17.871, Political Science Lab
Spring 2013
Problem set # 5: Multiple regression, sampling, and hypothesis testing

Handed out: April 8, 2013
Due back: April 22, 2013

1. We are interested in understanding the influence of three factors on attitudes toward gay marriage in 2012: party identification, the importance of religion in one's life, and marital status.

   The data are taken from the 2012 Cooperative Congressional Election Study.

   The following is how the variables of interest were measured:

   gay_marriage: dependent variable. Based on the answer to the question, "Do you favor or oppose allowing gays and lesbians to marry legally?" Coded so that 1 = favor, 0 = oppose.

   pew_religimp: Based on the answer to the question, "How important is religion in your life?" 0 = Not at all important, 1 = Not too important, 2 = Somewhat important, 3 = Very important

   democrat: party identification. -1 = Republican, 0 = Independent, +1 = Democrat.

   married: marital status. 1 = married, 0 = not married.

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

```
      Source |       SS       df       MS              Number of obs =   49197
-------------+------------------------------           F(  3, 49193) = 6539.93
       Model |  3495.76786     3  1165.25595           Prob > F      =  0.0000
    Residual |  8764.98894 49193  .178175532           R-squared     =  0.2851
-------------+------------------------------           Adj R-squared =  0.2851
       Total |  12260.7568 49196  .249222636           Root MSE      =  .42211

------------------------------------------------------------------------------
gay_marriage |      Coef.   Std. Err.
-------------+----------------------------------------------------------------
pew_religimp |  -.1666477   .0017669
    democrat |   .1876996   .0024084
     married |   -.059384   .0038789
       _cons |   .8733829   .0043432
------------------------------------------------------------------------------
```

   1a. Provide a substantive interpretation of each coefficient, including the intercept. (Remember that because the dependent variable is binary, we can interpret coefficients in terms of changes in probabilities.)

   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 sample size of 10. Second, we draw an infinite number of samples from this database, with sample size of 1,000.

   True or false:

   3a. We should expect the average of all the draws when the sample size of 10 to be different from the average of all the draws when the sample size is 1,000.

   3b. We should expect the standard deviation of all the averages produced by drawing a sample size of 10 to be 10 times greater than the standard deviation of all the averages produced by drdawing a sample size of 1,000.

3. The mean of SAT scores is supposed to be 500, with a standard deviation of 100. Suppose the distribution is normally distributed.

   4a. If we randomly choose one person's score, what is the probability it is greater than 700?
   4b. If we draw a random sample of 100 people, what is the probability the average score will be above 505?

4. The 2008 Current Population Survey Special Report on Voting and Registration interviewed 36,000 adults of voting age in November 2008, after the general election. 23,209 of them reported they had voted. Is this number "too high" or "too low," given the percentage of eligible voters we know participated in the 2008 presidential election (roughly 56% turnout)?

5. In the survey mentioned above, 11,109 of the 16,776 men and 13,164 of the 19,224 women said they voted. Run the t-test to see how likely it is that the fraction of men in the electorate voting equals the fraction of women voting.