HED 612

Homework #7

**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 HW7\_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**

Questions 1-6 will focus on interpreting and calculating confidence intervals for the effect of hours per week spent on homework (X= byhmwrk) on reading test scores (Y= bytxrstd). We will first clean these variables by creating “analysis versions” of the variables and then run the regression.

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 an independent variable (X) 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. Describe the dependent variable (read\_testscr) and independent variable of interest (hw\_hours) by running summary statistics and interpreting the mean of each variable.

Respondents, on average, have a reading test score of 50.53.

Respondents, on average, spend 10.31 hours a week on homework.

1. Run the regression in R using the ELS data. Interpret (using continuous X interpretation).

= .25966

On average, every additional hour per week spent on homework is associated with a 0.25966 point increase in reading test score.

1. Calculate 95% confidence interval for using and SE( ). Interpret this CI in words. Check your calculation using confint().

[0.2396876, 0.2796324]

*We are 95% confident that the population parameter 𝛽1 lies somewhere between 0.24 and 0.28*

**Questions about Interpreting Categorical X**

Questions 7-11 will focus on interpreting a categorical X variable. We will first clean a dummy variable of f1sex by creating an “analysis version” dummy variable and then run the regression for the effect of being female on reading test score in comparison to males (reference group).

1. Create a 0/1 dummy variable called female where 0 = male and 1=female

els <- els %>%

mutate(female= ifelse(f1sex==2, 1, 0))

1. Run regression of the effect of being female (X= female) on reading test score in comparison to males (Y= read\_testscr). *Note: convert the female variable to a factor variable first!*

els$female <- as.factor(els$female)

1. Write out the population regression model (label variables and identify what the reference and non-reference categories are)

Where:

* = reading test scores
* = 0/1 female,
  + - * 1 = female (non-reference group)
      * 0 = male (reference group)

1. Write out the OLS Prediction Line with estimates.
   1. What is the predicted value of reading score for females?
   2. What is the predicted value of reading score for males?

Female, X=1: = 51.2509

Male, X=0: =

1. What is the null and alternative hypothesis for the regression coefficient, , on female?
   1. Using an alpha-level of .05, do we reject the null hypothesis?
   2. Interpret in words.

; in other words, the difference in reading test scores for female identified respondents, in comparison to male identified respondents, is equal to zero

; in other words, the difference in reading test scores for female identified respondents, in comparison to male identified respondents, is not equal to zero

Yes, with a p-value of 0.000 that is well below our alpha-level of .05, we can reject the null hypothesis in favor of the alternative hypothesis.

* = the average difference in reading test scores for a female identified respondent as opposed to a male identified respondent is a 1.4547 increase in reading test scores