Midterm Exam

Chenghao Meng

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 (http://www.bu.edu/cas/files/2017/02/GRS-Academic-Conduct-Code-Final.pdf).

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.

```
# Load the data
supermarket <- read.csv(file="https://raw.githubusercontent.com/KratosMC/MA678-Miderterm-test/main/Su
permarket%20data-Chenghao%20Meng.csv",header=T)
head(supermarket)
```

	Supermarket <chr>	Fruit <int>	Vegetable <int>	Meat <int>	Egg <int>	Seafood <int>	Condiment <int>
1	Bravo	71	154	166	30	111	435
2	7Fresh	190	238	179	31	193	478

2 rows

```
# Load the packages
pacman::p_load(tidyverse,pwr,boot,arm)
```

- This dataset is about the number of category of food ingredients and condiments sold through online channels in two supermarket brands (Bravo and 7Fresh). Since their stock varies with location, the dataset only indicates the situation of these two supermarkets around my apartment.

- Because 7Fresh and Bravo are two different types of supermarket, 7Fresh is a smart retail and Bravo is a traditional supermarket brand, I am very interest in whether the type of supermarket will bring significant difference to the number of category in different food ingredients and condiments in general and in each types of fodd ingredients and condiments.

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.

```
# Re-organize the oringinal dataset for plot
market <- data.frame(smart=as.character(rep(c(0,1),6)),
                     category_number=c(supermarket$Fruit,supermarket$Vegetable,
                  supermarket$Meat,supermarket$Egg,
                  supermarket$Seafood,supermarket$Condiment),
                     category=c(rep(c("Fruit"),2),rep(c("Vegetable"),2),rep(c("Meat"),2),rep(c("Egg"
),2),rep(c("Seafood"),2),rep(c("Condiment"),2))
                  )

se <- market %>% group_by(category) %>% summarise(se=sd(category_number))
market <- left_join(market,se,by="category")

head(market)
```

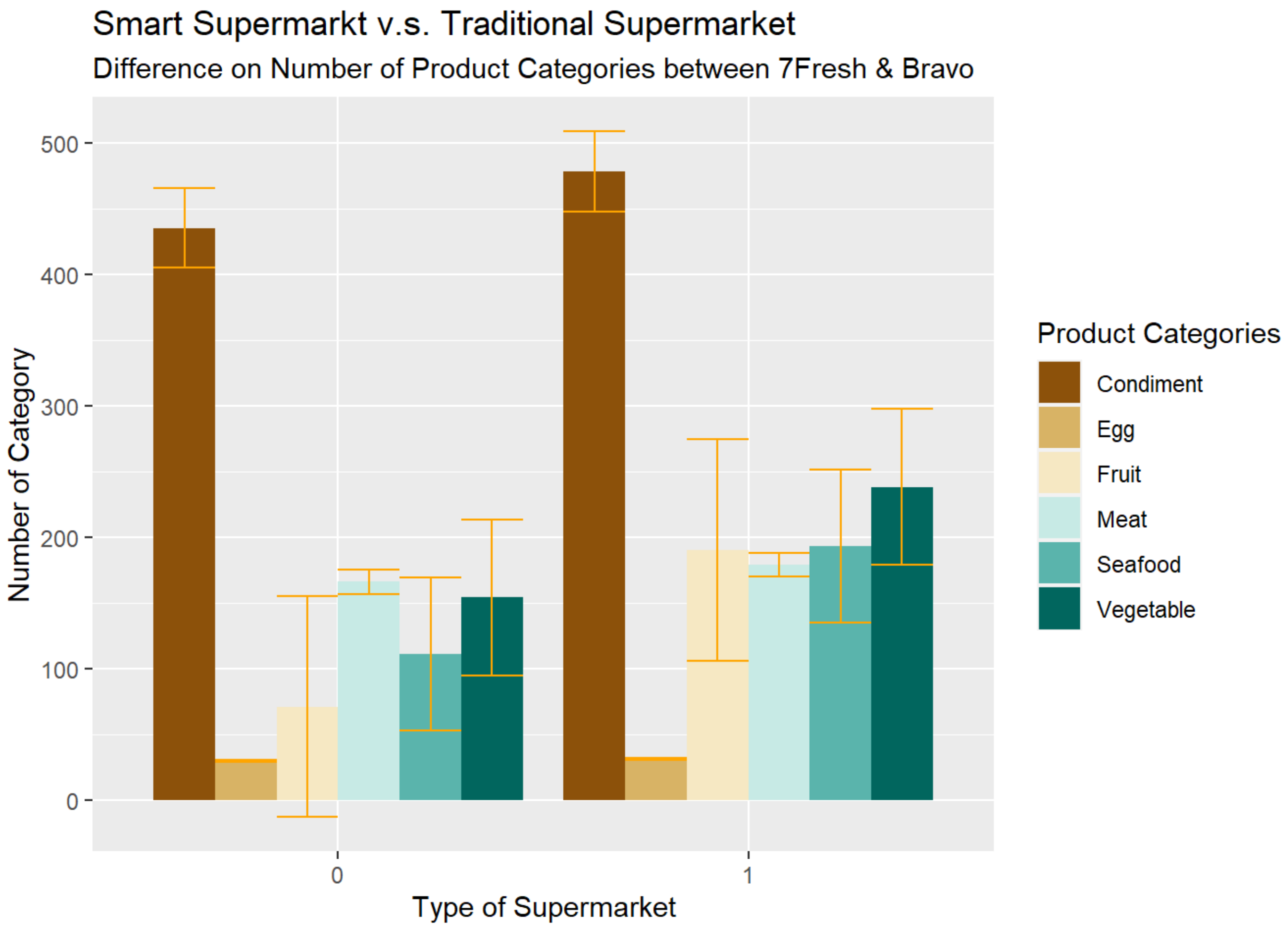
	smart <chr>	category_number <int>	category <chr>	se <dbl>
1	0	71	Fruit	84.145707
2	1	190	Fruit	84.145707
3	0	154	Vegetable	59.396970
4	1	238	Vegetable	59.396970
5	0	166	Meat	9.192388
6	1	179	Meat	9.192388

6 rows

- Now, the following is the bar plots of the dataset given.

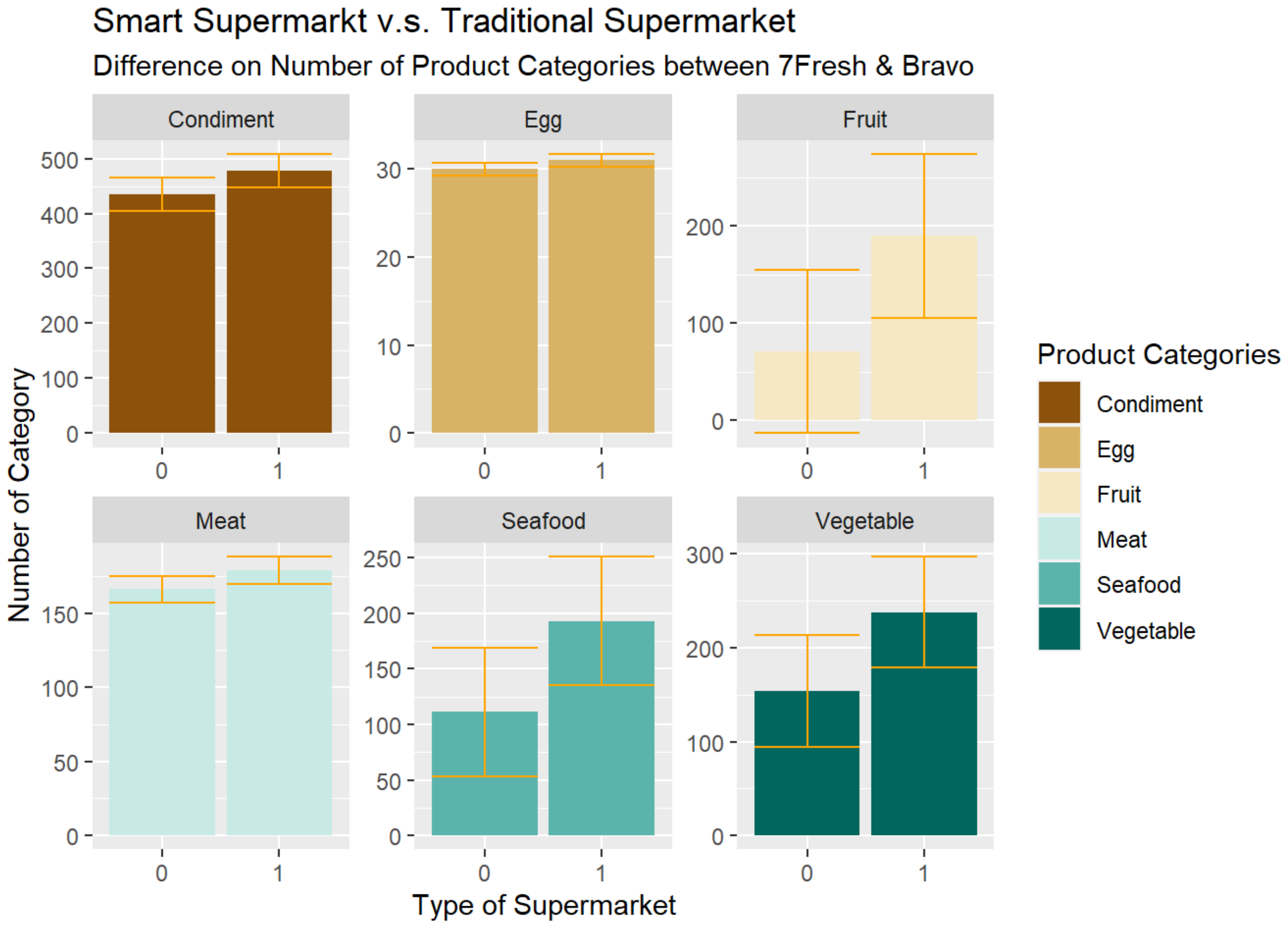
```
# Draw a bar plot
p <- ggplot(data=market,aes(x=smart,y=category_number,fill=category))+
  geom_bar(stat = "identity",position="dodge") +
  geom_errorbar(aes(x=smart, ymin=category_number-se, ymax=category_number+se), stat="identity", posi
tion="dodge",col="orange") +
  scale_fill_brewer(palette = "BrBG")

p <- p + xlab("Type of Supermarket") + ylab("Number of Category") +
  ggtitle("Smart Supermarkt v.s. Traditional Supermarket",subtitle = "Difference on Number of Product
Categories between 7Fresh & Bravo") + labs(fill="Product Categories") # Change the title of legend
p
```



- From the bar plot above, we can see that there are some difference between the number of categories on daily-consumed products. However, the difference is not apparent in general. As a result, a bar plot with facets is needed.

```
# Barplot with facet
p + facet_wrap(~category,scales="free")
```



From those facets, I can notice that there are big differences on number of category of *Fruit*, *Seafood* and *Vegetable*. Model establishment is needed for further analysis.

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.

- The two groups in my analysis is independent, so I will use two sample t-test.

```
# Infer the effect size
pwr.t.test(n=6,d=NULL,sig.level=0.05,power=0.8,type = "two.sample")
```

```
##
##      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
```

- Since *d* denotes the effect size, this effect size 0f 1.796 is larger than 1, which is abnormal. That means my sample size in each group is not enough.

```
mean_market <- market %>% group_by(smart) %>% summarise(average=mean(category_number),sd=sd(category_
number))

u1 <- mean_market$average[1]
u2 <- mean_market$average[2]
sig1 <- mean_market$sd[1]
sig2 <- mean_market$sd[2]
```

```
# Calculate the effect size-Cohen's d of the dataset
d_market <- (abs(u1-u2)/sqrt((sig1^2+sig2^2)/2))
# Calculate the sample size
pwr.t.test(n=NULL,d=d_market,sig.level=0.05,power=0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 101.8568
##               d = 0.3944569
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

- The result above indicate that there need to be 102 observations in each group of my study. However, there are only 6 observations in each group currently. As a result, my sample size is actually not enough for the problem at hand.
- Because the true value of effect size is unkown for most of the time, the use of external information to define the effect size is needed.

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.

- Because my independent variables are binary and my dependent variable is numeric and continuous, linear regression will be suitable for the analysis.
- Meanwhile, I would like to see the difference of the number of food and condiment categories in general and in each types of food and condiment. So, the inclusion of `category` column of the dataset is needed.

```
# Fit the model with lm
fit_lm <- glm(category_number~as.numeric(smart)+category,data=market)
summary(fit_lm)
```

```
##
## Call:
## glm(formula = category_number ~ as.numeric(smart) + category,
##     data = market)
##
## Deviance Residuals:
##     1      2      3      4      5      6      7      8      9     10     11
## -31.0   31.0  -13.5   13.5   22.0  -22.0   28.0  -28.0  -12.5   12.5    7.0
##    12
##  -7.0
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         428.00      24.71  17.322 1.17e-05 ***
## as.numeric(smart)    57.00      18.68   3.052 0.028368 *
## categoryEgg        -426.00      32.35 -13.168 4.51e-05 ***
## categoryFruit      -326.00      32.35 -10.077 0.000165 ***
## categoryMeat       -284.00      32.35  -8.779 0.000318 ***
## categorySeafood    -304.50      32.35  -9.412 0.000228 ***
## categoryVegetable  -260.50      32.35  -8.052 0.000478 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1046.6)
##
##     Null deviance: 218557  on 11  degrees of freedom
## Residual deviance:   5233  on  5  degrees of freedom
## AIC: 122.99
##
## Number of Fisher Scoring iterations: 2
```

Validation (10pts)

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

```
# K-fold Validation
# Model with smart and category as indicators
cv.glm(market,fit_lm,K=10)$delta[1]
```

```
## Warning in cv.glm(market, fit_lm, K = 10): 'K' has been set to 12.000000
```

```
## [1] 2511.84
```

```
cv.glm(market,glm(category_number~as.numeric(smart),data=market),K=10)$delta[1]
```

```
## Warning in cv.glm(market, glm(category_number ~ as.numeric(smart), data =
## market), : 'K' has been set to 12.000000
```

```
## [1] 25057.16
```

- Since `cv.glm` uses average MSE across each fold as the output, the choosen model's average MSE across each fold is 2511.84, which is obviously smaller than the model with `smart` as the only indictor with 25057.16 as the average MSE. That means my choice is more approriate.

Inference (10pts)

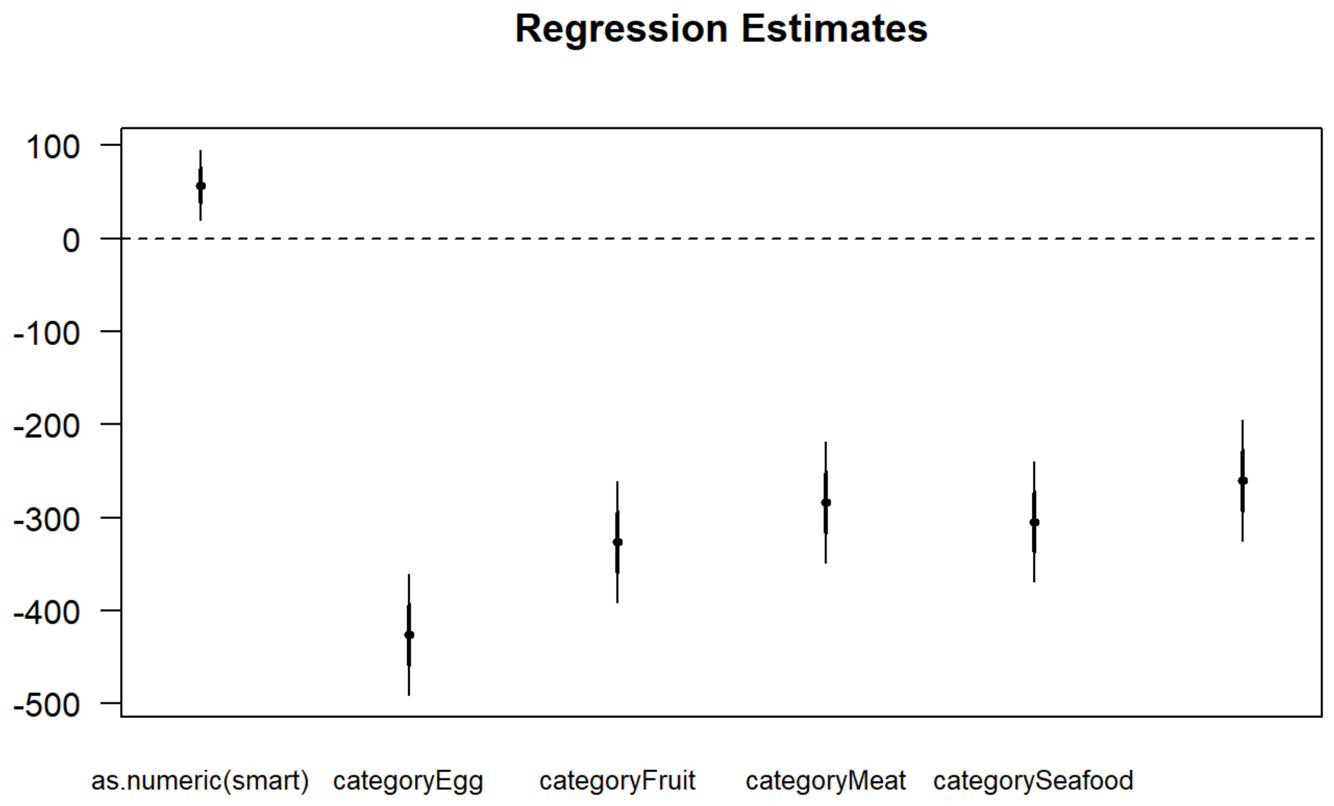
Based on the result so far please perform statistical inference to compare the comparison of interest.

```
# Confidence Interval
confint(fit_lm)
```

```
## Waiting for profiling to be done...
```

```
##                         2.5 %     97.5 %
## (Intercept)         379.57196  476.42804
## as.numeric(smart)    20.39185   93.60815
## categoryEgg        -489.40718 -362.59282
## categoryFruit      -389.40718 -262.59282
## categoryMeat       -347.40718 -220.59282
## categorySeafood    -367.90718 -241.09282
## categoryVegetable -323.90718 -197.09282
```

```
# Visualize the CI
coefplot(fit_lm, vertical=FALSE, var.las=1, frame.plot=TRUE)
```



Discussion (10pts)

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

- The result indicates that the types of supermarket do not bring significant difference to the number in each product category of interest in the supermarket around my apartment in general.

- However, when comparing number in different categories of food ingredients and condiments with same types of supermarket, the fitted model shows that there are 426 fewer types of **Egg** than **Condiment**, 326 fewer types of **Fruit** than **Condiment**, 284 fewer types of **Meat** than **Condiment**, 305 fewer types of **Seafood** than **Condiment** and 261 fewer types of **Vegetable** than **Condiment**,on average. Note that, the difference of their comparison to number of types of condiment is significant according to p-value.

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.

- Concerns and Limitations:
    - 1. I have only investigate two supermarket around my apartment,even though Bravo is a traditional one and 7Fresh is a smart retail, the result of the identification for the difference brought by supermarket type might not be solid due to the limited number within one supermarket brand and the limited number of supermarket brand.
    - 2. Even though the p-value in the model is significant, my observations in the dataset is very small, and as a result, the conclusion generated from the model is not solid and robust.
    - 3. Standard of deciding which one is smart retail and which one is traditional supermarket is not that specific.
    - 4. My analysis seems to be underpowered.
- Future Opportunity

  I may investigate more supermarket brands to make my dataset a larger one to include more smart retail supermarket and traditional supermarket. And I may also clarify the standard of smart retail supermarket.

Comments or questions

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

- When I explore the Internet, I find that Cohen’s d is either 0.2, 0.5 or 0.8. So, if we use the common Cohen’s d to compute sample size with designated power of 0.8 and sig.level of 0.05, don’t we all get the same sample size?

- I think you may inform us that we need to use the data in the data collection assignment to perform certain analyses based on the model we learned, so that we may have more structured dataset.