Midterm Exam

Lin Zhou

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.

```
library(MASS)
library(pwr)
library(ggplot2)
library(arm)
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.11-2, built: 2020-7-27)
```

```
## Working directory is /Users/nadou/Desktop
```

```
library(nnet)
data1 <- read.csv("data.csv")
head(data1)
```

```
##   Index Age Occupation AWT CN IR MUA WOE TUP
## 1     1   2          1   2  1  2   2   1   4
## 2     2   2          5   3  1  2   1   5   2
## 3     3   2          1   1  2  2   1   5   4
```

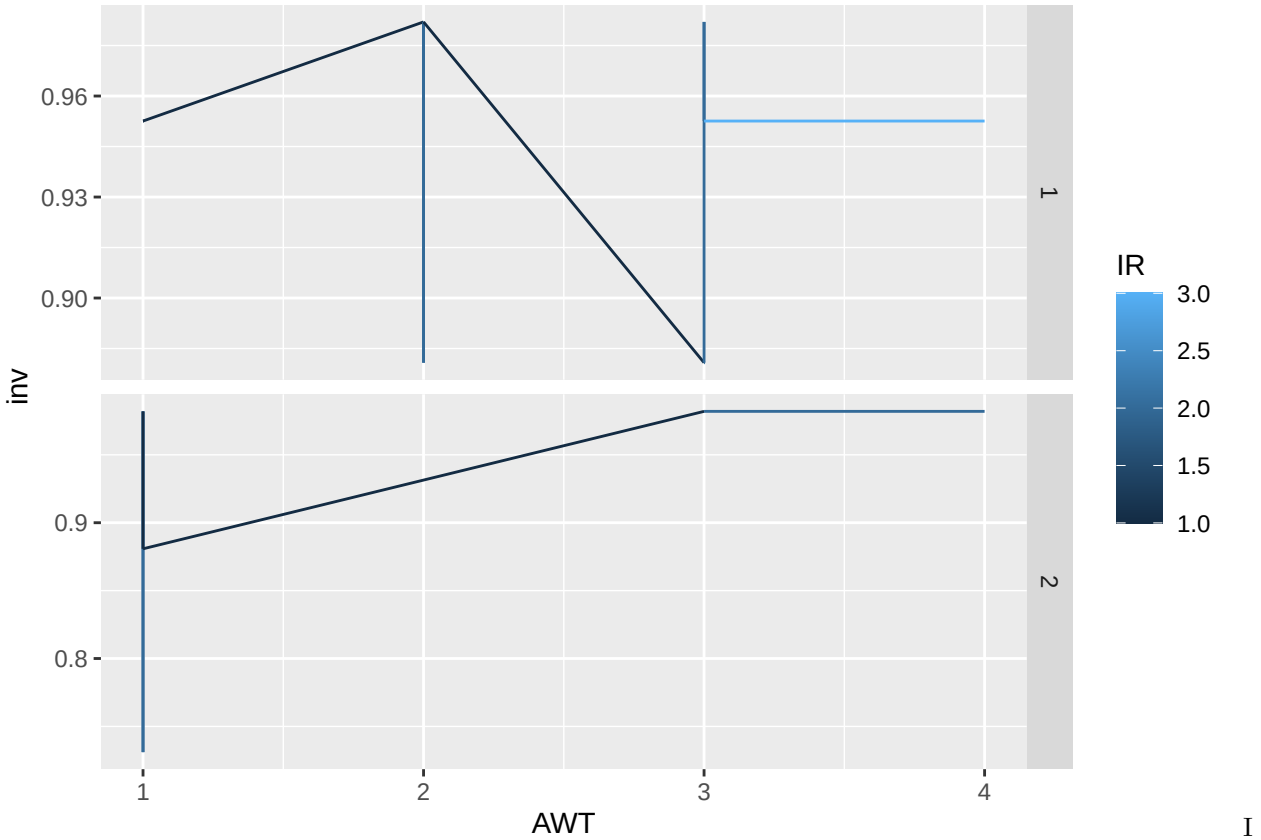
```
## 4      4   2          2   1  1  1   3   2   3
## 5      5   4          9   2  1  2   1   4   3
## 6      6   2          2   4  2  1   4   5   4
```

The response variable of the data is TUP "Average cell phone using time of individuals per day". I am going to explore different factors which might affect the cell phone using time like age, occupation, average working time, whether in relationship with others, whether paying attention to current news and so on. (All the variables are categorical variables). I want to focus on the connections between TUP and AWT(average working time per day), CN(whether interested in current news), IR(whether in relationship with others) because for other predictive variables, the sample is too small so that I can't extract useful information through the dataset.

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.

```
inv <- invlogit(data1$TUP)
ggplot(data1, aes(x = AWT, y = inv, colour = IR)) +
  geom_line() +
  facet_grid(CN ~., scales = "free")
```



I planned to present the data as graphs with y-axis as frequency, x-axis as TUP, color as IR, and in 2 graphs one is CN =1, the other is CN =2. But when I clear up all situations, I found there are total 96 situations which can not be presented on graphs for there are only 20 observations.

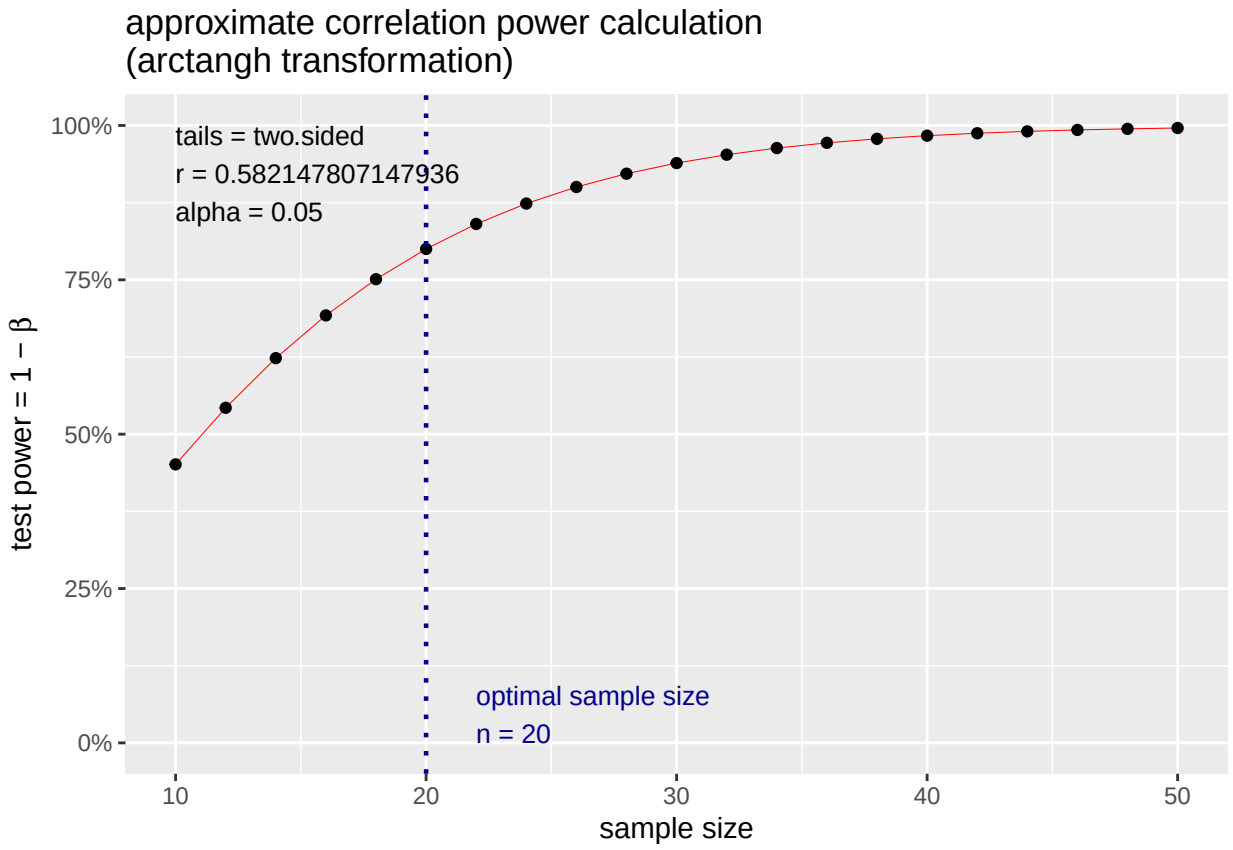
Power Analysis (10pts)

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_power <- pwr.r.test(n = 20, sig.level = 0.05, power = 0.8, alternative = "two.sided")
r_power
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 20
##               r = 0.5821478
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
plot(r_power)
```



The effect size is 0.582 which is r = 0.582, suggests the points are relatively close to fitted line. But in this case, I don't think the sample size is large enough for us to fit and analysis. From my understanding, the effect size corresponds to true difference between populations. The value of effect size represents the probability that this difference would be significant. The effect size tells us whether the difference we observed is the true difference in reality. Thus, if we use the effect size from the fitted model, the comparison makes no sense.

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.

```
fit1 <- polr(data = data1, factor(TUP) ~ IR + AWT + CN)
summary(fit1)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = factor(TUP) ~ IR + AWT + CN, data = data1)
##
## Coefficients:
##       Value Std. Error t value
## IR  -0.8000     0.9047 -0.8842
## AWT  0.5139     0.5432  0.9461
## CN   0.6250     1.1178  0.5592
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -2.5366  2.4208    -1.0478
## 2|3 -0.6930  2.2103    -0.3135
## 3|4  0.8915  2.2255     0.4006
##
## Residual Deviance: 46.41688
## AIC: 58.41688
```

I pick multinomial regression model since the outcome of data is ordinal categorical(cell phone using time less than 2 hours as 1, 2-4 hours as 2 and so on). I focus on three predictive variables IR(whether in relationship with others), AWT(average working time per day), CN(whether pay attention to current news). I use the polr function which is used to fit a logistic regression model to an ordered factor response.As a result, the link function should be logit.

Validation (10pts)

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

```
fit2 <-  multinom(factor(TUP) ~ IR + AWT + CN, data = data1)
```

```
## # weights:  20 (12 variable)
## initial  value 27.725887
## iter  10 value 18.648356
## iter  20 value 17.974091
## iter  30 value 17.970678
## final  value 17.970655
## converged
```

```
summary(fit2)
```

```
## Call:
## multinom(formula = factor(TUP) ~ IR + AWT + CN, data = data1)
##
## Coefficients:
##   (Intercept)        IR      AWT         CN
## 2   -1.202551 -11.65368 18.57724   2.247187
```

```
## 3   26.894168 -10.98912 17.87642 -24.655163
## 4   -2.097249 -11.79839 18.72036   3.351468
##
## Std. Errors:
##   (Intercept)       IR      AWT       CN
## 2    103.8616 251.1065 75.40046 264.8601
## 3    117.5155 251.1082 75.40490 117.5152
## 4    103.8546 251.1045 75.39968 264.8588
##
## Residual Deviance: 35.94131
## AIC: 59.94131
```

If the dataset is large, I would like to use cross validation. But the sample size is small, as a result, I choose to fit the data into different models and compare the AIC of these models to check whether it is the best one among these assumed models.

Inference (10pts)

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

```
coef1 <- coef(summary(fit1))
```

```
##
## Re-fitting to get Hessian
```

```
p <- pnorm(abs(coef1[, "t value"]), lower.tail = FALSE) * 2
coef1 <- cbind(coef1, "p value" = p)
ci <- confint(fit1)
```

```
## Waiting for profiling to be done...
##
## Re-fitting to get Hessian
```

```
confint.default(fit1)
```

```
##
## Re-fitting to get Hessian
```

```
##          2.5 %    97.5 %
## IR  -2.5731484 0.9732212
## AWT -0.5507336 1.5786154
## CN  -1.5657857 2.8157949
```

```
exp(cbind(OR = coef(fit1), ci))
```

```
##            OR      2.5 %    97.5 %
## IR  0.4493453 0.06623648  2.538595
## AWT 1.6718669 0.61996975  5.737715
## CN  1.8682546 0.21317659 19.947077
```

CN: For people who do not pay attention to current news, the odds of time of using cellphone per day(high time scale versus low time scale) is multiplied 1.868 times of times of using cellphone of those who pay attention to current news, holding constant all other variables.
AWT: For people who have more average working time per day(one scale higher), the odds of time of using cellphone per day(high time scale versus low time scale) is multiplied 1.672 times of times of using cellphone of those who have less average working time, holding constant all other variables. IR: For people who do not have relationship with others, the odds of time of using cellphone per day(high time scale versus low time scale) is multiplied 0.449 times of times of using cellphone of those who have in relationship with others, holding constant all other variables.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result. From the model, I can not find strong connections between cellphone using time and IR, AWT, CN. The scope of the confidence interval is extremely wide. There are three possible reasons. The first one is that the sample size is too small which can not provide us with enough information and data to show the connection. The second one is that there might exist more suitable model for this data. The last one is actually there does not exist any connections between this response variables and predictive variables.

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. The model does not fit well because of the limitation of the sample. Need to expand the scope of the investigated population since they are not diversified. 2. Moreover, in data collection part, from my perspective, it's better to set the outcome variable continuous. 3. Besides, the predictive variables do not seem to be independent based on the results and common sense, maybe I need to do principal component analysis to explore the latent structure among variables.

Reference: arm package: https://cran.r-project.org/web/packages/arm/index.html
pwr package:https://cran.r-project.org/web/packages/pwr/pwr.pdf
ggplot2 package:https://cran.r-project.org/web/packages/ggplot2/index.html
nnet package:https://cran.r-project.org/web/packages/nnet/nnet.pdf
MASS package:https://cran.r-project.org/web/packages/MASS/index.html

Comments or questions

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