17.871 Political Science Lab
Spring 2014
Problem set # 5: Multiple regression, sampling, and hypothesis testing

Handed out: April 9, 2014
Due back:  April 23, 2014

1.  We are interested in understanding how people feel about the expansion of voting rights to Americans during the twentieth century.  One of the things the courts did in the 1960s and 1970s was rule that college students had a constitutional right to choose whether they vote back home or where they are living while going to college.

    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:

    **studentsgood:** dependent variable.  Based on the answer to the question, "Did allowing college students to choose whether they would vote at home or where they were going to college, (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 0,1 interval.

    **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.

    **fraudscale:** 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.

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

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

```
. reg studentsgood01 age01 age01sq liberal01 dem01 fraudscale voted [aw=weight]
(sum of wgt is   6.9754e+02)

      Source |       SS       df       MS              Number of obs =      698
-------------+------------------------------           F(  6,   691) =    19.36
       Model |  8.75513579      6   1.4591893          Prob > F      =   0.0000
    Residual |  52.0694938    691  .075353826          R-squared     =   0.1439
-------------+------------------------------           Adj R-squared =   0.1365
       Total |  60.8246296    697  .087266326          Root MSE      =   .27451


-----------------------------------------
 studentsg~01 |      Coef.   Std. Err.
-------------+---------------------------
       age01 |  -.7660128    .1534894
     age01sq |   .7387247    .1650827
   liberal01 |   .0914473    .0501546
       dem01 |   .1295636    .0302423
  fraudscale |  -.1268015    .0379206
     voted12 |  -.0914447    .0296263
       _cons |   .9173865    .0531573
-----------------------------------------
```

      1a.  Provide a substantive interpretation of each coefficient.

      1b.  Calculate the 95% confidence interval around each coefficient.

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 10, 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 1,000, 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 $45,596, 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 $3,450, 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 from Bahamians is the same as the average answer from Americans.)

4.  In the 2013 CCES survey that formed the basis of question 1, 78.5% of respondents reported they "definitely" voted in the 2012 presidential election. The number of respondents was 1,000. The best estimate of turnout in the 2012 election, from official election statistics (to get the numerator) and Census Bureau population estimates (to get the denominator) is that 58.7% of voting-eligible Americans voted in 2012.  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.)