HED 612

Homework #13

**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**:

*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 HW13\_lastname.R within lecture13 subfolder

*About the data* ***(we’re using two different data sets)!***

We will be using two different datasets that are part of the “AER” R package.

The MA Schools dataset contains data on test performance, school characteristics and student demographic backgrounds for schools in Massachusetts. The data are district-wide averages for public elementary school districts in 1998-99.

The Resume Names dataset contains data about resume, call-back and employer information for 4,870 fictitious resumes.

***Part 1: Logarithms***

*For the next set of questions, we will use* ***MA School Data*** *to investigate the effect of district average income per capita on test scores.*

* *X= income, district-average per capita income in $000s*
* *Y=* score8, *8th grade composite score (math + English + science) on Massachusetts Comprehensive Assessment System (MCAS) test*

1. Within the R script created above: load the AER and tidyverse libraries; check that your directory is set to the R project; open the MA Schools data.

data("MASchools")

1. Investigate the effect of district average income per capita (X= income) on district average 8th grade student MCAS scores (Y= score8)by creating a scatterplot of the data. Add a linear model to the scatterplot. By just looking at the scatterplot, does the linear model fit the data well? Why or why not? (hint: use R syntax from the Lecture 12/13 R script to create the plot)

MASchools %>% ggplot(aes(x=income, y=score8)) + geom\_point() +

stat\_smooth(method = "lm")

The data does not fit the data well as the value of the effect of income on test scores seems to differ across various ranges of income.

1. Write out the population regression equation for the ***Linear-Log Model*** for effect of district average income per capita (X= income) on district average 8th grade student MCAS scores (Y= score8)(hint: using the logarithm of X**)**.

Where Y= Student test scores, X1= log of income

1. Run the *Linear-Log Model* for the population regression equation in #3. (hint: you need to use the log() function within your lm() function).
   1. Write out the OLS prediction line with estimates.
   2. Interpret the coefficient

A 1% increase in district income per capita is associated with a 0.65407 (0.01\*65.407) point increase in district average test scores.

* 1. Is the coefficient significant at the 0.000, 0.001, 0.01, or 0.05 level?

The coefficient is significant at the 0.000 level.

* 1. What is the change in district average 8th grade MCAS scores for a change in $16k to $17k in average district income per capita?

– ()

– ()

– ()

– () = 3.9653

A change from $16k to $17k in district income per capita students is associated with a 3.9653 decrease in district average test scores.

***Part 2: Linear Probability Model***

*For the next set of questions, we will use* ***Resume Name Data*** *to investigate the effect of ethnic-sounding names on resumes on the probability of receiving a call back job interview.*

* *X= ethnicityv2,* variable indicating ethnicity “attached” to name on resume
  + *0=* White-sounding name [reference group]
  + *1=* African-American sounding name.
* Y=callv2, variable indicating whether the applicant called back?
  + 0= did not receive a call back
  + 1= did receive a call back

1. Within the R script created above: open the Resume Names data.

data("ResumeNames")

1. Create “clean” (0/1 coding) dummy variables of Y and X variables for analysis. Convert the independent variable of interest into a factor variable.

ResumeNames <- ResumeNames %>%

mutate(ethnicityv2 = ifelse(ethnicity=="afam", 1, 0))

ResumeNames$ethnicityv2 <- as.factor(ResumeNames$ethnicityv2)

ResumeNames <- ResumeNames %>%

mutate(callv2 = ifelse(call=="yes", 1, 0))

1. Write out the population regression equation for the ***Linear Probability Model*** for the effect of ethnic sounding names on resumes (X= ethnicityv2) on receiving a call back job interview (Y= callv2)(hint: X is categorical**)**.

Where Y= receiving a call back, X1= 0/1 ethnic sounding name

1. Run the *Linear Probability Model* for the population regression equation in #3.
   1. Write out the OLS prediction line with estimates.
   2. Interpret the coefficient

On average, a resume with an ethnic sounding name as opposed to a white sounding name is associated with a 3.2% decrease in the probability of receiving a call back interview.

* 1. Is the coefficient significant at the 0.000, 0.001, 0.01, or 0.05 level?

coefficient is significant at the 0.000 level.