HED 612

Homework #10

**Directions**:

* Write your name on this document
* If the questions below ask you to execute R commands, then copy all R syntax (indicated via Courier Font) into the R script
* The R script should have a #comment indicating what number question the R syntax refers to for this assignment
* *Submit your answers file along with your R script to the D2L Dropbox*

**Before you begin**:

*Download the ELS Data [If you did not complete this during class]*:

* Create a new data folder called “els”
  + hed612 >>> data >>> els
* Download the ELS Dataset from D2L (under Datasets)
  + Place the “els-stu-by-f2-select-vars.dta” dataset into the “els” folder you created in the previous step

*Create a new R Script for this homework assignment*

* Open the RProject you created last week (should be in your main hed612 folder)
* Once the RStudio window opens, within the R project session, open a new R Script
  + files >>> New File… >>> R Script
* Save the file as HW10\_lastname.R within lecture7 subfolder

*About the data*

The Education Longitudinal Study of 2002 (ELS) is a nationally representative, longitudinal study of 10th graders in 2002 and 12th graders in 2004. It follows students throughout secondary and postsecondary years. The study surveys students, their parents, math and English teachers, and school administrators. Variables capture information on student’s social background, home educational support system, school and classroom characteristics, employment, and academic outcomes.

**Questions about Confidence Intervals**

For these multivariate regression questions we are trying to understand what is the effect of hours per week spent on homework (X1= byhmwrk) on reading test scores (Y= bytxrstd), controlling for whether a student is an English is the student’s native language (X2= f1stlang) and student’s socioeconomic index composite score (X3= f1ses1).

Questions 1-5 will ask you to first clean some of these variables by creating “analysis versions.”

1. Within the R script created above: load the haven, tidyverse, labelled libraries; check that your directory is set to the R project; open the ELS data.

els <- read\_dta("data/els/els-stu\_by-f2-select-vars.dta")

1. Create a dependent variable (Y) called read\_testscr that recodes missing observations in bytxrstd as true NA rather than -8 using the following R syntax.

els <- els %>%

mutate(read\_testscr = ifelse(bytxrstd==-8, NA, bytxrstd))

1. Create the main independent variable of interest (X1) called hw\_hours that recodes missing observations in byhmwrk as true NA rather than -9/-8/-4 and top-coded 97/98/99 values to maximum hours reported using the following R syntax.

els <- els %>%

mutate(hw\_hours = recode(as.integer(byhmwrk),

`-9` = NA\_integer\_,

`-8` = NA\_integer\_,

`-4` = NA\_integer\_,

`97`= 26L,

`98` = 21L,

`99` = 26L))

1. Our first control variable, (X2= f1stlang), is clean, has no missing observations, and just needs to be created into a factor version of the original categorical variable to be ready for analysis.

attributes(els$f1stlang)

els %>% count(f1stlang)

els$f1stlang\_fac <- as.factor(els$f1stlang)

1. Our second control variable, (X3= f1ses1), has missing observations coded as -8. We need to convert these observations to true missing “NA” observations in R to get this continuous variable ready for analysis.

attributes(els$ f1ses1)

els %>% count(f1ses1)

els <- els %>%

mutate(f1ses1\_v2 = ifelse(f1ses1==-8, NA, f1ses1))

1. Run a regression with hw\_hours as the only independent variable and read\_testscr as the dependent variable.
   1. Write out the population regression model

Where:

= reading test scores

= hours of homework completed each week

* 1. Write out the OLS regression line (with and without estimates).
  2. Interpret

= 0.26; A one-hour increase in hours of homework completed each week is associated with, on average, a 0.26-point increase in reading test score

* 1. What is the predicted reading test score for a student that spends 10 hours per week on homework?

1. Run a regression with hw\_hours as the main independent variable of interest; read\_testscr as the dependent variable; with f1stlang\_fac and f1ses1\_v2 as covariates.
   1. Write out the population regression model

Where:

= reading test scores

= hours of homework completed each week

= 0/1 English is first language

(0= reference group, English is not first language)

(1= non-reference group, English is first language)

= SES composite score

* 1. Write out the OLS regression line (with and without estimates).
  2. Interpret

= 0.19; A one-hour increase in hours of homework completed each week is associated with, on average, a 0.19-point increase in reading test score, holding first language and SES constant.

* 1. What is the predicted reading test score for a student that spends 10 hours per week on homework, whose English is their native language, and has an SES index score of 1.5?

58.52

1. Based on the measures of Adjusted and Standard Error of the Regression, is the model in Question 6 or Question 7 better fitting? *[See Lecture 5 if you need to refresh on interpreting measures of model fit]*

The simple regression in Model 1 has an of 0.04243, indicating that homework hours alone explains 4% of the variation in reading test scores. On the other hand, the multivariate regression model in Model 2 has an of 0.2094, indication homework hours, first language, and SES explain 21% of the variation in reading test scores. Accordingly, Model 2 is the better fitting model.

The simple regression in Model 1 has a SER of 9.623, indicating that the average distance between reading test score of a random student and the reading test score predicted by the OLS regression is 9.623 points. On the other hand, the multivariate regression model in Model 2 has a SER of 8.743, indicating that the average distance between reading test score of a random student and the reading test score predicted by the OLS regression is 8.743 points. Because the distance between a random student’s reading test score and our prediction is closer in Model 2, Model 2 is the better fitting model.