Homework 2

JP

1/14/2020

# Homework 2

https://rpubs.com/rslbliss/r\_mlm\_ws

https://rpsychologist.com/r-guide-longitudinal-lme-lmer

Google Search https://www.google.com/search?q=two+level+random+intercept+model+r&oq=two+level+random+intercept+model+r&aqs=chrome..69i57j33.18238j0j4&sourceid=chrome&ie=UTF-8

https://mran.microsoft.com/snapshot/2017-04-11/web/packages/sjPlot/vignettes/sjplmer.html

# install.packages('readstata13')  
  
# install.packages('stargazer')  
  
getwd()

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

# get file directory path  
  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.3

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

## v ggplot2 3.1.1 v purrr 0.3.2  
## v tibble 2.1.1 v dplyr 0.8.1  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'stringr' was built under R version 3.5.2

## Warning: package 'forcats' was built under R version 3.5.3

## -- 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 3.5.3

library(betas)  
library(lme4)

## Warning: package 'lme4' was built under R version 3.5.3

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':  
##   
## expand

library(psych)

## Warning: package 'psych' was built under R version 3.5.3

##   
## Attaching package: 'psych'

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

library(knitr)

## Warning: package 'knitr' was built under R version 3.5.3

library(kableExtra)

## Warning: package 'kableExtra' was built under R version 3.5.3

##   
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':  
##   
## group\_rows

data <- read.dta13("E:/UO/R Projects/SOC 613/data/AddHealth.dta")  
  
data <- data %>%   
 janitor::clean\_names() %>%   
 dplyr::select(aid, schoolid,  
 bmi\_w1,  
 sex,  
 parent\_highestedu,  
 age\_w1)

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
* 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.)
* 1 = male
* 2 = female
* parent\_highestedu = highest educational attainment of either parent/guardian/parent-figure of adolescents in the sample:
* 1 = Less than high school (no HS degree)
* 2 = Completed high school or equivalent
* 3 = Some college (no degree)
* 4 = College degree or more
* age\_w1 = age in years (continuous measure) reported at wave 1

1. Draw diagrams using the Goldstein notation we learned in class to represent the following data structures:
2. Children (level 1) nested in Households (level 2) which are nested in Counties (level 3) which in turn are nested in States (level 4).
3. Repeated surveys of the same neighborhoods over time. (Hints: who lives in a given neighborhood changes over time, and we are interested in how the neighborhood changes, not in how people change. Imagine for this example that the outcome of interest was a property of the neighborhood itself, such as population density.).
4. Repeated surveys of the same respondents over time.
5. Repeated surveys of the same respondents over time, and those respondents are nested in workplaces. Assume the workplace someone is nested in does not change over time.
6. For each case, draw within one graphical model the regression line(s) for a continuous outcome (workplace chemical exposure) and the continuous predictor (years of education):
7. No overall relationship between years of education and level of workplace chemical exposure. Substantial differences in workplace chemical exposure between six workplaces, but the relationship between education and exposure does not vary by workplace.
8. A negative overall relationship between years of education and level of workplace chemical exposure (i.e., more years of education is associated with reduced workplace chemical exposure). Substantial differences in workplace chemical exposure between six workplaces, but the relationship between education and exposure does not vary by workplace.
9. A negative overall relationship between years of education and level of workplace chemical exposure (i.e., more years of education is associated with reduced workplace chemical exposure). Substantial differences in workplace chemical exposure between six workplaces, AND the relationship between education and exposure DOES vary by workplace.
10. Consider your own research interests and use an example that interests you to answer these questions.
11. Draw the diagram for a two-level model using Goldstein’s notation. Be sure to specify what the units are at level 1 and level 2.

states & counties

1. Give at least two examples of variables that are associated with EACH level.

states = median household income & percentage inactive

counties = prevalence of liquor stores & rates of violent crime

1. Write a two-level random intercept model based on this example, including each of the steps we covered (micro model, macro model, combined model). Be sure to specify what the outcome is and include at least one predictor (x variable) associated with level 1. Do not attempt to include a variable associated with level 2 (this is more advanced and we will cover it later).

Outcome is BMI and the one predictor is rates of violent crime

Micro model: within-state, between-counties

Macro model: between-counties

Combined model:

1. Using the Add Health data we will consider the case of adolescents (aid) nested in schools (schoolid). The outcome of interest is BMI at wave 1 (bmi\_w1). In all steps if you are asked to write a model be sure to include the specification of the random effect terms. Also if you are asked to write a multilevel model be sure to include all steps (micro model, macro model, combined model).

data <- data %>%   
 mutate(parent\_highestedu = recode(parent\_highestedu, '1' = 'less\_than\_hs',  
 '2' = 'hs\_grad',  
 '3' = 'some\_college',  
 '4' = 'college\_degree'),  
 sex = recode(sex, '1' = 'male',  
 '2' = 'female')) %>%   
 mutate(parent\_highestedu = relevel(as.factor(parent\_highestedu), ref = 'college\_degree'),  
 sex = relevel(as.factor(sex), ref = 'male'))

1. Model 1: Write a single-level random intercept model for bmi\_w1 that includes the following predictors: female, age\_w1, and parent\_highestedu (reference level = college degree or more). What critical assumption does this model make that we could address in a multilevel model?

The critical assumption is

1. Model 2: Write a two-level random intercept model for bmi\_w1 that includes NO predictor variables (intercept only). Write interpretations (in words) for every parameter (betas, residuals, and sigma-squares) in the model (be sure to include interpretations for the beta parameter that is the outcome in the Macro model).

Micro model: (within-school, between-adolescent)

Macro model: (between-school)

Combined model:

Interpretations: \beta\_0 is the BMI for an adolescent (across all schools) \beta\_{0j} is the BMI for school j \epsilon\_{0ij} is the residual differntial BMI for adolescent i in school j \mu\_{0j} is the residual differential BMI for an adolescent in school j \sigma^2\_{\mu0} is the between-school variance of BMI \sigma^2\_{\epsilon0} is the within-school/between-person variance of BMI

1. Model 3: Write a two-level random intercept model for bmi\_w1 that includes the following predictors: female, age\_w1, and parent\_highestedu (reference level = college degree or more). Write interpretations (in words) for every parameter (betas, residuals, and sigma-squares) in the model (be sure to include interpretations for the beta parameter that is the outcome in the Macro model).

Micro model: (within-school, between-adolescent)

Macro model: (between-school)

Combined model:

Interpretations: \beta\_0 is the BMI for an adolescent (across all schools) \beta\_{0j} is the BMI for school j \epsilon\_{0ij} is the residual differntial BMI for adolescent i in school j \mu\_{0j} is the residual differential BMI for an adolescent in school j \sigma^2\_{\mu0} is the between-school variance of BMI \sigma^2\_{\epsilon0} is the within-school/between-person variance of BMI

1. Fit models 1, 2 and 3 using appropriate commands in STATA (reg or mixed) or in the software of your choice, and using the AddHealth.dta data set.

model1 <- lm(bmi\_w1 ~ sex + + age\_w1 + parent\_highestedu, data = data)  
summary(model1)

##   
## Call:  
## lm(formula = bmi\_w1 ~ sex + +age\_w1 + parent\_highestedu, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.785 -2.852 -0.959 1.757 33.868   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.67354 0.49554 31.629 < 2e-16 \*\*\*  
## sexfemale -0.26222 0.10990 -2.386 0.0171 \*   
## age\_w1 0.41908 0.03115 13.455 < 2e-16 \*\*\*  
## parent\_highesteduhs\_grad 0.60629 0.14107 4.298 1.75e-05 \*\*\*  
## parent\_highesteduless\_than\_hs 1.25930 0.19798 6.361 2.15e-10 \*\*\*  
## parent\_highestedusome\_college 0.53859 0.13833 3.894 9.98e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.324 on 6204 degrees of freedom  
## (294 observations deleted due to missingness)  
## Multiple R-squared: 0.03884, Adjusted R-squared: 0.03807   
## F-statistic: 50.14 on 5 and 6204 DF, p-value: < 2.2e-16

confint(model1)

## 2.5 % 97.5 %  
## (Intercept) 14.7020993 16.64497343  
## sexfemale -0.4776567 -0.04678839  
## age\_w1 0.3580223 0.48013961  
## parent\_highesteduhs\_grad 0.3297541 0.88283555  
## parent\_highesteduless\_than\_hs 0.8711929 1.64740883  
## parent\_highestedusome\_college 0.2674218 0.80976783

lm.beta::lm.beta(model1)

##   
## Call:  
## lm(formula = bmi\_w1 ~ sex + +age\_w1 + parent\_highestedu, data = data)  
##   
## Standardized Coefficients::  
## (Intercept) sexfemale   
## 0.00000000 -0.02973583   
## age\_w1 parent\_highesteduhs\_grad   
## 0.16801965 0.06054383   
## parent\_highesteduless\_than\_hs parent\_highestedusome\_college   
## 0.08507614 0.05494145

mse <- 4.324^2  
mse

## [1] 18.69698

model2 <- lmer(bmi\_w1 ~ 1 + (1|schoolid), data = data, REML = FALSE)  
summary(model2)

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: bmi\_w1 ~ 1 + (1 | schoolid)  
## Data: data  
##   
## AIC BIC logLik deviance df.resid   
## 36382.8 36403.1 -18188.4 36376.8 6288   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.6637 -0.6634 -0.2015 0.4359 7.2195   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## schoolid (Intercept) 1.033 1.016   
## Residual 18.511 4.302   
## Number of obs: 6291, groups: schoolid, 132  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 22.346 0.106 210.8

confint(model2)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.8516529 1.208471  
## .sigma 4.2275293 4.379530  
## (Intercept) 22.1333922 22.553766

# stargazer(model2, type = 'text', ci = TRUE, ci.level = 0.95, single.row = TRUE)  
# qf(p = .95, df1 = 1, df2 = 6490)

model3 <- lmer(bmi\_w1 ~ sex + + age\_w1 + parent\_highestedu + (1|schoolid), data = data, REML = FALSE)  
  
summary(model3)

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: bmi\_w1 ~ sex + +age\_w1 + parent\_highestedu + (1 | schoolid)  
## Data: data  
##   
## AIC BIC logLik deviance df.resid   
## 35759.7 35813.6 -17871.9 35743.7 6202   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.5716 -0.6598 -0.2034 0.4250 7.4050   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## schoolid (Intercept) 0.49 0.700   
## Residual 18.20 4.266   
## Number of obs: 6210, groups: schoolid, 132  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 16.27495 0.55167 29.501  
## sexfemale -0.26860 0.10905 -2.463  
## age\_w1 0.38547 0.03507 10.991  
## parent\_highesteduhs\_grad 0.41312 0.14641 2.822  
## parent\_highesteduless\_than\_hs 0.94181 0.20577 4.577  
## parent\_highestedusome\_college 0.42366 0.14021 3.022  
##   
## Correlation of Fixed Effects:  
## (Intr) sexfml age\_w1 prnt\_hghstdh\_ prn\_\_\_  
## sexfemale -0.144   
## age\_w1 -0.973 0.044   
## prnt\_hghstdh\_ -0.070 0.002 -0.048   
## prnt\_hghs\_\_ -0.021 0.007 -0.066 0.342   
## prnt\_hghstds\_ -0.096 0.015 -0.023 0.461 0.329

confint(model3)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.5506991 0.86769797  
## .sigma 4.1909541 4.34259945  
## (Intercept) 15.1864005 17.37239836  
## sexfemale -0.4823595 -0.05483376  
## age\_w1 0.3157683 0.45459542  
## parent\_highesteduhs\_grad 0.1234018 0.70242910  
## parent\_highesteduless\_than\_hs 0.5331566 1.34972397  
## parent\_highestedusome\_college 0.1479657 0.69928508

anova(model3)

## Analysis of Variance Table  
## Df Sum Sq Mean Sq F value  
## sex 1 165.43 165.43 9.0919  
## age\_w1 1 2343.41 2343.41 128.7886  
## parent\_highestedu 3 438.90 146.30 8.0404

1. What is the MSE from the single-level model? What parameter from the multilevel models does this most closely relate to? Compare the MSE from the single-level model to the equivalent parameter in Model 3—is it different, and if so then how so? What might account for any difference?

Conceptual Questions: a. What does it mean if a school has a negative value for �"#? What does it mean if that value is positive?

1. What does it mean if an individual adolescent has a negative value for �"%#? What does it mean if that value is positive?