Midterm Exam

Chi 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.

I found an expired bottle of hand cream,so I had an idea to measure the length of each squeeze.Each time I squeeze,I mark the farthest distance with a pencil, and then erase the mark for the next measurement. I used right hand and left hand to make a comparison that on average which handwould make a further "squeeze". Also, I would like to know that whether there is a relationship between the squeeze of left hand and right hand. I measure 5 times each hand,and total 10 observations.

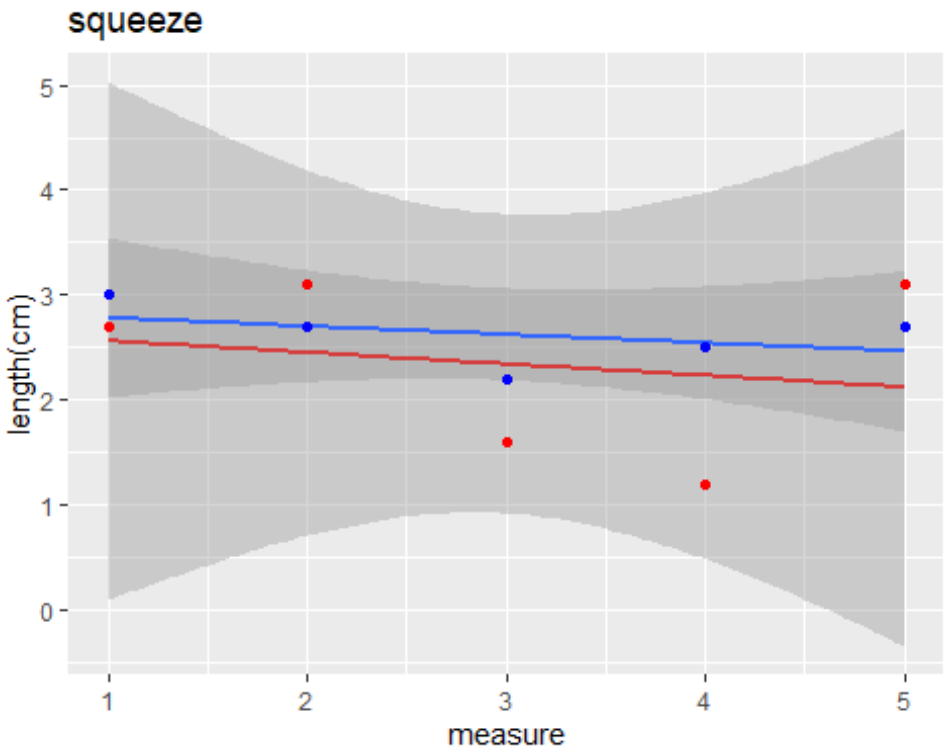
```
squeeze <- read.csv("data collection.csv")
```

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)
pic<-ggplot()+geom_smooth(data=squeeze, aes(x=X,y=left),method="lm",col
or="red")+
geom_smooth(data=squeeze, aes(x=X,y=right),method="lm")+
geom_point(squeeze,mapping=aes(x=X,y=left),
color="red")+geom_point(squeeze,mapping=aes(x=X,y=right),
color="blue")+labs(x = "measure", y = "length(cm)", title = "squeeze")
pic

## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



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.

When I using 80% power, sample size =5, I got inferred effect size 2.024439 from T test. To determine my sample size is enough, I calculated Cohen d = -0.4245349 (small). Performing the T test, for a desired power of 80%, hypothesized effect size of -0.4245349, I need a sample at least 88 per group. For the reason why I should Not use the effect size from the fitted model is, the published results tend to be overestimates. Interventions are often tested on people where they will be most effective, and effects will be smaller in the general population. What is more, the magnitude of the effect size will be vastly overstated if it is published(type M error).

```
#install.packages("pwr")
library(pwr)

## Warning: package 'pwr' was built under R version 4.0.3

pwr.t.test(n=5, power = 0.8, sig.level =0.05)

##
##      Two-sample t test power calculation
##
##               n = 5
##               d = 2.024439
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group

#install.packages("effsize")
library(effsize)

## Warning: package 'effsize' was built under R version 4.0.3

l <- c(squeeze$left[!is.na(squeeze$left)])
r <- c(squeeze$right[!is.na(squeeze$right)])
cohen.d(l,r)

##
## Cohen's d
##
## d estimate: -0.4245349 (small)
## 95 percent confidence interval:
##     lower     upper
## -1.899317  1.050247

pwr.t.test(d=-0.4245349, power =0.8, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 88.06904
##               d = 0.4245349
```

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

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.

In my data, the left and right are both measured randomly with continuous scale, not categorical, therefore I decided to use linear regression model in my case. Linear regression helps me to find out what extent there is a linear relationship between the left and right hand's "squeeze".

```
fit <-lm(squeeze$right~squeeze$left)
summary(fit)
```

```
##
## Call:
## lm(formula = squeeze$right ~ squeeze$left)
##
## Residuals:
##        1        2        3        4        5
##  0.29885 -0.09132 -0.25319  0.13697 -0.09132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.0925     0.3502   5.975  0.00938 **
## squeeze$left   0.2254     0.1418   1.590  0.21006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2509 on 3 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4573, Adjusted R-squared:  0.2764
## F-statistic: 2.528 on 1 and 3 DF,  p-value: 0.2101
```

Validation (10pts)

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

We can see the Residual standard error is around 0.25 which indicates that my regression model is valid. The R squared is around 0.457 which is not quite small. Considering this is a data measured of human behavior, therefore less than 0.5 is pretty normal. The F-statistic is 2.528, which can also indicates to reject the null

hypothesis that there is no relationship between two variables. The following small error from accuary function can also show the validation of the model.

```
pre <- predict(fit,newdata=squeeze,type='response')
print(pre)

##        1        2        3        4        5        6        7
## 2.701149 2.791315 2.453193 2.363027 2.791315       NA       NA

library(forecast)

## Warning: package 'forecast' was built under R version 4.0.3

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

accuracy(pre, squeeze$right)

##                     ME      RMSE       MAE        MPE     MAPE
## Test set 1.77633e-16 0.1943486 0.1743295 -0.5664521 6.742697
```

Inference (10pts)

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

If the length of right hand squeeze differed by 1 cm, the length of left hand squeeze will differ by by 0.2254 cm, on average. The intercept is significant which means if the right hand is 2.09, the left hand is significantly different from 0. Also, t value and F-statistic show that there is a relationship between left and right hand's squeeze. However, since our result is not significant, we fail to reject that null hypothesis that there is no relation. But all of the results shows that on average, right hand can squeeze further than left hand.

```
summary(fit)

##
## Call:
## lm(formula = squeeze$right ~ squeeze$left)
##
## Residuals:
##        1        2        3        4        5
##  0.29885 -0.09132 -0.25319  0.13697 -0.09132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.0925     0.3502   5.975  0.00938 **
## squeeze$left   0.2254     0.1418   1.590  0.21006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2509 on 3 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4573, Adjusted R-squared:  0.2764
## F-statistic: 2.528 on 1 and 3 DF,  p-value: 0.2101
```

Discussion (10pts)

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

To sum up, there are severl coefficients and plot show that generally, my right hand can squeeze further than my left hand.What is more, even though the result of my regression model is not significant, there are still many evidences show there is a relationship between the left and right hand's squeeze.

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.

Since I just collected a simple data, there is not many details I can get from it. However, when I went through these steps, I found I was not quite familiar with dealing some issues such as to infer effect size in t test, or explain the reason why I choosing linear regression model,or how do I validate the model etc. This exam gives me a chance to find these issues, so I can fix them before I truly encounter similar problems in the future. I would check webs or ask someone who knows the answer to fix the problem in further study. For the problems I have met, I will try to memory the key points to avoid them.

Comments or questions

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