17.871, Political Science Lab
Spring 2012
Problem set # 4: Multiple regression, sampling, and hypothesis testing

Handed out: April 4, 2012
Due back: April 12, 2012

1. We are interested in understanding public reaction to the Obama candidacy in 2008, and in particular, how much race played a part in voters feeling proud that Obama was a candidate. We suspect that African Americans will express greater pride than voters of other races. But, we also suspect that Democrats may express greater pride, as well. Because African American tend to be overwhelmingly (but not entirely) Democrats, it would be interesting to know how much of African American pride in Obama's candidacy is due to race, and how much is due to partisanship. To disentangle the separate effects of race from partisanship, we resort to multiple regression.

   The data are taken from the 2008 American National Election Study.

   The following is how the three variables of interest were measured:

   proud: Based on the answer to the question "Now we would like to know something about the feelings you have toward Barack Obama. Has Barack Obama — because of the kind of person he is, or because of something he has done — made you feel proud?" 1 = yes, 0 = no.

   democrat: party identification. 0 = strong Republican, 1 = weak Republican, 2 = Republican-leaning Independent, 3 = Pure Independent, 4 = Democratic-leaning Independent, 5 = weak Democrat 6 = strong Democrat.

   black: 1 = respondent identifies as African American, 0 otherwise.

   The following is the variance-covariance matrix between the relevant variables in the regression:

	proud	democrat	black
proud	0.246711		
democrat	0.462936	4.01059	
black	0.081344	0.328053	0.188347

1a. Calculate both the *bivariate* and *multivariate* regression coefficients with **proud** as the dependent variable and **democrat** and **black** as the independent variable(s). (In other words, calculate four different coefficients, two in separate bivariate regressions, and one in a single multivariate regression.)

1b. Explain why the bivariate and multivariate regression coefficients are different (assuming that they are, in fact, different).

2. Students of public opinion are interested in how voters develop perceptions of the policy world. One interesting question is whether people believe crime is getting better or worse and how that relates to the types of information people receive.

People who read newspapers tend to believe crime is not as much of a problem as people who don't read newspapers. People who watch lots of game shows tend to believe crime is more of a problem than people who don't watch game shows.

Consider the following series of regressions. The dependent variable in each case is "crimeworse," which is coded 1 if the respondent believes crime has gotten worse in the past year, 0 if it has stayed the same, and -1 if it has gotten better. The independent variables are "newspaper" (the number of times each week the respondent reads the newspaper) and "wheel," the number of times in the past month the respondent watched the game show "Wheel of Fortune." (The data are taken from the 2000 American National Election Study.)

	(1)	(2)	(3)
newspaper	-0.036 (0.010)	—	-0.037 (0.010)
wheel	—	0.050 (0.017)	0.053 (0.017)
intercept	0.062 (0.045)	-0.059 (0.029)	0.071 (0.045)
n	863	834	834
r^2	.015	.010	.027
rmse	0.83	0.83	0.83

What does this pattern of regression coefficients tell us about the correlation between newspaper reading and watching Wheel of Fortune?

3. True or false: The average of 10 random draws from a distribution with population mean μ is 10 times more variable than the average of 1,000 random draws from the same distribution.

4. If we gave everyone in America an IQ test, we would get an average of 100 with a standard deviation of 16. It would be normally distributed. If we draw a sample of 100 people,

   4.a what is the probability that the *average* of the sample will be more than 105?

   4.b what fraction of the sample should we expect to be between 90 and 110?

   4.c what fraction of the sample should we expect to be above 140?

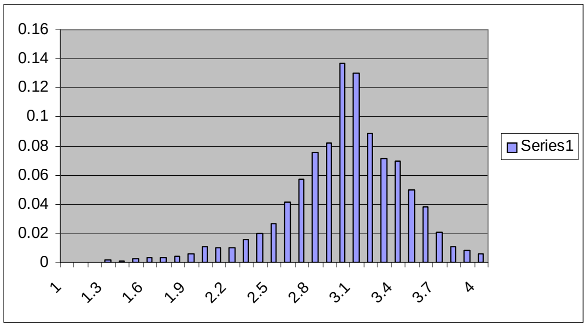
5. You are conducting research into how much money MIT undergraduates spent on their spring break trip. You send out a survey that asks people to tell you, in ranges, how much they spent. Here are the results:

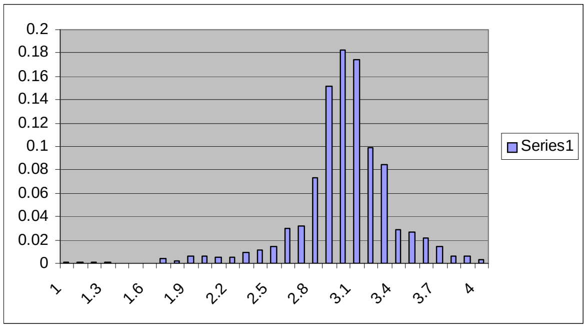
$0	150
$0-$100	50
$100-$200	40
$200-$500	30
$500-$1000	20
More than $1000	5

Provide your estimate of the average cost of a spring break trip and the standard error associated with it. (Hint: you need to be explicit about how to handle the five observations the last category.)

6. Early in March 2012 the Pew Research Center Poll reported that 46% of respondents opposed same-sex marriage. The sample size was 1,200 people. What is the 95% confidence interval around this result?

7. The following two histograms show the distribution of grade point averages of two universities. Both have a mean GPA of 3.0 and standard deviation of 0.7. If you sampled 100 students from University A and 100 students from University B, how would you expect the averages and standard errors of the two samples differ?





8. 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)?

9. 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.