Homework 7

JP

2/25/2020

Homework Assignment #7 In this assignment you will be working with two data sets. The first is the data analyzed by Gelman and Hill (2007) on the number of times the police stopped individuals for questioning or searching on the streets of New York City. They used data collected by the police over a 15-month period in 1998-1999 in which the stops were classified by ethnicity, type of suspected crime, and the precinct (area) in which the stop occurred. Look forthe data file “police.dta” under the “Files” tab on Canvas.

police <- read.dta13("E:/UO/R Projects/SOC 613/data/police.dta")

You will alsobe 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.”

health <- read.dta13("E:/UO/R Projects/SOC 613/data/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 eitherhand 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 police.dtavariables you will be using in this assignment.

* eth= ethnicity (1=black, 2=Hispanic, 3=white)
* crime= type of crime (1=violent crimes; 2=weapons offenses; 3=property crimes; 4=drug crimes)
* stops= number of police stops over 15-month period in 1998-1999
* pop= population size of each ethnic group in the precinct

The following is a basic description of Add Healthvariables 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
* bmi\_w1= a continuous measure of body mass index (BMI) calculated from adolescent height and weight in wave 1 of the 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.)o1 = maleo2 = female
* age\_w1= age in years (continuous measure) reported at wave 1

## Tasks:

1)Briefly explain the difference between an Incidence Rate and an Incidence Rate Ratio. Define each clearly, using examples if you wish.

Answer: An incidence rate is the probability of something over a period of time. For example, the amount of newly diagnosed cases of a disease within a year. Also referred to as intensity.

Incidence rate ratio is the comparison of incidence rates between two groups.

2)In the “police” <data:Our> main research question is whether racial/ethnic minorities were stopped more frequently by police in NYCfor suspected violent crimesduring the study period(crime=1).

Write a two-level Random Intercepts Poisson model for the outcome “stops.” In this model, nest racial/ethnic groups(level 1) within precincts(level 2). Include “eth” as a FE predictor in the model and specify “white” as the reference group. Since different racial/ethnic groups have different population sizes in each precinct, be sure to use “pop” as an offset in your model (i.e., natural log of “pop”).When you write themodel, be sure to include all steps (micro, macro, combined).

Micro:

Macro 1:(Intercept)

Combined:

3)Fit the model in STATA and request the “betas” in the form of Incidence Rates and Incidence Rate Ratios.

\*Hint: be sure to drop cases for other types of suspected crimes (i.e., keep only when crime ==1).

colnames(police)

## [1] "stops" "pop" "arrests" "precinct" "eth" "crime"   
## [7] "totp" "prblack" "prhisp"

police <- police %>%   
 filter(crime == 1) %>%   
 mutate(eth = as.factor(eth),  
 pop = as.numeric(pop),  
 eth = recode(eth, '1' = 'black',  
 '2' = 'hispanic',  
 '3' = 'white'),  
 eth = relevel(as.factor(eth), ref = 'white'))  
  
model <- lme4::glmer(stops ~ eth + (1 | precinct) + offset(log(pop)),  
 family = poisson(link = 'log'), data = police)  
summary(model)

## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: poisson ( log )  
## Formula: stops ~ eth + (1 | precinct) + offset(log(pop))  
## Data: police  
##   
## AIC BIC logLik deviance df.resid   
## 5072.9 5086.5 -2532.4 5064.9 221   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -10.0016 -1.6420 0.0077 2.7286 23.4459   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## precinct (Intercept) 0.601 0.7752   
## Number of obs: 225, groups: precinct, 75  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.25868 0.09146 -79.37 <0.0000000000000002 \*\*\*  
## ethblack 3.06127 0.02129 143.78 <0.0000000000000002 \*\*\*  
## ethhispanic 2.08324 0.02156 96.61 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) ethblc  
## ethblack -0.187   
## ethhispanic -0.176 0.792

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

## est 2.5 % 97.5 %  
## (Intercept) -7.258682 -7.437934 -7.079430  
## ethblack 3.061270 3.019538 3.103001  
## ethhispanic 2.083242 2.040978 2.125506

rtab\_model1\_wald <- exp(ctab\_model1\_wald)  
  
# Answer: Incidence rates & ratios  
print(rtab\_model1\_wald, digits = 3)

## est 2.5 % 97.5 %  
## (Intercept) 0.000704 0.000588 0.000842  
## ethblack 21.354652 20.481827 22.264672  
## ethhispanic 8.030460 7.698135 8.377132

4)Questions about the results:

a.What is the interpretation of the variance parameter STATA estimated?

Answer: There was a large amount of variation (.60) in the number of stops between precincts.

b.What is the interpretation of the constant?

Answer: The constant is the expected number of stops per white person in preccincts was .0007 stops in a median precinct.

c.Explain your findings related to our main research question using interpretations of appropriate beta parameters.

Answer: The expected number of stops in precincts was 21.35 times higher for blacks than for whites, and 8.03 tiems higher for hispanics than for whites.

5)In the Add Health [data: Our](data:Our) main research question is whether there is residual heteroscedasticity in the relationship between age \_w1 and BMI in wave 1.

a.Model 1: Write a two-level linear Random Slopesmodel nesting adolescents (level 1) in schools (level 2). Include “female” gender and “age\_w1” as FE predictors of interestand treat “age\_w1” as a random slope variable. You may assume that the level 2 covariance between random intercepts and slopes = 0.

Micro:

Macro 1:(Intercept)

Macro 2: (Slope)

Combined:

b.Model 2: Write a two-level linearRandom Slopesmodel nesting adolescents (level 1) in schools (level 2). Include “female” gender and “age\_w1” as FE predictors of interest, and treat “age\_w1” as a random slope variable. In this model, also allow for level 1 heteroscedasticity with the “age \_w1” variable.You may assume that the level 2 covariance between random intercepts and slopes = 0 and you may assume that the level 1 covariance = 0.

Micro:

Macro 1:(Intercept)

Macro 2: (Slope)

Combined:

c.Provide interpretations of the two individual-level variance parameters.

Answer: The two individual-level variance parameters are the variation within schools or between adolescents. Secondly, to examine heteroscedasticity, the second individual-level variance is the variation within schools by age of adolescents.

d.Explain what it means to treat age in three ways in this model (as a FE predictor, as a random variable at level 2, and as a random variable at level 1). In other words, why include all three?

Answer: Including the age three different ways allows for the researchers to address: how age is associated with BMI overall (as a fixed effect), how age is associated with BMI across different schools (as a random slope), and how age may vary between adolescents within each school on BMI (random variable at level 1).

6)Fit both models in STATA. Store the results using “estimates store”commands as you fit the models.

colnames(health)

## [1] "aid" "schoolid" "age\_w1"   
## [4] "race" "sex" "family\_income"   
## [7] "family\_paybills" "family\_assistancea" "family\_assistanceb"  
## [10] "family\_assistancec" "family\_assistanced" "family\_assistancee"  
## [13] "family\_assistancef" "multpov\_w1" "parent\_highestedu"   
## [16] "hrs\_worked" "pay\_perhour" "bmi\_w1"   
## [19] "smoke\_30days\_w1" "mj\_30days\_w1" "hrs\_sleep"   
## [22] "gen\_health" "good\_health" "cesd"   
## [25] "smoke\_30days\_w2" "bmi\_w2" "mj\_30days\_w2"   
## [28] "bmi\_w3" "smoke\_30days\_w3" "mj\_30days\_w3"   
## [31] "smoke\_30days\_w4" "bmi\_w4" "multpov\_w4"

health <- health %>%   
 dplyr::select(bmi\_w1, age\_w1, sex, schoolid, aid)  
  
str(health)

## 'data.frame': 6504 obs. of 5 variables:  
## $ bmi\_w1 : num 40.3 20.6 NA 24.1 31.4 ...  
## $ age\_w1 : num 17 18 15 18 19 13 11 14 13 14 ...  
## $ sex : num 2 2 1 1 2 1 1 1 1 1 ...  
## $ schoolid: num 178 108 115 167 142 278 278 224 256 185 ...  
## $ aid : chr "57100270" "57101310" "57103171" "57103869" ...  
## - attr(\*, "datalabel")= chr ""  
## - attr(\*, "time.stamp")= chr ""  
## - attr(\*, "formats")= chr "%8s" "%12.0g" "%12.0g" "%12.0g" ...  
## - attr(\*, "types")= int 8 255 255 255 255 255 255 255 255 255 ...  
## - attr(\*, "val.labels")= Named chr "" "" "" "" ...  
## ..- attr(\*, "names")= chr "" "" "" "" ...  
## - attr(\*, "var.labels")= chr "RESPONDENT IDENTIFIER" "School ID" "Age at Wave 1 Interview" "Race / Ethnicity" ...  
## - attr(\*, "version")= int 111  
## - attr(\*, "label.table")= list()  
## - attr(\*, "expansion.fields")= list()  
## - attr(\*, "byteorder")= int 2  
## - attr(\*, "orig.dim")= int 6504 33

health\_model1 <- nlme::lme(bmi\_w1 ~ age\_w1 + sex,  
 random = list(schoolid = pdDiag(~ age\_w1)),  
 data = health,  
 na.action = na.omit,  
 method = 'ML')  
  
  
opt <- nlmeControl(maxIter = 50, opt = 'nlm')  
  
health\_model2 <- nlme::lme(bmi\_w1 ~ age\_w1 + sex,  
 random = list(schoolid = pdDiag(~ age\_w1)),  
 weights = varIdent(form = ~ 1 | age\_w1),  
 data = health,  
 na.action = na.omit,  
 method = 'ML',  
 control = opt)

## Warning in logLik.reStruct(object, conLin): Singular precision matrix in  
## level -1, block 1

## Warning in logLik.reStruct(object, conLin): Singular precision matrix in  
## level -1, block 4

## Warning in logLik.reStruct(object, conLin): Singular precision matrix in  
## level -1, block 13

summary(health\_model1)

## Linear mixed-effects model fit by maximum likelihood  
## Data: health   
## AIC BIC logLik  
## 36263.92 36304.4 -18125.96  
##   
## Random effects:  
## Formula: ~age\_w1 | schoolid  
## Structure: Diagonal  
## (Intercept) age\_w1 Residual  
## StdDev: 0.7611361 0.0002194813 4.276025  
##   
## Fixed effects: bmi\_w1 ~ age\_w1 + sex   
## Value Std.Error DF t-value p-value  
## (Intercept) 16.705930 0.5775704 6157 28.924492 0.0000  
## age\_w1 0.396247 0.0351829 6157 11.262468 0.0000  
## sex -0.272822 0.1086506 6157 -2.510999 0.0121  
## Correlation:   
## (Intr) age\_w1  
## age\_w1 -0.947   
## sex -0.322 0.041  
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -2.6449933 -0.6613791 -0.2139515 0.4251470 7.3554194   
##   
## Number of Observations: 6291  
## Number of Groups: 132

summary(health\_model2)

## Linear mixed-effects model fit by maximum likelihood  
## Data: health   
## AIC BIC logLik  
## 36239.85 36347.8 -18103.92  
##   
## Random effects:  
## Formula: ~age\_w1 | schoolid  
## Structure: Diagonal  
## (Intercept) age\_w1 Residual  
## StdDev: 0.7197526 0.01480077 4.262458  
##   
## Variance function:  
## Structure: Different standard deviations per stratum  
## Formula: ~1 | age\_w1   
## Parameter estimates:  
## 16 14 13 18 17 15 19   
## 1.0000000 0.9116710 0.9374813 1.0334199 1.0044682 1.0820790 1.1584940   
## 12 20 21 11   
## 1.0189303 0.9286994 0.9998891 1.6046800   
## Fixed effects: bmi\_w1 ~ age\_w1 + sex   
## Value Std.Error DF t-value p-value  
## (Intercept) 16.554263 0.5739515 6157 28.842616 0.0000  
## age\_w1 0.404069 0.0351861 6157 11.483772 0.0000  
## sex -0.260816 0.1079159 6157 -2.416846 0.0157  
## Correlation:   
## (Intr) age\_w1  
## age\_w1 -0.947   
## sex -0.322 0.041  
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -2.6796274 -0.6630158 -0.2089098 0.4302554 6.9926133   
##   
## Number of Observations: 6291  
## Number of Groups: 132

# install.packages('nlmeControl')  
anova(health\_model1, health\_model2)

## Model df AIC BIC logLik Test L.Ratio p-value  
## health\_model1 1 6 36263.92 36304.4 -18125.96   
## health\_model2 2 16 36239.85 36347.8 -18103.92 1 vs 2 44.07278 <.0001

7)Using a LR test, compare the two models –was it “worth” allowing for heteroscedasticity in this case? Describe what these results mean.

Answer: It was worth it to investigate heteroscedasticity within schools. By addressing that there is heteroscedasticity within schools, this then provides evidence that findings of variation between schools may actually be attributed to differences in age between adolescents and within schools. So seeing variation in the slopes of age on BMI between schools may actually be attributed to differences within schools.

Final Project: Createtable “shells”that you willeventuallyfill with results fromyour final project. Tables shells are set up as you would present your results, but “empty”of results at this pointbecause you have not yet run your models.This is an opportunity to think ahead to what results you will want to include and discuss in your papers. Typically,“Table 1” includes demographic data about your sample, as well as information about the data structure (number of level 1 and level 2 units, number of level 1 units per level 2 units, etc). Tables 2 (and so on) will likely include the results of the various models you fit. Use this exercise to picture what comparisons you will wish to highlight. Consider whether you want to include things in the table like ICCs and likelihood ratios.

demographic <- data.frame(Variable = c('Obesity Rates',  
 'Percent Rurality',  
 'Median Household Income',  
 'Population',  
 'Violent Crime Rates',  
 'Physical Inactivity'),  
 Overall = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2011 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2012 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2013 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2014 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2015 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2016 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2017 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2018 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'),  
 Year\_2019 = c('XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX',  
 'XX'))  
Demographic Table

## Variable Overall Year\_2011 Year\_2012 Year\_2013 Year\_2014  
## 1 Obesity Rates XX XX XX XX XX  
## 2 Percent Rurality XX XX XX XX XX  
## 3 Median Household Income XX XX XX XX XX  
## 4 Population XX XX XX XX XX  
## 5 Violent Crime Rates XX XX XX XX XX  
## 6 Physical Inactivity XX XX XX XX XX  
## Year\_2015 Year\_2016 Year\_2017 Year\_2018 Year\_2019  
## 1 XX XX XX XX XX  
## 2 XX XX XX XX XX  
## 3 XX XX XX XX XX  
## 4 XX XX XX XX XX  
## 5 XX XX XX XX XX  
## 6 XX XX XX XX XX

models <- data.frame(Variable = c(' ',  
 'Fixed effect estimates',  
 'Intercept',  
 'Adult Obesity',  
 'Percent Rurality',  
 'Median Household Income',  
 'Population',  
 'Release Year',  
 'Violent Crime',  
 'Random effect estimates',  
 'County',  
 'Violent Crime',  
 'Fit statistics',  
 'AIC/BIC',  
 'Intraclass correlation coefficient',  
 'County'),  
 Model1 = c('No Predictors',  
 ' ',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 ' ',  
 'value and CI',  
 ' ',  
 ' ',  
 'fit values',  
 ' ',  
 'ICC'),  
 Model2 = c('County-Level Predictors',  
 ' ',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 ' ',  
 'value and CI',  
 ' ',  
 ' ',  
 'fit values',  
 ' ',  
 'ICC'),  
 Model3 = c('Random Slopes',  
 ' ',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 'value and CI',  
 ' ',  
 'value and CI',  
 'value and CI',  
 ' ',  
 'fit values',  
 ' ',  
 ' '))

Model Table

models

## Variable Model1  
## 1 No Predictors  
## 2 Fixed effect estimates   
## 3 Intercept value  
## 4 Adult Obesity   
## 5 Percent Rurality   
## 6 Median Household Income   
## 7 Population   
## 8 Release Year value  
## 9 Violent Crime   
## 10 Random effect estimates   
## 11 County value   
## 12 Violent Crime   
## 13 Fit statistics   
## 14 AIC/BIC fit values  
## 15 Intraclass correlation coefficient   
## 16 County ICC  
## Model2 Model3  
## 1 County-Level Predictors Random Slopes  
## 2   
## 3 value value   
## 4 value value   
## 5 value value   
## 6 value value   
## 7 value value   
## 8 value value   
## 9 value value   
## 10   
## 11 value value   
## 12 value   
## 13   
## 14 fit values fit values  
## 15   
## 16 ICC