Midterm Exam

AOYI LI

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.

```
#load data
study_room<-read.csv(file="studyroom.csv")
study_room
```

```
##            group choice freq dist timesp price drink privr
## 1       graduate      1    3    5      3    12     1     1
## 2       graduate      1    3   30      3     9     1     1
## 3       graduate      0    0    0      3     1     1     1
## 4       graduate      1    1   15      2     6     1     1
## 5       graduate      1    3   30      3     6     1     1
## 6       graduate      0    0    0      1     1     0     0
## 7       graduate      0    0    0      3     1     0     1
## 8       graduate      1    2    0      5     6     1     1
## 9       graduate      0    0   15      5     1     0     1
## 10 undergraduate      0    0    0      1     3     1     1
```

```
## 11 undergraduate      1    1   30      3     6     0     1
## 12 undergraduate      1    5   30      3     9     1     1
## 13 undergraduate      0    0    0      2     3     1     1
## 14 undergraduate      1    1    0      5     3     0     1
## 15 undergraduate      0    0   15      3     6     1     1
## 16 undergraduate      0    0   15      1     3     1     1
## 17 undergraduate      1    5   15      3    12     0     1
## 18 undergraduate      1    1   15      2     3     1     1
## 19 undergraduate      0    0    5      3     6     1     1
## 20 undergraduate      1    3    0      2     3     1     1
## 21 undergraduate      1    2   15      2     3     1     1
## 22 undergraduate      0    0    0      5     1     1     1
## 23 undergraduate      0    0    5      5     3     1     1
## 24       working      1    1   15      3     6     1     1
## 25       working      1    1   15      1     3     1     1
## 26       working      1    3   30      3     6     1     1
## 27       working      0    0   15      2     3     1     1
## 28       working      1    1   30      3     9     1     1
## 29       working      1    1   30      3     9     1     1
```

```
study_room$group<-as.factor(study_room$group)
study_room$drink<-as.factor(study_room$drink)
study_room$privr<-as.factor(study_room$privr)
```

#Data Description Group: Current status(undergraduate, graduate,working). Choice: Whether they would go to paid study room outside schools or not.(1:yes, 0:no) Freq: How many times they would go to the study room per week. Dist: The maximum acceptable distance from their home/school to the study room. Timesp: Average time spend in the study room. Price: The maximum acceptable price per hour. Drink: Whether they would buy drinks offered by the study room or not.(1:yes, 0:no) Privr: Whether they would choose a private room to study or not. (1:yes, 0:no)

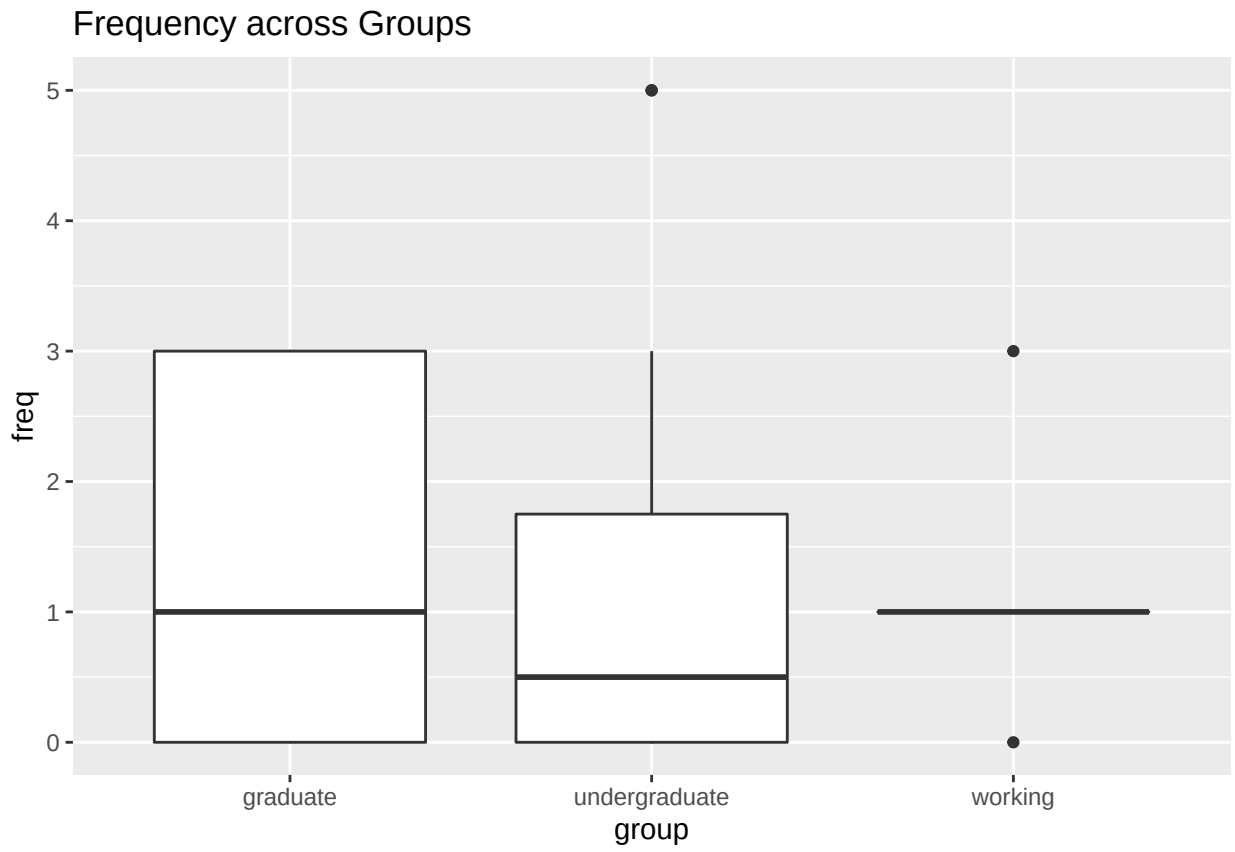
I plan to start my own paid study room, but site selection and pricing are the biggest problem that I am facing now. So, the data is collected to understand the market demand for study room outside schools and the differences of the demands for different groups. I created a online survey to collect data from my friends and their friends about how many times they would go to a paid study room in one week, how much they are willing to pay per hour and what is their maximum acceptable distance from their home/school to the study room.

The main goal here is to compare the differences of the frequencies for each group(undergraduate,graduate and working) and to understand the factors that will affect the frequencies.

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.

```
#boxplot
study_room%>%ggplot(aes(group,freq))+
  geom_boxplot()+ggtitle("Frequency across Groups")
```




```
#highest acceptable price across groups
ggplot(aes(price,freq,col=group),data=study_room)+geom_point()+geom_smooth(se=FALSE)+ggtitle("Highest A
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.945

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 5.055

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 25

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.945

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 5.055

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 6.0048e-017
```

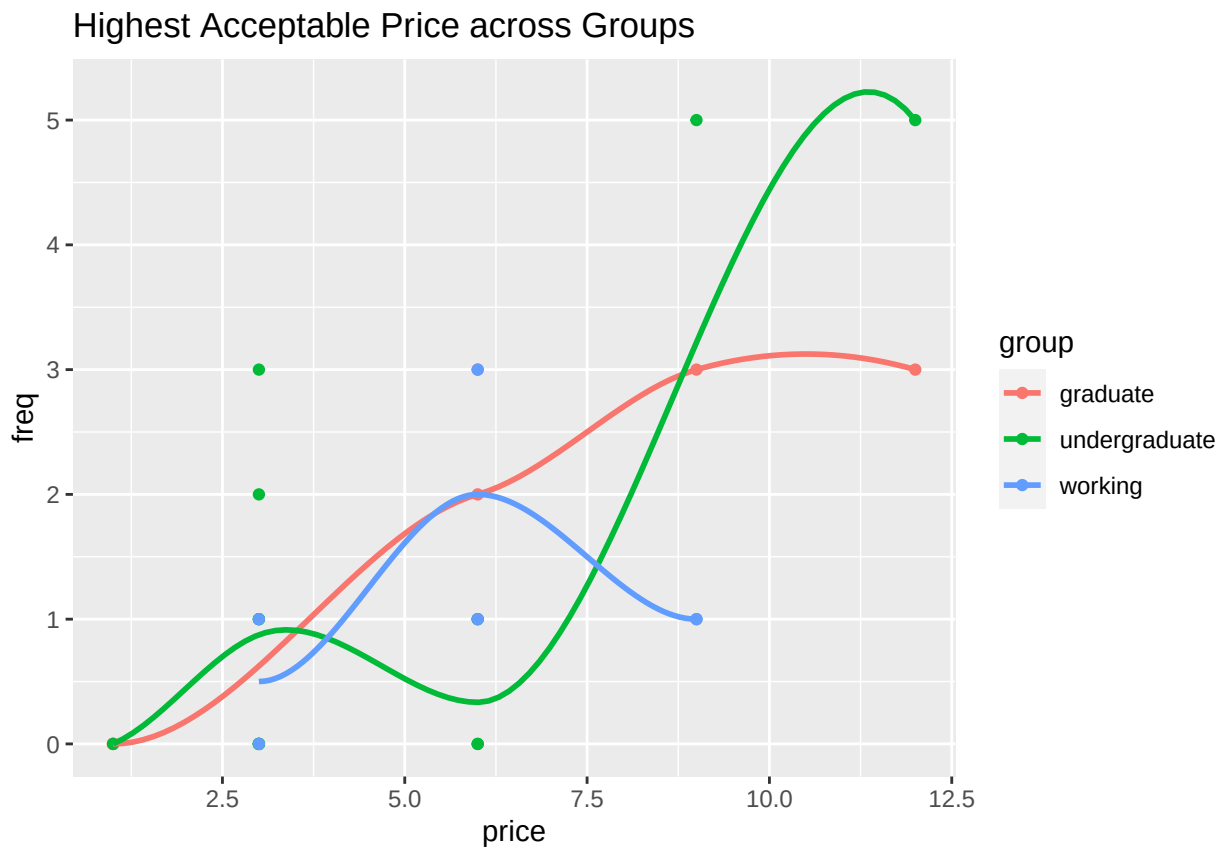
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 9

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 2.97

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 3.03

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 9.1809
```



From the boxplot, we can find that there is not a clear difference across each group. From the other plot, we can see that there might be a positive relationship between price and frequency.

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
##2 sample t test
#graduate and undergraduate
pwr.t2n.test(n1=9,n2=14,d=NULL,sig.level=0.05,power=0.8)
```

```
##
##      t test power calculation
##
##              n1 = 9
##              n2 = 14
##               d = 1.255321
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```r
#undergraduate and working
pwr.t2n.test(n1=14,n2=6,d=NULL,sig.level=0.05,power=0.8)
```

```
##
##      t test power calculation
##
##              n1 = 14
##              n2 = 6
##               d = 1.445628
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```r
#graduate and working
pwr.t2n.test(n1=9,n2=6,d=NULL,sig.level=0.05,power=0.8)
```

```
##
##      t test power calculation
##
##              n1 = 9
##              n2 = 6
##               d = 1.597641
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```r
##general linear model test
pwr.f2.test(u=2,v=26,sig.level=0.05,power=0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 26
##              f2 = 0.3736753
##       sig.level = 0.05
##           power = 0.8
```

Cohen suggests that d value of 0.2,0.5 and 0.8 represent small, medium, and large effect sizes respectively. Since all 3 test have d greater than 1, the difference between 2 means is larger than 1 standard deviation. Since the sample size is less than 50, it tends to over-inflate results. And the general linear model test has f2 = 0.37, indicates large effect size, which means we only need a small sample size. But my sample size is only 29, which might not be enough for the problem. From the tests, the effect size is too large, the effect size from the fitted model will cause some error. Thus, it should not be used.

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.

```
#original model
fit<-lm(freq~dist+timesp+price+privr+group+drink,data=study_room)
summary(fit)
```

```
##
## Call:
## lm(formula = freq ~ dist + timesp + price + privr + group + drink,
##     data = study_room)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78422 -0.61794 -0.02269  0.66068  2.42161
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.25595    1.22678  -0.209   0.8367
## dist                0.02393    0.02592   0.923   0.3665
## timesp             -0.03955    0.21649  -0.183   0.8568
## price               0.29550    0.08610   3.432   0.0025 **
## privr1              0.15570    1.55887   0.100   0.9214
## groupundergraduate  0.00691    0.54748   0.013   0.9900
## groupworking       -0.80986    0.72447  -1.118   0.2762
## drink1             -0.13566    0.65212  -0.208   0.8372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.206 on 21 degrees of freedom
## Multiple R-squared:  0.5214, Adjusted R-squared:  0.3619
## F-statistic: 3.269 on 7 and 21 DF,  p-value: 0.01647
```

I used the simple linear model will all predictors in the table to fit the original model, but the output shows that the model does not have a good fit with only one significant variable and low adjusted R square.

```
#final model
ols_step_best_subset(fit)
```

```
##              Best Subsets Regression
## -------------------------------------------------
## Model Index    Predictors
```

```
## ------------------------------------------------
##      1         price
##      2         price group
##      3         dist price group
##      4         dist price group drink
##      5         dist timesp price group drink
##      6         dist timesp price privr group drink
## ------------------------------------------------
##
##                                                  Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC        SBIC        SBC        MSEP
## ----------------------------------------------------------------------------------------------------------
##   1        0.4733      0.4538      0.3856    -1.8878     92.5678    11.1586     96.6697    36.0952
##   2        0.4986      0.4384      0.3327    -0.9961     95.1427    12.5498    101.9792    35.7388
##   3        0.5201      0.4401      0.3095     0.0590     95.8699    14.3037    104.0736    35.6293
##   4        0.5207      0.4165       0.227     2.0339     97.8352    17.0372    107.4063    37.1340
##   5        0.5212      0.3906      0.2003     4.0100     99.8023    19.7782    110.7407    38.7778
##   6        0.5214      0.3619        -Inf     6.0000    101.7885    22.5332    114.0942    40.6051
## ----------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
fit1<-lm(freq~dist+price+group,data=study_room)
summary(fit1)
```

```
##
## Call:
## lm(formula = freq ~ dist + price + group, data = study_room)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81913 -0.69276  0.01683  0.57050  2.43121
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.32737    0.49826  -0.657  0.51742
## dist                0.02479    0.02389   1.038  0.30974
## price               0.29281    0.07914   3.700  0.00112 **
## groupundergraduate  0.01772    0.48276   0.037  0.97102
## groupworking       -0.82070    0.64512  -1.272  0.21550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.129 on 24 degrees of freedom
## Multiple R-squared:  0.5201, Adjusted R-squared:  0.4401
## F-statistic: 6.502 on 4 and 24 DF,  p-value: 0.001081
```

Then I tried a simple variable selection method to find a better subset of variables. From the result we can see that the best model is a bivariable model with outcome frequency and predictor price. However, I want to see the coefficients for different groups and distance might be a significant predictor when sample size becomes large. In addition, I find out that the third model is better than the second one and does not have too much difference between the first one. Thus, I chose the third model as the final model, including dist, group and price as predictors, to predict to outcome.

Validation (10pts)

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

```
#Cross validation
fit_11<-stan_glm(freq~dist+price+group,data=study_room,refresh=0)

loo_11<-loo(fit_11)
```

```
## Warning: Found 1 observation(s) with a pareto_k > 0.7. We recommend calling 'loo' again with argumen
```

```
print(loo_11)
```

```
##
## Computed from 4000 by 29 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo    -48.6 3.6
## p_loo         5.4 1.2
## looic        97.3 7.3
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     24    82.8%   905
##  (0.5, 0.7]   (ok)        4    13.8%   591
##    (0.7, 1]   (bad)       1     3.4%   347
##    (1, Inf)   (very bad)  0     0.0%   <NA>
## See help('pareto-k-diagnostic') for details.
```

```
fit_12<-glm(freq~dist+price+group,data=study_room)
cv.glm(study_room,fit_12)$delta
```

```
## [1] 1.518923 1.510059
```

Since elpd_loo is the estimated log score along with a standard error representing uncertainty, elpd_loo here is not a large number, and the cv.glm result which is 1.51 is also small, which means the model is fine.

```
#compare models
library(broom)
glance(fit)%>%select(adj.r.squared,sigma,AIC,BIC,p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC p.value
##           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1         0.362  1.21  102.  114.  0.0165
```

```
glance(fit1)%>%select(adj.r.squared,sigma,AIC,BIC,p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC p.value
##           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1         0.440  1.13  95.9  104. 0.00108
```

```
#marginal plot
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

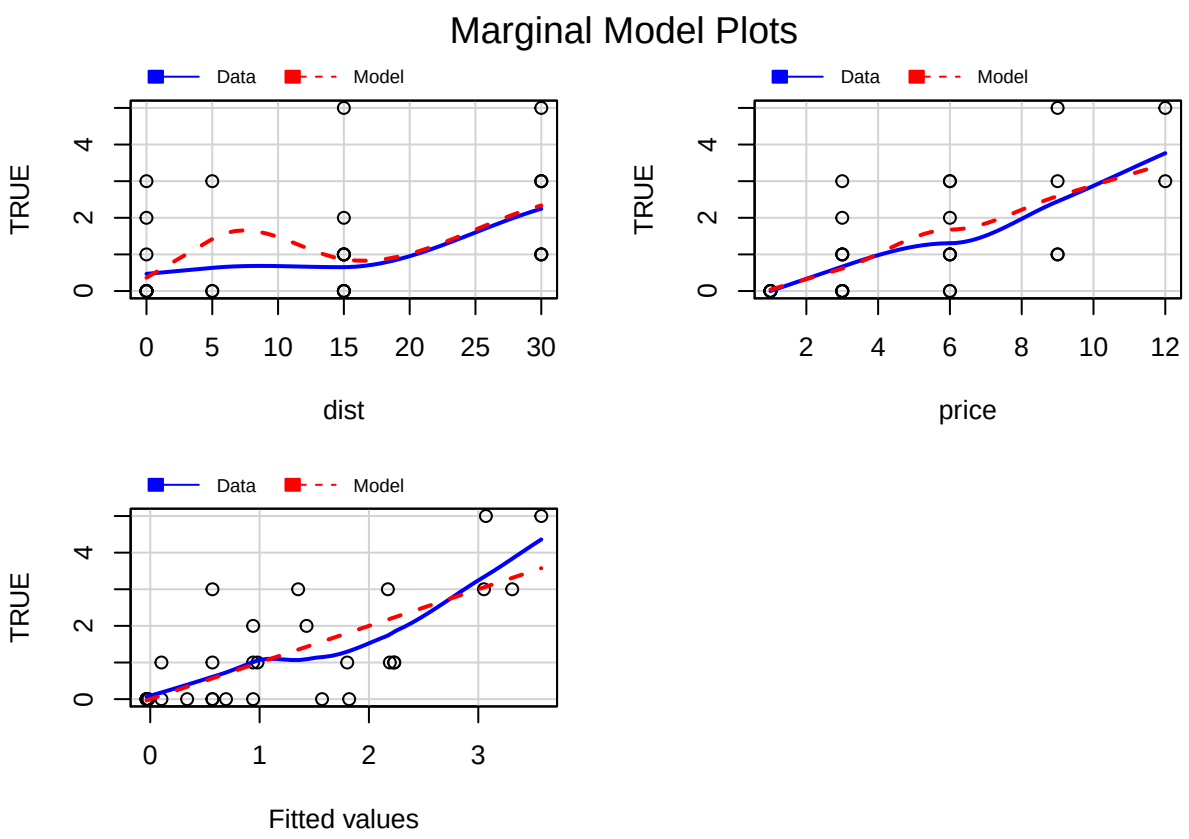
```
## The following object is masked from 'package:rstanarm':
##
##     logit
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

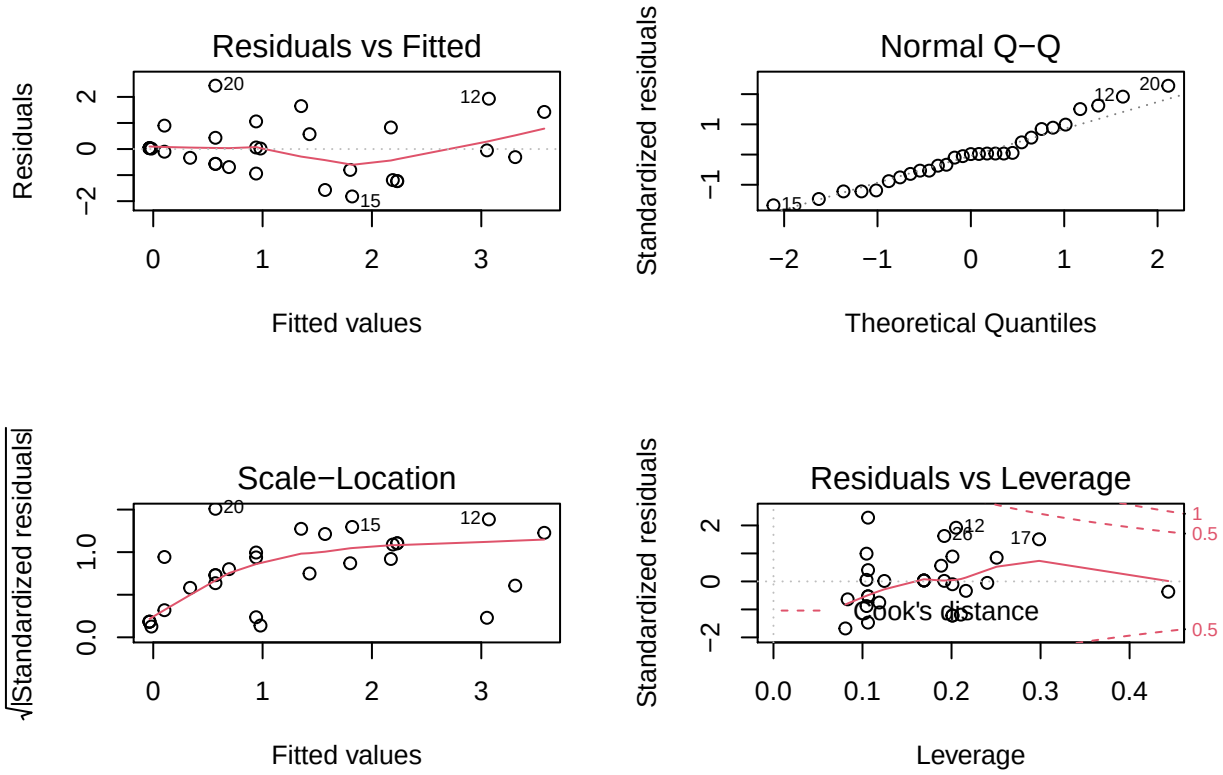
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
mmps(fit1, sd = FALSE,
    smooth=TRUE, key=TRUE)
```

```
## Warning in mmps(fit1, sd = FALSE, smooth = TRUE, key = TRUE): Interactions and/
## or factors skipped
```




```
#diagnostic plot
par(mfrow = c(2, 2))
plot(fit1)
```



By comparing the final model and the original model, we can easily conclude that the final model is a better model than the original one, since it has higher adj.R square, lower AIC, etc,. In addition, 2 lines on the marginal plot are approximately to the same line, the points on the residual plot are quite random around 0 and Q-Q Plot is approximate linear, indicating that the model is not bad.

Inference (10pts)

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

```
#compare groups
study_room%>%
  group_by(group)%>%
  summarise(means=mean(freq, na.rm=T), sds=sd(freq, na.rm=T), n=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

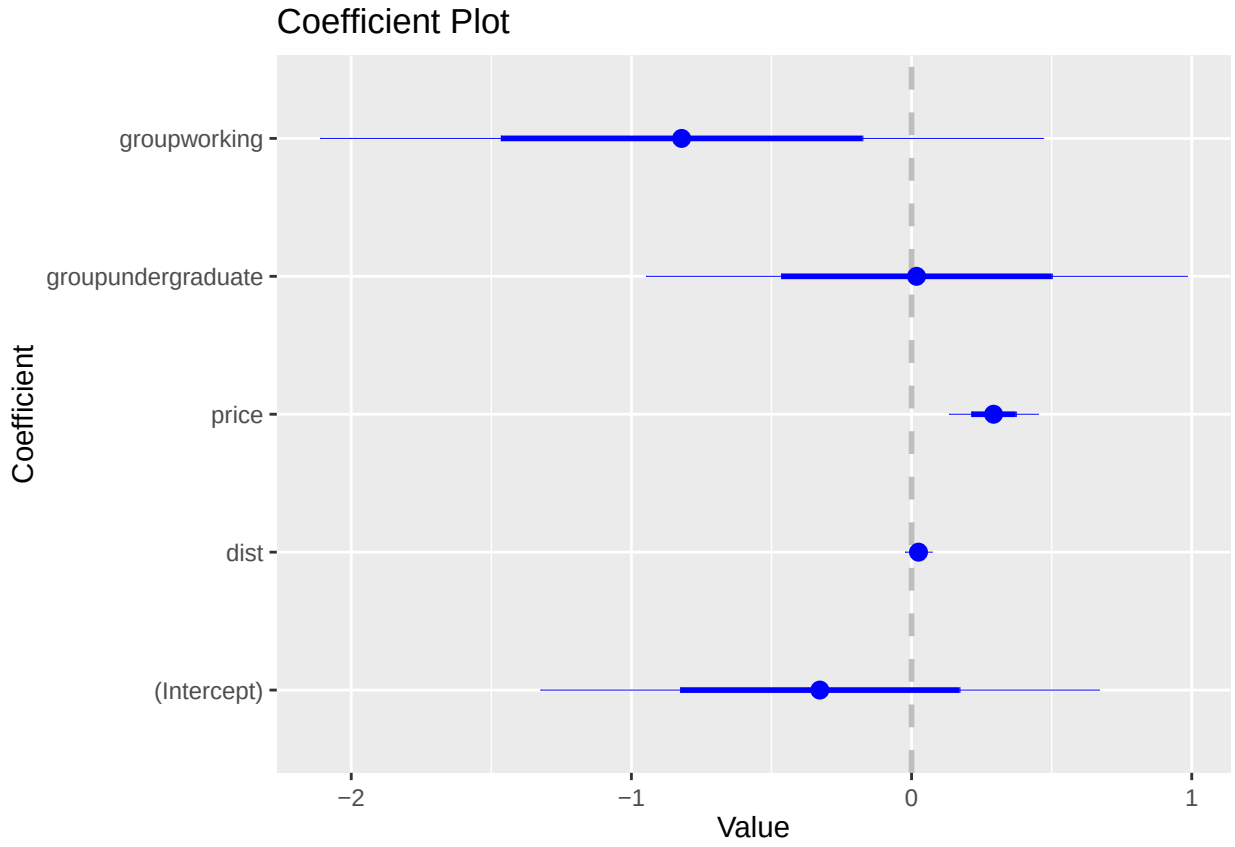
```
## # A tibble: 3 x 4
##   group         means   sds     n
##   <fct>         <dbl> <dbl> <int>
## 1 graduate       1.33 1.41      9
## 2 undergraduate  1.29 1.82     14
## 3 working        1.17 0.983     6
```

```
#CI
confint(fit_12)
```

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

##                           2.5 %     97.5 %
## (Intercept)         -1.30394193 0.64920915
## dist                -0.02203495 0.07162331
## price                0.13769416 0.44792678
## groupundergraduate -0.92847314 0.96391642
## groupworking        -2.08510635 0.44370864
```

```
coefplot(fit_12,frame.plot=TRUE)
```



From the result we can see that means for different groups only have tiny differences.

All variables have the confidence intervals across 0 expect price, which means only the variable price is statistically significant on 95% confidence level.

Discussion (10pts)

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

In order to understand what factors will influence the frequency, I chose general linear regression model with price, dist and group as predictors to do the analysis. Although, price is the only significant variable in this model, I still chose to include other 2 variables because they might have some association with the outcome in the real life. From the model result, the coefficient of price shows that if the price increase by 1, the frequency will increase by 0.29, which is not really understandable. It can be easier to understand in the opposite direction which is that people who plan to go to the study room several times in a week might be willing to pay more per hour.

From the group comparison analysis, we can conclude that, there is no difference for the number of times a person will go to the study room in a week across each group. It is not a result what I expect to get, and I will collect more data to do further analysis.

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.

My concerns are that the result for the whole analysis is not what I expect. For example, I think distance might be a significant predictor for the number of times people would go to the study room per week. And the result for the relationship between price and frequency is also weird, because I do not think people are willing to spend that much money for one hour study in the study room. These problems might cause by the sample size and data collection bias. Because there are only 29 data points in total for 3 groups of people, it might be too small to do the analysis. In addition, I collect the data from the my friends and their friends, which means those people might have similar living environment, education background, age range and consumption level. Thus, the data might not be random sampled.

In the future, I will collect more data from wide range of people and try to find more related variables to get a better regression model and understandable result.

Comments or questions

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

I have a question which is that I feel my whole analysis is weird because it seems that all predictors have no direct relationship with the outcome. Is that reasonable to do the regression analysis based on this data?