chapter3

Rong Guang

2022-11-14

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, 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 overuse 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' on-campus and off-campuse 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 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 overuse

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
```

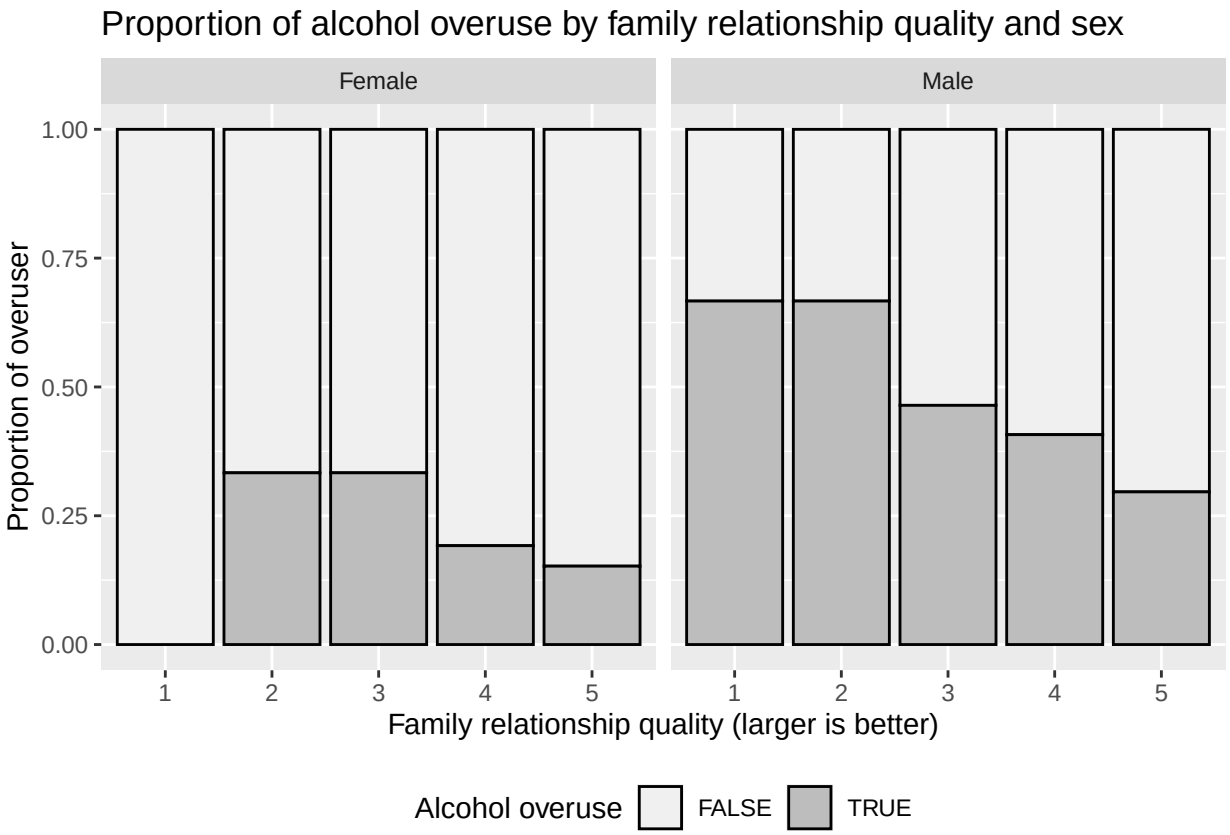
```
?count
```

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 overuser",
       title = "Proportion of alcohol overuse by family relationship quality and sex")+
  theme(legend.position = "bottom")+
  guides(fill=guide_legend(title = "Alcohol overuse"))+
  scale_fill_discrete(labels = c("FALSE" = "Non-overuser", "TRUE" = "Overuser"))+
  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 overuse decreases, except for female from a very bad (level 1) relationship family, which had a proportion of over-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 overuse

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-campus and on-campus school performance, respectively, and hence they are selected as candidate predictors.

3.2.1 exploring the association between on-campus performance and alcohol overuse

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 over-users and non-over-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 overuser",
       y = "Freuqncy of class absences",
       title = "Frequency of class absences by alcohol overuse and gender")+
  theme(legend.position = "none")+
  scale_x_discrete(labels = c("FALSE" = "Non-overuser", "TRUE" = "Overuser"))
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 on.campus.performance according to the hypothesis of current study.

```
alc <- alc %>%
  mutate(on.campus.performance = absences)
```

3.2.2 exploring the association between off-campus performance (study time) and alcohol overuse

3.2.2.1 numerically explore the association

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

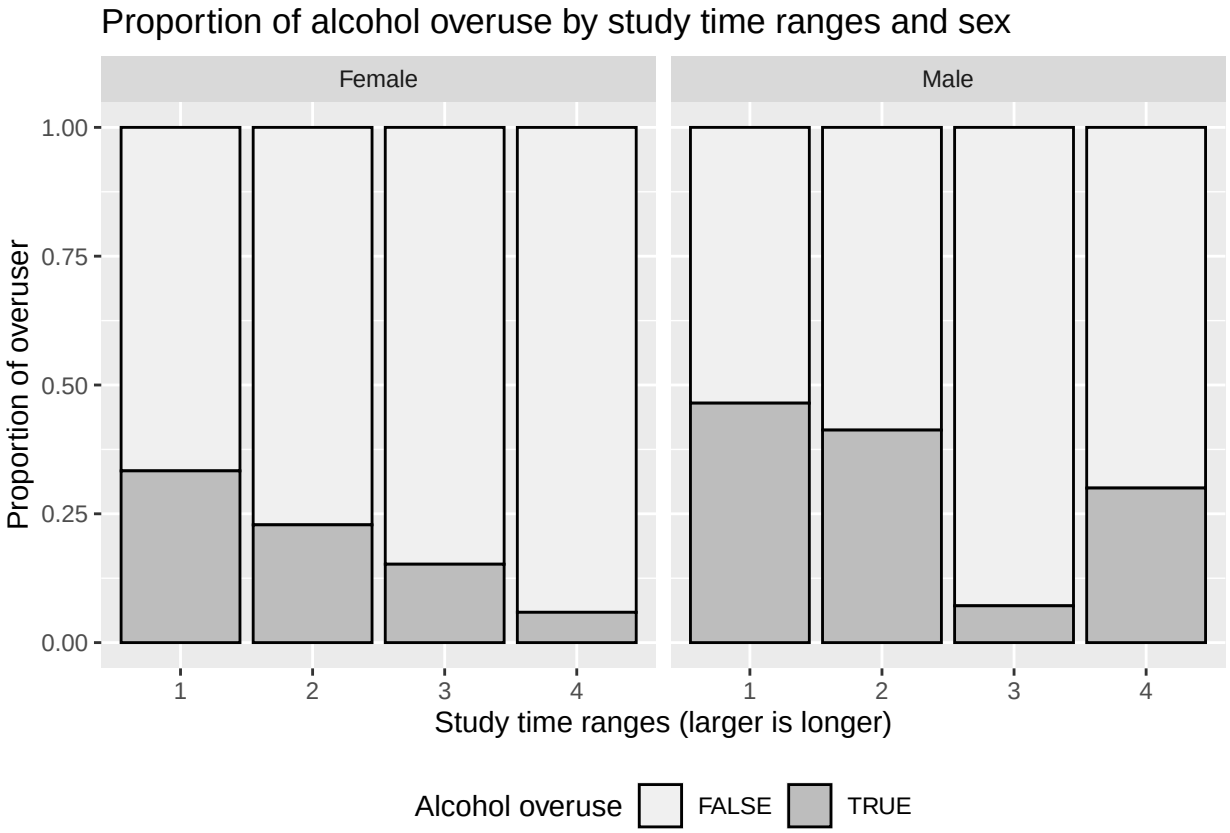
```
## # A tibble: 8 x 3
##   high_use studytime     n
##   <lgl>        <dbl> <int>
## 1 FALSE            1    56
## 2 FALSE            2   128
## 3 FALSE            3    52
```

```
## 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 over-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 overuser",
       title = "Proportion of alcohol overuse by study time ranges and sex")+
  theme(legend.position = "bottom")+
  guides(fill=guide_legend(title = "Alcohol overuse"))+
  scale_fill_discrete(labels = c("FALSE" = "Non-overuser", "TRUE" = "Overuser"))+
  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 overuse 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 over-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.campus.performance according to the hypothesis.

3.2.3 re-code the variable of study length

```
alc <- alc %>%
  mutate(off.campus.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 overuse

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

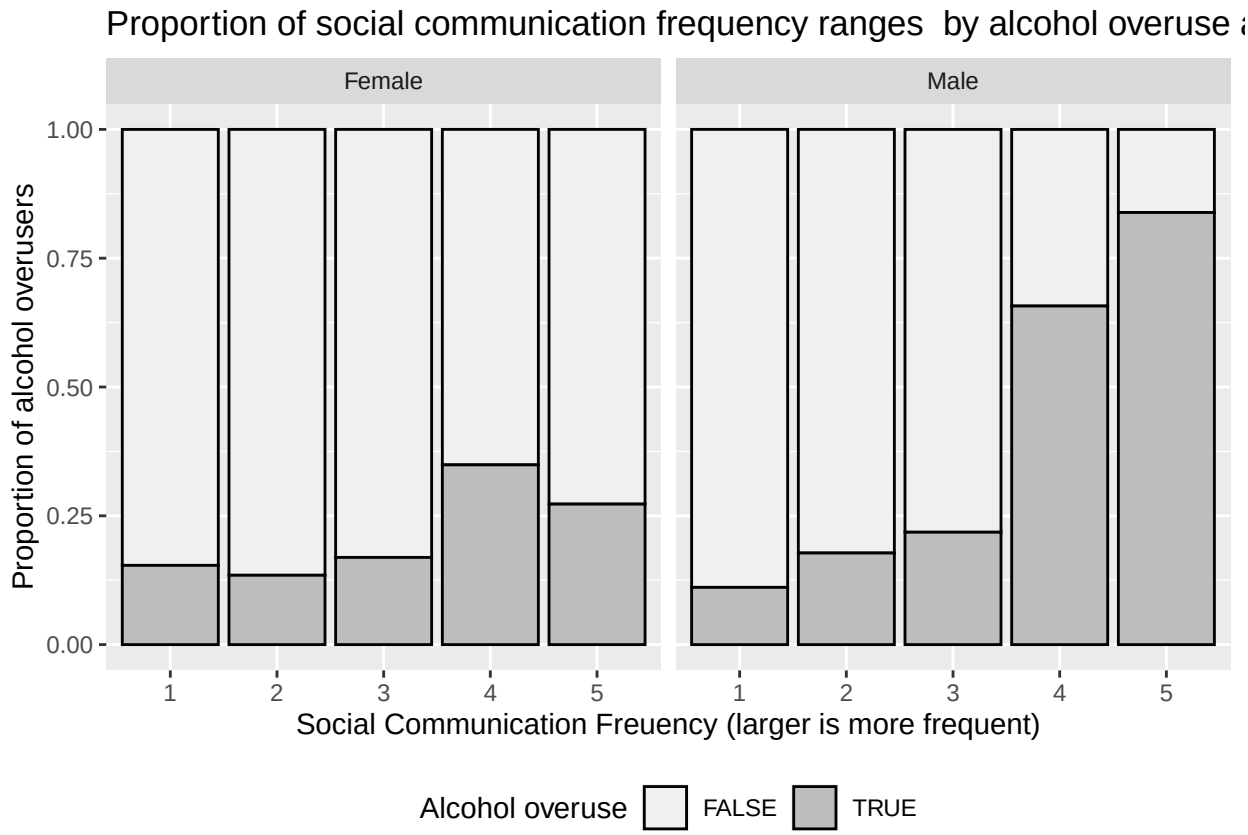
```
alc %>% count(high_use, goout)
```

```
## # 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 overusers",
       title = "Proportion of social communication frequency ranges  by alcohol overuse and sex")+
  theme(legend.position = "bottom")+
  guides(fill=guide_legend(title = "Alcohol overuse"))+
  scale_fill_discrete(labels = c("FALSE" = "Non-overuser", "TRUE" = "Overuser"))+
  scale_fill_brewer(palette = "Greys")

p4
```



According to the bar plot, the proportion of alcohol over-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.campus.performance + on.campus.performance,
summary(fit1)
```

```
##
## Call:
## glm(formula = high_use ~ family.quality:sex + social:sex + off.campus.performance +
##     on.campus.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.campus.performanceModerate study -0.44754    0.31133  -1.438  0.15058
## off.campus.performanceLong study     -1.03288    0.43290  -2.386  0.01704 *
## on.campus.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 overuse. 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.campus.performance =
           case_when(off.campus.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.campus.performance + on.campus.performance,
summary(fit2)
```

```
##
## Call:
## glm(formula = high_use ~ family.quality:sex + social:sex + off.campus.performance +
##     on.campus.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.campus.performanceModerate to long study -0.58019    0.30223  -1.920
## on.campus.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.campus.performanceModerate to long study  0.05489 .
## on.campus.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 overuse, except for off-campus 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.campus.performanceModerate to long study  0.56  0.31   1.02
## on.campus.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 overuse 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 over-user. Participants who have one more time of absence from class will on average have 1.07 times more odds being an alcohol over-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 over-user. For male students, every one unit of family relationship quality increase would lead to 0.71 times less odds being alcohol over-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 over-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 over-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)
```

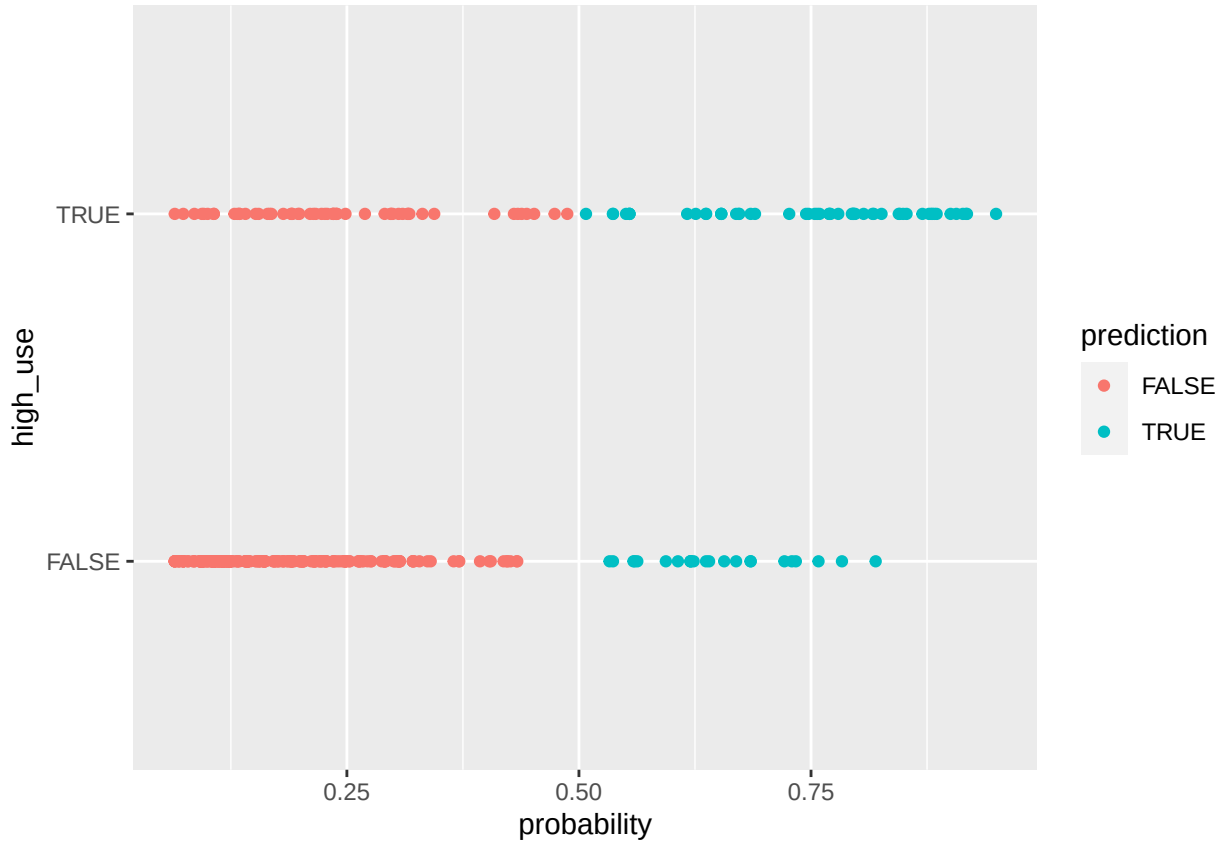
4.4.2.1 cross tabulation of predcition versus the actual values

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

Among 259 participants who are not alcohol over-users, our model correctly predicts 236 (91%) of them. Among 111 participants who are alcohol over-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)) +
  geom_point()
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   132  127
##    TRUE     43   68
```

Among 259 participants who are not alcohol over-users, random guess correctly guesses 126 (49%) of them. Among 111 participants who are alcohol over-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 over-users is roughly equal to random guess, indicating the model is better applied in predicting who is a non-alcohol-over-user.

4.5 cross validation

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 (Bonus)

5.1 preparation

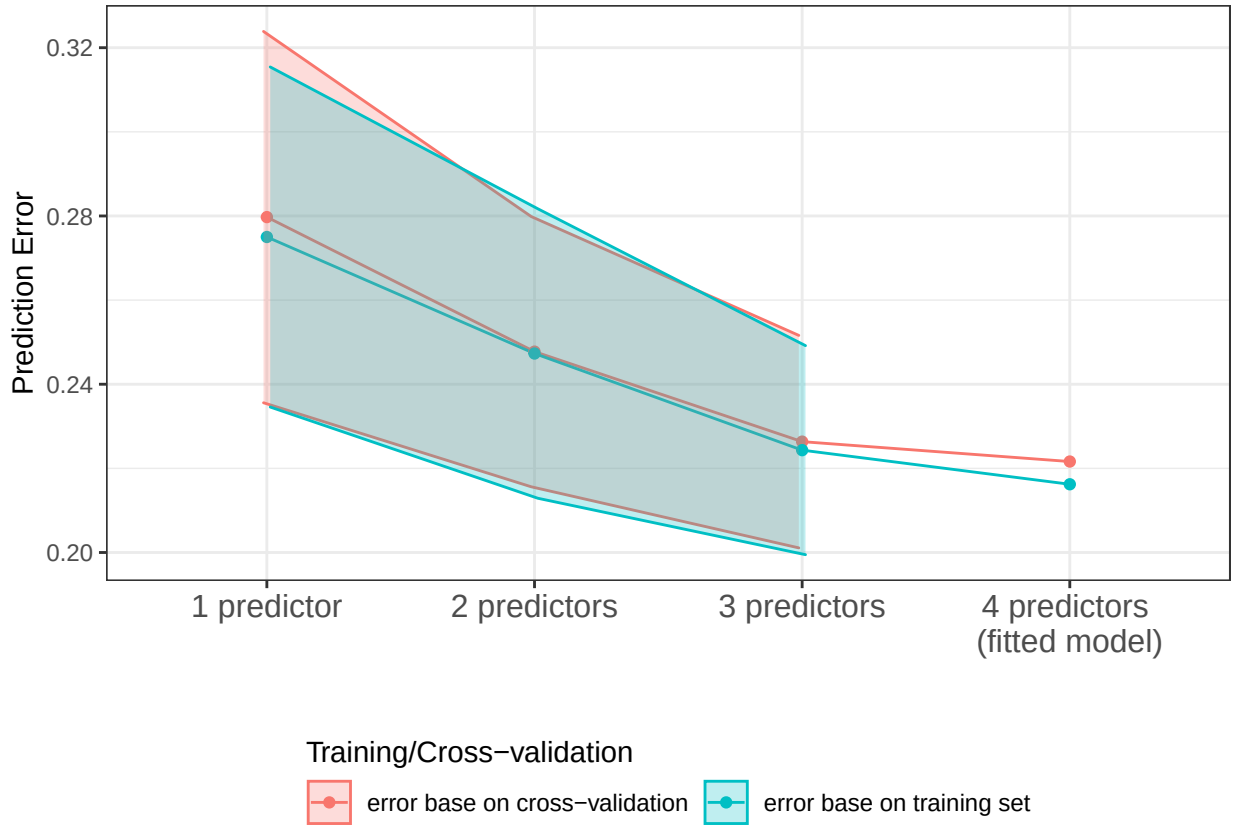
```
library(utils)#install.packages("utils") library for generating all possible combinations of n elements

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

#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 mattrices 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
```

```r
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
```

```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, color = tag)) +
  geom_line()+
  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 = "bottom", axis.text.x = element_text(size=12)) +
  labs(x =  "", y = "Prediction Error")+
```

```
  scale_x_discrete(labels = c("1 predictor", "2 predictors", "3 predictors", "4 predictors \n(fitted mo
  scale_fill_discrete(labels = c("pred_cv" = "error base on cross-validation", "pred_training" = "error
  scale_color_discrete(labels = c("pred_cv" = "error base on cross-validation", "pred_training" = "erro
```



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.