Midterm Exam

Haoran Su

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 question I want to discuss is how long people use phones daily. I set up two groups, workers and students to see which group spend more time on phones. I made a survey to collect data about their daily usage of phones. I put up the survey link to my friends and post it online in a survey help website to get more responses. Out of the surveys I collected, 5 of them is unreliable as most blankets are blank while submitted. So I deleted those 5 lines. Explanations of columns: Group: 1 implies they are in the group of students, 2 implies that they are in the group of working. Charge: Times of charging the phone in one day. Use: duration of phone using in one day Open: times of opening the phone and unlocking it in one day.

It would be very interesting to see the condition of phone usage of the workers and students.*

```
# change the volumn types, omit the "NA" rows
phone.raw <- read.csv("Data collection.csv")
phone<-na.omit(phone.raw)
phone<-select(phone,Group, Charge,Use,Open)
#when the blanket is "n" in "Use" volumn, which implies that the form filler feels using phone many tim
Open.max=max(phone$Open)
phone$Open<- as.numeric(phone$Open)
```

```
## Warning: NAs introduced by coercion
```

```
for(i in 1:row_number(phone)){
  if (phone[i,4]=="n") { phone[i,4]<-Open.max}
  else if (phone[i,4]=="na") {phone[i,4]<- NULL}
  if (phone[i,2]==1) { phone[i,2]<-0}
  else if (phone[i,2]==2) {phone[i,2]<-1}
}
```

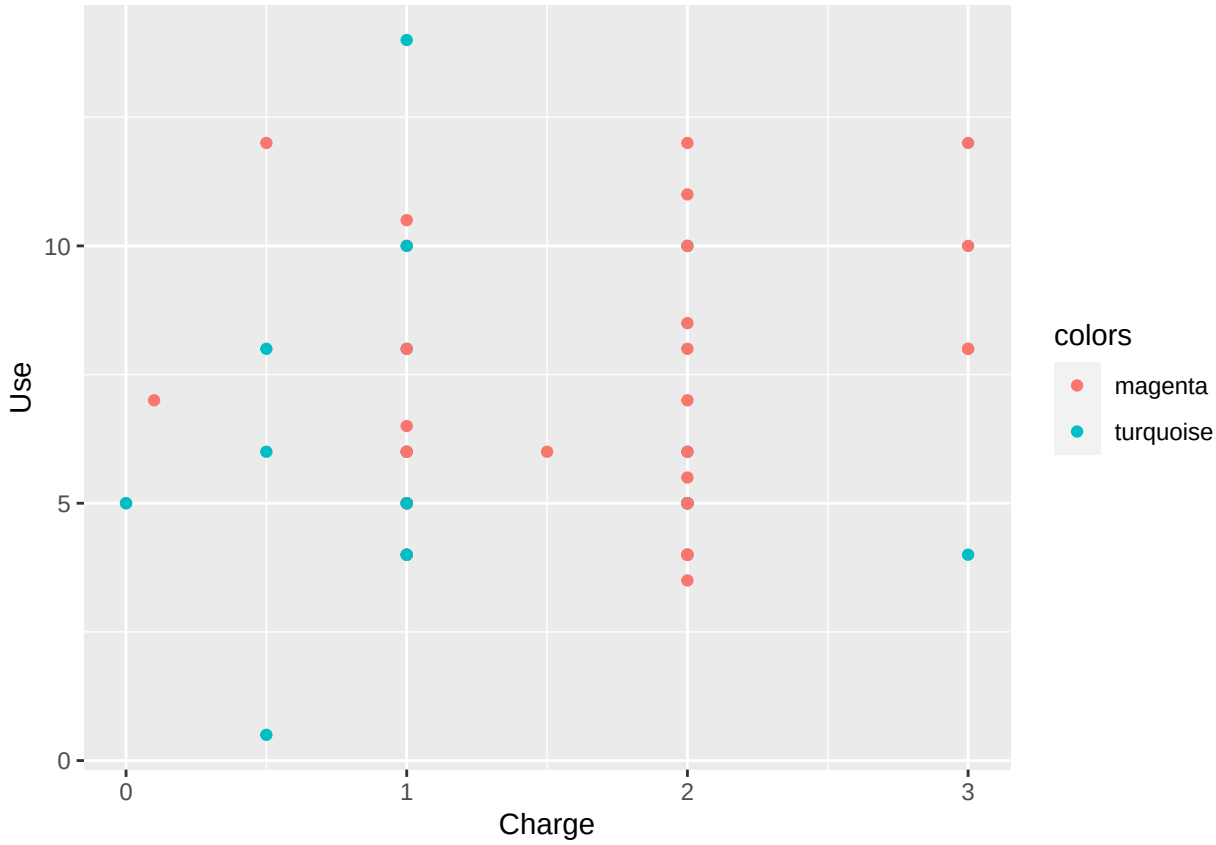
```
## Warning in 1:row_number(phone): numerical expression has 276 elements: only the
## first used
```

```
phone<-na.omit(phone)
#Group: 0 implies they are in the group of students, 1 implies that they are in the group of working.
```

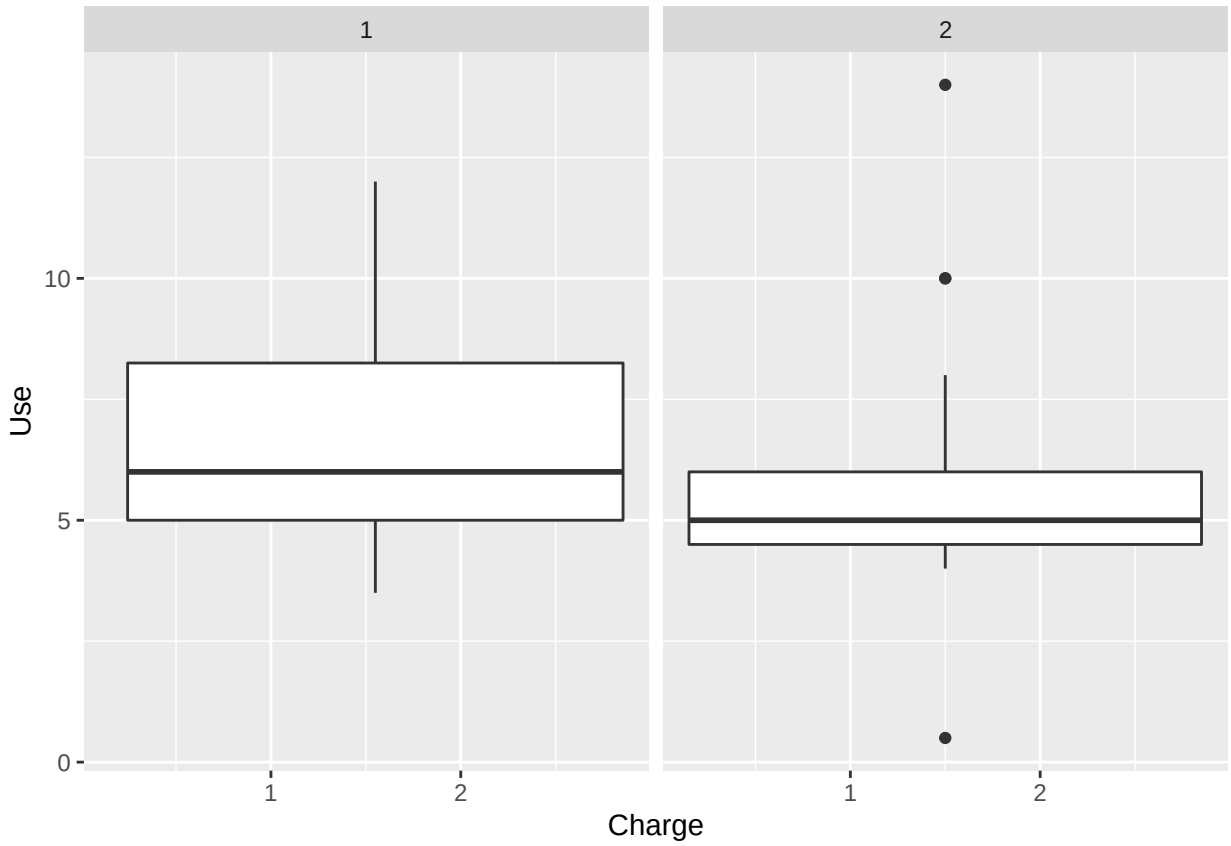
EDA (10pts)

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.

```
colors = ifelse(phone$Group==1,"magenta","turquoise")
ggplot(data=phone) +geom_point(aes(y=Use,x=Charge,col=colors))
```




```
ggplot(data=phone, aes(y=Use,x=Charge,group=Group)) +
    geom_boxplot()+
  facet_wrap(~Group)
```




```
by(phone, phone$Group,describe)
```

```
## phone$Group: 1
##        vars  n  mean    sd median trimmed   mad min max range skew kurtosis
## Group     1 35  1.00  0.00      1    1.00  0.00 1.0   1   0.0  NaN      NaN
## Charge    2 35  1.66  0.73      2    1.64  1.48 0.1   3   2.9 0.11    -0.66
## Use       3 35  6.93  2.58      6    6.72  2.97 3.5  12   8.5 0.64    -0.91
## Open      4 35 34.57 25.16     30   30.83 19.27 1.0 100  99.0 1.30     1.26
##          se
## Group  0.00
## Charge 0.12
## Use    0.44
## Open   4.25
## ------------------------------------------------------------
## phone$Group: 2
##        vars  n  mean    sd median trimmed   mad min max range skew kurtosis
## Group     1 23  2.00  0.00      2    2.00  0.00 2.0   2   0.0  NaN      NaN
## Charge    2 23  1.20  0.73      1    1.18  0.74 0.0   3   3.0 0.49    -0.32
## Use       3 23  5.80  2.72      5    5.53  1.48 0.5  14  13.5 1.16     1.81
## Open      4 23 19.61 13.69     20   17.53 14.83 3.0  60  57.0 1.37     1.67
##          se
## Group  0.00
## Charge 0.15
## Use    0.57
## Open   2.85
```

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.

```
phone%>% group_by(Group)%>%summarise(mean=mean(Charge), sd=sd(Charge))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   Group  mean    sd
##   <int> <dbl> <dbl>
## 1     1  1.66 0.729
## 2     2  1.20 0.735
```

```
phone%>% group_by(Group)%>%summarise(mean=mean(Use), sd=sd(Use))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   Group  mean    sd
##   <int> <dbl> <dbl>
## 1     1  6.93  2.58
## 2     2  5.80  2.72
```

```
phone%>% group_by(Group)%>%summarise(mean=mean(Open), sd=sd(Open))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   Group  mean    sd
##   <int> <dbl> <dbl>
## 1     1  34.6  25.2
## 2     2  19.6  13.7
```

```
d1=abs(1.74-1.32)/sqrt(0.70^2+0.65^2)
d2=abs(6.92-5.80)/sqrt(2.58^2+2.72^2)
d3=abs(34.57-19.61)/sqrt(25.16^2+13.69^2)
pwr.t.test(d=d1, power=0.8, sig.level = 0.05, type = 'two.sample')
```

```
##
##      Two-sample t test power calculation
##
##               n = 82.17426
##               d = 0.4396761
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr.t.test(d=d2, power=0.8, sig.level = 0.05, type = 'two.sample')
```

```
##
##      Two-sample t test power calculation
##
##               n = 176.8491
##               d = 0.2987485
```

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

```
pwr.t.test(d=d3, power=0.8, sig.level = 0.05, type = 'two.sample')
```

```
##
##      Two-sample t test power calculation
##
##               n = 58.52282
##               d = 0.5222852
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

As the test power calculation result shows, the optimized size is n=82,176,58 for each group. The sample size I have is not enough for the problem to discuss. As the data I could collected is limited so far, although the size is not enough, I would use the current data to do the analysis.

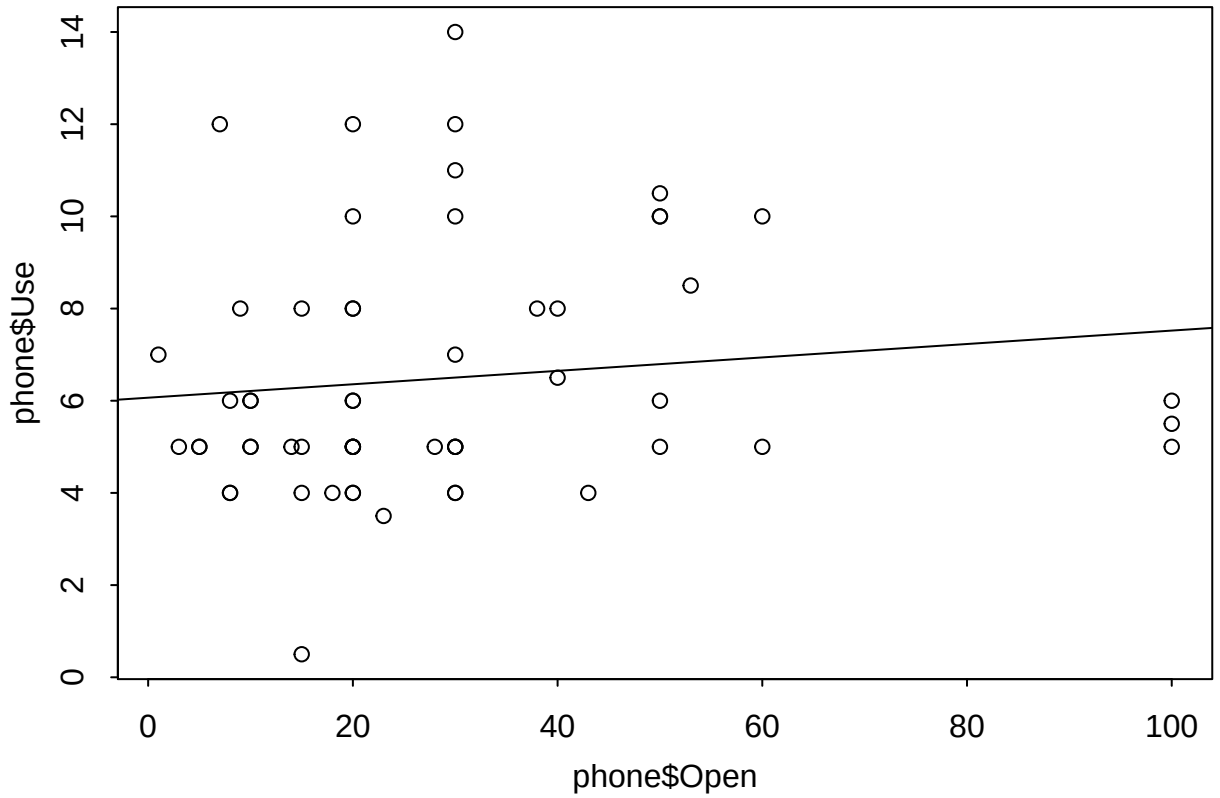
Modeling (10pts)

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.

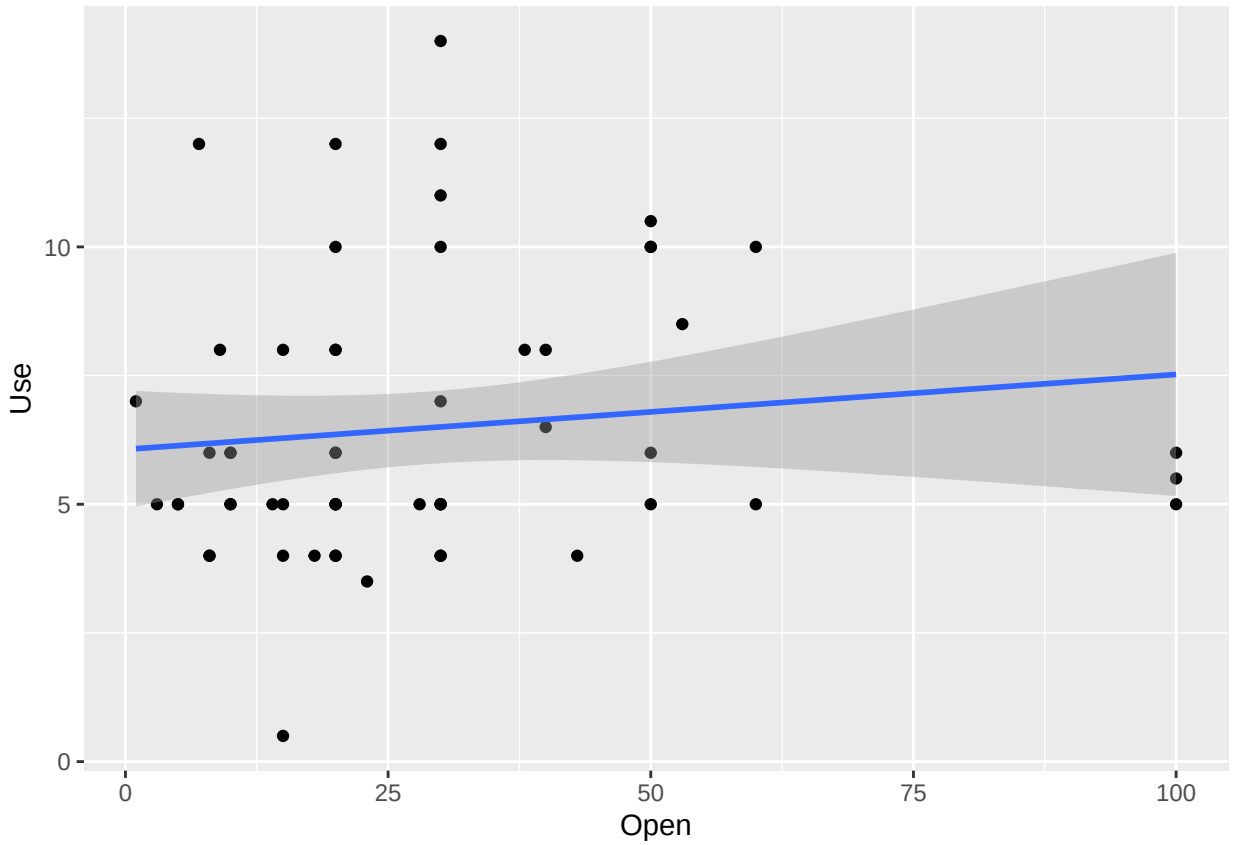
```
model1<-lm(Use~Open, data=phone)
print(model1)
```

```
##
## Call:
## lm(formula = Use ~ Open, data = phone)
##
## Coefficients:
## (Intercept)         Open
##     6.06574      0.01456
```

```
par(mar=c(3,3,2,1), mgp=c(2,.7,0), tck=-.01)
plot(y=phone$Use, x=phone$Open)
abline(coef(model1))
```



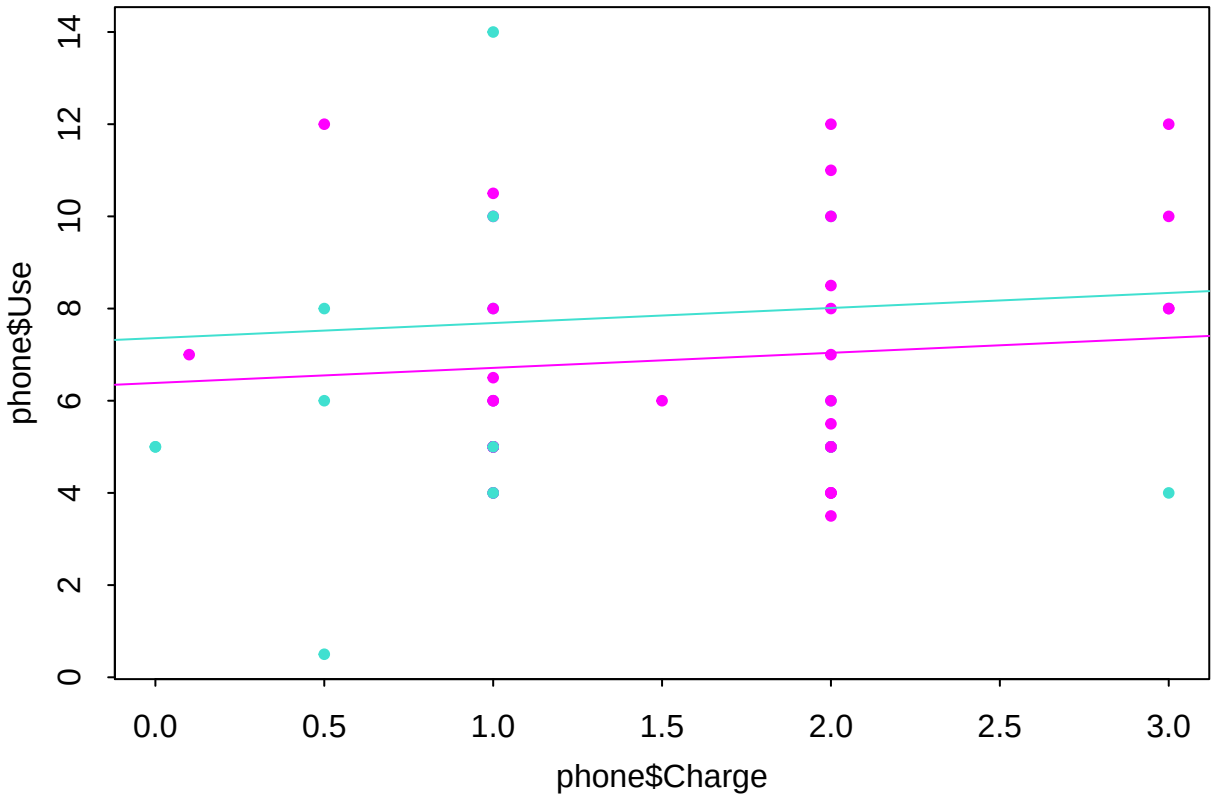

```
ggplot(data = phone, aes(x=Open,y = Use))+ geom_point()+
geom_smooth(formula = "y~x", method = "lm")
```




```
model2<-lm(Use~Charge+Group, data=phone)
print(model2)
```

```
##
## Call:
## lm(formula = Use ~ Charge + Group, data = phone)
##
## Coefficients:
## (Intercept)       Charge        Group
##      7.3587       0.3267      -0.9725
```

```
b_hat <- coef(model2)
colors = ifelse(phone$Group==1,"magenta","turquoise")
par(mar=c(3,3,2,1), mgp=c(2,.7,0), tck=-.01)
plot(y=phone$Use,x=phone$Charge,pch=20,col=colors)
abline(b_hat[1] + b_hat[3],b_hat[2],col="magenta")
abline(b_hat[1],b_hat[2],col="turquoise")
```



Validation (10pts)

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

```
AIC(model1)
```

```
## [1] 282.7245
```

```
AIC(model2)
```

```
## [1] 282.5666
```

I use AIC for model validation. The model with the lower AIC and BIC score is preferred so model2 is preferred as AIC(model2) is slightly lower .

Inference (10pts)

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

```
coef(model2)
```

```
## (Intercept)      Charge       Group
##   7.3586898   0.3267381  -0.9725035
```

Looking at the coefficients, the slope of 7.36 means that if a person does not charge the phone in one day than on average he would spend 7.36 hours on phone daily. If the person is a student, they are expected to spend 0.97 hours less on phones every day comparing to an employed person. The slope coefficient 0.33 means that with one more charging attempt in a day, on average the person would spend 0.33 hours more on phone daily.

Discussion (10pts)

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

Comparing the group of students and workers, the students on average charge phones 1.75 times a day, while workers charge their phone 1.33 times a day. Students generally spend more time on phones than workers, while the maximum using time of students is 12, less than the maximum using time of 14 in the group of workers. Comparing the standard deviations, there is not a large difference about their charging times and using time. The distribution of times of opening the phone in one day for students is much more decentralized. As the model fitted is not considered well so I will not draw many conclusions from that.

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.

First of all, the data collected is not enough for the 80% poter test comparison. Not so many people came and filled out the survey. Maybe the posting of surveys could be longer the next time to collect more samples. Secondly, data is not so valid. 30% of data is collected from one-by-one sending between my friends. 70% of the data is filled out by surveys on online forum. All the data collected from my friends is in the group of students and most if the data from online forum is in the group of working people. So there might be bias of the distribution of samples. *Also, I used the techniques of "giving points" on the survey distribution platform in order to get more samples in a limited time. From this source, I observed that many people gave a round number about phone using time and did not check out the real time shown by the apps on phone. This made a bias of fake data.* Then there is no significant result in the model I fit. I could only draw some conclusions of the distribution of plots. It might be for the bad design of the questions I asked in the survey. The design for data collection is important from the beginning.

Comments or questions

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

**I would like to have some suggestions about how to deal with "n", "na" in the surveys. I did not deleted them all as I don't have so many samples if they are deleted. For "n", I think it might be the implication of using the phones for too many times so I set it as the maximum number of the using data.