Midterm Exam

Your Name

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.

-This dataset is about the usage time of computer for 6 people on Oct.25th.2020. The usage time is categorized into work and entertainment, and the unit of time is hour.

-The reason why I collect this data is that I want to know how many hours my friends spent on computer on Sunday and if there is any difference of usage time between person and person

```r
#data
computer<-read.csv("C:/Users/LXD/Documents/R/data collection.csv")
head(computer)
```

```
##   people Entertainment Work
## 1  Irish             6    0
## 2    Bob             2    6
## 3 Stella             1    0
## 4 Regyna             1    7
## 5   Chen             6    4
## 6  Aaron             4    2
```

```r
#package
pacman::p_load(tidyverse,pwr,boot,arm)
```

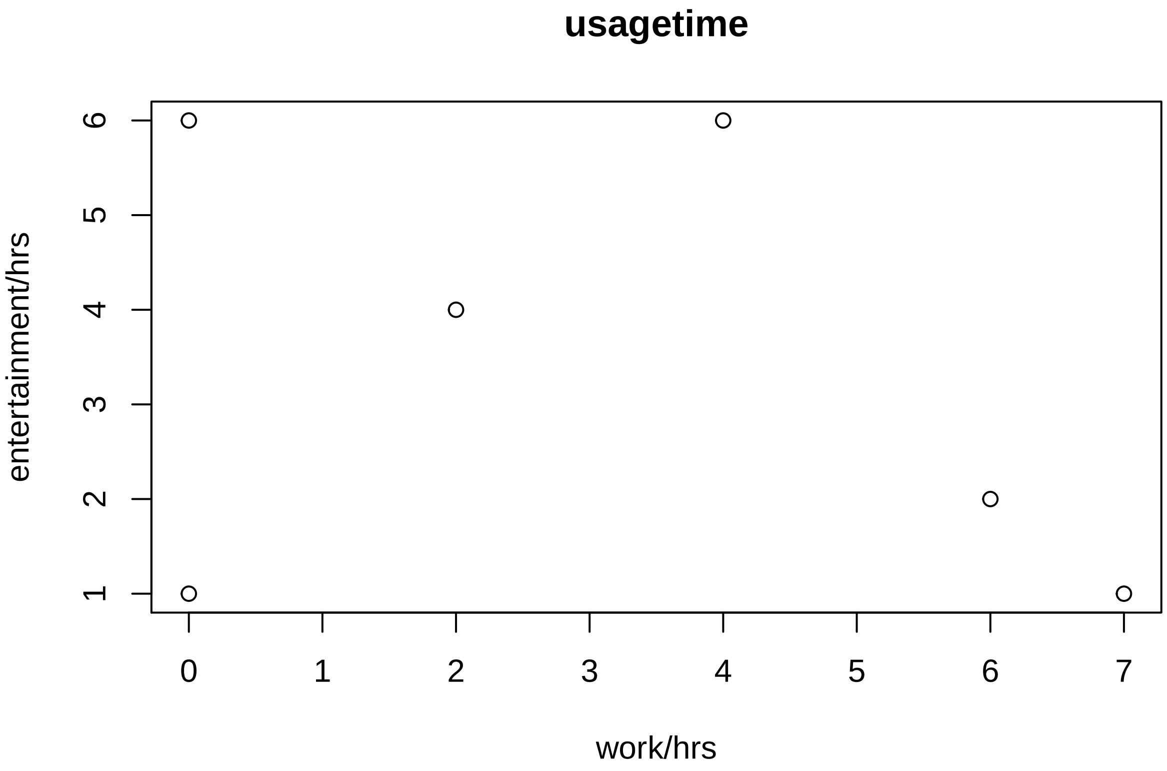
EDA (10pts)

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.

```
computer1 <- data.frame(usage=as.character(rep(c(0,0,0,0,0,0,1,1,1,1,1,1),1)),
                    time=c(computer$Entertainment,computer$Work),
name=c(("i"),c("b"),c("s"),c("r"),c("c"),c("a"),c("i"),c("b"),c("s"),c("r"),c("c"),c("a"))
                  )
head(computer1)
```

```
##   usage time name
## 1     0    6    i
## 2     0    2    b
## 3     0    1    s
## 4     0    1    r
## 5     0    6    c
## 6     0    4    a
```

```
fit2<- plot(x=computer$Work,y=computer$Entertainment,xlab="work/hrs",ylab = "entertainment/hrs",main =
```



```
view(fit2)
```

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.

- Since the two groups are not independent and they are related, I choose pwr.r.test. The result shows that the sample size should be at least 123. Therefore, my sample size is not enough.

```
# calculate effective size by func.

pwr.t.test(n=6,sig.level=0.05,power=0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 6
##               d = 1.795541
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# calculate effective size on my own


u1 <- mean(computer1$time[computer1$usage==1])
u2 <- mean(computer1$time[computer1$usage==0])
s1 <- sd(computer1$time[computer1$usage==1])
s2 <- mean(computer1$time[computer1$usage==0])

effective_size <- (abs(u1-u2)/sqrt((s1^2+s2^2)/2))
effective_size
```

```
## [1] 0.05260244
```

```
# calculate sample size

pwr.t.test(d=effective_size,sig.level=0.05,power=0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 5674.117
##               d = 0.05260244
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

- the result shows the sample size should be at least 160,and my sample size is only 6.Not enough.

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.

For my model, the outcome variable is continuous, and my predicted variable has binary variable and

```
fit_lm <- glm(time~as.numeric(usage)+name,data=computer1)
summary(fit_lm)
```

```
##
## Call:
## glm(formula = time ~ as.numeric(usage) + name, data = computer1)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  2.9167  -2.0833   0.4167  -3.0833   0.9167   0.9167  -2.9167   2.0833
##       9       10       11       12
## -0.4167   3.0833  -0.9167  -0.9167
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.083e+00  2.377e+00   1.297    0.251
## as.numeric(usage) -1.667e-01  1.797e+00  -0.093    0.930
## nameb              1.000e+00  3.112e+00   0.321    0.761
## namec              2.000e+00  3.112e+00   0.643    0.549
## namei              3.535e-15  3.112e+00   0.000    1.000
## namer              1.000e+00  3.112e+00   0.321    0.761
## names             -2.500e+00  3.112e+00  -0.803    0.458
##
## (Dispersion parameter for gaussian family taken to be 9.683333)
##
##     Null deviance: 72.250  on 11  degrees of freedom
## Residual deviance: 48.417  on  5  degrees of freedom
## AIC: 66.794
##
## Number of Fisher Scoring iterations: 2
```

Validation (10pts)

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

```
cv.glm(computer1,fit_lm,K=9)$delta[1]
```

```
## Warning in cv.glm(computer1, fit_lm, K = 9): 'K' has been set to 12.000000
```

```
## [1] 23.24
```

```
cv.glm(computer1,glm(time~as.numeric(usage)+name,data=computer1),K=9)$delta[1]
```

```
## Warning in cv.glm(computer1, glm(time ~ as.numeric(usage) + name, data =
## computer1), : 'K' has been set to 12.000000
```

```
## [1] 23.24
```

The two value of k is same, the model is kind of appropraite.

Inference (10pts)

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

```
confint(fit_lm)
```

```
## Waiting for profiling to be done...
##                        2.5 %   97.5 %
## (Intercept)       -1.574875 7.741542
## as.numeric(usage) -3.687941 3.354608
## nameb             -5.099027 7.099027
## namec             -4.099027 8.099027
## namei             -6.099027 6.099027
## namer             -5.099027 7.099027
## names             -8.599027 3.599027
```

Discussion (10pts)

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

The result shows that different person will not affect the usage of time on computer for work or entertainment

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. my dataset is too small, I think I need to do more research.

2.Also, the conclusion of my analysis is not reliable for many reasons such as the data is too limited.

3. in the future, I think I need more information about how to select the model.
4. Think more about the interesting question.

Comments or questions

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