HED 612

Homework #3

**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**:

*RE\_DOWNLOAD the General Social Survey Data*:

* Create a new data folder called “gss”
  + hed612 >>> data >>> gss
* Download the California Dataset from D2L (under Datasets)
  + Place the “GSS2018.Rdata” dataset into the “gss” 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 HW3\_lastname.R within lecture2 subfolder

*About the data*

Since 1972, GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

# Questions related to GSS2018.Rdata

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 GSS data.

load("data/gss/GSS2018.Rdata")

For the remaining questions, we will explore the relationship between respondents’ income (realrinc) and socioeconomic index (sei10). *If you did not re-download the data from D2L you will not have these variables in your gss dataset!*

* realrinc is self-reported income (this is the continuous version of the categorical income variable we worked with in class).
* sei10 is a score from 0 to 100 (100 indicating higher SES) calculated based on respondents’ occupation, occupation prestige, income, and education (see [GSS Codebook G](http://gss.norc.org/documents/codebook/GSS_Codebook_AppendixG.pdf)).

1. In R, calculate the sample mean for realrinc and then for sei10. Interpret the mean for each variable in your own words. *Example code for this question can be found in Lecture 2’s r script.*
2. In R, generate a scatter plot where X= realrinc and Y= sei10. Interpret the plot in your own words. (Hint: Are there discontinuities? Outliers? Just try your best on this!)

gss %>% ggplot(aes(x=realrinc, y=sei10)) + geom\_point()

1. In R, generate a scatter plot where X= realrinc and Y= sei10 for observations where respondents report less than $150,000 in realrinc (hint use filter). Add a prediction line (hint use the stat\_mooth option). Interpret the relationship between realrinc and sei10. Is it positive, negative, or no relationship?
2. What is a residual? Imagine that we observe a survey respondent with an income of $25,00 and a socioeconomic index of 30 (observed value of sei10). The mean socioeconomic index of all respondents reporting an income of $25,000 is 45 (predicted value of sei10). What is the value of the residual for sei10 for this observation?
3. In your own words, why do usually prefer running a correlation rather than the covariance for understanding the relationship between two variables?
4. In R, calculate a correlation where X= realrinc and Y= sei10. Interpret the type and strength of the correlation.

gss %>% summarise(cor(realrinc, sei10, use = "complete.obs"))

1. In R, calculate a correlation where X= realrinc and Y= sei10 for observations where respondents report less than $150,000 in realrinc (hint use filter). Interpret the tyoe and strength of the correlation.
2. Do correlations differ for all respondents in the sample versus for those that report less than $150,000? How do they differ? Why do you think they differ (hint: scatterplot in Q3 is helpful in understanding this)?