IODS course project

Rong Guang

Contents

Introduction to Open Data Science - Course Project

```
# This is a so-called "R chunk" where you can write R code.

date()
```

```
## [1] "Thu Nov 17 09:22:04 2022"
```

Chapter 1: Start me up

1. Thoughts about the course

1.1 My feeling about the course

I am feeling very good, since I finally start using GitHub from scratch. **All my colleagues** are using GitHub for collaboration, but I do not even have any idea what it is. This course is very timely. Hopefully after this course, I can join their collaboration loop effectively.

1.2 What do I expect to learn

I do not have any particular expectation in the statistical knowledge I am going to learn, because anything about statistics is interesting to me. Actually, I hope in the course I could have chance of exercising **GitHub collaboration in practical manner**. For example, my colleagues always mention the creation of some 'branches' of some "root" data sets, which I still do not understand very clearly what is it.

2. reflect on my learning with the *R for Health Data Science book*

2.1 How did it work as a "crash course" on modern R tools and using RStudio?

Luckily, I just finished taking Prof. Kimmo's another course "Quantitative Research Skills", where I got chance to walk through several chapters of *R for Health Data Science book*. This way I suppose I could make some comments on the book retrospectively. I would say this is the best book for novice R users to start off. The author did the instruction from an industry perspective (health science), which avoid getting too deep in math but focuses on the R as a statistical tool for solving industry problem itself. When I tried to start with R years ago, one of the biggest barrier is for every statistical feature I want to realize, R always present a multitude of pathways to realize it (for example, to compute a new variable for a data set, you can you ".[,c] <-", or ".$c<-", or " %>% " + "mutate"), other source of material/textbook always tried to feed you with all these possibilities, which makes things complicated. This book just doesn't. For each function to execute, it only show you one approach. This is just what most non-statistics-majors want!

2.2 Which were your favorite topics?

Data wrangling, definitely (although there is no specific chapter for it, but it scatters throughout the whole book). Most descriptive, inferential and exploratory statistics are so easy to realize in R. By the merit of R, for these purposes, all I need is first, know what I am going to do (base on my stat knowledge), and then search in Google about which package(s) are for doing that, and finally install it, read the instruction&example and then do it. Very mechanical. There is never a big challenge. To me, data wrangling (more than dozens of variable) requires some real effort, and hence more interesting. And there is never an end. Today I realize a wrangling task in three lines of codes, maybe tomorrow I could come up with or stumble over a way that only requires two, or even one. This process is super exciting.

2.3 Which topics were most difficult?

Data wrangling, of course according to my story above, followed by logistic regression. For logistic regression, I have several big unsolved problem on mind. They are: **a.** considering there is not an alternative of linear regression's r square for logistic regression, how do I evaluate the variance that my model could explain? **b.** after modeling additive effect, model modification is recommended to follow. It is not difficult to create a main effect plot to observe candidate multiplicative variables if **y** is continuous. But is it also applicable to a binary **y** ? However, I suppose this is mainly due to my lack of relevant stat knowledge, having nothing to do with R language.

3. My GitHub Repo

Please find it Here: https://github.com/rg450318262/IODS-project

4. Token

Since I have generated the token before the new task is published (That is the only way I get my machine connected to GitHub), I will not generate it again in case any overlapping problems.

This is the end of chapter 1

Chapter 2: Regression and model validation

Describe the work you have done this week and summarize your learning.

- Describe your work and results clearly.
- Assume the reader has an introductory course level understanding of writing and reading R code as well as statistical methods.
- Assume the reader has no previous knowledge of your data or the more advanced methods you are using.

```
date()
```

```
## [1] "Thu Nov 17 09:22:04 2022"
```

1 Preparing

1.1 read the data set

```
library(tidyverse)
learn <- read_csv(file = "data/learning2014.csv")
```

1.2 Code categorical data

```
learn <- learn %>%
  mutate(gender = gender %>%
           factor() %>%
           fct_recode("Female" = "F",
                      "Male" = "M"))
```

1.3 explore the data set

```
#explore dimensions
dim(learn)
```

```
## [1] 166   7
```

The data set has 166 observations of 7 variables.

```
#explore structure
str(learn)
```

```
## tibble [166 x 7] (S3: tbl_df/tbl/data.frame)
##  $ gender  : Factor w/ 2 levels "Female","Male": 1 2 1 2 2 1 2 1 2 1 ...
##  $ age     : num [1:166] 53 55 49 53 49 38 50 37 37 42 ...
```

```
##  $ attitude: num [1:166] 3.7 3.1 2.5 3.5 3.7 3.8 3.5 2.9 3.8 2.1 ...
##  $ deep    : num [1:166] 3.58 2.92 3.5 3.5 3.67 ...
##  $ stra    : num [1:166] 3.38 2.75 3.62 3.12 3.62 ...
##  $ surf    : num [1:166] 2.58 3.17 2.25 2.25 2.83 ...
##  $ points  : num [1:166] 25 12 24 10 22 21 21 31 24 26 ...
```

The data set has six numeric (integer type) variables and one categorical (binary) variable.

1.4 describe the variables

Under the funding of *International Survey of Approaches to Learning*, 183 Finnish students who took the course "Introduction to Social Statistics" during 2014 fall participated in a survey about their learning, resulting in a data set with 32 variables and 166 observations (due to missing data points, the sample size is smaller than 183). The current data set for analysis is a convenient subset of it. It includes variables about the participants' demographic characteristics such as age and sex, as well as the final points they got for certain exam (could possibly be statistics). It also includes 4 psychological dimensions including study attitude (reflecting their motivation to the subject), deep learning score (reflecting how well their learning style fits into the deep learning type), surface learning score (reflecting how well their learning style fits into the surface learning type) and strategy learning score (reflecting how well their learning style fits into the strategic learning type).

1.4.1 describe the coninuous variable

```
library(tidyverse)
library(finalfit) # a package introduced in RHDS book.
                  #The "gg_glimpse" function could give nice descriptive
                  #statistics for both types of variables.
library(DT) # show table in a html-based neat view.
ff_glimpse(learn)$Continuous %>% datatable() # descriptive statistics for
```

Show 10 entries Search:

	label	var_type	n	missing_n	missing_percent	mean	sd	min	quartile_25	median	quartile_75	max
age	age	<dbl>	166	0	0.0	25.5	7.8	17.0	21.0	22.0	27.0	55.0
attitude	attitude	<dbl>	166	0	0.0	3.1	0.7	1.4	2.6	3.2	3.7	5.0
deep	deep	<dbl>	166	0	0.0	3.7	0.6	1.6	3.3	3.7	4.1	4.9
stra	stra	<dbl>	166	0	0.0	3.1	0.8	1.2	2.6	3.2	3.6	5.0
surf	surf	<dbl>	166	0	0.0	2.8	0.5	1.6	2.4	2.8	3.2	4.3
points	points	<dbl>	166	0	0.0	22.7	5.9	7.0	19.0	23.0	27.8	33.0

Showing 1 to 6 of 6 entries Previous 1 Next

```
#categorical data shown in html-based data table view.
```

According to their distribution shapes visualized in section 1.5 (next section), non-normally distributed variables were reported as median and Q1-Q3; roughly normal variables will be reported as mean±sd.

The age of the participants was between 17 and 55 years old (median: 22; Q1-Q3:21,27 years old). Their exam points were 22.7±5.9. Their deep learning scores were 3.7±0.6. Their surface learning scores were 2.8±0.5. And their strategic learning scores were 3.1±0.8.

1.4.2 describe the categorical variable

```
ff_glimpse(learn)$Categorical %>% datatable() # descriptive statistics for categorical data shown in ht
```

Show 10 entries Search:

	label	var_type	n	missing_n	missing_percent	levels_n	levels	levels_count	levels_percent
gender	gender	<fct>	166	0	0.0	2	"Female", "Male"	110, 56	66, 34

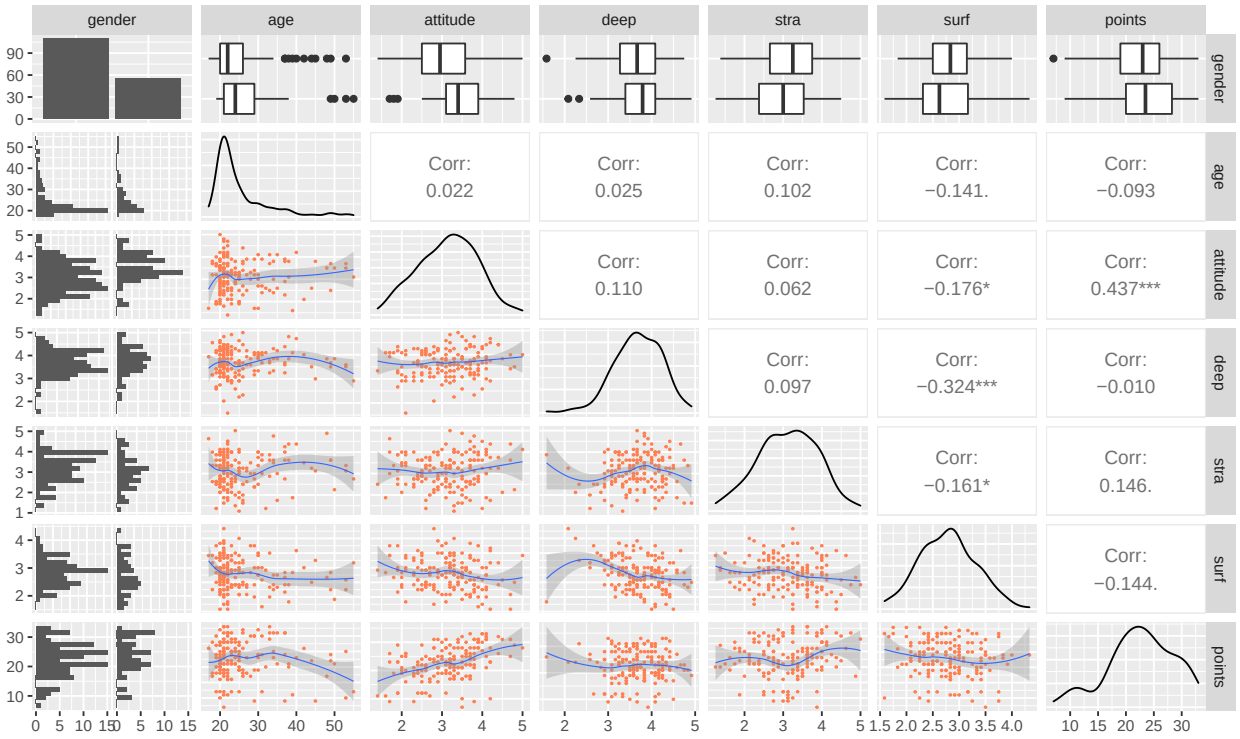
Showing 1 to 1 of 1 entries Previous 1 Next

Among the 166 participants, 110 (66%) were female and 56 (34%) were male. According to a 2021 statistics, Finnish universities had a male:female ratio of 1:1.2, indicating female in our sample is over-represented.

1.5 visualize the data set

```
library(GGally)
library(ggplot2)
library(tidyverse)
#create a self-defined function so that correlation matrix produced by ggpairs could
#show LOWESS smoothing with scatter plot.
my_fn <- function(data, mapping, method="loess", ...){ #require the input of two
       #arguments: data and mapping; arguments method is set to be "Loess"
       #for more information about using Loess to check function form, please
       #go to (https://thestatsgeek.com/2014/09/13/checking-functional-form-in-logistic-regression-usin
      p <- ggplot(data = data, mapping = mapping) +  #call ggplot function
```

```
      geom_point(size = 0.3, color = "coral") +   #call point graph, reduce the
        #size, turn color to coral, for better visualization
      geom_smooth(size = 0.3, method=method, ...) # fit Loess regression
      p  #print the result
}
# create an plot matrix with ggpairs()
ggpairs(learn,
        lower= list(combo = wrap("facethist", bins = 20),
                    continuous = my_fn) #call self-defined function "my_fn"
        )
```



According to the visualization, it is found that the distribution of age is right-skewed; other numeric variables, though with slight skewness, can be roughly treated as normal distribution. All of the values of numeric variables did not show any remarkable difference between males and females. Variables "points", "attitude" and "deep" have 1 to 3 out-liers, respectively, and age has quite a number of out-liers. By examining the raw data, no evidence of mistaken record was detected. These out-liers were thus kept for analysis. Using variable "points" as reference, variable "attitude" showed a significant linear correlation(r=0.437). Although the correlation coefficient between age and points is only -0.093, the LOESS smoothing has shown there might be a quadratic relationship between them.

2. Fitting the model

2.1 variable selection

According to the visualization in section 1.4, age (as polynomial form due to its non-linearity with the outcome) and attitude were used to fit the model that predicts exam points. Although no noticeable effect of gender was observed, it also entered the model for it being adopted as an important factor for predicting exam points in a multitude of publications.

2.2 fitting

```
fit1 <- learn %>%  #using attitude, the polynomial age and gender to predict exam points
  lm(points ~ attitude  + poly(age, 2, raw =T) + gender, data = .) #ploy() is to
    #include 2nd order function form, where "2" means the order
summary(fit1) # summarize the results
```

```
##
## Call:
## lm(formula = points ~ attitude + poly(age, 2, raw = T) + gender,
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1258  -3.1673   0.5261   3.6243  10.3486
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -4.246278   5.618253  -0.756 0.450873
## attitude                3.774918   0.576603   6.547  7.5e-10 ***
## poly(age, 2, raw = T)1  1.094988   0.345856   3.166 0.001849 **
## poly(age, 2, raw = T)2 -0.017766   0.005188  -3.424 0.000782 ***
## genderMale             -0.615038   0.894211  -0.688 0.492569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.148 on 161 degrees of freedom
## Multiple R-squared:  0.256,  Adjusted R-squared:  0.2375
## F-statistic: 13.85 on 4 and 161 DF,  p-value: 9.947e-10
```

The results showed that except for gender, other variables all had significant predicting effect (all p<0.01). Besides, F-statistics ($p < 0.01$) had rejected the null that the response variable cannot be represented as a function of any of the predictor variables, indicating the model is valid. Adjusted R-squared showed that the model explained 23.75% of the variance of exam points. However, in the next step I further reduced the model complexity by removing insignificant variable base on the rule of parsimony.

2.3 removing insignifiant predictor

The model was fit again by removing gender.

```
fit2 <- learn %>%
  lm(points ~ attitude  + poly(age, 2, raw =T), data = .)#ploy() is to
    #include 2nd order function form, where "2" means the order
summary(fit2)
```

```
##
## Call:
## lm(formula = points ~ attitude + poly(age, 2, raw = T), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -16.904  -3.290   0.293   3.594  10.342
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -3.651986   5.542377  -0.659 0.510882
## attitude                3.656205   0.549269   6.656 4.13e-10 ***
## poly(age, 2, raw = T)1  1.068947   0.343218   3.114 0.002180 **
## poly(age, 2, raw = T)2 -0.017435   0.005158  -3.380 0.000907 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.139 on 162 degrees of freedom
## Multiple R-squared:  0.2538, Adjusted R-squared:   0.24
## F-statistic: 18.36 on 3 and 162 DF,  p-value: 2.634e-10
```

The results showed that all variables had significant predicting effect (all $p<0.01$). Besides, F-statistics ($p < 0.01$) had rejected the null that the response variable can not be represented as a function of any of the predictor variables, indicating the model is valid. Adjusted R-squared showed that the model explained 24% of the variance of exam points, which slightly outperformed the previous model. I took this model as the final model for model diagnostics.

In the final model, variable "attitude" has a coefficient of 3.65, indicating for 1 unit of attitude increase, the exam points is expected to increase 3.65, after controlling for other factors. The first order term of age has an coefficient estimate of 1.06, indicating that, overall, for every 1 unit increase of age, the exam points is expected to increase 1.06, after controlling for other factors. For the second order term of age, an estimated coefficient of -0.017 indicated for different value ranges of age, the effect on exam points might be significantly different. This auto-interaction might lead to -0.017 decrease in exam points across these ranges, after controlling for the other factors.

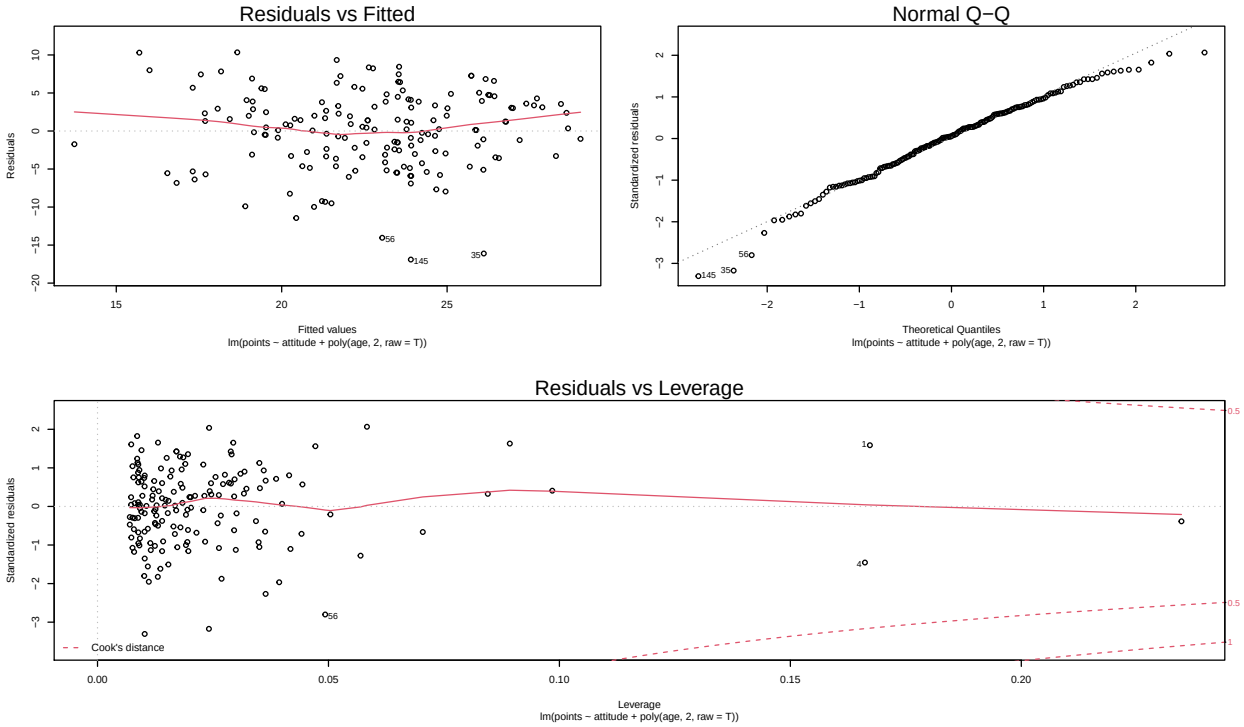
The practical explanation for these coefficient might be *a.* attitude reflects the motivation of study and higher motivation will lead to better exam performance; *b.* statistics requires quite a bit of domain knowledge (economics, health, psychology..), logic reasoning and math foundations. Older students might have advantage in these aspects. *c.* However, this advantage will see a ceiling effect at around 30 years old (according to the graph above), and due to the aging and family burden, students over 30 years old might start to become less and less competitive in stat learning over time.

3. Model diagnostic

In model diagnostic, some of the assumptions (linearity and normality) of linear regression were checked. Besides, observations with high influence will be examined in this section.

3.1 diagnostic plots

```
par(cex = 0.5,fig=c(0,0.5,0.5,1)) #set the coordinate of picture 1
plot(fit2, which = 1)  #plot diagnostic picture 1

par(cex = 0.5,fig=c(0.5,1,0.5,1), new=TRUE)#set the coordinate of picture 2
plot(fit2, which = 2) #plot diagnostic picture 2

par(cex = 0.5, fig=c(0,1,0,0.5), new=TRUE)#set the coordinate of picture 3
plot(fit2, which = 5)#plot diagnostic picture 3
```



Residuals vs fitted plot (upper left) showed the data points are randomly scattered around the dotted line of y = 0, and the fitted line (red) is roughly horizontal without distinct patterns or trends, indicating a linear relationship. The linearity assumption of linear regression is met.

The QQ plot (upper right) showed most of the points plotted on the graph lies on the dashed straight line, except for the lower and upper ends, where some points deviated from the line, indicating the distribution might be slightly skewed. Considering the fact that in large sample size the assumption of linearity is almost never perfectly met, I see the assumption of normality as being approximately met.

3.2 other linear model assumptions

The assumption of independence requires no relation between the different observations. I do not have information of how this study was designed, hence not being able to make any conclusion. However, I could imagine how hard it took to meet it here, since including students taking courses from different lecturers or different lecturer groups would lead to violation of it. On the other hand, if the results were from students of one same lecturer (or lecturer group), it might take several semesters to collect such a large sample or might take students from several different classes/majors in one semester, either way the assumption was violated.

Homoscedasticity is another assumption to check. However, considering it is better evaluated by fitted values against root of standardized residuals (the #3 in plot() function), which is not required to produce in the current assignment, I did not further dig into it. By looking into its rough substitute plot "residual vs fitted" (upper left, above), no obvious heteroscedasticity was detected.

3.3 influential observations

Influential observations were shown in the bottom plot, where the red dashed line indicate cook's distance. Cook's distance is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis. It measures the effect of deleting the observation for each given observation. In the plot, points, if there is any, outside the red dashed line are believed to have high influence. The graph for current

model showed no points outside the line. The plot also showed the case numbers of 3 data points with the largest cook's distances, which are #1, #4 and #56. However, there are also other rules of thumbs for the cutoff, which are stricter. They include using an absolute value of 1, or using 4/n (n is sample size), or using 4×(the mean of the cooks distance for the whole sample). I did not further report them since it is not required in this assignment. I did it somewhere else for fun. If you are interested, please go to a r markdown file named "Supplement_Codes.html" or "Supplement_Codes.Rmd" under my "IODS-project" folder.

of cook's distance,

where x is the index number of our sample and y is the cook's distance score for each observation. This is to evaluate, if there's any, the data points being tremendously influential to the coefficient estimate. Cook's distance is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis. It measures the effect of deleting the observation for each given observation. There is no consensus on the cutoff for being influential using this indicator. The rules of thumbs include using an absolute value of 1, or using 4/n (n is sample size), or using 4×(the mean of the cooks distance for the whole sample). The plot showed the case numbers of 3 data points with the largest cook's distances, which are #1, #4 and #56. I did not further report them since it is not required in this assignment. I did it somewhere else for fun. If you are interested, please go to a r markdown file named "Supplement_Codes.html" or "Supplement_Codes.Rmd" under my "IODS-project" folder.

This is the end of chapter 2
****

Here we go again...

a test to see if it works

Chapter 3: Logistic regression

1 Preparing

1.1 read the data set

```
library(tidyverse)
alc <- read_csv(file = "data/alc.csv")
```

1.2 check the data set

```
glimpse(alc)
```

```
## Rows: 370
## Columns: 35
## $ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP",~
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F",~
## $ age        <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15,~
## $ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",~
## $ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "LE~
## $ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T",~
```

```
## $ Medu       <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,~
## $ Fedu       <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3,~
## $ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "servic~
## $ Fjob       <chr> "teacher", "other", "other", "services", "other", "other", ~
## $ reason     <chr> "course", "course", "other", "home", "home", "reputation", ~
## $ guardian   <chr> "mother", "father", "mother", "mother", "father", "mother",~
## $ traveltime <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,~
## $ studytime  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,~
## $ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no", "n~
## $ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes",~
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye~
## $ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "yes~
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye~
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes",~
## $ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "no"~
## $ famrel     <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,~
## $ freetime   <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1,~
## $ goout      <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3,~
## $ Dalc       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
## $ Walc       <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3,~
## $ health     <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5,~
## $ failures   <dbl> 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0,~
## $ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes", ~
## $ absences   <dbl> 5, 3, 8, 1, 2, 8, 0, 4, 0, 0, 1, 2, 1, 1, 0, 5, 8, 3, 9, 5,~
## $ G1         <dbl> 2, 7, 10, 14, 8, 14, 12, 8, 16, 13, 12, 10, 13, 11, 14, 16,~
## $ G2         <dbl> 8, 8, 10, 14, 12, 14, 12, 9, 17, 14, 11, 12, 14, 11, 15, 16~
## $ G3         <dbl> 8, 8, 11, 14, 12, 14, 12, 10, 18, 14, 12, 12, 13, 12, 16, 1~
## $ alc_use    <dbl> 1.0, 1.0, 2.5, 1.0, 1.5, 1.5, 1.0, 1.0, 1.0, 1.0, 1.5, 1.0,~
## $ high_use   <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
```

2 Hypothesis

2.1 introduction

Despite the health risk and public harm associated with heavy drinking, alcohol is the most commonly used substance in developed countries (Flor and Gakidou 2020). A large scale longitudinal study has identified Finland as the only Nordic country whose alcohol-attributable harms has increased (Room et al. 2013). Given that alcohol consumption typically starts in late adolescence or early adulthood (Lees et al. 2020), measures to detect alcohol misuse among young people, especially college students, should be a top public health priority. Identifying a comprehensive set of early life factors associated with college students' alcohol use disorders could be an important starting point.

2.2 literature review

College students typically spend a tremendous amount of time with their family members, emphasizing the influence of family quality on any type of habit acquisitions. Evidence has shown family relationship quality is strongly correlated with early alcohol use (Kelly et al. 2011; Brody and Forehand 1993), and the effect is interactive with gender (Kelly et al. 2011). Since studying also comprises an important part of college life, it is important to evaluate how college life and alcohol use interact with each other. An 21 year follow-up of 3,478 Australian since they were child has found level of academic performance predicts their drinking problems, independently of a selected group of individual and family con-founders (Hayatbakhsh et al. 2011).

College students start to build up their social networks. An increased exposure to social communications is reasonably expected among them, which might incur alcohol involvement. A survey has found typical social drinking contexts were associated with men's average daily number of drinks and frequency of drunkenness, indicating social communications, interacted with gender, might have influence on college students' alcohol usage(Senchak, Leonard, and Greene 1998).

2.3 Hypothesis

According to the literature review, 4 potential early-life factors is identified to predict excessive alcohol usage among college students. They are *a.* family relationship quality (interactive with gender); *b.* school performance; *c.* social communication (interactive with gender). I herein proposed a 3-factor alcohol high-use model for college students and test it using a secondary data set collected for other purposes.

In the data set, variables including gender, quality of family relationships ("famrel"), number of school absences ("absences"), weekly study time ("studytime") and frequency of going out with friends ("goout") could be candidate indicators for the current model. The variable "gender" and "famrel"'s relevance to the predictors are self-explanatory. School performance includes college students' in-class and off-class performance, which could be reflected by variables "absences" and "studytime", respectively. Variable "goout" captures the involvement of social activity, which is a good indicator to social communication. Note that base on the well-reported evidence introduced above, gender will not enter the model independently. Instead, it will comprise interaction terms with family relationship quality and social communication, respectively, and then enter the model.

3 Data exploration

3.1 exploring the association between faimly relationship quality and alcohol high-use

The variable "famrel" in original data set elicited quality of family relationships (numeric: from 1 - very bad to 5 - excellent). In the current analysis, it is selected as a candidate predictor for the model to reflect the same idea–quality of family relationship.

3.1.1 numerically explore the association

```
alc %>% count(high_use, famrel)
```

```
## # A tibble: 10 x 3
##    high_use famrel     n
##    <lgl>     <dbl> <int>
##  1 FALSE         1     6
##  2 FALSE         2     9
##  3 FALSE         3    39
##  4 FALSE         4   128
##  5 FALSE         5    77
##  6 TRUE          1     2
##  7 TRUE          2     9
##  8 TRUE          3    25
##  9 TRUE          4    52
## 10 TRUE          5    23
```

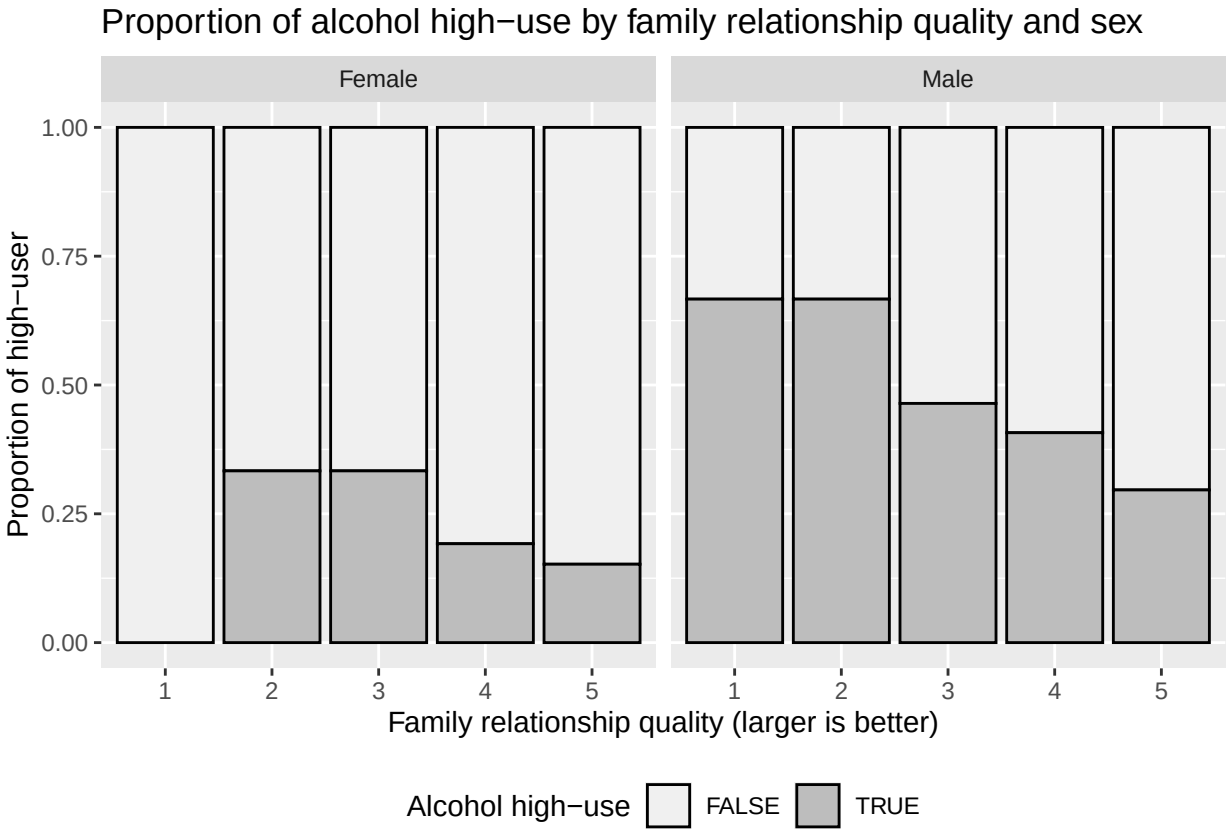
```
?count
```

```
## Help on topic 'count' was found in the following packages:
##
##   Package               Library
##   dplyr                 /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##   plyr                  /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##
##
## Using the first match ...
```

It is found the absolute sample of participants with very bad (level 1) and/or bad (level 2) family quality is very small in number (n = 8). Caution should be taken about the potential large error.

3.1.2 graphically explore the association

```
sex.labs <- c("Female", "Male")
names(sex.labs) <- c("F", "M")
p1 <- alc %>%
  ggplot(aes(x = factor(famrel), fill = high_use)) +
  geom_bar(position = "fill", color ="black") +
  facet_wrap(~sex,
             labeller = labeller(sex = sex.labs)) +
  labs(x = "Family relationship quality (larger is better)",
       y = "Proportion of high-user",
       title =
         "Proportion of alcohol high-use by family relationship quality and sex")+
  theme(legend.position = "bottom")+
  guides(fill=guide_legend(title = "Alcohol high-use"))+
  scale_fill_discrete(labels = c("FALSE" = "Non-high-user",
                                 "TRUE" = "high-user"))+
  scale_fill_brewer(palette = "Greys")

p1
```



The value of the variable "famrel" includes numbers from 1 - very bad to 5 - excellent. In the current study, I presume that the intervals between each consecutive pair of value is consistent, and hence see it as a numeric variable.

According to the bar plot of proportion, the hypothesis of using the current variable in model fitting is validated. It is found that with the increasing of family relationship quality, the proportion of alcohol high-use decreases, except for female from a very bad (level 1) relationship family, which had a proportion of high-users at zero. However, this low proportion suffers from a risk of error due to the small sample in the level (n = 8). The result should be interpreted with caution.

To facilitate understanding, the variable's name will be changed to family.quality according to the hypothesis.

3.1.3 re-code the variable of family relationship quality

```
alc <- alc %>%
  mutate(family.quality = famrel)
```

3.2 exploring the association between school performance (absences) and alcohol high-use

The variable "studytime" in original data set captured participants' weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours). The variable "absences" in original data set captured participants' number of school absences (numeric: from 0 to 93). It is presumed in the current analysis that they reflect off-class and in-class school performance, respectively, and hence they are selected as candidate predictors.

3.2.1 exploring the association between in-class performance and alcohol high-use

3.2.1.1 numerically explore the association

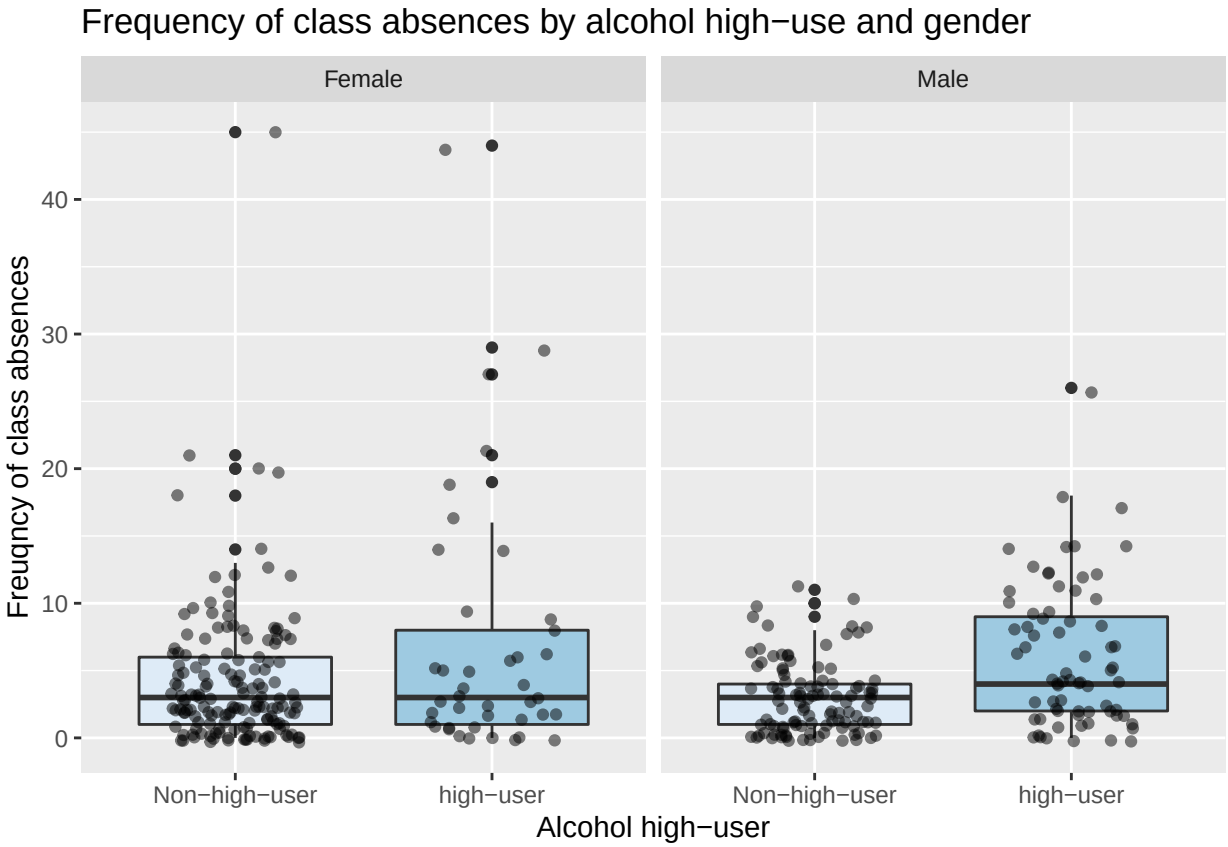
```
alc %>% group_by(high_use, sex) %>%
  summarise(mean = mean(absences),
            sd = sd(absences),
            sampleSize = n())
```

```
## # A tibble: 4 x 5
## # Groups:   high_use [2]
##   high_use sex    mean    sd sampleSize
##   <lgl>    <chr> <dbl> <dbl>      <int>
## 1 FALSE    F      4.25  5.29        154
## 2 FALSE    M      2.91  2.67        105
## 3 TRUE     F      6.85  9.40         41
## 4 TRUE     M      6.1   5.29         70
```

From the table, it is found that the frequency of class absences differed greatly between alcohol high-users and non-high-users, indicating its validity in entering the model.

3.2.1.2 graphically explore the association

```
p2 <- alc %>%
  ggplot(aes(x = high_use, y = absences, fill = high_use)) +
  geom_boxplot() +
  geom_jitter(width=0.25, alpha=0.5)+
  facet_wrap(~sex, labeller = labeller(sex = sex.labs)) +
  scale_fill_brewer(palette = "Blues")+
  labs(x = "Alcohol high-user",
       y = "Freuqncy of class absences",
       title =
         "Frequency of class absences by alcohol high-use and gender")+
  theme(legend.position = "none")+
  scale_x_discrete(labels = c("FALSE" = "Non-high-user",
                              "TRUE" = "high-user"))
p2
```



The box plot showed similar information to the previous table. No noticeable difference in proportions of absences can be observed between genders, and hence their interaction would not be considered in fitting the model.

###3.2.1.3 rename the variable

To facilitate understanding, the name of variable "absences" will be changed to in.class.performance according to the hypothesis of current study.

```
alc <- alc %>%
  mutate(in.class.performance = absences)
```

3.2.2 exploring the association between off-class performance (study time) and alcohol high-use

3.2.2.1 numerically explore the association

```
alc %>% count(high_use, studytime)
```

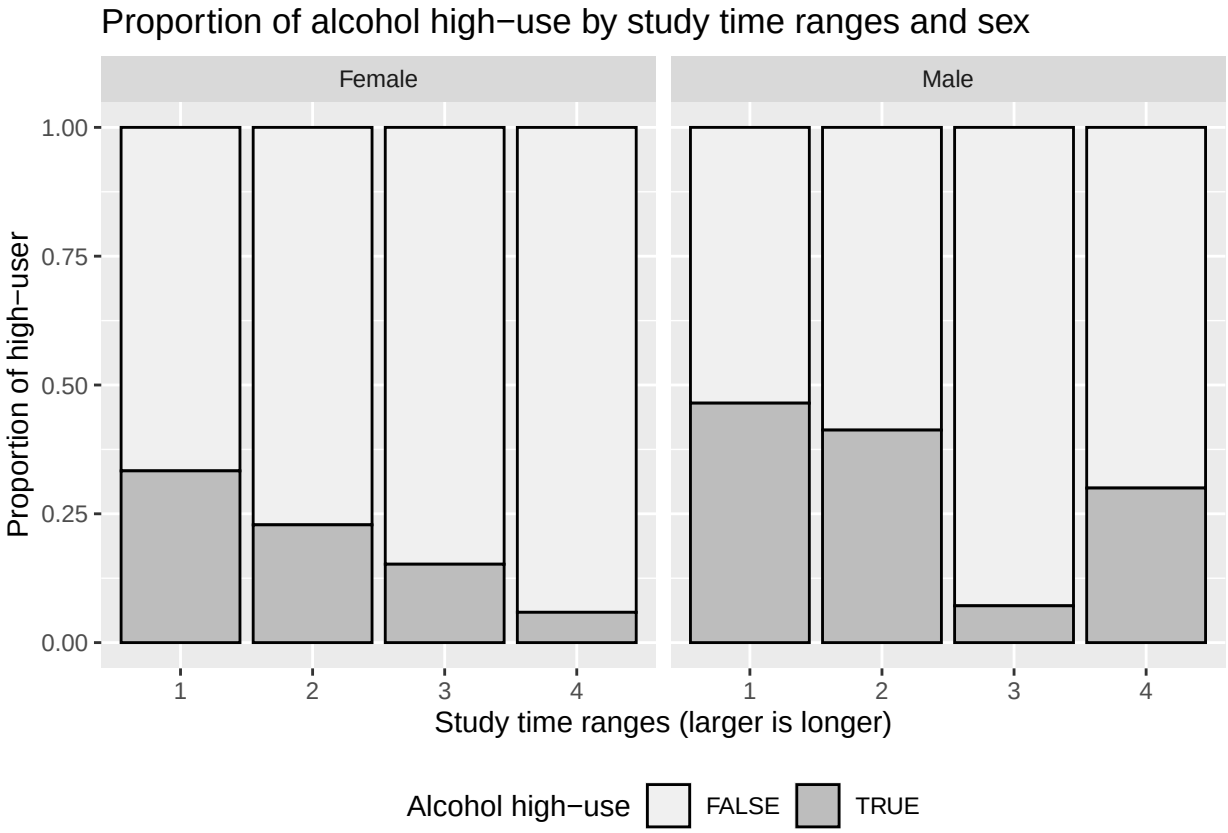
```
## # A tibble: 8 x 3
##   high_use studytime     n
##   <lgl>        <dbl> <int>
## 1 FALSE            1    56
## 2 FALSE            2   128
## 3 FALSE            3    52
## 4 FALSE            4    23
```

```
## 5 TRUE             1    42
## 6 TRUE             2    57
## 7 TRUE             3     8
## 8 TRUE             4     4
```

From the table, it is found the sample of participants with long and very long (level 4 and 5) study time in alcohol high-user group is very small in number (n = 12). Caution should be taken about the potential large error.

3.2.2.2 graphically explore the association

```
p3 <- alc %>%
  ggplot(aes(x = factor(studytime), fill = high_use)) +
  geom_bar(position = "fill", color = "black") +
  facet_wrap(~sex,
             labeller = labeller(sex = sex.labs)) +
  labs(x = "Study time ranges (larger is longer)",
       y = "Proportion of high-user",
       title = "Proportion of alcohol high-use by study time ranges and sex")+
  theme(legend.position = "bottom")+
  guides(fill=guide_legend(title = "Alcohol high-use"))+
  scale_fill_discrete(labels = c("FALSE" = "Non-high-user",
                                 "TRUE" = "high-user"))+
  scale_fill_brewer(palette = "Greys")
p3
```



The levels of study time ranges include 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours. Their intervals are inconsistent, and hence it is not appropriate to enter a model as numeric variables, and it will be transformed into a categorical variable.

According to the bar plot of proportion, it is found that with the increasing of study time, the proportion of alcohol high-use decreases, indicating its validity in entering the model. However, male with long study time (level 3) is an exception, which had the lowest proportion of high-users across the levels. Notably, this low proportion suffers from a risk of error due to the small sample in the level. To address the risk of error, the levels of study time ranges will be re-coded as Long study(original level 3 + original level 4), Moderate study(original level 2) and Light study (original level 1). Besides, no noticeable difference in proportions of study length can be observed between genders, and hence their interaction would not be considered in fitting the model.

To facilitate understanding, the name of variable "studytime" will be changed to off.class.performance according to the hypothesis.

3.2.3 re-code the variable of study length

```
alc <- alc %>%
  mutate(off.class.performance =
           case_when(studytime == 3 |studytime == 4~"Long study",
                     studytime == 2~"Moderate study",
                     studytime == 1~"Light study") %>%
           factor(levels = c("Light study", "Moderate study", "Long study")))
```

3.3 exploring the association between social communication frequency and alcohol high-use

The variable "goout" in original data set captured participants' frequency of going out with friends (numeric: from 1 - very low to 5 - very high). It is presumed in the current analysis that it reflects social involvement, and hence it is selected as a candidate predictor.

3.3.1 numerically explore the association

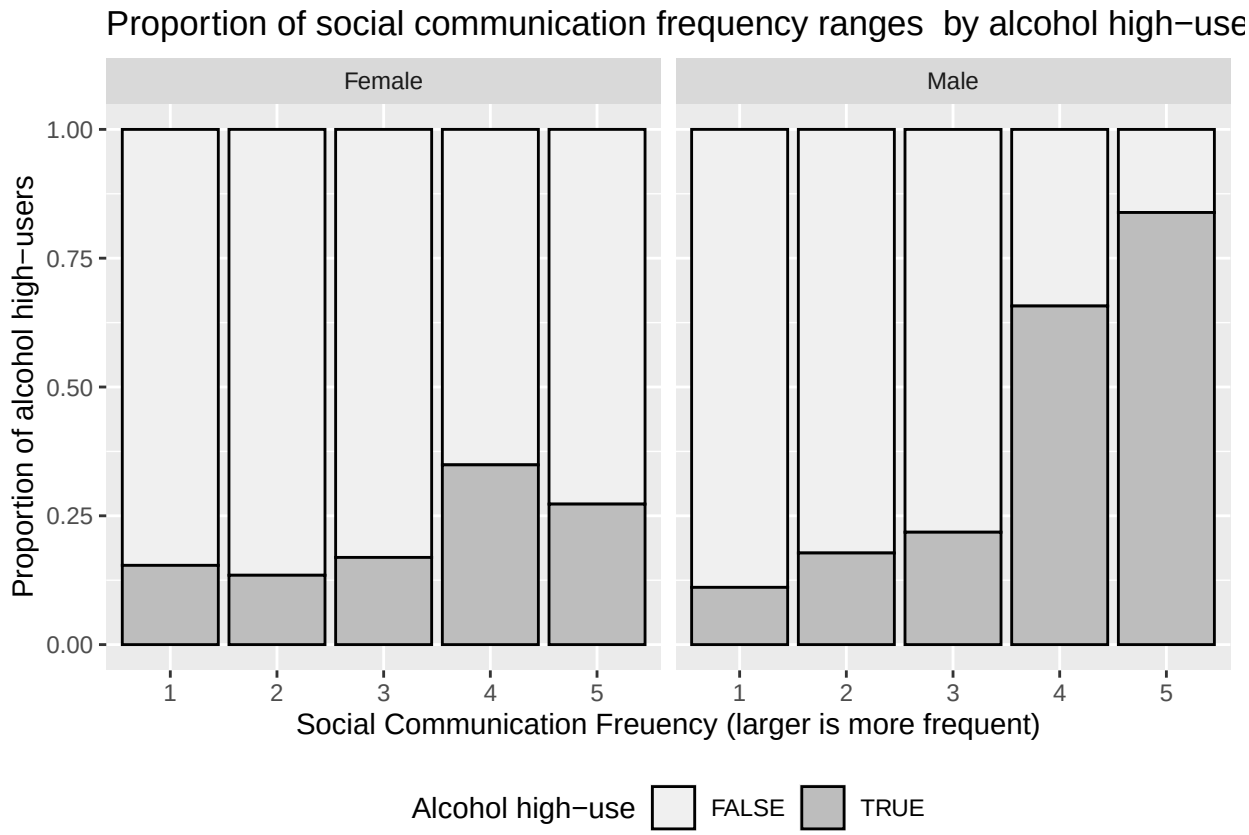
```
alc %>% count(high_use, goout)
```

```
## # A tibble: 10 x 3
##    high_use goout     n
##    <lgl>    <dbl> <int>
##  1 FALSE        1    19
##  2 FALSE        2    82
##  3 FALSE        3    97
##  4 FALSE        4    40
##  5 FALSE        5    21
##  6 TRUE         1     3
##  7 TRUE         2    15
##  8 TRUE         3    23
##  9 TRUE         4    38
## 10 TRUE         5    32
```

From the table, it is found the sample of participants having very low frequency of social communication (level 1) is small in number (n = 21). Caution should be taken about the potential large error.

3.3.2 graphically explore the association

```
p4 <- alc %>%
  ggplot(aes(x = factor(goout), fill = high_use)) +
  geom_bar(position = "fill", color = "black") +
  facet_wrap(~sex,
             labeller = labeller(sex = sex.labs)) +
  labs(x = "Social Communication Freuency (larger is more frequent)",
       y = "Proportion of alcohol high-users",
       title =
         "Proportion of social communication frequency ranges  by alcohol high-use and sex")+
  theme(legend.position = "bottom")+
  guides(fill=guide_legend(title = "Alcohol high-use"))+
  scale_fill_discrete(labels = c("FALSE" = "Non-high-user",
                                 "TRUE" = "high-user"))+
  scale_fill_brewer(palette = "Greys")

p4
```



According to the bar plot, the proportion of alcohol high-users changed tremendously across different levels of social communication, indicating good validity of our model hypothesis about this variable. There is a clear borderline between social communication levels 1-3 and levels 4-5, though the difference is varied across genders. The levels are hence re-coded into two–Infrequent (original level 1-3) and Frequent (original level 4+5). Its interaction with sex will also be considered in fitting the model. This corresponds to the finding of previous evidence(Senchak, Leonard, and Greene 1998).

3.2.3 re-code the variable of social communication

```
alc <- alc %>%
  mutate(social = goout>3)

alc <- alc %>%
  mutate(social = social %>%
           factor() %>%
           fct_recode("Frequent" = "TRUE",
                      "Infrequent" = "FALSE"))
```

4 Model fitting

4.1 fitting base on the original hypothesis

```
fit1 <- glm(high_use~ family.quality:sex + social:sex + off.class.performance + in.class.performance, da
summary(fit1)
```

```
##
## Call:
## glm(formula = high_use ~ family.quality:sex + social:sex + off.class.performance +
##     in.class.performance, family = "binomial", data = alc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8260  -0.6946  -0.4982   0.6392   2.3077
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -0.05927    0.59970  -0.099  0.92127
## off.class.performanceModerate study -0.44754    0.31133  -1.438  0.15058
## off.class.performanceLong study     -1.03288    0.43290  -2.386  0.01704 *
## in.class.performance                 0.06651    0.02294   2.899  0.00374 **
## family.quality:sexF                 -0.37948    0.15263  -2.486  0.01291 *
## family.quality:sexM                 -0.34314    0.14677  -2.338  0.01939 *
## sexF:socialFrequent                  0.92783    0.38133   2.433  0.01497 *
## sexM:socialFrequent                  2.50131    0.39566   6.322 2.58e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 452.04  on 369  degrees of freedom
## Residual deviance: 350.52  on 362  degrees of freedom
## AIC: 366.52
##
## Number of Fisher Scoring iterations: 4
```

All of the hypothesized predictors have at least one level being significant in the model. Comparing to light study participants, moderate study participants is not significant in predicting alcohol high-use. Hence,this variable will be dichotomized into Light study and moderate to long study for better model performance and parsimony of levels. The reason why it is not dichotomized into long study and moderate to short study is because the sample of long study category is extremely small, risking introducing error in our model.

4.2 re-code variable with insignificant levels

```
alc <- alc %>%
  mutate(off.class.performance =
           case_when(off.class.performance == "Light study"~ "Light study",
                     TRUE~ "Moderate to long study") %>%
           factor(levels = c("Light study",
                             "Moderate to long study")))
```

4.3 fitting the model again

```
fit2 <- glm(high_use~ family.quality:sex + social:sex + off.class.performance + in.class.performance, da
summary(fit2)
```

```
##
## Call:
## glm(formula = high_use ~ family.quality:sex + social:sex + off.class.performance +
##     in.class.performance, family = "binomial", data = alc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8509  -0.6950  -0.4982   0.6364   2.3410
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                              -0.06960    0.59922  -0.116
## off.class.performanceModerate to long study -0.58019    0.30223  -1.920
## in.class.performance                      0.07155    0.02275   3.145
## family.quality:sexF                      -0.40471    0.15192  -2.664
## family.quality:sexM                      -0.34358    0.14686  -2.339
## sexF:socialFrequent                       1.01880    0.37639   2.707
## sexM:socialFrequent                       2.51453    0.39586   6.352
##                                          Pr(>|z|)
## (Intercept)                               0.90753
## off.class.performanceModerate to long study  0.05489 .
## in.class.performance                      0.00166 **
## family.quality:sexF                       0.00772 **
## family.quality:sexM                       0.01931 *
## sexF:socialFrequent                       0.00679 **
## sexM:socialFrequent                      2.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 452.04  on 369  degrees of freedom
## Residual deviance: 352.88  on 363  degrees of freedom
## AIC: 366.88
##
## Number of Fisher Scoring iterations: 4
```

Now all of the hypothesized predictors are significant in predicting alcohol high-use, except for off-class performance, which has a a p value of 0.05489, being very close to 0.05. An increase in sample size would very possibly make it significant. I hence keep this predictor in the model. Consequently, fit2 will be our final model.

4.4 interpreting the model results

4.4.1 transforming the coeficients to ORs

```
OR <- coef(fit2) %>% exp()
CI <- confint(fit2) %>% exp()
ORCI <- cbind(OR,CI)
print(ORCI, digits = 2)
```

```
##                                                OR 2.5 % 97.5 %
## (Intercept)                                  0.93  0.29   3.02
## off.class.performanceModerate to long study  0.56  0.31   1.02
## in.class.performance                         1.07  1.03   1.13
## family.quality:sexF                          0.67  0.49   0.90
## family.quality:sexM                          0.71  0.53   0.94
## sexF:socialFrequent                          2.77  1.33   5.86
## sexM:socialFrequent                         12.36  5.84  27.75
```

Our hypothesis that *a.* family relationship quality (interactive with gender); *b.* school performance; *c.* social communication (interactive with gender) could be predictors for alcohol high-use among college students is justified. According to the final model, comparing to participants who study less than 5 hours per week, those who study more than 5 hours have on average 0.55 times less odds to be an alcohol high-user. Participants who have one more time of absence from class will on average have 1.07 times more odds being an alcohol high-user. These findings about the predictive effect of academic performance on alcohol use is consistent with previous evidence(Hayatbakhsh et al. 2011).For female college students, every one unit of family relationship quality increase would lead to 0.66 times less odds being alcohol high-user. For male students, every one unit of family relationship quality increase would lead to 0.71 times less odds being alcohol high-user. These indicate the predictive effects of family relationship on alcohol use are present and different across genders. This finding is consistent with previous evidence (Kelly et al. 2011). For female college students, comparing to students who do not have social involvement frequently, those who usually have social engagement have 2.77 times more odds of being alcohol high-users. For male students, this effect is also present but the effect size goes as high as 12.36 times more odds of being alcohol high-users. These indicate the predictive effects of social engagement on alcohol use are present and tremendously different across genders. This finding is consistent with previous evidence(Senchak, Leonard, and Greene 1998).

4.4.2 exploring predictions

```
prob <- predict(fit2, type = "response")

alc <- alc %>%
  mutate(probability = prob)

alc <- alc %>%
  mutate(prediction = probability>0.5)
table(high_use = alc$high_use, prediction = alc$prediction)
```

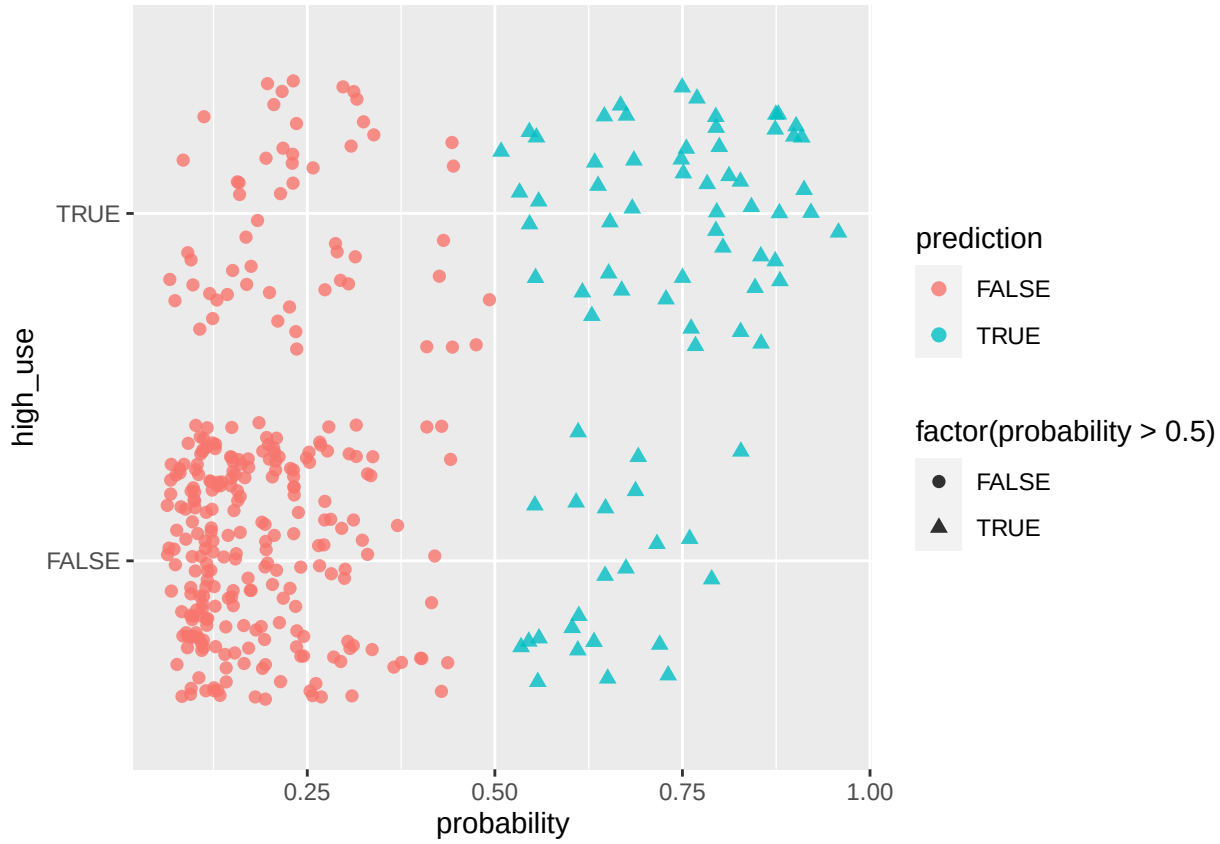
4.4.2.1 cross tabulation of predcition versus the actual values

```
##         prediction
## high_use FALSE TRUE
##    FALSE   236   23
##    TRUE     57   54
```

Among 259 participants who are not alcohol high-users, our model correctly predicts 236 (91%) of them. Among 111 participants who are alcohol high-users, our model correctly predicts 54 of them (49%) of them. In all, among the 370 predicts, 80(21.6%) were inaccurate.

4.4.2.2 scatter plot of the precition versus the actual values

```
library(dplyr); library(ggplot2)

p5 <- alc %>%
  ggplot(aes(x = probability,
             y = high_use,
             color = prediction,
             shape =factor(probability>0.5))) +
  geom_point(position = position_jitter(0.01),
             alpha = 0.8, size =2)
p5
```



4.4.2.3 comparing the model to the performance of random guess

```
random.guess <- runif(n= nrow(alc), min = 0, max = 1)
alc <- alc %>%
  mutate(random.guess = random.guess)
alc <- alc %>%
  mutate(prediction.guess = random.guess>0.5)
table(high_use = alc$high_use, prediction = alc$prediction.guess)
```

```
##         prediction
## high_use FALSE TRUE
##    FALSE   130  129
##    TRUE     57   54
```

Among 259 participants who are not alcohol high-users, random guess correctly guesses 126 (49%) of them. Among 111 participants who are alcohol high-users, random guess correctly guesses 55 of them (49%) of them. In all, among the 370 predicts, 181(49%) were inaccurate. Our model shows a tremendously better overall performance than random guess. However, its effect on correctly predicting the alcohol high-users is roughly equal to random guess, indicating the model is better applied in predicting who is a non-alcohol-high-user.

4.5 cross validation (Bonus)

4.5.1 define loss function

```
# define a loss function (average prediction error)
loss_func <- function(class, prob) {
  n_wrong <- abs(class - prob) > 0.5
  mean(n_wrong)
}
```

4.5.2 compute prediction error base on traing data set

```
# compute the average number of wrong predictions in the (training) data
training.error.full <- loss_func(alc$high_use, alc$probability)
training.error.full
```

```
## [1] 0.2162162
```

The prediction error rate is 21.6%, outperforming the model in Exercise Set 3, which had about 26% error.

4.5.3 compute prediction error base on 10-fold cross validation

```
# 10-fold cross-validation
set.seed(16)
library(boot)
cv <- cv.glm(data = alc, cost = loss_func, glmfit = fit2, K = 10)
cross.val.error.full <- cv$delta[1]
cross.val.error.full
```

```
## [1] 0.2216216
```

According to the result of 10 fold cross-validation, the model has an average error rate of 22.2%, a bit larger than the results from training model, but the error rate is still notably lower than the model in Exercise.

5 Observing the relationship between prediction error and number of predictors (Super Bonus)

5.1 preparation

```
library(utils)#install.packages("utils") library for generating all possible combinations of n elements

#pass the name of 4 predictors for our final model into an object
used.predictor <- c("family.quality:sex",
                    "social:sex",
                    "off.class.performance",
                    "in.class.performance")
```

5.2 generating prediction error for all possible combinations of subsets of selected predictors

```
#define a list "mylist"
mylist <- list()

#define a matrix with 2 rows and 6 columns, "ct.error", which means
#cross-validation and training error
ct.error <- matrix(nrow=2, ncol = 6)

#start a loop that generate all possible combinations of the 4 used predictors,
#the combination could have 1-4 elements. Four each number of element, start a
#loop (i in 1:4); Within the loop, another loop is used to pass all the prediction
#error results from cross validation and training data set into a matrix. Each
#Matrix will have two rows saving results of cv and training data set, respectively.
#The number of columns will be dependent on how many combinations will be produced,
#with the maximum number being 6 (number of possible combinations of  2 predictors).
#Base on the number of i, 4 matrices will be generated, and saved in mylist.

for(i in 1:4){
  combinations <- combn(used.predictor, i)
   all.formula.text <- apply(combinations, 2,
                             function(x)paste("high_use~",
                                              paste(x, collapse = "+")))
  for(j in 1:length(all.formula.text)){
    all.formula <- as.formula(all.formula.text[j])
    model <- glm(all.formula, data = alc, family = "binomial")
    cv <- cv.glm(data = alc, cost = loss_func, glmfit = model, K =10)
    ct.error[1,j] <- cv$delta[1]
    alc <- mutate(alc, probability = predict(model, type = "response"))
```

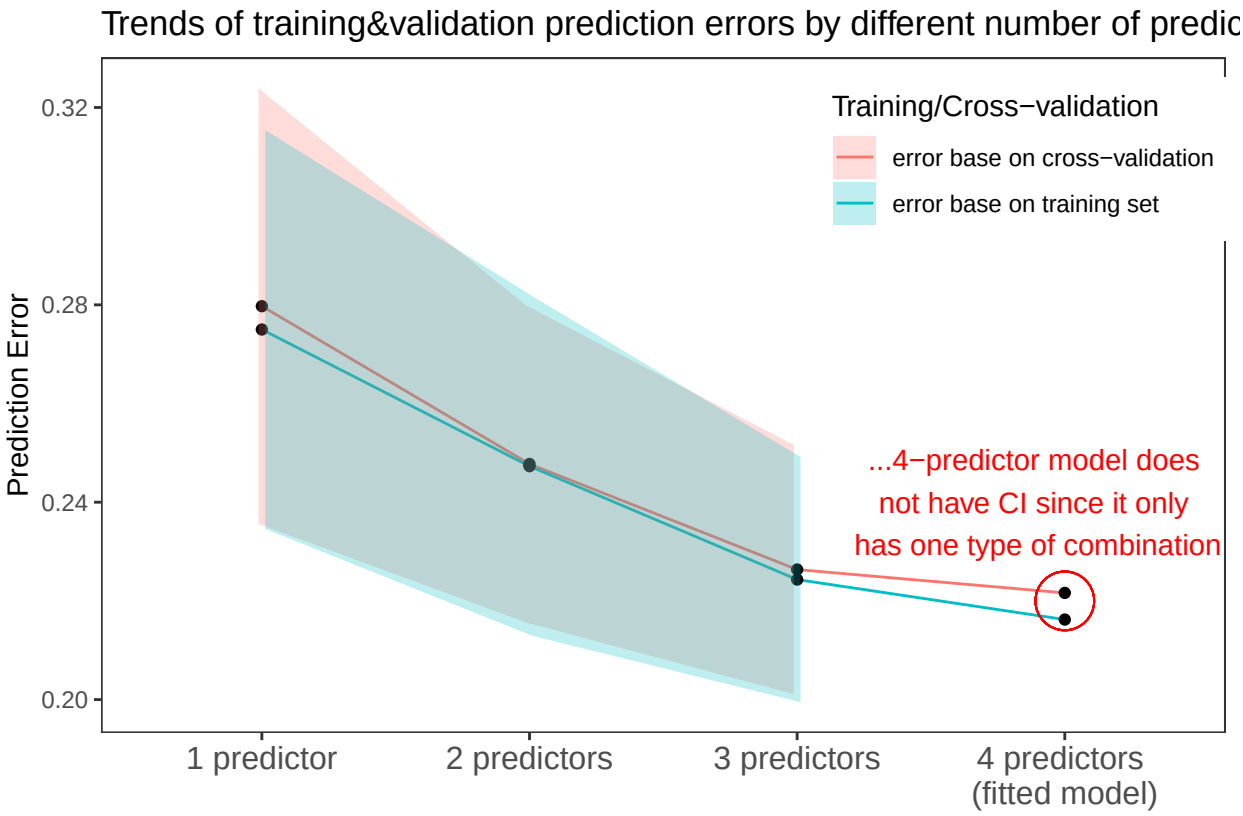
```
    ct.error[2,j] <- loss_func(alc$high_use, alc$probability)
  }
   mylist[[i]] <- ct.error
   ct.error <- matrix(nrow=2, ncol = 6)
}

#collapse the 4 matrices in mylist into 4 data frames.
for(w in 1:4){
 assign(paste0("df",w), as.data.frame(mylist[[w]]))
  }

#merge the 4 data frames into 1 by row.
all.error <- rbind(df1,df2,df3,df4) #name the data set as all.error

#add a new column in all.error, which reflects if the result of this row is
#from cross validation or training set

tag <- rep(c("pred_cv", "pred_training"), times = 4)

#add another new column in all.error, which reflects if the result of this row is
#base on 1, 2, 3 or 4 predictors.
predictor_number <- rep(c(1,2,3,4), each = 2)

all.error <- all.error %>%
  mutate(tag = tag,
         predictor_number = predictor_number)

#calculate the mean and sd for each row.
#note that the rows base on 4 predictor will not have sd, since there is only
#one combination.
all.error <- all.error %>%
  mutate(mean = rowMeans(select(.,V1:V6), na.rm = T),
         sd = apply(.[,1:6], 1, function(x)sd(x, na.rm=T)))
#check the all.error data set
all.error
```

```
##          V1        V2        V3        V4        V5        V6           tag
## 1 0.3135135 0.2135135 0.3000000 0.2918919        NA        NA       pred_cv
## 2 0.2972973 0.2135135 0.3000000 0.2891892        NA        NA pred_training
## 3 0.2135135 0.3027027 0.2675676 0.2135135 0.2108108 0.2783784       pred_cv
## 4 0.2108108 0.3054054 0.2648649 0.2135135 0.2054054 0.2837838 pred_training
## 5 0.2162162 0.2108108 0.2648649 0.2135135        NA        NA       pred_cv
## 6 0.2135135 0.2081081 0.2621622 0.2135135        NA        NA pred_training
## 7 0.2216216        NA        NA        NA        NA        NA       pred_cv
## 8 0.2162162        NA        NA        NA        NA        NA pred_training
##   predictor_number      mean         sd
## 1                1 0.2797297 0.04503604
## 2                1 0.2750000 0.04124759
## 3                2 0.2477477 0.04014825
## 4                2 0.2472973 0.04299761
## 5                3 0.2263514 0.02577033
## 6                3 0.2243243 0.02535360
## 7                4 0.2216216         NA
```

```
## 8               4 0.2162162         NA
```

5.3 plotting the trends of training&validation prediction errors by different number of predictors

```
#plot all.error
#the error ribbon is 95% confidence interval
#4 predictors (the fitted model) do now have a error range because there is only one combination
all.error %>% ggplot(aes(x = factor(predictor_number), y = mean, group = tag), color = tag) +
  geom_line(aes(color = tag))+
  geom_point()+
  geom_ribbon(aes(ymin = mean-1.96*sd/sqrt(rowSums(!is.na(select(all.error,V1:V6)))),
                  ymax = mean+1.96*sd/sqrt(rowSums(!is.na(select(all.error,V1:V6)))),
                  fill = tag), alpha =0.25,
              position = position_dodge(0.05))+
  guides(fill = guide_legend(title = "Training/Cross-validation", title.position = "top"),
         color = guide_legend(title = "Training/Cross-validation", title.position = "top"))+
  theme_bw()+
  theme(legend.position = c(0.82,0.85), axis.text.x = element_text(size=12),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x =  "", y = "Prediction Error",
       title = "Trends of training&validation prediction errors by different number of predictors")+
  scale_x_discrete(labels = c("1 predictor", "2 predictors",
                              "3 predictors",
                              "4 predictors \n(fitted model)"))+
  scale_fill_discrete(labels = c("pred_cv" = "error base on cross-validation",
                                 "pred_training" = "error base on training set"))+
  scale_color_discrete(labels = c("pred_cv" = "error base on cross-validation",
                                  "pred_training" = "error base on training set"))+
  annotate("text",
           label =
             "↓4-predictor model does \nnot have CI since it only \nhas one type of combination",
           x = 3.9, y = 0.24, color = "red")+
  geom_point(aes(x=4, y= 0.22), shape = 1, size = 10, color = "red")
```



According to the plot, the prediction error of our 4-predictor final model is low comparing to the mean prediction errors of the possible combinations of either of the 1-,2-, and 3- predictor models. However, this goodness is only statistically significant comparing to 1- predictor models.(falls with 95%ci of 2- and 3-predictor models). This might be due to the possible combinations of 4 predictors is very small in number, resulting in large error ranges.

Supplementary of Chapter 3: some inspiration from the super bous task (This is doing for practicing, not part of assignment Chapter 3)

Inspired by the bonus task, where different <4 number of predictors' influence on prediction error was observed, I started to get interested in how different number of random combinations of predictors added to the final model would affect the error. There are almost 30 variables that were not used in the final model. Twenty-three of them do not have direct relationship with the entered predictors, and hence they were selected to be a free predictor pool. One to 15 different predictors were randomly selected from the pool, each with 100 random repetitions (if all possible combinations of the number of predictors <100, then all possible combination will be used), resulting in 1423 models. The error rate base on training dataset and 10 fold cross validation were computed and ploted in a line chart. 95% confidence interval for each number of added predictors were also calculated and visualized.

The reason why only 15 maximum added predictors will be used instead of all 23 predictors is because the current sample size could not faithfully support model with more than 19 predictors, according to a rule of thumb that for each predictor used in model, a sample of 20 is required.

Preparing the predictor pool

```r
# The predictors used in final 4-factor model
fixed.predictor <- c("family.quality:sex",
                     "social:sex",
                     "off.class.performance",
                     "in.class.performance")

# The variables not used in final model
not.used.predictor <- c("sex", "famsize", "studytime", "famrel", "Dalc",
                        "Walc", "G1", "G2", "G3", "alc_use", "high_use",
                        "family.quality", "social", "probability",
                        "prediction", "random.guess", "prediction.guess",
                        "goout")

#The set of free predictor pool
free.predictor<- setdiff(names(alc), fixed.predictor)
free.predictor<- setdiff(free.predictor, not.used.predictor)
```

Building a loop that generates the result of 1423 models with 5~19 predictors (4 predictors in final model are fixed)

```r
mylist <- list()

ct.error <- matrix(nrow=2, ncol = 100)


for(i in 1:15){
  combinations <- combn(free.predictor, i)
  if(choose(23,i)>100){
    ss = 100
  }else{
    ss = choose(23,i)
  }
  for(j in 1:ss){
    rn <- round(runif(1,min = 1, max = choose(23,i)), 0)
    sample.comb <- combinations[,rn]
    formula.text <- paste(
      "high_use ~ family.quality:sex + social:sex + off.class.performance + in.class.performance+",
      paste(sample.comb, collapse = "+"))
    model <- glm(formula.text, data = alc, family = "binomial")
    cv <- cv.glm(data = alc, cost = loss_func, glmfit = model, K =10)
    ct.error[1,j] <- cv$delta[1]
    alc <- mutate(alc, probability = predict(model, type = "response"))
    ct.error[2,j] <- loss_func(alc$high_use, alc$probability)
  }
  mylist[[i]] <- ct.error
  ct.error <- matrix(nrow=2, ncol = 100)
}
```

Collapsing the results into different data frame and merge them

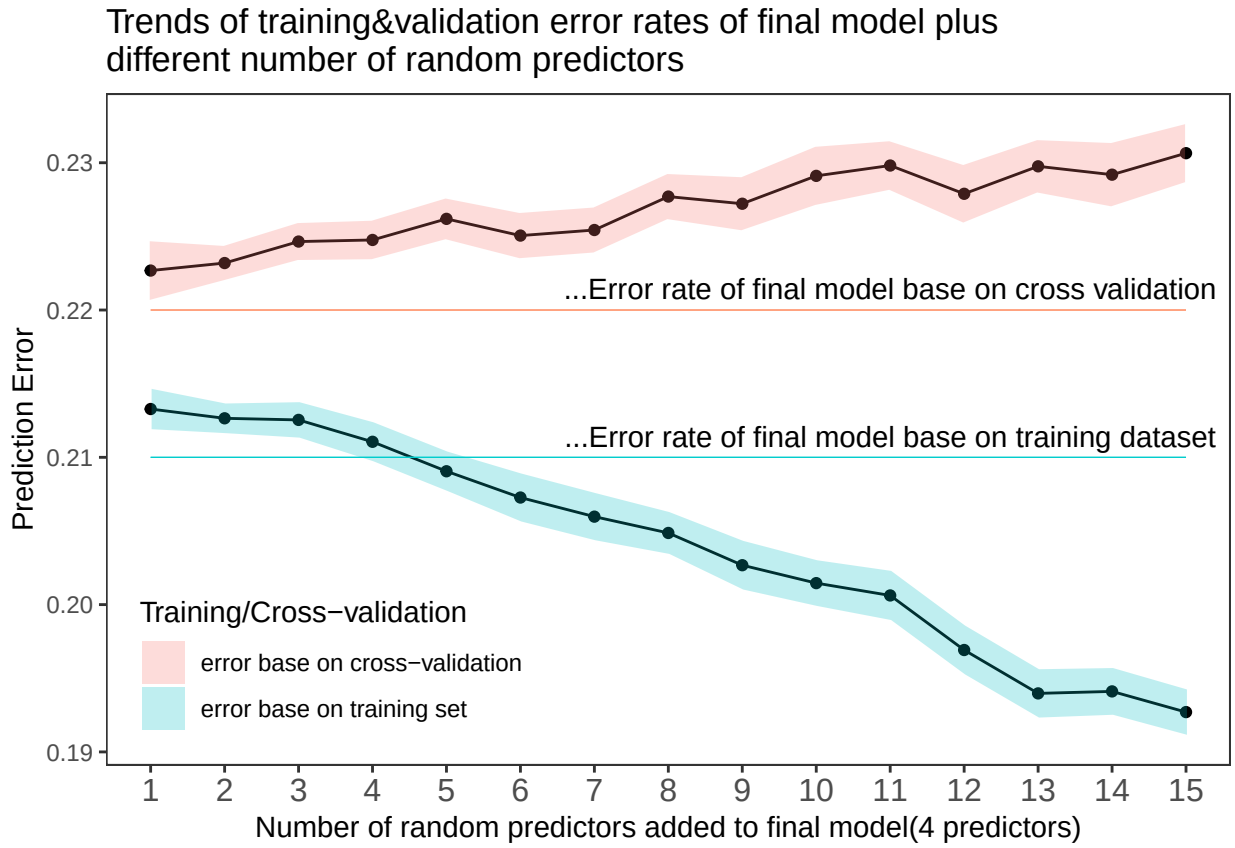
```r
for(w in 1:15){
 assign(paste0("df",w), as.data.frame(mylist[[w]]))
  }
```

```r
all.error <- rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12,df13,df14,df15)

tag <- rep(c("pred_cv", "pred_training"), times = 15)

#add another new column in all.error, which reflects if the result of this row is
#base on 1-15 predictors.
predictor_number <- rep(c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15), each = 2)

all.error <- all.error %>%
  mutate(tag = tag,
         predictor_number = predictor_number)

#calculate the mean and sd for each row.
all.error <- all.error %>%
  mutate(mean = rowMeans(select(.,V1:V100), na.rm = T),
         sd = apply(.[,1:100], 1, function(x)sd(x, na.rm=T)))
```

Plotting

```r
#plot all.error
#the error ribbon is 95% confidence interval
#4 predictors (the fitted model) do now have a error range because there is only one combination
all.error %>% ggplot(aes(x = factor(predictor_number), y = mean, group = tag)) +
  geom_line()+
  geom_point()+
  geom_ribbon(aes(ymin =
                    mean-1.96*sd/sqrt(rowSums(!is.na(select(all.error,V1:V100)))),
                  ymax =
                    mean+1.96*sd/sqrt(rowSums(!is.na(select(all.error,V1:V100)))),
                  fill = tag), alpha =0.25,
                position = position_dodge(0.05))+
  guides(fill = guide_legend(title = "Training/Cross-validation",
                             title.position = "top"),
         color = guide_legend(title = "Training/Cross-validation",
                              title.position = "top"))+
  theme_bw()+
  theme(legend.position = c(0.2,0.15),
        axis.text.x = element_text(size=12),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x =  "Number of random predictors added to final model(4 predictors)",
       y = "Prediction Error",
       title="Trends of training&validation error rates of final model plus \ndifferent number of rando
  scale_fill_discrete(labels = c("pred_cv" = "error base on cross-validation",
                                 "pred_training" = "error base on training set"))+
  scale_color_discrete(labels = c("pred_cv" = "error base on cross-validation",
                                  "pred_training" = "error base on training set"))+
  geom_line(aes(y = 0.22), color = "coral", size = 0.2, alpha = 0.8)+
  geom_line(aes(y = 0.21), color = "cyan3", size = 0.2, alpha =1)+
 annotate("text",
          label = "↓Error rate of final model base on cross validation",
          x = 11, y = 0.22, vjust = -0.5)+
 annotate("text",
```

```
        label = "↓Error rate of final model base on training dataset",
        x = 11, y = 0.21, vjust = -0.5)
```



It is found in the plot that the prediction error rate of the 4-predictor final model by cross validation is always lower than the mean prediction error (and their lower ends of confidence interval) of the final model plus 1 to 15 randomly selected predictors, indicating the goodness of our final model.

It is also interesting to observe that the more predictors introduced to the model, the error rate by training data set keeps decreasing, indicating more predictors produce better models. However, the results of error rate by cross validation show an opposite effect, where the error rates generally increase with more predictors (though some flucutations are present). Put together, it can be infered that measuring model error rates using training data set itself would lead to over-estimation of the model goodness when more predictors enter the model.

Reference

Brody, Gene H., and Rex Forehand. 1993. "Prospective Associations Among Family Form, Family Processes, and Adolescents' Alcohol and Drug Use." *Behaviour Research and Therapy* 31 (6): 587–93. https://doi.org/10.1016/0005-7967(93)90110-g.

Flor, Luisa Socio, and Emmanuela Gakidou. 2020. "The Burden of Alcohol Use: Better Data and Strong Policies Towards a Sustainable Development." *The Lancet Public Health* 5 (1): e10–11. https://doi.org/10.1016/s2468-2667(19)30254-3.

Hayatbakhsh, Mohammad Reza, Jake M. Najman, William Bor, Alexandra Clavarino, and Rosa Alati. 2011. "School Performance and Alcohol Use Problems in Early Adulthood: A Longitudinal Study." *Alcohol* 45 (7): 701–9. https://doi.org/10.1016/j.alcohol.2010.10.009.

Kelly, Adrian B., John W. Toumbourou, Martin OFlaherty, George C. Patton, Ross Homel, Jason P. Connor, and Joanne Williams. 2011. "Family Relationship Quality and Early Alcohol Use: Evidence for Gender-Specific Risk Processes." *Journal of Studies on Alcohol and Drugs* 72 (3): 399–407. https://doi.org/10.15288/jsad.2011.72.399.

Lees, Briana, Lindsay R. Meredith, Anna E. Kirkland, Brittany E. Bryant, and Lindsay M. Squeglia. 2020. "Effect of Alcohol Use on the Adolescent Brain and Behavior." *Pharmacology Biochemistry and Behavior* 192 (May): 172906. https://doi.org/10.1016/j.pbb.2020.172906.

Room, Robin, Kim Bloomfield, Gerhard Gmel, Ulrike Grittner, Nina-Katri Gustafsson, Pia Mäkelä, Esa Österberg, Mats Ramstedt, Jürgen Rehm, and Matthias Wicki. 2013. "What Happened to Alcohol Consumption and Problems in the Nordic Countries When Alcohol Taxes Were Decreased and Borders Opened?" *The International Journal of Alcohol and Drug Research* 2 (1): 77–87. https://doi.org/10.7895/ijadr.v2i1.58.

Senchak, Marilyn, Kenneth E. Leonard, and Brian W. Greene. 1998. "Alcohol Use Among College Students as a Function of Their Typical Social Drinking Context." *Psychology of Addictive Behaviors* 12 (1): 62–70. https://doi.org/10.1037/0893-164x.12.1.62.