HED 612

Homework #5

**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 CA Data [If you did not complete this during class]*:

* Create a new data folder called “ca”
  + hed612 >>> data >>> ca
* Download the California Dataset from D2L (under Datasets)
  + Place the “caschool-v2” dataset into the “ca” 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 HW4\_lastname.R within lecture2 subfolder

*About the data*

The CA Schools dataset contains data on test performance, school characteristics and student demographic backgrounds. The data used here are from all 420 K-6 and K-8 districts in California with data available for 1998 and 1999. Each observation indicates one school district.

# Questions related to caschools-v2.dta

Our research question for this homework assignment is: What is the effect of percent of English language learners on district average income?

1. Within the R script created above: load the haven, tidyverse, labelled, and ggplot2 libraries; check that your directory is set to the R project; open the California data.

caschool<- read\_dta("data/ca/caschool-v2.dta")

1. Describe the dependent variable (avginc) and independent variable of interest (el\_pct) by running summary statistics and interpreting the mean of each variable. *Note: average district income is already in thousands!*

The district average income for the 420 school districts in the dataset is $15.317 (in $000s).

The average percent of English Language Learner enrollment across the 420 school districts in the dataset is 15.8% .

1. What is the definition of in words? Write out the mathematical formula for (just in terms of ESS and TSS, no need to include formulas for generating ESS and TSS)

𝑅2 measures the fraction of the variance in Y that is explained by X (and is not already explained by sample mean, 𝑌̄).

1. What is the definition of *Standard Error of the Regression* (SER)? Write the mathematical formula for SER. (just in terms of SSR, no need to include formula for generating SSR)

SER is, on average, how far away an actual observed value of Y is from the predicted value of Y for a random observation

SER =

1. Run a regression of the effect of percent of English language learners (X= el\_pct) on district average income (Y= avginc).

See R Script

1. Interpret the value for the regression in Q5. Calculate based on formula in Q3. [hint: you’ll need to run regression and use the anova() function]

= 0.09451; Percentage of English language learners explains 9.5% of the variation in district average income (not already explained by )

= = 0. 09450806

1. Interpret the SER value for the regression in Q5. Calculate SER based on on formula in Q4. [hint: you’ll need to run regression and use the anova() function]

SER = 6.884; Average distance between “average district income” of a random district and the “average district income” predicted by the OLS regression is 6.884

SER =

1. Using R, generate the standard deviation of district average income (avginc). What does the value of the standard deviation of district average income (avginc)in comparison to SER in Q7 imply about how much better the regression model predicts district average income than the sample mean of district average income?

Given the standard deviation of district average income is 7.22589 and SER is 6.884, the regression model does not result in a much better improvement in predicting district average income than the sample mean.