)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      total ~ gender + age
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 30
##  predictors:   3
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept) 11.4    5.5  4.7  11.5  18.3
## gender1     -0.6    1.4 -2.2  -0.5   1.2
## age         -0.2    0.2 -0.4  -0.2   0.1
## sigma        3.7    0.5  3.1   3.6   4.4
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 7.2    1.0  6.0   7.2   8.4
```

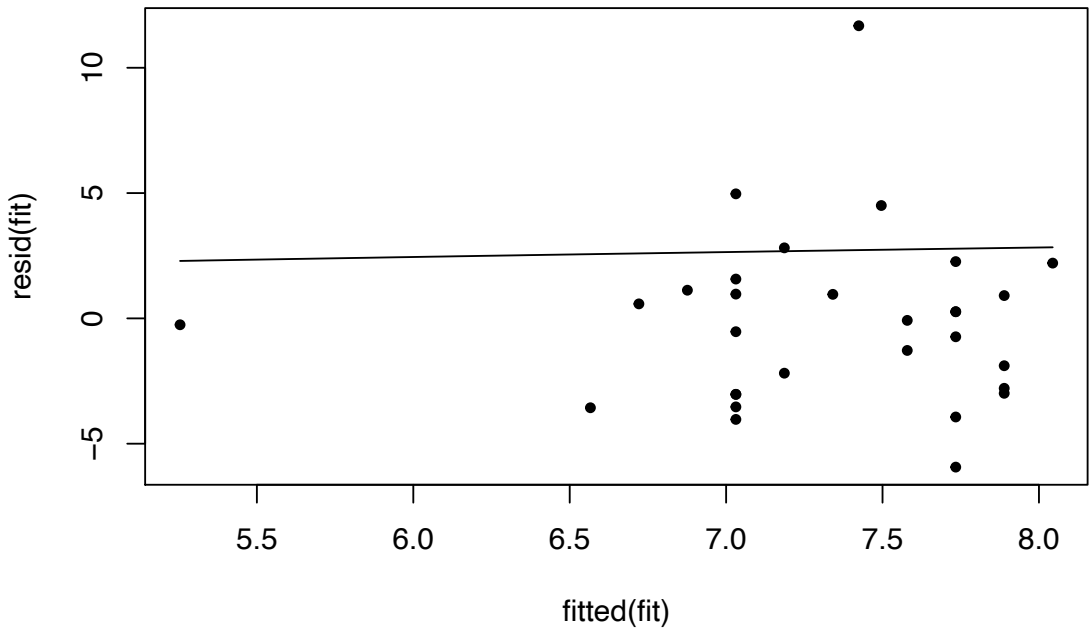
```
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.1  1.0  4070
## gender1       0.0  1.0  4576
## age           0.0  1.0  4140
## sigma         0.0  1.0  3504
## mean_PPD      0.0  1.0  4096
## log-posterior 0.0  1.0  1768
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

Since the predictor is the binary variable and the outcome variable is continuous variable, so I choose the linear regression.

Validation (10pts)

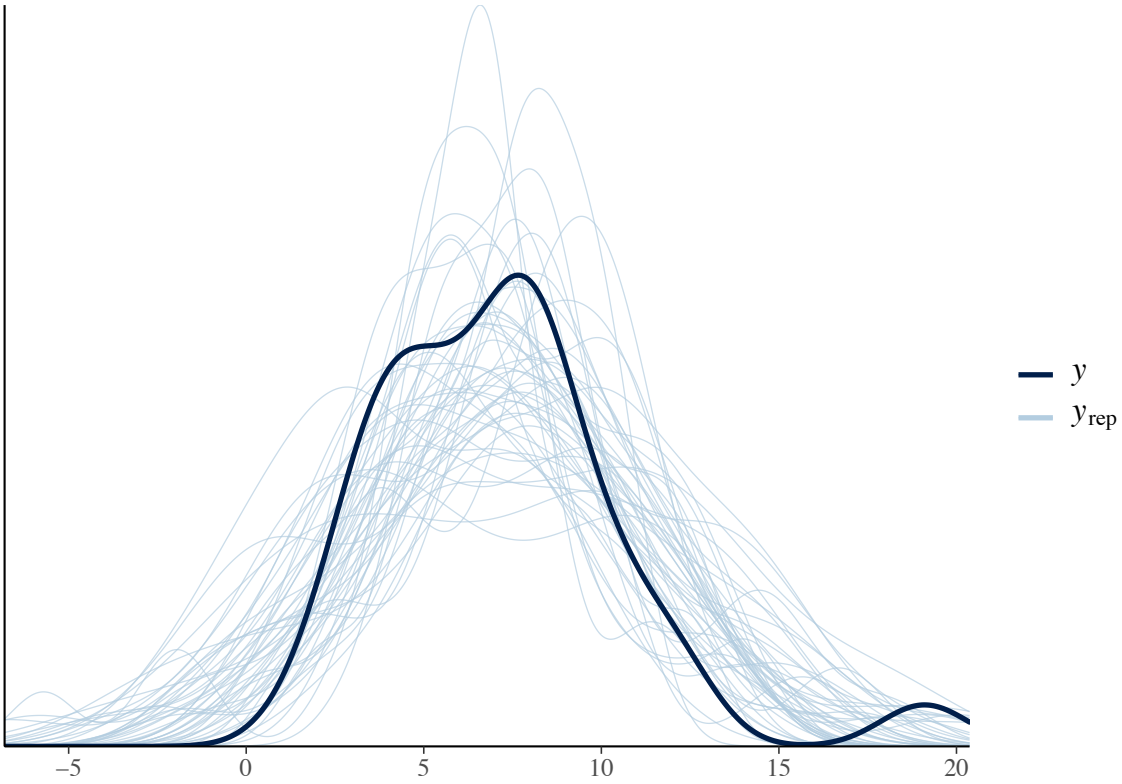
Please perform a necessary validation and argue why your choice of the model is appropriate.

```
plot(fitted(fit),resid(fit),pch=20)
curve(sqrt(x),add=T)
```



We can see some of our points is above the line, which probably indicates some level of overdispersion.

```
pp_check(fit)
```



Most of the expected values are distributed around the actual value, covering and very close to the actual value, so the model is effective.

Inference (10pts)

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

```
posterior_interval(fit)
```

```
##                      5%        95%
## (Intercept)  2.5261229 20.5848756
## gender1     -2.7406681  1.7167173
## age         -0.5136784  0.1885134
## sigma        2.9210812  4.6340961
```

Discussion (10pts)

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

In my model, age increase one unit, the total hours spent on smartphone will decrease 0.2 hours, when other variable remain unchanged. Compared with the person who has same age, we can conclude that the hours that female spent on different smartphone apps will be 0.6 hours more than the male.

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.

1.sample size: Since the sample size is so small, I cannot conclude that how the gender and age play a role in the time spent on different smartphone apps. I will send out more survey to people.

2.In my dataset, the person who did the survey are my peers mostly, the data are not representative. I will send out more survey to different age groups of the people.

3.This time, I tried many regression model to find the best fit model, but the results are all not ideal. I will try to find the other regression model that I will learned later to check whether they are suited for this datasetor not.

Comments or questions

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