Midterm Exam

Kunyu Liu

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.

My data is about people's willingess of buying cosmetics among Sephora, my variables are listed below: gender age income(in dollars) willingness(people's willingness of Buying cosmetics,the data is proportional) holiday season(binary variable, whether one buy cosmetics among Sephora during holiday season or not) on sale(binary variable, whether one buy cosmetics among Sephora during Sephora sale)

I did my survey with people in different age in my dataset, which includes high school student(my friends who attend high school in the US, and who have a younger sister or brother who attends the high school), undergraduate school student(people I asked are all in Urbana-Champaign), people who have work experience(at the age of 20-30), and people aged more than 30 years. My total observations is 20.

My comparison interest is people's willingness of buying cosmetics in different age groups, and attribution of people's willingness of buying cosmetics among Sephora.

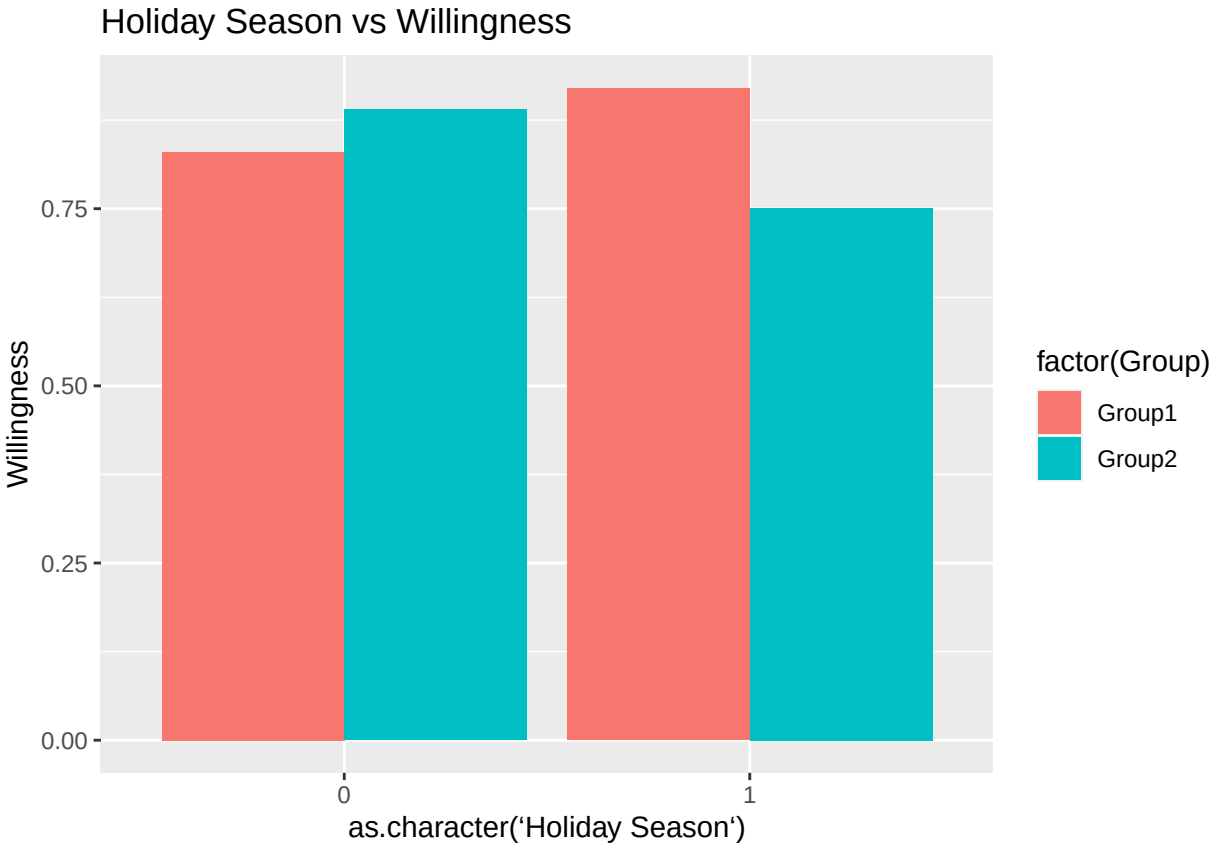
```
library("readxl")
#load data
data = read_excel("data1.xlsx")
attach(data)
```

```r
# I split my dataset into two groups, group1 is people in a younger age, group2 is people in an older a
group1 = data[1:10,]
group2 = data[11:20,]
data$Group = c(rep("Group1",times=10),rep("Group2",times=10))
```

EDA (10pts)

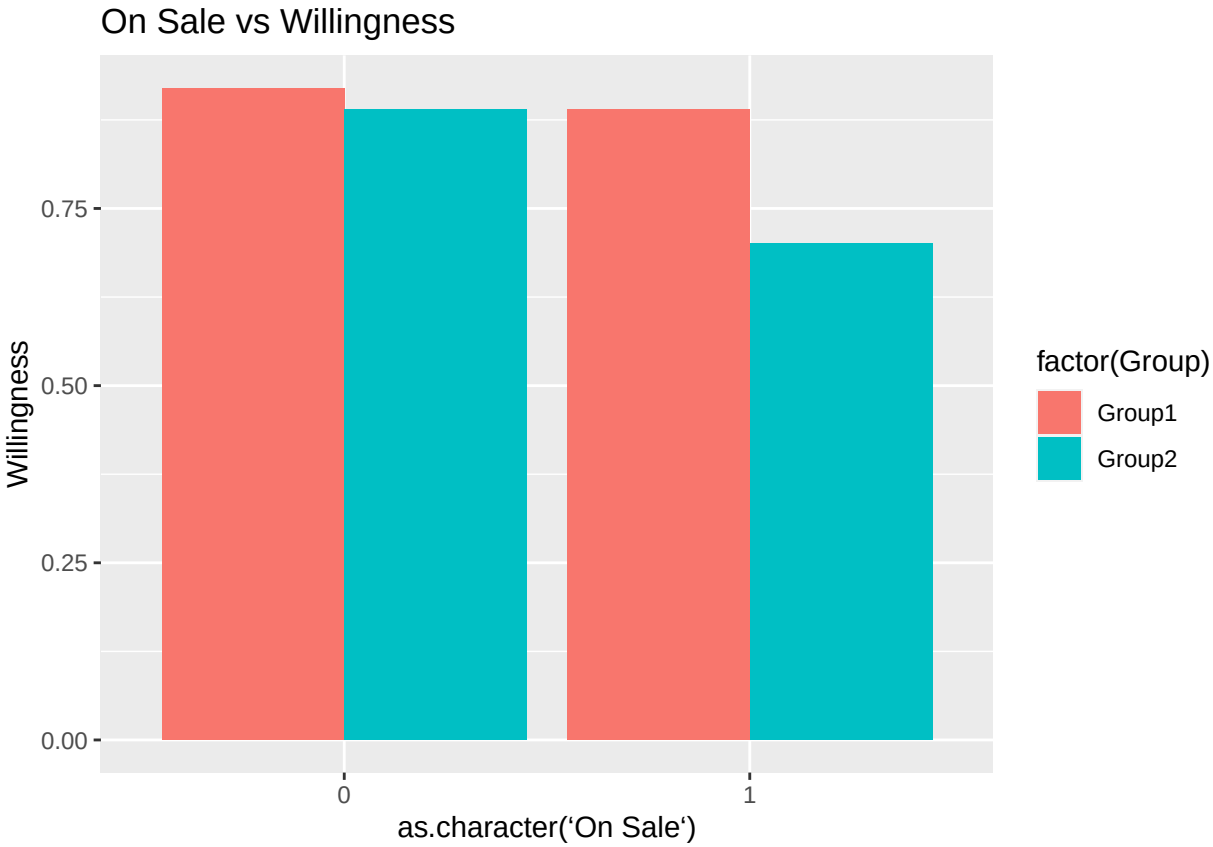
Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```r
ggplot(data = data) +
geom_bar(aes(x = as.character(`Holiday Season`), y = Willingness, fill = factor(Group)),
           position = position_dodge(), stat = "identity") +
  ggtitle("Holiday Season vs Willingness")
```



From the Graph which compare holiday season and people's willingness to buy cosmetics, I can find that group 1 which is younger people, tend to have more willingness to buy cosmetics among Sephora during holiday season, group 2 which is older people, tend to have less willingness to buy in holiday season.

```r
ggplot(data = data) +
geom_bar(aes(x = as.character(`On Sale`), y = Willingness, fill = factor(Group)),
           position = position_dodge(), stat = "identity") +
  ggtitle("On Sale vs Willingness")
```



From the Graph which compare on sale and people's willingness to buy cosmetics, I can find that group 1 and group 2 has almost same willingess in not buying cosmetics during Sephora on sale, and group 1 which is younger people has more willingness to buy cosmetics during Sephora on sale compared with group2.

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.

```
library("pwr")
pwr.t.test(
  n = 10,
  d = NULL,
  sig.level = 0.05,
  power = 0.8,
  type = "two.sample"
)
```

```
##
##      Two-sample t test power calculation
##
##               n = 10
##               d = 1.324947
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
##
## NOTE: n is number in *each* group
```

Based on the above result, I have a d equal to 1.32 when use 80% cutoff to do the power analysis which is too large. In order to solve this, I use the equation

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

to calculate the effect size first, and then find the enough sample size.

```
mean(group1$Willingness)
```

```
## [1] 0.817
```

```
# Calculate d
n1 = length(group1$Willingness)
n2 = length(group2$Willingness)
var1 = var(group1$Willingness)
var2 = var(group2$Willingness)
sdpool = sqrt(((n1 - 1) * var1 + (n2 - 1) * var2)/(n1 + n2 + 2))
sdpool
```

```
## [1] 0.1181332
```

```
d = abs(mean(group1$Willingness) - mean(group2$Willingness))/sdpool
```

After calculate the enough effect size, I then use power.t.test function to find enough sample size when use 80% cutoff to do the power analysis.

```
pwr.t.test(
  n = NULL,
  d = 0.1181332,
  sig.level = 0.05,
  power = 0.8,
  type = "two.sample"
)
```

```
##
##      Two-sample t test power calculation
##
##               n = 1125.806
##               d = 0.1181332
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Based on the calculation above, the sample size in each group need to be 1126, so it can be prove that my sample size is too small.

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.

I pick the linear regression model which is best fits my data and fits my model, because my resonponse variable is a continuous variable, so the linear model is the best fit model.

```
mod = lm(Willingness ~ as.numeric(Gender) + Age + `Income(Dollars)` + as.numeric(`Holiday Season`) + as
summary(mod)
```

```
##
## Call:
## lm(formula = Willingness ~ as.numeric(Gender) + Age + `Income(Dollars)` +
##     as.numeric(`Holiday Season`) + as.numeric(`On Sale`), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22473 -0.07471 -0.01196  0.06155  0.21554
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.066e+00  2.433e-01   4.380 0.000629 ***
## as.numeric(Gender)           1.053e-01  8.994e-02   1.170 0.261396
## Age                         -1.511e-02  1.594e-02  -0.948 0.359102
## `Income(Dollars)`            1.246e-07  1.410e-05   0.009 0.993075
## as.numeric(`Holiday Season`) 3.309e-02  7.449e-02   0.444 0.663654
## as.numeric(`On Sale`)       -1.080e-01  8.024e-02  -1.346 0.199716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1285 on 14 degrees of freedom
## Multiple R-squared:  0.5147, Adjusted R-squared:  0.3414
## F-statistic:  2.97 on 5 and 14 DF,  p-value: 0.04939
```

Except the intercept, none of the variable is significant at 95% level of significance.

Validation (10pts)

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

```
# Do the cross validation with stan
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())
##
## Attaching package: 'rstanarm'
## The following object is masked from 'package:coefplot':
##
##     invlogit
## The following object is masked from 'package:boot':
```

```
##
##     logit
```

```
mod1 = stan_glm(Willingness ~ Gender + Age + `Income(Dollars)` + `Holiday Season` + `On Sale`, data = da
summary(mod1,digits=3)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      Willingness ~ Gender + Age + `Income(Dollars)` + `Holiday Season` +
##     `On Sale`
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 20
##  predictors:   6
##
## Estimates:
##                      mean   sd     10%    50%    90%
## (Intercept)         1.045  0.244  0.740  1.042  1.356
## Gender              0.100  0.093 -0.016  0.100  0.215
## Age                -0.014  0.016 -0.033 -0.014  0.006
## `Income(Dollars)`   0.000  0.000  0.000  0.000  0.000
## `Holiday Season`    0.033  0.080 -0.067  0.032  0.134
## `On Sale`          -0.102  0.085 -0.208 -0.102  0.003
## sigma               0.136  0.027  0.105  0.132  0.172
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 0.726  0.045 0.670 0.725 0.781
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                     mcse  Rhat  n_eff
## (Intercept)         0.006 1.001 1781
## Gender              0.002 1.001 2178
## Age                 0.000 1.001 1597
## `Income(Dollars)`   0.000 1.001 1606
## `Holiday Season`    0.001 1.000 2921
## `On Sale`           0.002 1.001 2239
## sigma               0.001 1.001 1684
## mean_PPD            0.001 1.000 3107
## log-posterior       0.063 1.000 1261
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
# Use LOO Cross Validation
loo(mod1)
```

```
## Warning: Found 3 observation(s) with a pareto_k > 0.7. We recommend calling 'loo' again with argumen
```

```
##
## Computed from 4000 by 20 log-likelihood matrix
##
```

```
##          Estimate  SE
## elpd_loo      6.8 3.5
## p_loo         7.5 2.0
## looic       -13.6 7.0
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     13    65.0%   432
##  (0.5, 0.7]   (ok)        4    20.0%   409
##    (0.7, 1]   (bad)       2    10.0%   95
##    (1, Inf)   (very bad)  1     5.0%   20
## See help('pareto-k-diagnostic') for details.
```

As the results show, neither the value of `elpd_loo` and `looic` is closer to 0, so the model may have some test error.

```
# Then, I use K-Fold Cross Validation
kfold(mod1)
```

```
## Fitting model 1 out of 10

## Fitting model 2 out of 10

## Fitting model 3 out of 10

## Fitting model 4 out of 10

## Fitting model 5 out of 10

## Fitting model 6 out of 10

## Fitting model 7 out of 10

## Fitting model 8 out of 10

## Fitting model 9 out of 10

## Fitting model 10 out of 10

##
## Based on 10-fold cross-validation
##
##            Estimate  SE
## elpd_kfold      7.4 3.3
## p_kfold          NA  NA
## kfoldic       -14.8 6.5
```

The kfold cross validation has the same result as I use loo cross validation, the value of `elpd_kfold` is also not close to 0.

```
# Then, I use glm instead of stan_glm, and use the cv.glm function for LOO Cross Validation
library("boot")
mod2 = glm(Willingness ~ Gender + Age + `Income(Dollars)` + `Holiday Season` + `On Sale`, data = data)
cv.glm(data, mod2, K=10)$delta[1]
```

```
## [1] 0.03059783
```

Based on the calculation, the value is small, so the expected prediction error will be lower. Thus, the linear model is an appropriate model for my analysis.

Inference (10pts)

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

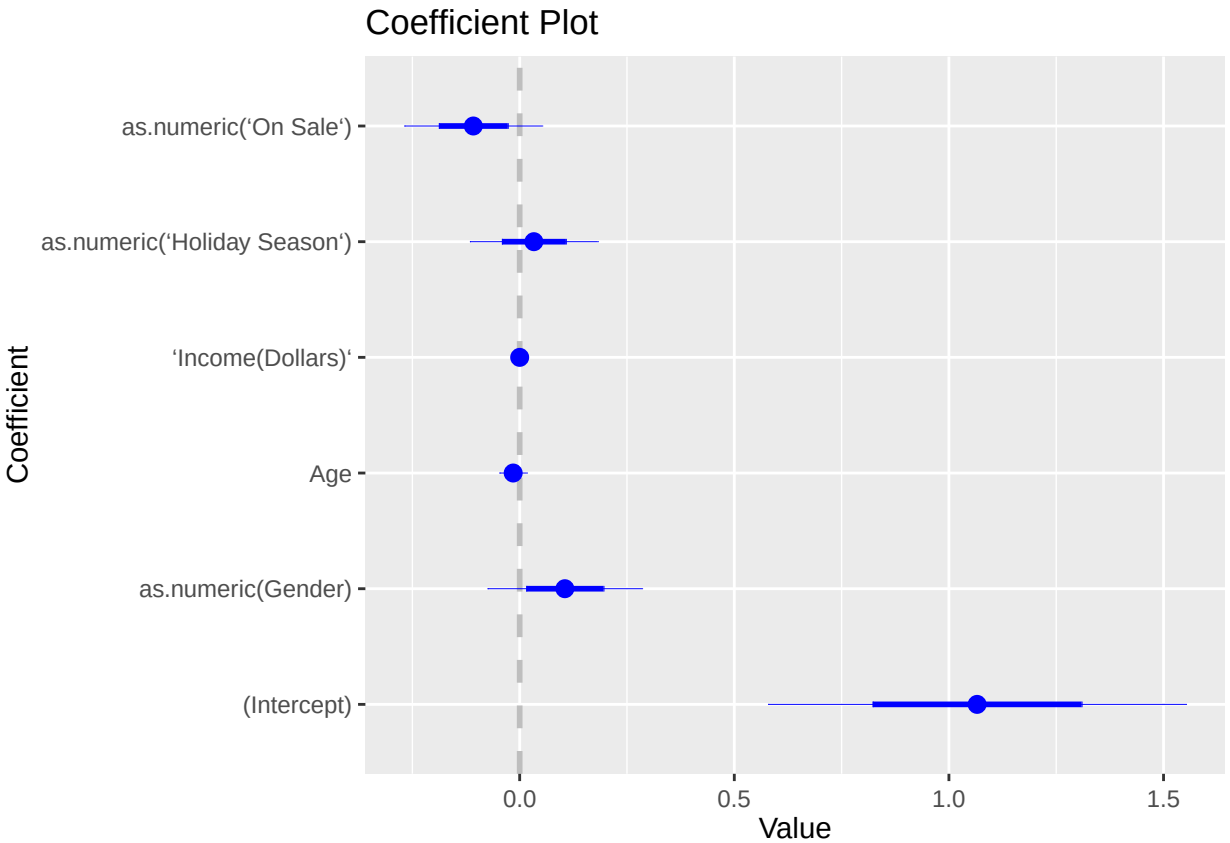
```
# Use `confint()` to calculate the confidence intervals
confint(mod)
```

First, I do the Classical Statistical Inference

```
##                                     2.5 %       97.5 %
## (Intercept)                   5.437895e-01 1.587576e+00
## as.numeric(Gender)           -8.763912e-02 2.981595e-01
## Age                          -4.929728e-02 1.907140e-02
## `Income(Dollars)`            -3.011397e-05 3.036312e-05
## as.numeric(`Holiday Season`) -1.266750e-01 1.928599e-01
## as.numeric(`On Sale`)        -2.800797e-01 6.409659e-02
```

From the results, the confidence intervals of all the predictors contain zero, which means these variables are not statisticlly significant different at the 0.05 level.

```
# visualize coefficient estimates together with their corresponding confidence intervals
library("coefplot")
coefplot(mod, vertical=FALSE, var.las=1, frame.plot=TRUE)
```



Above is the confidence interval plot, cause my p-value is 0.05, which is small, so it is hard to indicate which might be the case of confidence interval for a certain predictor.

P-Value and Statistical Significance The p value of my model is 0.05, so the model is statistically significant at 95% level of significance. The result of p-value which relative to some null hypothesis indicate no effect present. However, it cannot be conclude the result is stable because of p-value is less than 0.05. Thus, although the p-value is an important value in statistical analysis, we cannot make a conclusion only by looking at p-value.

Discussion (10pts)

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

In conclusion, my overall model is statistically significant at 95% level of significance, but variables in my model doesn't have strong significance, and this also explains we cannot use p-value as the only measurement of whether the scientific conclusions or business decisions is significant or not. The reason for that is my sample size is too small, and I think it is the reason for the low statistical power as well. I think there is some correlation between age and people's willingness of buying cosmetics among Sephora, but my research has some limitations.

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.

The biggest concern in my analysis is the sample size, because it is a big problem in scientific research, which may reduce the power of study. In this analysis, I only have 20 observations, which may increases margin of error. In my future study, I need to increase my sample size, add more sample with people in different age groups, and to see how the results will change. Also, some data in the high school student group is historical data, I interviewed my friends who attend high school in the US in the past, so the income and willingness may be differ if I asked people who attend high school currently. In the future, I also need to change this part of historical data into people who attend high school currently.

Comments or questions

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