17.871 Political Science Lab
Spring 2015
Problem set # 5: Multiple regression, sampling, and hypothesis testing

Handed out: April 13, 2015
Due back: April 22, 2015

1. We are interested in understanding how people feel about the expansion of voting rights to Americans in recent years. One of the ways that voting rights have been expanded is by requiring that non-English speakers be allowed to receive language assistance when they vote.

   The data are taken from the MIT module of the 2013 Cooperative Congressional Election Study.

   The following is how the variables of interest were measured:

   englishhelpgood: dependent variable. Based on the answer to the question, "Did requiring voting assistance for non-English speakers… (1) diminish elections a lot, (2) diminish elections somewhat, (3) improve elections somewhat, or (4) improve elections a lot?" This variable was then recoded to lie in the [1,4] interval as indicated by the response categories.

   age01: calculated as 2013 – birth year, and then to lie in the [0,1] interval. (In other words, the oldest respondents were 84 years old, and were coded as 1; the youngest respondents were 18 years old, and were coded as 0; everyone else is rescaled proportionately within the interval.)

   age01sq: the square of **age01**

   liberal01: ideology coded 1 = very conservative, 2 = conservative, 3 = moderate, 4 = liberal, 5 = very liberal. This variable was then recoded to lie in the [0,1] interval.

   dem01: party identification coded 1 = Democrat, 0.5 = neither Democrat or Republican, 0 = Republican.

   fraudscale01: a scale produced from several items on the survey. The scale is coded in the [0,1] interval, so that 1 means you think voting fraud happens all the time, and 0 means you think that voting fraud never happens.

   votedin12: dummy variable equal to 1 if the respondent reported "definitely voted" in 2012, 0 otherwise.

   Three dummy variables to code race:
   - **black** = 1 if the respondent identified as African American, 0 otherwise
   - **hisp** = 1 if the respondent identified as Hispanic, 0 otherwise

- **asian** = 1 if the respondent identified as Asian-American, 0 otherwise
- **other_race** = 1 if the respondent identified as any other racial group except white, 0 otherwise

The following is a portion of the Stata printout from the regression.

```
reg englishhelpgood age01 age01sq liberal01 dem01 fraudscale01 votedin12 black hisp
asian other_race [aw=weight]
(sum of wgt is   7.9249e+02)

      Source |       SS       df       MS              Number of obs =     788
-------------+------------------------------           F( 10,   777) =   23.08
       Model |  234.213276    10  23.4213276           Prob > F      =  0.0000
    Residual |  788.508188   777  1.01481105           R-squared     =  0.2290
-------------+------------------------------           Adj R-squared =  0.2191
       Total |  1022.72146   787  1.29951901           Root MSE      =  1.0074

--------------------------------------
englishhel~d |      Coef.   Std. Err.
-------------+------------------------
       age01 |  -1.901979   .5187674
     age01sq |   1.546851   .5452445
   liberal01 |    .193919   .0428526
       dem01 |   .4977591   .1025577
fraudscale01 |  -.6948561   .1398234
   votedin12 |  -.1935157   .0925268
       black |   .2844007   .1153371
        hisp |   .4029849   .1433924
       asian |   .7126826    .206912
  other_race |  -.0924905   .1985293
       _cons |   2.155843   .1740834
--------------------------------------
```

1a. Provide a substantive interpretation of each regression coefficient.

1b. Calculate the 95% confidence interval around the first three slope coefficients (age01, age01sq, and liberal01)

2. Suppose we had a database of all the household incomes in the United States. First, we draw an infinite number of samples from this database, with a sample size of 100, and calculate the average household income for each sample, and save them all in a database called "Sample 1". Second, we draw an infinite number of samples from this same database of household incomes in the United States, with sample size of 10,000, and calculate the average household income for each sample, and save them all in a database called "Sample 2."

2a. If the average of all the values in Sample 1 is $54,596, would you expect the average of all values in Sample 2 to be (a) greater than (b) less than, or (c) the same as Sample 1? Why?

2b. If the standard error of Sample 1 is $3,950, what would you expect the standard error of Sample 2 to be?

3. Suppose you sampled 1,000 Americans and asked them if they had a favorable personal opinion of Barack Obama. You have respondents give answers on a four-point scale, with "very favorable" = 4 and "very unfavorable" = 1. You calculate the average answer to the question and then calculate the 95% confidence interval around the answer. You are interested in knowing whether residents of the Bahamas have the same opinion of Obama as Americans do. What is the sample size of Bahamians you would have to draw in the Bahamas, if you wanted the 95% confidence interval of the answer *from Bahamians* to equal that of the 95% confidence interval from Americans? (Assume that the average answer, and the standard deviation, from Bahamians is the same as the average answer, and standard deviation, from Americans.)

4. In the 2014 Cooperative Congressional Election Study, 63.2% of respondents reported they "definitely" voted in the 2012 presidential election. The number of respondents was 56,200. The best estimate of turnout in the 2014 election, from official election statistics is that 35.9% of voting-eligible Americans voted in 2014. Calculate the *t*-ratio to measure how far the CCES turnout rate is from the actual turnout rate. (Note that you will need to calculate the variance of the CCES turnout estimate from what you have been taught about variances of proportions.)