Final Writeup

Phuc Nguyen, Ezinne Nwankwo, Frances Hung

2/19/2020

Data Cleaning

Due to the structure of the survey (students skip certain questions based on their answers), there are a lot of missing values induced in the data for certain variables. We impute zero values for NAs if the variables are ordinal and if NAs are naturally lowest in the ordinal structure. Otherwise, we create indicator variables for whether a person answered NA or not for each variable in question. We additionally drop true missing values.

Exploratory Data Analysis

We look at drinking patterns casually by visualizing some survey question responses from Section C, which is about drinking patterns in college. By looking at histograms, we can see that drinking at off-campus parties and bars results in the highest number of drinks taken. The reasons students cited as most important for drinking were getting drunk and as a reward. This suggests that drinking location and social groups may be important predictors of drinking behavior.

For our dependent variable measuring drinking behavior, we choose to use `drinkcat`, a variable created by researchers which classifies drinking behavior into four ordinal risk categories. The sexual assault index variable draws from three questions in the survey: D5_H (unwanted sexual advance), D5_I (date rape/sexual assault), E15 (intoxicated, unable to consent).

Methods

We use Structural Equation Modeling (SEM) for modeling causal relationships between information collected via the survey and our dependent variables `drinkcat` and `sexassault_ind`. The SEM package we use in R, `lavaan`, uses partial least squares regression to find a linear model relating the independent and dependent variables through latent factors. We standardize all of the variables we use in order to make our coefficients interpretable and comparable in the final model.

We create the following latent factors from existing survey question variables (Fig. 1). Since these factors must be correlated, interpretable, and continuous (due to our use of the `lavaan` package), we ensure that the survey question variables included in each factor are uniformly ordinal in an interpretable way. Our two factors related to students' pre-college lives are Family Education/Drinking (G14-17) and High School Drinking Behavior (G9-11). The external-policy related factors are College Alcohol Policy (B1, B2, B9), College Alcohol Education (B7, B8), and City Alcohol Policy (B11-13). The rest of the latent factors have to do with students' college experiences and opinions: Academic Wellbeing (F1, F4), Personal Wellbeing (F6), Support for Stricter Policies (D3), Opinions on # Appropriate Drinks (D1), Time Spent on Activities (F5), and General Interests for College Experience (split into activities involving and not involving partying/Greek life, A10).

In order to understand the relationships between drinking behaviors and sexual assault, we fit another latent factor model with mediation on the indicator of being a victim of sexual assault given family education/drinking background, high school, college drinking behavior, and participations in parties. Mediation analysis allows us to estimate how much of the association between a risk factor, such as participation in parties, and the risk of sexual assault is mediated by the victim's drinking behavior.

Results

Drinking Behavior

Table **??** shows the estimates and 95% confidence intervals for the coefficents of latent variables in the drinking behavior model. We accept the model as its RMSEA 95% CI of [0.053, 0.054] is around the conventional threshold of 0.05 for a good fit (CITE). Since the survey answers generally did not have physical units, we standardized the variables to compare the size of the estimated coefficients. The dependent variable, i.e. the rating of levels of drinking, also does not have physical units. Thus, we will interpret these coefficents in terms of the relative magnitude and direction of their association with heavier drinking. High school drinking history is most correlated with more drinking in college. The latent factor Parties summarizing the importance of fraternity/sorority, parties and athletics to students (A10), with a lower value corresponding to more importance, has the second largest association. In other words, students who participate more in the mentioned activities tend to drink more. The variable Communities measures the importance of other activities such as volunteering, religion, arts, activism and academic (A10), has the third largest association. Students who care less about these activities tend to drink more. Fourthly, students who perceive a larger number of drink appropriate tend to drink more. Other latent variables with significant but smaller associations with heavier drinking in college include Personal Wellbeing, Family Education/Drinking and Support for More Lenient Policy. Specifically, students who are generally more happy, whose families approve of drinking, or who support more lenient drinking policies on campus tend to drink more. There is not enough evidence to establish a relationship between stricter college drinking policy and drinking behaviors. We find a small association between more exposure to alcohol/drinking education and heavier drinkers. We hypothesize that students who already have drinking problems or schools with more prevalent drinking culture might consequently have more educational programs on the issue.

Risk of Sexual Assault

Sensitivity Analysis

We can compare our SEM model for college drinking severity to an elastic net model created using the `glmnet` package. Both of these models carry out dimension reduction, so it is interesting and reassuring to note that the results of both seem to corraborate one another. From the elastic net model, the most important predictors for heavy drinkers are A10_F(parties, -), B6_D (times they were part of drinking group which was asked to be quieter/less disruptive, +), D1_B (opinion on appropriate amount for off-campus bar drinking, +), D9_B (percentage of friends who are binge drinkers, +), and G11 (HS number of binges, +). This supports the results of our latent factor model, which finds strong correlations of college drinking with drinking attitudes, emphasis on partying, and high school drinking.

References

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819368/

http://lavaan.ugent.be/index.html

An Introduction to Structural Equation Modeling. J.J. Hox. University of Amsterdam/Utrecht University

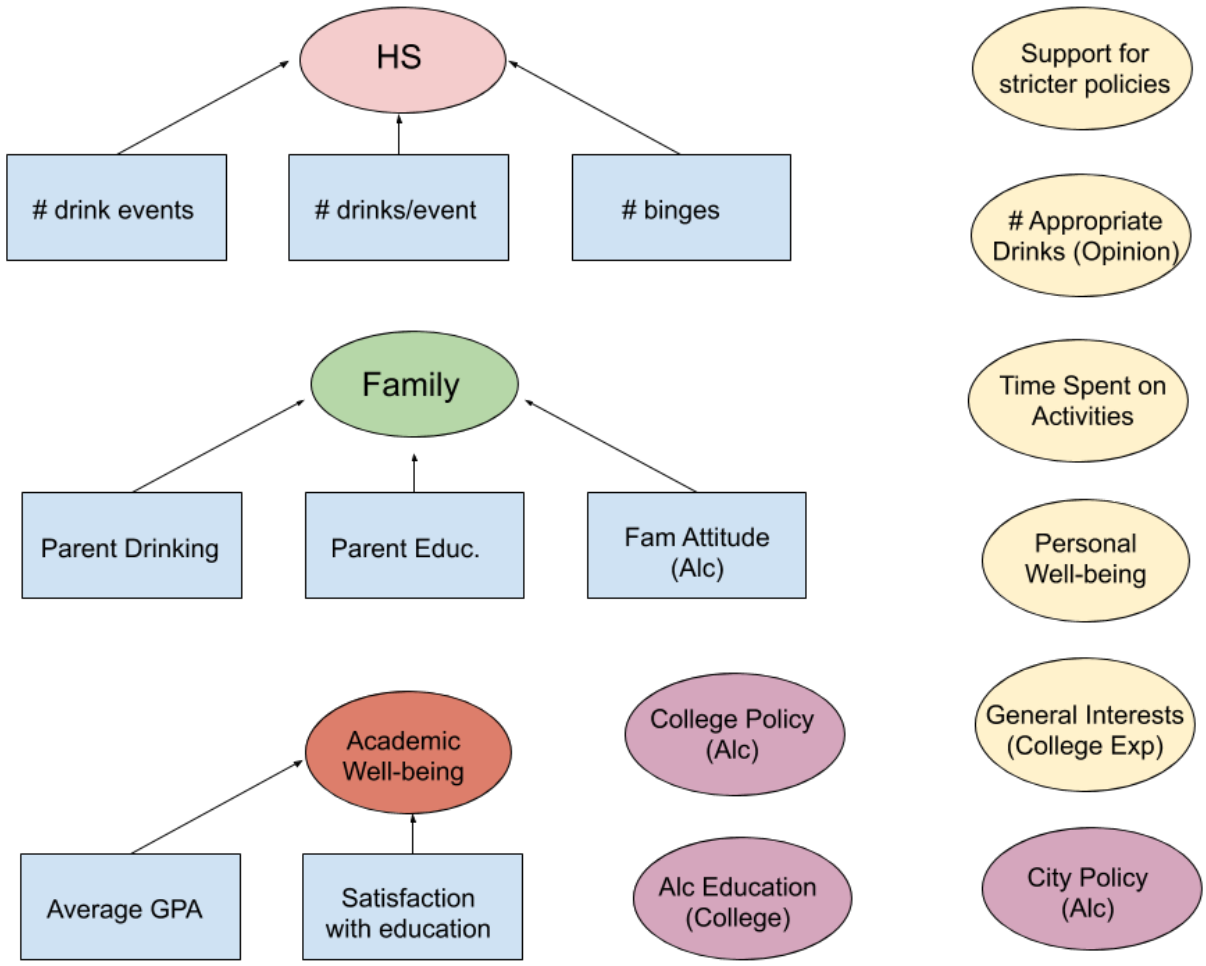

Figure 1: Figure shows latent factors (ovals) in the SEM for drinking behaviors. Squares show summary of actual questions in the survey.

Figures

Appendix

Data Cleaning

Reading In Data

Split file by number of columns for each variable in Record_layout.txt file. Spaces are missing values.

```
read_cas <- function(folder, file, recordfile) {
  skip <- grep("--------", readLines(unz(folder, recordfile)))
  record_layout <- read.table(unz(folder, recordfile),
                              fill = TRUE, skip = skip, header = FALSE)
  colnames(record_layout) <- c("variable_name", "start_col", "end_col", "type")
  record_layout <- record_layout %>% filter(!is.na(end_col))
  widths <- as.numeric(as.character(record_layout$end_col)) - as.numeric(as.character(record_layout$sta
  cas <- read.fwf(unz(folder, file), widths = widths, header    = FALSE,
                  col.names = record_layout$variable_name)
  return(cas)
}
cas97 <- read_cas(folder = "Harvard_CAS_1997.zip",
                  file = "Harvard_CAS_1997/DS0001/03163-0001-Data.txt",
                  recordfile = "Harvard_CAS_1997/DS0001/03163-0001-Record_layout.txt")
```

```
summary_na<-function(df) {
  na_count <- colSums(apply(df, 2, is.na))
  data.frame(name = colnames(df), na_pct = na_count/nrow(df)) %>% arrange(desc(na_pct))
}
```

The survey asks students to skip some questions on purpose. We manually go through the survey to impute these values and drop remaining missing entries.

Section A

```
# See which vars have lots of NA's
cas97 %>%
  dplyr::select(which(grepl("^A", colnames(.)))) %>%
  summary_na()

# Fill in NA's that mean 0
cas97 <- cas97 %>%
  mutate(A8_answered = cas97 %>%
                       dplyr::select(which(grepl("^A8_", colnames(.)))) %>%
                       rowSums(na.rm = TRUE)) %>%
  mutate(A8_1 = ifelse(is.na(A8_1) & A8_answered > 0, 0, A8_1),
         A8_2 = ifelse(is.na(A8_2) & A8_answered > 0, 0, A8_2),
         A8_3 = ifelse(is.na(A8_3) & A8_answered > 0, 0, A8_3),
         A8_4 = ifelse(is.na(A8_4) & A8_answered > 0, 0, A8_4)) %>%
  # A6 should be dummified
  mutate(ones = 1) %>%
  tidyr::pivot_wider(names_from = A6,
                     values_from = ones,
                     values_fill = list(ones = 0),
```

```
                      names_prefix = "A6_") %>%
  # Some NA in A7 means student lived off campus
  mutate(A7 = ifelse(is.na(A7) & A6_5 == 1, 0, A7)) %>%
  # Drop A4, A5 that are transfer questions, otherwise have to dummify
  dplyr::select(-A4, -A5, -A8_answered, -A8)

# How many complete cases left?
cas97 %>%
  dplyr::select(which(grepl("^A", colnames(.)))) %>%
  filter(complete.cases(.)) %>%
  nrow()
```

Section B

No skipped questions

For B10, an indicator is an option, fill in zero for than NA

```
# Which columns in this section have lots of NA?
cas97 %>%
  dplyr::select(which(grepl("^B", colnames(.)))) %>%
  summary_na()

# Fill in NA that actually means 0
cas97 <- cas97 %>%
  mutate(B10_answered = cas97 %>%
                        dplyr::select(which(grepl("^B10_", colnames(.)))) %>%
                        rowSums(na.rm = TRUE)) %>%
  mutate(B10_1 = ifelse(is.na(B10_1) & B10_answered > 0, 0, B10_1),
         B10_2 = ifelse(is.na(B10_2) & B10_answered > 0, 0, B10_2),
         B10_3 = ifelse(is.na(B10_3) & B10_answered > 0, 0, B10_3),
         B10_4 = ifelse(is.na(B10_4) & B10_answered > 0, 0, B10_4),
         B10_5 = ifelse(is.na(B10_5) & B10_answered > 0, 0, B10_5),
         B10_6 = ifelse(is.na(B10_6) & B10_answered > 0, 0, B10_6),
         B10_7 = ifelse(is.na(B10_7) & B10_answered > 0, 0, B10_7),
         B10_8 = ifelse(is.na(B10_8) & B10_answered > 0, 0, B10_8)) %>%
  dplyr::select(-B10_answered)

# How many complete cases left?
cas97 %>%
  dplyr::select(which(grepl("^B", colnames(.)))) %>%
  filter(complete.cases(.)) %>%
  nrow()
```

Section C

Section D

```
# Which columns in this section have lots of NA?
cas97 %>%
  dplyr::select(which(grepl("^D\\d", colnames(.)))) %>%
  summary_na()
```

```
cas97 <- cas97 %>%
  dplyr::select(-D7_1, -D7_2, -D7_3, -D7_4)

# How many complete cases left?
cas97 %>%
  dplyr::select(which(grepl("^D\\d", colnames(.)))) %>%
  filter(complete.cases(.)) %>%
  nrow()
```

Section E

Possible groups to consider: Female vs Male, Younger vs Older than 21 years old

```
# Which columns in this section have lots of NA?
cas97 %>%
  dplyr::select(which(grepl("^E\\d", colnames(.)))) %>%
  summary_na()

cas97 <- cas97 %>%
  # Fill NA resulted from dummifying vars E23
  mutate(E23_answered = cas97 %>%
                        dplyr::select(which(grepl("^E23_", colnames(.)))) %>%
                        rowSums(na.rm = TRUE)) %>%
  mutate(E23_1 = ifelse(is.na(E23_1) & E23_answered > 0, 0, E23_1),
         E23_2 = ifelse(is.na(E23_2) & E23_answered > 0, 0, E23_2),
         E23_3 = ifelse(is.na(E23_3) & E23_answered > 0, 0, E23_3)) %>%
  # Fill E27 with 0 for people less than 21 years old
  mutate(E27_A = ifelse(is.na(E27_A) & AGELT21 == 1, 0, E27_A)) %>%
  dplyr::select(-E27_B, -E27_C) %>%
  # Combine E24, 25, 26: Did you get seriously injured within 6 hours of drinking: fill 0 in NA E25
  replace_na(list(E25 = 0)) %>%
  # E10-12 skipped if never had sex. Can combine E9 and E11, drop E10, E12
  mutate(E9n11 = ifelse(is.na(E11) & E9 == 1, 0, E11 + 1)) %>%
  dplyr::select(-E9, -E10, -E12, -E11) %>%
  # E13-15 skipped if male
  replace_na(list(E13 = 0, E14 = 0, E15 = 0)) %>%
    #mutate(sex_assualt = ifelse(D4_H > 1 | D4_I > 1 | E15 > 1,max(D4_H,D4_I,E15, na.rm = TRUE), 1))
  dplyr::select(-E24, -E26, -E23_answered, -E23) %>%
  #create sexual assault due to drinking variable
  mutate(sex_assualt_ind = ifelse(D4_H > 1 | D4_I > 1 | E15 > 0, 1, 0))
```

Section F

Everything seems fine

```
# Which columns in this section have lots of NA?
cas97 %>%
  dplyr::select(which(grepl("^F", colnames(.)))) %>%
  summary_na()

# Remove F70/30FRND
cas97 <- cas97 %>%
  dplyr::select(-F70FRND, -F30FRND)
```

Section G

```
cas97 <- cas97 %>%
  mutate(G3_answered = cas97 %>%
                       dplyr::select(which(grepl("^G3_", colnames(.)))) %>%
                       rowSums(na.rm = TRUE)) %>%
  mutate(G3_1 = ifelse(is.na(G3_1) & G3_answered > 0, 0, G3_1),
         G3_2 = ifelse(is.na(G3_2) & G3_answered > 0, 0, G3_2),
         G3_3 = ifelse(is.na(G3_3) & G3_answered > 0, 0, G3_3),
         G3_4 = ifelse(is.na(G3_4) & G3_answered > 0, 0, G3_4),
         G3_5 = ifelse(is.na(G3_5) & G3_answered > 0, 0, G3_5),
         ) %>%
  dplyr::select(-G3_answered, -G3) %>%
  mutate(ones=1) %>%
  pivot_wider(names_from = G4,
              values_from = ones,
              values_fill = list(ones=0),
              names_prefix = "G4_") %>%  #indicators for religion
  mutate(G14_NA = as.numeric(G14==5),
         G14_none = as.numeric(G14==6),
         G15_NA = as.numeric(G15==5),
         G15_none = as.numeric(G15==6)) %>% #indicators for don't know/NA, parental drinking
  #indicator for no family agreement about drinking
  mutate(G16_none=as.numeric(G16==4)) %>%
  # Change "I don't know" into 0
  mutate(G14 = ifelse(G14 == 8, 0, G14),
         G15 = ifelse(G14 == 8, 0, G15),
         G16 = ifelse(G16 == 1, 0, G16),
         G16 = ifelse(G16 == 4, 1, G16)
         )
# Which columns in this section have lots of NA?
cas97 %>%
  dplyr::select(which(grepl("^G\\d", colnames(.)))) %>%
  summary_na()
```

Here we check for the number of complete cases after the preprocessing step above:

```
cas97 %>%
  dplyr::select(which(grepl("^[[:upper:]]\\d", colnames(.)))) %>%
  dplyr::select(which(!grepl("^C\\d", colnames(.)))) %>% # Remove Section C for now
  filter(complete.cases(.)) %>%
  dim()
```

We only have about 7100 observations for 212 variables

```
mean(is.na(cas97$DRINKCAT))
```

There are few missing values for the response DRINKCAT

Multicolinearity

```
cas97_noc <- cas97 %>%
  dplyr::select(which(grepl("^[[:upper:]]\\d", colnames(.)))) %>%
  dplyr::select(which(!grepl("^C\\d", colnames(.)))) %>% # Remove Section C for now
  filter(complete.cases(.)) %>%
  dplyr::select(-A6_NA)
```

```
sort(apply(cas97_noc, 2 , sd))
```

Drop A6_NA in complete cases, because the standard deviation is 0.

Let's check the correlation between the variables

```
plot_cormat <- function(X) {
  cormat <- round(cor(X), 2)
  melted_cormat <- reshape2::melt(cormat)
  ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() + scale_fill_gradient2(low = "blue", mid = "white", high = "red")
}

plot_cormat(cas97_noc)
```

```
cormat <- round(cor(cas97_noc), 2)
covvec <- cormat[upper.tri(cormat, diag = FALSE)]
hist(covvec, breaks = 50)
reshape2::melt(cormat) %>%
  filter(Var1 != Var2) %>%
  arrange(value)

check_cor <- function(df, desc = FALSE) {
  cormat <- round(cor(df), 2)
  if (desc) {
    reshape2::melt(cormat) %>%
    filter(Var1 != Var2) %>%
    arrange(desc(value))
  } else {
    reshape2::melt(cormat) %>%
    filter(Var1 != Var2) %>%
    arrange(value)
  }

}
```

EDA Graphs

Elastic Net Model