Homework 6

JP

2/14/2020

Homework Assignment #6 In this assignment you will again be working with a subset of the publicly available version of the data from the National Longitudinal Study of Adolescent to Adult Health (Add Health). This data can be found on the course website in Canvas under the “Files” tab. Look for data file “AddHealth.dta.” Use the STATA commands handout (also available in the Assignments tab) to help you complete this assignment. On the due date please hand in a hardcopy of your STATA output and answers to all questions in this assignment. You can either hand in a document with your answers (e.g., Word doc) with STATA output attached or you can incorporate answers to the questions directly into the STATA output in the form of comments. I prefer typed assignments in Times New Roman size 12 or Arial size 11. The following is a basic description of variables you will be using in this assignment. § aid = a unique id number assigned to each adolescent respondent § schoolid = a unique id number assigned to each school in the data set § smoke\_30days\_w1 = a (semi) continuous measure of the number of days the student smoked cigarettes in the past 30 days § cesd = a continuous measure of depression (score on the CESD survey) § sex = adolescent sex (it is unclear whether this survey item was more closely measuring sex or gender. For our purposes assume it measures gender.) o 1 = male o 2 = female § parent\_highestedu = highest educational attainment of either parent/guardian/parent-figure of adolescents in the sample: o 1 = Less than high school (no HS degree) o 2 = Completed high school or equivalent o 3 = Some college (no degree) o 4 = College degree or more § age\_w1 = age in years (continuous measure) reported at wave 1

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.0 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readstata13)

## Warning: package 'readstata13' was built under R version 3.6.2

library(lme4)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(optimx)

## Warning: package 'optimx' was built under R version 3.6.2

library(lmerTest)

## Warning: package 'lmerTest' was built under R version 3.6.2

##   
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':  
##   
## lmer

## The following object is masked from 'package:stats':  
##   
## step

library(dfoptim)  
  
options(max.print = 99999)  
options(scipen = 999)  
  
getwd()

## [1] "E:/UO/R Projects/SOC 613/scripts"

set.seed(232020)  
  
data <- read.dta13("E:/UO/R Projects/SOC 613/data/AddHealth.dta")  
  
data <- data %>%   
 dplyr::select(aid,   
 schoolid,   
 smoke\_30days\_w1,  
 cesd,  
 sex,  
 parent\_highestedu,  
 age\_w1) %>%   
 rename(student = aid,  
 school = schoolid,  
 smoke = smoke\_30days\_w1,  
 depression = cesd,  
 parent\_ed = parent\_highestedu,  
 age = age\_w1)  
  
colnames(data)

## [1] "student" "school" "smoke" "depression" "sex"   
## [6] "parent\_ed" "age"

Tasks: 1) Briefly explain the difference between “binary” and “binomial” outcomes.

Answer: Binary outcomes are when data are 0’s and 1’s and the outcome is the probability of having something (yes) for an individual. Binomial outcomes are cells of proporitions in a group of people.

1. Briefly explain the difference between “Odds” and “Odds Ratio.”

Answer: Odds are the values of probabilities of yes divided by the probabilities of no. Odds ratios are the odds of one group divided by the odds of the other group to make comparisons.

1. Smoking:
2. Generate a variable called “smoker” which = 1 if the student smoked on 1 or more days in the past 30 during wave 1, and = 0 if not.

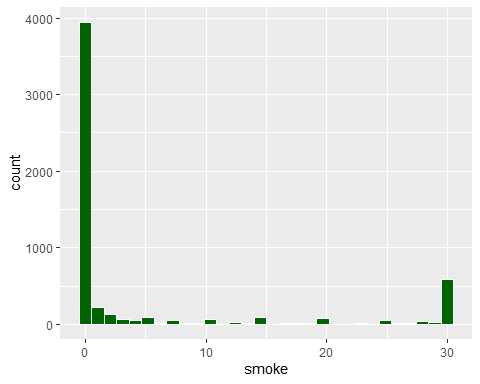
describe(data$smoke, na.rm = TRUE)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 5591 4.92 10.08 0 2.4 0 0 30 30 1.85 1.7  
## se  
## X1 0.13

data %>%   
ggplot(aes(smoke)) +  
 geom\_histogram(color = 'white', fill = 'darkgreen')

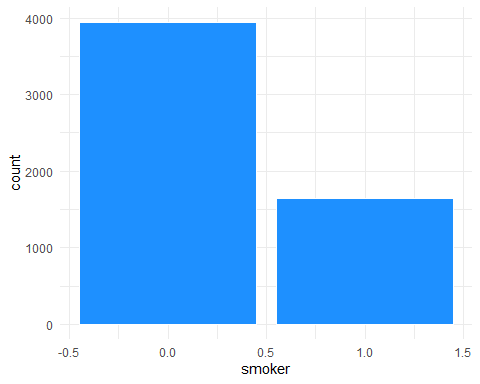
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 913 rows containing non-finite values (stat\_bin).



# Answer: 3A  
data <- data %>%   
 mutate(smoker = case\_when(smoke <= 0 ~ 0,  
 smoke > 0 ~ 1))  
  
data %>%   
ggplot(aes(smoker)) +  
 geom\_bar(color = 'white', fill = 'dodgerblue') +  
 theme\_minimal()

## Warning: Removed 913 rows containing non-finite values (stat\_count).



# Answer 3B  
data %>%   
 drop\_na(smoker) %>%   
 group\_by(smoker) %>%   
 summarize(n = n()) %>%   
 mutate(freq = n/sum(n))

## # A tibble: 2 x 3  
## smoker n freq  
## <dbl> <int> <dbl>  
## 1 0 3944 0.705  
## 2 1 1647 0.295

1. Obtain statistics on “smoker” – what proportion of respondents are past-30 day smokers in wave 1?

Answer: .29 or 29% of respondents are past smokers.

1. Is “smoker” a binary or binomial outcome?

Answer: Smoker is a binary outcome.

1. Model 1: Write a logistic Random Intercepts model, where adolescents (level 1) are nested in schools (level 2). The outcome of interest is “smoker.” Include the following FE predictors: “female” and “parent education” (reference level = college degree or more). When writing the model be careful to include all steps (micro, macro, combined) and with the final combined model be sure to specify whether the model is Bernoulli or binomial, and what the variance at each level is.

Micro:

Macro 1:(Intercept)

1. Using STATA, fit Model 1 and request the “betas” be provided in the form of Odds and Odds Ratios. Then answer the following questions using the results:

data <- data %>%   
 mutate(parent\_ed = recode(parent\_ed, '1' = 'less\_than\_hs',  
 '2' = 'hs\_grad',  
 '3' = 'some\_college',  
 '4' = 'college\_degree'),  
 sex = recode(sex, '1' = 'male',  
 '2' = 'female')) %>%   
 mutate(parent\_ed = relevel(as.factor(parent\_ed), ref = 'college\_degree'),  
 sex = relevel(as.factor(sex), ref = 'male'))  
  
model1 <- glmer(smoker ~ sex + parent\_ed + (1 |school),   
 data = data,   
 family = binomial(link = "logit"))  
summary(model1)

## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: binomial ( logit )  
## Formula: smoker ~ sex + parent\_ed + (1 | school)  
## Data: data  
##   
## AIC BIC logLik deviance df.resid   
## 6464.6 6504.3 -3226.3 6452.6 5507   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.2067 -0.6784 -0.5058 1.1205 2.9877   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## school (Intercept) 0.3739 0.6115   
## Number of obs: 5513, groups: school, 132  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.28822 0.08733 -14.750 < 0.0000000000000002 \*\*\*  
## sexfemale -0.08228 0.06167 -1.334 0.18212   
## parent\_edhs\_grad 0.49409 0.08407 5.877 0.00000000418 \*\*\*  
## parent\_edless\_than\_hs 0.38160 0.11974 3.187 0.00144 \*\*   
## parent\_edsome\_college 0.40689 0.08140 4.999 0.00000057771 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) sexfml prnt\_dh\_ prn\_\_\_  
## sexfemale -0.365   
## prnt\_dhs\_gr -0.474 0.004   
## prnt\_dlss\_\_ -0.361 0.010 0.383   
## prnt\_dsm\_cl -0.480 0.022 0.510 0.362

# cc\_model1 <- confint(model1, parm="beta\_")   
cc\_model1\_wald <- confint(model1, parm = "beta\_", method = "Wald")  
# ctab\_model1 <- cbind(est=fixef(model1), cc\_model1)  
ctab\_model1\_wald <- cbind(est = fixef(model1), cc\_model1\_wald)  
  
ctab\_model1\_wald

## est 2.5 % 97.5 %  
## (Intercept) -1.28822047 -1.4593938 -1.11704709  
## sexfemale -0.08228076 -0.2031478 0.03858623  
## parent\_edhs\_grad 0.49408711 0.3293065 0.65886776  
## parent\_edless\_than\_hs 0.38160468 0.1469090 0.61630037  
## parent\_edsome\_college 0.40689191 0.2473461 0.56643769

rtab\_model1\_wald <- exp(ctab\_model1\_wald)  
  
# Answer: Odds & Odds Ratios  
print(rtab\_model1\_wald, digits = 3)

## est 2.5 % 97.5 %  
## (Intercept) 0.276 0.232 0.327  
## sexfemale 0.921 0.816 1.039  
## parent\_edhs\_grad 1.639 1.390 1.933  
## parent\_edless\_than\_hs 1.465 1.158 1.852  
## parent\_edsome\_college 1.502 1.281 1.762

1. What is the Odds being a smoker among males with a parent who has completed a college degree or more?

Answer: The odds of being a smoker that is a male with a parent with a college degree or more is .28

1. How might we interpret the Odds Ratio associated with the lowest education level (less than high school degree)?

Answer: Adolescents with parents that have less than a high school degree had 1.47 times the odds of reporting having smoked in the last 30 days than adolescents with parents that have a parent with a college degree or more.

1. Regardless of whether or not it is statistically significant, how would we interpret the Odds Ratio for “female” in this model?

Answer: Females had .92 times the odds of reporting having smoked in the last 30 days compared to male adolescents.

1. Is the OR for “female” statistically significant, and what does this mean?

Answer: The odds ratio for females was not statistically significant from males. This means that the odds of female adolescents reporting smoking is not significantly different from the odds of male adolescents reporting smoking.

1. Model 2: Write a linear Random Coefficients model for the continuous outcome CESD, where adolescents (level 1) are nested in schools (level 2). Include the following FE variables: female, parent education (reference level = college degree or more), and age\_w1. In addition, treat “parent education” as a random coefficient. Be sure to include all steps (micro, macro, Combined model) and to include any level 2 and/or level 1 variance-covariance matrices. When writing the model include covariances in the variance-covariance matrix.

Macro 1:(Intercept)

Macro 2a: (Slope for less than high school)

Macro 2b: (Slope for hs grad)

Macro 2c: (Slope for some college)

Combined:

Level 2 Variances: College degree or more

Level 2 Variances: less than high school

Level 2 Variances: high school degree

Level 2 Variances: some college

1. Using STATA, fit Model 2 but make the assumption that all covariances = 0. Then answer the following questions using the results.
2. What is the between-school variance in the reference category (those with parent education = college degree or more)? Show any formulas you use.

parent\_dum <- dummy.code(data$parent\_ed)  
  
data <- data.frame(data, parent\_dum)  
  
# colnames(data)  
# str(data)  
  
model2 <- lmer(depression ~ hs\_grad + less\_than\_hs + some\_college +  
 sex + age + (hs\_grad + less\_than\_hs + some\_college || school),  
 data = data,  
 REML = FALSE,  
 control = lmerControl(optimizer = 'Nelder\_Mead'))  
summary(model2)

## Linear mixed model fit by maximum likelihood . t-tests use  
## Satterthwaite's method [lmerModLmerTest]  
## Formula: depression ~ hs\_grad + less\_than\_hs + some\_college + sex + age +   
## (hs\_grad + less\_than\_hs + some\_college || school)  
## Data: data  
## Control: lmerControl(optimizer = "Nelder\_Mead")  
##   
## AIC BIC logLik deviance df.resid   
## 42352.1 42426.5 -21165.0 42330.1 6395   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.9965 -0.7034 -0.1872 0.5084 5.6774   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## school (Intercept) 0.55287 0.7436   
## school.1 hs\_grad 0.12640 0.3555   
## school.2 less\_than\_hs 3.16423 1.7788   
## school.3 some\_college 0.07629 0.2762   
## Residual 42.68472 6.5334   
## Number of obs: 6406, groups: school, 132  
##   
## Fixed effects:  
## Estimate Std. Error df t value Pr(>|t|)  
## (Intercept) 3.33869 0.79982 1313.32004 4.174 0.000031857319  
## hs\_grad 1.54653 0.22042 219.25142 7.016 0.000000000028  
## less\_than\_hs 2.66819 0.36807 79.77951 7.249 0.000000000238  
## some\_college 0.74319 0.21234 240.79061 3.500 0.000554  
## sexfemale 1.44821 0.16465 6380.46053 8.796 < 0.0000000000000002  
## age 0.43716 0.05071 1130.48262 8.620 < 0.0000000000000002  
##   
## (Intercept) \*\*\*  
## hs\_grad \*\*\*  
## less\_than\_hs \*\*\*  
## some\_college \*\*\*  
## sexfemale \*\*\*  
## age \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) hs\_grd lss\_t\_ sm\_cll sexfml  
## hs\_grad -0.081   
## less\_thn\_hs -0.031 0.270   
## some\_colleg -0.103 0.444 0.266   
## sexfemale -0.150 0.002 0.005 0.016   
## age -0.975 -0.038 -0.042 -0.018 0.044

as.data.frame(VarCorr(model2))

## grp var1 var2 vcov sdcor  
## 1 school (Intercept) <NA> 0.55286924 0.7435518  
## 2 school.1 hs\_grad <NA> 0.12640213 0.3555308  
## 3 school.2 less\_than\_hs <NA> 3.16423238 1.7788289  
## 4 school.3 some\_college <NA> 0.07628729 0.2762015  
## 5 Residual <NA> <NA> 42.68472205 6.5333546

Answer: The variation between schools for adolescents with parents that have a college degree or more is .553

Level 2 Variances: College degree or more

1. What is the between-school variance among those whose parents have education = some college? Show any formulas you use.

Answer: The variation between schools for adolescents with parents that have a some college is 1.237

1. What is the between-school variance among those whose parents have education = high school degree? Show any formulas you use.

Answer: The variation between schools for adolescents with parents that have a high school degree is 13.209

Level 2 Variances: high school degree

1. What is the between-school variance among those whose parents have education = less than high school? Show any formulas you use.

Answer: The variation between schools for adolescents with parents that have a less than a high school degree is .679

Level 2 Variances: less than high school

1. Which group (those with parent education of college degree plus, some college, high school degree, or less than high school degree) has the largest betweenschool variance in CESD?

Answer: The most variation in depression scores appears to be adolescents with parents that have a high school degree.

1. On Canvas there is a file called “Final Project Worksheet.” Download this and fill it out to the best of your abilities. When you submit your homework, include a printed copy of this document (so I can provide some feedback) AND email me a copy of it. The document is in Word, so you should be able to fill it out electronically. I would like the electronic copy so I can begin to keep records about the final project you are proposing. We will schedule one-on-one meetings after this assignment is due, and I will use this worksheet to formally approve your final project proposal. If you already know you will not be doing a final project, there is no need to complete this worksheet. If you are not certain, it may make sense to complete it just in case so you can think through the possibility.