Midterm Exam

Haoyu Wang

11/7/2020

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.

```
Smoking <- read_csv("Smoking.csv")
summary(Smoking)
```

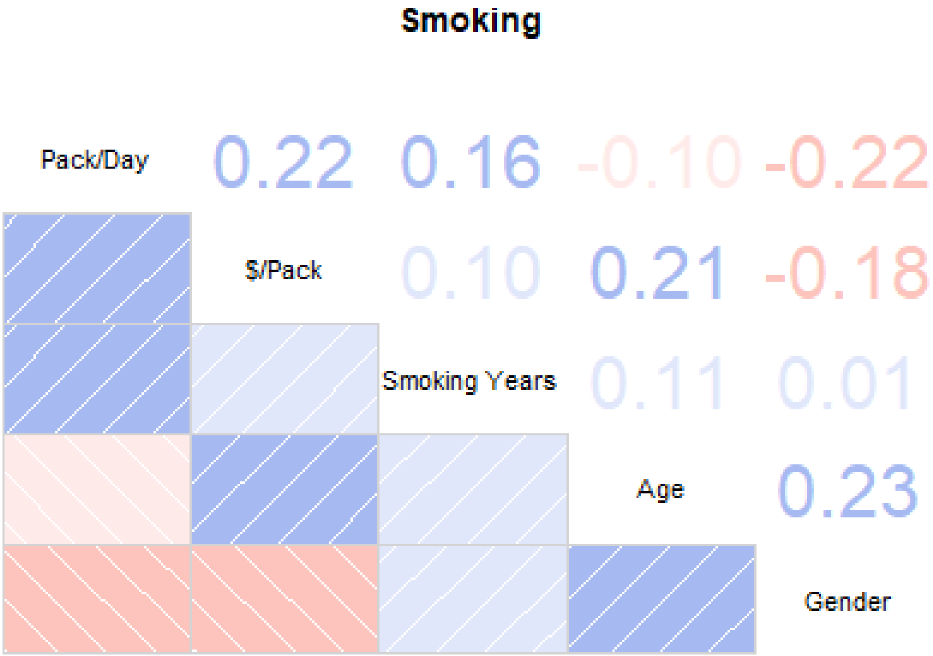
```
##       No.             City               Gender           Age       
##  Min.   : 1.00   Length:26          Min.   :1.000   Min.   :12.00  
##  1st Qu.: 7.25   Class :character   1st Qu.:1.000   1st Qu.:22.00  
##  Median :13.50   Mode  :character   Median :1.000   Median :23.00  
##  Mean   :13.50                      Mean   :1.269   Mean   :24.19  
##  3rd Qu.:19.75                      3rd Qu.:1.750   3rd Qu.:26.75  
##  Max.   :26.00                      Max.   :2.000   Max.   :46.00  
##  Smoking Years      Pack/Day         $/Pack     
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.00  
##  1st Qu.: 4.250   1st Qu.:0.500   1st Qu.:19.00  
##  Median : 6.500   Median :1.000   Median :23.00  
##  Mean   : 7.904   Mean   :1.271   Mean   :24.73  
##  3rd Qu.:10.000   3rd Qu.:2.000   3rd Qu.:26.50  
##  Max.   :20.000   Max.   :7.000   Max.   :75.00  
```

The data selected for this time are Some cities in China questions about smokers, with the main variables being sex, age, smoking age, smoking degree and amount spent. The main concern of this experiment is *whether the cost of smoking is related to sex, smoking years and smoking degree.*

EDA (10pts)

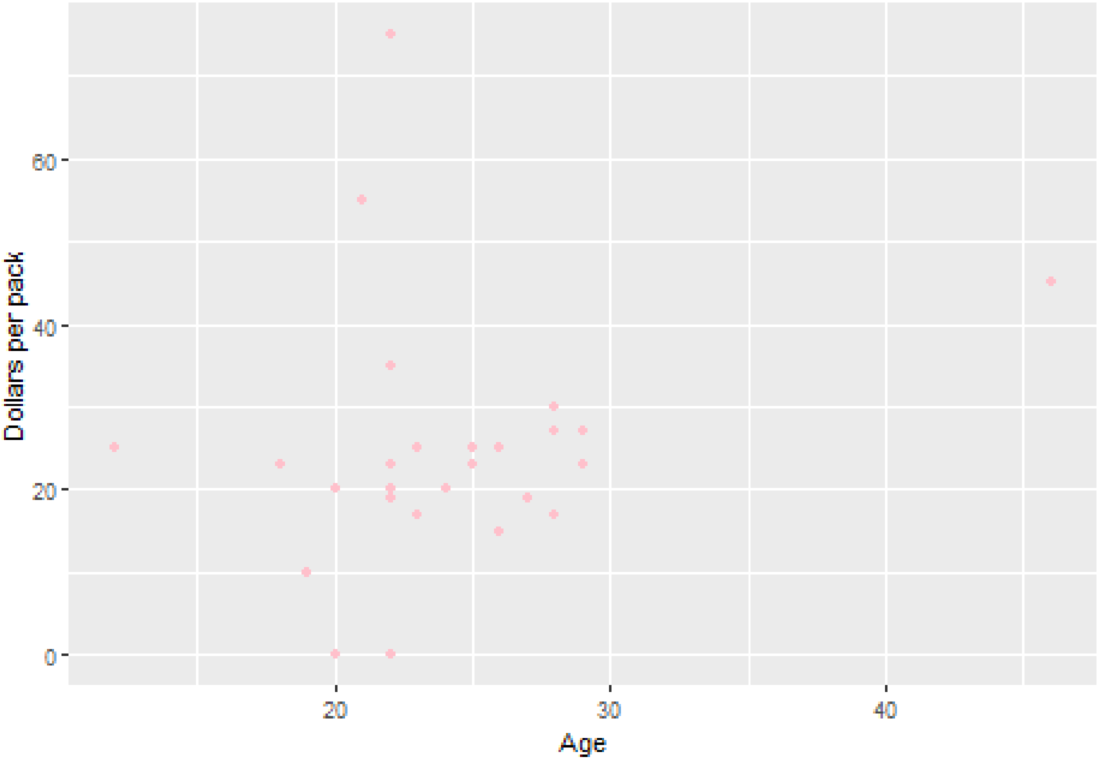
Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```
Smoking = Smoking[, 3:7]

cor=cor(Smoking)

corrgram(Smoking, order=TRUE,
         upper.panel=panel.cor, main="Smoking")
```

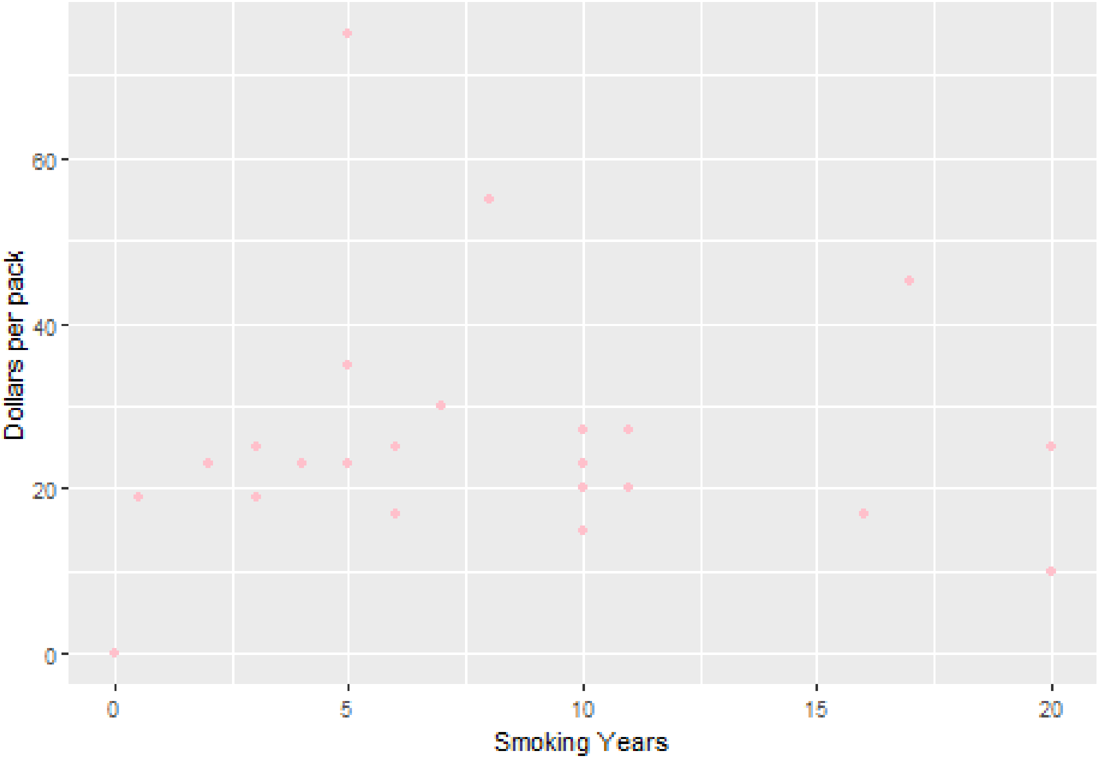


It can be seen from the correlation coefficient diagram that there is a positive relationship between the amount, age, daily smoking amount and smoking years, and a negative relationship with gender.

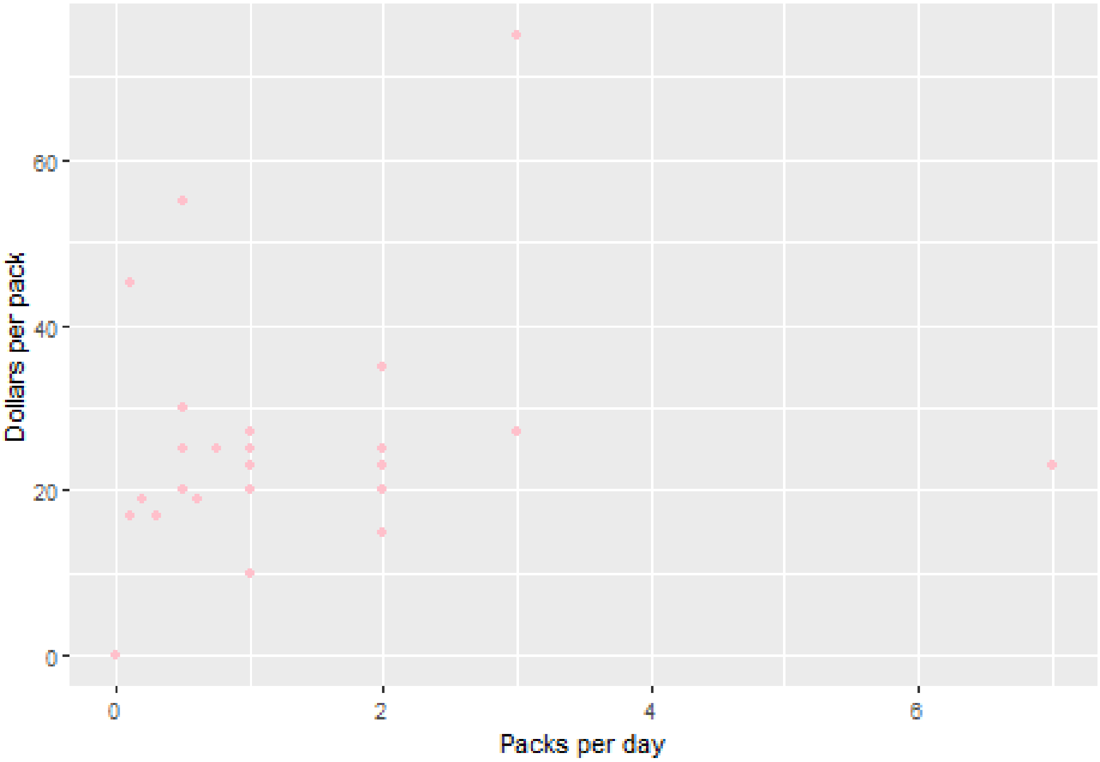
```
ggplot(Smoking) +
        geom_point(aes(x = Smoking$Age, y = Smoking$`$/Pack`), color = "pink") +
        labs(x = "Age", y = "Dollars per pack")
```



```
ggplot(Smoking) +
        geom_point(aes(x = Smoking$`Smoking Years`, y = Smoking$`$/Pack`), color = "pink") +
        labs(x = "Smoking Years", y = "Dollars per pack")
```



```
ggplot(Smoking) +
        geom_point(aes(x = Smoking$`Pack/Day`, y = Smoking$`$/Pack` ), color = "pink") +
        labs(x = "Packs per day", y = "Dollars per pack")
```



It can be seen from the scatter plot of the amount spent on smoking and the explanatory variables that there is a linear relationship between the amount spent on smoking and the number of years, age and daily smoking. In general, the linear regression model can be established in this experiment.

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.

```
power.t.test(n =length(Smoking),power=.80)
```

```
##
##      Two-sample t test power calculation
##
##               n = 5
##           delta = 2.024438
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The results of power test showed that the significant level was 0.05 and the delta value was 2.024438. It can *be seen from the scatter diagram and correlation coefficient analysis of the previous explanatory variables and the explained variables that although there is also a certain linear relationship between the explanatory variables and the explained variables, this relationship is weak. If the fitting model is used to test, there is a big gap between the test effect and the actual situation.*

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.

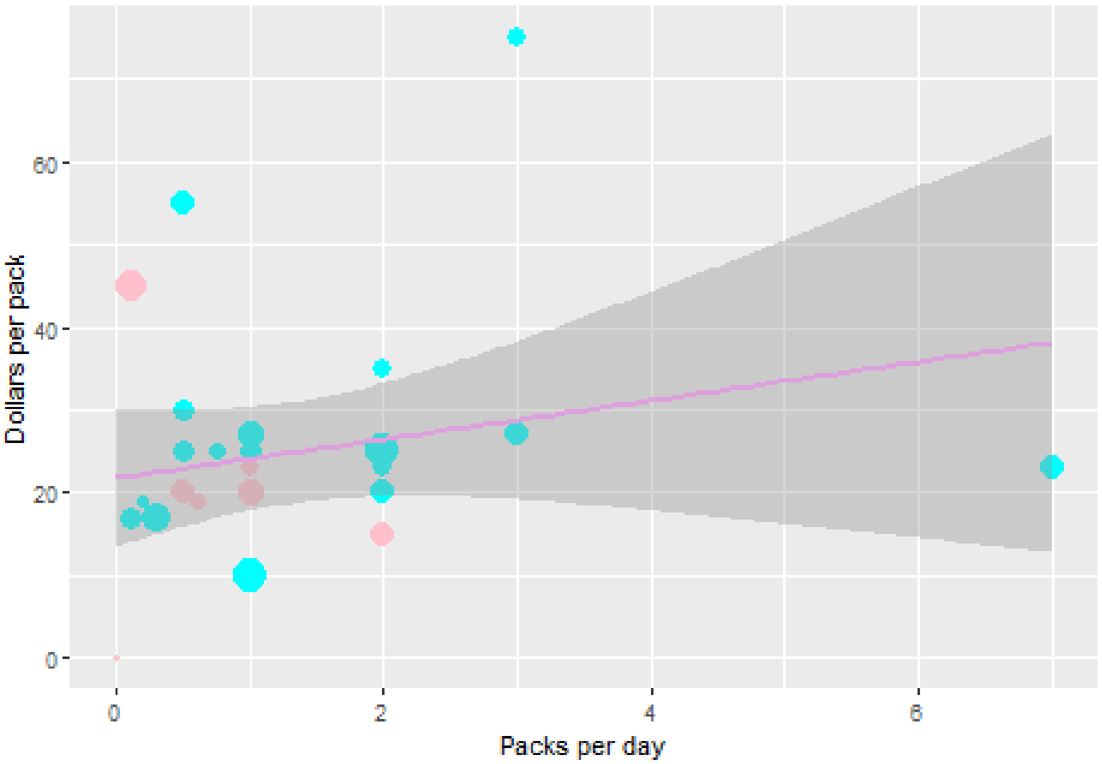
```
fit_model = lm(Smoking$`$/Pack` ~ Smoking$Gender + Smoking$Age + Smoking$`Smoking Years`+
                    Smoking$`Pack/Day`, data = Smoking)
summary(fit_model)
```

```
##
## Call:
## lm(formula = Smoking$`$/Pack` ~ Smoking$Gender + Smoking$Age +
##     Smoking$`Smoking Years` + Smoking$`Pack/Day`, data = Smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.060  -6.628  -2.045   2.987  46.692
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              12.5766    15.3156   0.821    0.421
## Smoking$Gender           -6.7596     7.1481  -0.946    0.355
## Smoking$Age               0.7122     0.5406   1.317    0.202
## Smoking$`Smoking Years`   0.1061     0.5571   0.190    0.851
## Smoking$`Pack/Day`        2.0976     2.2039   0.952    0.352
##
## Residual standard error: 15.37 on 21 degrees of freedom
## Multiple R-squared:  0.1439, Adjusted R-squared:  -0.0192
## F-statistic: 0.8822 on 4 and 21 DF,  p-value: 0.4914
```

```
confint(fit_model, level = 0.95)
```

```
##                               2.5 %    97.5 %
## (Intercept)             -19.2738229 44.427074
## Smoking$Gender          -21.6247602  8.105656
## Smoking$Age              -0.4121847  1.836499
## Smoking$`Smoking Years`  -1.0524104  1.264582
## Smoking$`Pack/Day`       -2.4857237  6.680918
```

```
ggplot(Smoking, aes(x = Smoking$`Pack/Day`, y = Smoking$`$/Pack`)) +
         geom_point(aes(color = Gender, size = `Smoking Years`))+
         scale_color_gradient(low = "cyan",high = "pink") +
         labs(x = "Packs per day", y = "Dollars per pack") +
         theme(legend.position="none") +
         geom_smooth(method = "lm", se = T, formula = y ~ x, col = "plum")
```



Validation (10pts)

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

First of all, in the data set selected in this experiment, the explained variables are continuous variables, so it is not suitable to use classification models such as Loglist models. Secondly, through the previous correlation coefficient and scatter plot, we can see that there is a certain linear relationship between the explained variables and the explained variables. Therefore, this experiment thinks that it is appropriate to choose a linear model.

Inference (10pts)

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

From the regression results, the statistical relationship between the explanatory variables and the explained variables is not significant, which indicates that the relationship between the explanatory variables and the amount spent by smokers is not very clear. The possible reason is that the sample size of this experiment is small.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result. *Men spend relatively more on smoking than women. In addition, age, length of smoking and daily smoking were positively correlated with the amount of smoking. But this positive relationship is more expensive, but the impact on smoking costs is slowly increasing.*

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. *In this case, although the linear regression model is used, the model results are deviated because of the interference of multiple collinearity, heteroscedasticity and autocorrelation in linear regression* *theory. Therefore, in the future study and practice should pay attention to identify and solve these problems. Make the model results more in line with the actual needs.*

Comments or questions

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