Midterm Exam

Jinzhe Zhang

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.

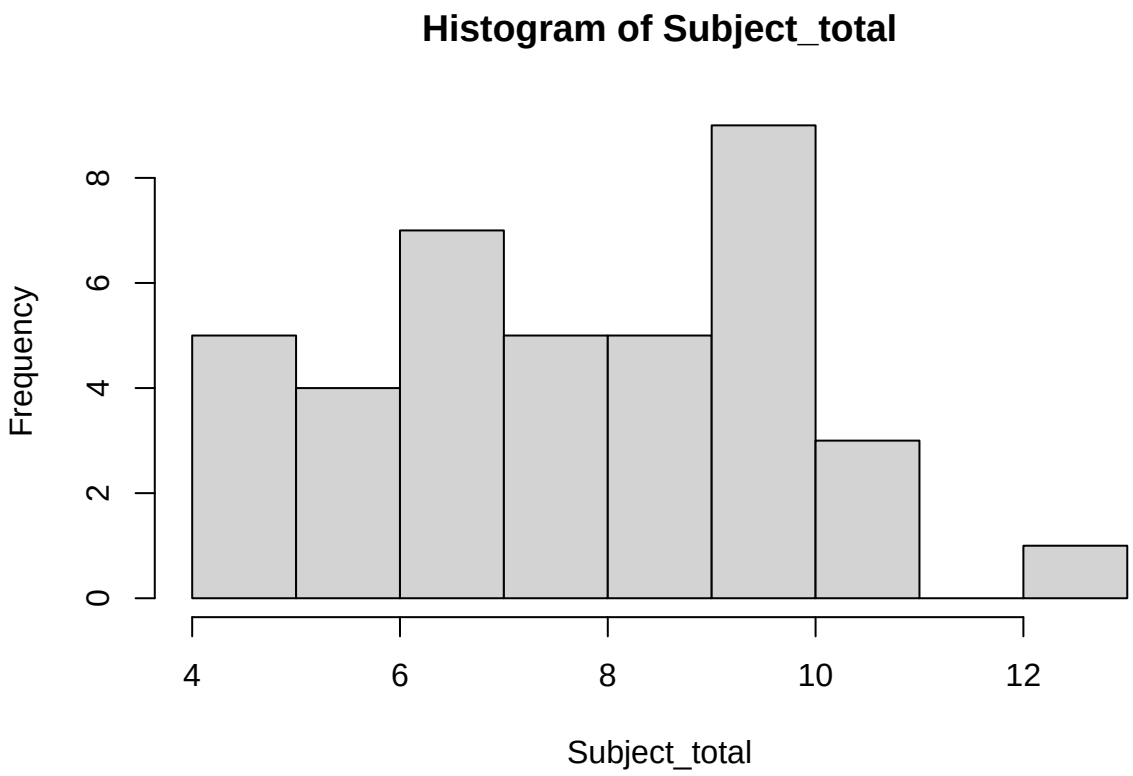
```
Data=read.csv("MA678-Data Collection.csv")
Data=Data[,-6:-9]

Subject_total=sample(1,length(Data$Age),replace = TRUE)

for (i in (1: length(Data$Age))){

  Subject_total[i]=Data[i,2]+Data[i,3]+Data[i,4]+Data[i,5]
}

hist(Subject_total)
```




```
#Explanation of the data:
#Basic background: Chinese people need to pass four subjects to get the driver license.
#This data was collected from the 40 persons who have passed the driver license examination.
#Subject One and subject Four test people driving and traffic rules by multiple choices on the website.
#The columns show the age of those people and  the times of each subject they took to pass the driver l


#The comparison of interest:
#Is it harder for aged people who are divided into three group by age(age<26,27-35,>36) to take the dir
```

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.

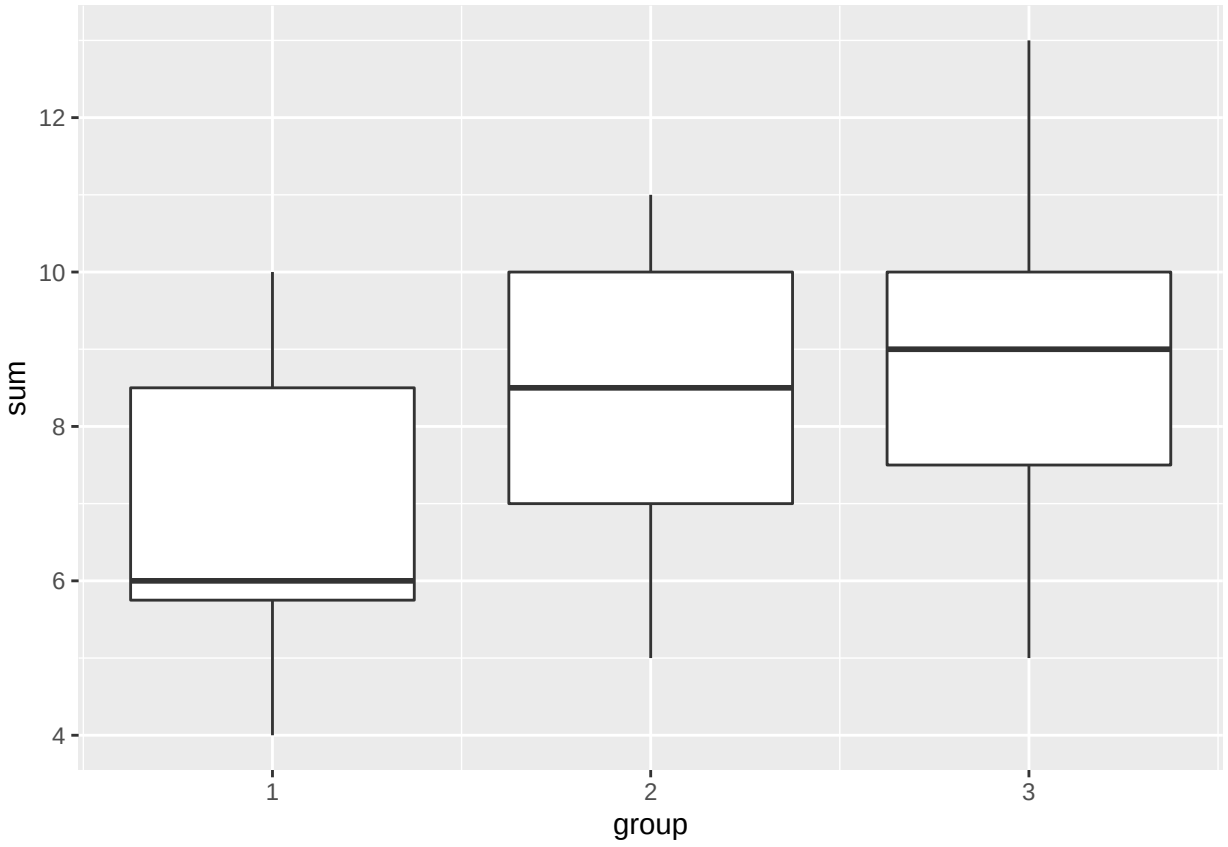
```
library(ggplot2)

# Here I group the people by different age.
#if group=1 means the person age is under 26
#if group=2 means the person age is between 27 and 35.
#if group=3 means the person age is above 35

#For the EDA part,
#I used the boxplot to show the times of four subjects of each person took by different age groups.
#Based on this boxplot, we can see the the group that age under 26 does take the least times of tests w


data.age=Data$Age
Data$group=c(1,1,2,3,3,1,1,3,3,3,1,2,3,1,3,3,1,1,3,2,2,3,3,3,2,1,1,3,1,3,3,3,3,2,2,2,3,1,3)
```

```r
Data$sum=rowSums(Data[2:5])
ggplot(Data,aes(x=group,y=sum,group=group))+geom_boxplot()
```



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.

```r
library(pwr)
pwr.t.test(n=15,d=NULL,sig.level=0.05,power=0.8,type = "two.sample")
```

```
##
##      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
```

```r
#Apparently, the level of effect size that is 1.059797 which is greater than one,.
# It notices me that the sample size is not enough for this problem.
#The reason why I should not use effect size is that the sample size of some groups is lower than 20,
#which means the effect size is actually overstated.
```

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.

```
library(rstanarm)
```

```
## Loading required package: Rcpp
## This is rstanarm version 2.21.1
## - See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!
## - Default priors may change, so it's safest to specify priors, even if equivalent to the defaults.
## - For execution on a local, multicore CPU with excess RAM we recommend calling
##   options(mc.cores = parallel::detectCores())
```

```
fit <- stan_glm(Data$sum ~ factor(Data$group), data=Data,x=TRUE,y=TRUE,family= gaussian(),refresh=0)
print(fit)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      Data$sum ~ factor(Data$group)
##  observations: 39
##  predictors:   3
## ------
##                     Median MAD_SD
## (Intercept)         6.9    0.6
## factor(Data$group)2 1.5    0.9
## factor(Data$group)3 2.0    0.7
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 2.0    0.2
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
# The reason I choose stan_glm regression which family is default as gaussian() as the model is that wh
```

Validation (10pts)

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

```
library(lmvar)
cv.lm(fit,k=3)
```

```
## Mean absolute error        :  1.784822
## Sample standard deviation  :  0.484821
##
## Mean squared error         :  4.49642
## Sample standard deviation  :  2.029094
##
## Root mean squared error    :  2.085747
## Sample standard deviation  :  0.4681043
```

```
# By doing the cross validation for my fitted model, the Mean absolute error is around 1,7 and the sd i
```

Inference (10pts)

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

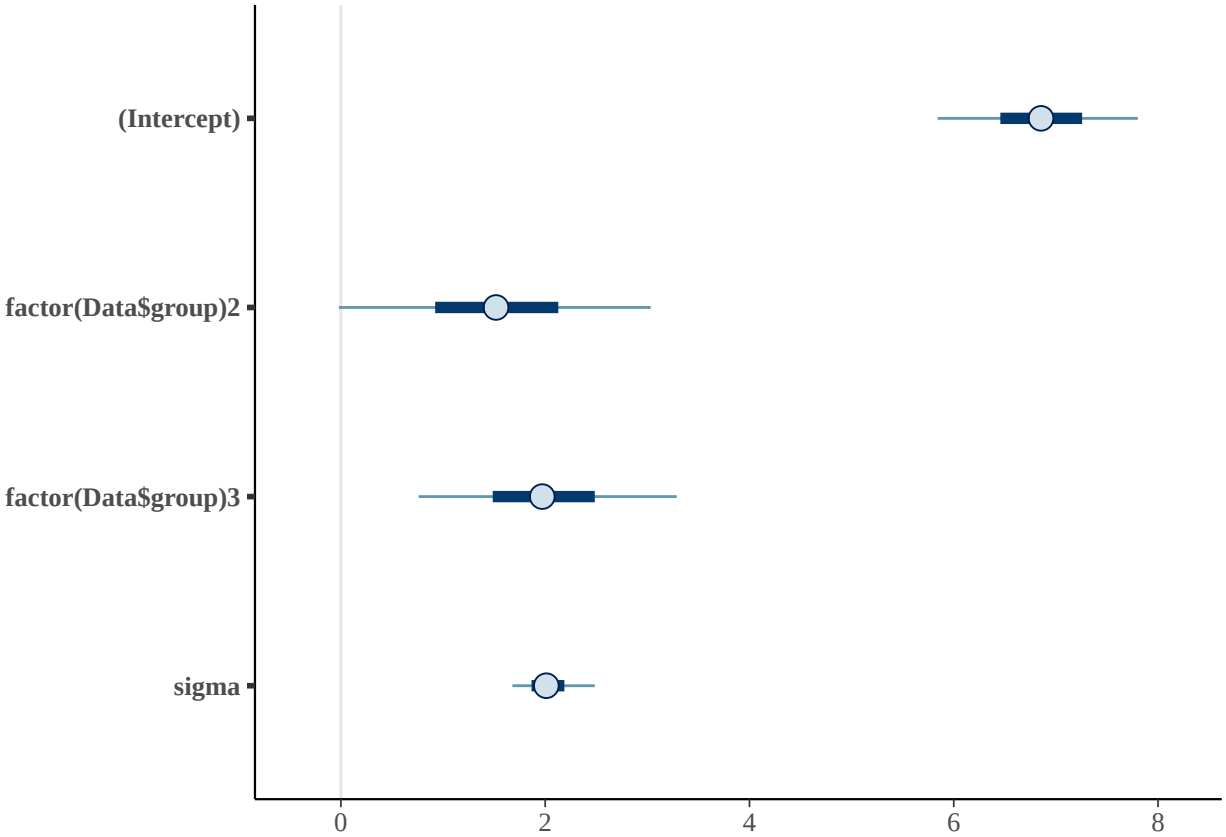
```
library(bayesplot)
```

```
## This is bayesplot version 1.7.2
## - Online documentation and vignettes at mc-stan.org/bayesplot
## - bayesplot theme set to bayesplot::theme_default()
##    * Does _not_ affect other ggplot2 plots
##    * See ?bayesplot_theme_set for details on theme setting
```

```
sims=as.matrix(fit)
posterior_interval(fit)
```

```
##                              5%      95%
## (Intercept)          5.84335248 7.802658
## factor(Data$group)2 -0.02003084 3.033425
## factor(Data$group)3  0.76177398 3.287176
## sigma                1.67889147 2.485475
```

```
mcmc_intervals(sims)
```




```
# Based on the result of posterior_interval function, the confidence interval for each coefficient is l
```

Discussion (10pts)

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

Based on my fitted model, the coefficients group 3 has the largest coefficient. Furthermore, people are harder to get the driver license because the coefficients for the model are all positve. For example the person in group 2 (age27-35) will take more 1.5 times of tests than the person in group1(age <26)

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 concerns is that there should be more predictors included in the model. The other one is the sample size is too small to estimate the result correctly Because the age of person may just have correlation with times of subjects people took, but may not the casual factor. To fix this problem, I am going to include more predictors such as the education background, income and the time the person spent on practicing.

Comments or questions

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