**Homework #6**
***** Due Tuesday, March 26, 2013 at the beginning of lecture (9:30 a.m.)*****

You will have lab problems and written problems for this assignment.

You will be using the data set `wage_guar.dta` located at Beachboard to answer the Lab Problems. Please work on the STATA code during lab so you can ask questions of the lab instructor. For the lab problems, I want you to submit your `log` file and your **typed** answers in complete sentences. You do not need to submit your `do` file.

The answers to the written problems are to be neatly written out or typed. The answers to anything that asks you to explain should also be in complete sentences.

Please submit everything in one package in the order assigned in this homework.

## Lab Problems

**L1.** The dataset `wage_guar.dta` contains data on 1,289 individuals interviewed in March 1995 for the *Current Population Survey (CPS)* by the U.S. Census BureauThis question looks at the determinants of their wages. Create a do file `wage95cps.do` and a log file `wage95cpsdlog` for this problem. Submit the log file with your answers.

**a.** Create the following variables and then save data as `wage95cps.dta`

`lnwage` : defined as the log(wage)
`exp_squared` : defined as exper squared
`college` : defined as having education of 16 years or above
`some_college` : defined as having education of 12 or above but less than 16
`no_college` : defined as having education of less than 12
`female_exp` : interact the female and the experience variables
`lnexp` : defined as the log(exper)
`annual_wage` : defined as the annual wage if the person works full time at his/her hourly wage. So, 8 hours a day, 1289 days a year.

**b.** Run the following regressions to look at the determinants of wages. After you estimate the first model (i), use the Breusch-Pagan test to test for heteroskedasticity. If you find that the errors are heteroskedastic, use robust standard errors to correct for heteroskedasticity by re-estimating that model and then go ahead and use robust standard errors for the rest of the models.

i. Dependent variable: `wage`; explanatory variables are `female nonwhite union education exper`
ii. Dependent variable: `annual_wage`; explanatory variables are `female nonwhite union education exper`
iii. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union education exper`

iv. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union education lnexp`
v. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union education exper exp_squared`
vi. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union college some_college exper`

**c.** Run a regression omitting `union` and `nonwhite` from model (vi) from part b. This is the restricted model that can be used to test if union and nonwhite are jointly significant. Save the RSS as `rss_r` using `ereturn list` and the `scalar` command.

**d.** Now, run model (vi) from part b again. Save the RSS for this model as `rss_u` using `ereturn list` and the `scalar` command.

**e.** Now, calculate the F-statistic manually using the saved scalars from parts c and d the STATA scalar command: `sca fstat=…` Display the f-statistic by typing `sca list fstat`

**f.** Run model (vi) from part b again. Use the STATA command `test` to test for joint significance of `union` and `nonwhite` in this model.

**g.** Now, create a special variable for use in testing if the coefficients on `college` and `some_college` are the same. This variables is `all_college = college+some_college`

**h.** Run a regression such that: dependent variable: `lnwage`; explanatory variables are `female nonwhite union college all_college exper`

**i.** Now, run the model (vi) from part b again. Use the STATA command `test` to test whether the coefficients on `college` and `some_college` are the same in this model.

**j.** Test for multicollinearity of the model (vi) from part b by using the `vif` command.

**Questions**

**Q1.** Explain (giving one or two reasons) why we might want to transform a variable using log (Note: This is natural log).

Q2. Use your results from L1 for this problem.
a. Is there a problem with heteroskedasticity? Explain how you know.
b. For each of the models below (which correspond to the models from part b of L1, interpret the coefficients identified here.
i. Dependent variable: `wage`; explanatory variables are `female nonwhite union education exper.` Interpret the coefficient on `exper`
ii. Dependent variable: `annual_wage`; explanatory variables are `female nonwhite union education exper.` Interpret the coefficient on `exper`

iii. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union education exper` Interpret the coefficient on exper

iv. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union education lnexp.` Interpret the coefficient on `lnexp`

v. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union education exper exp_squared -` Interpret the effect of experience on wages, using `exper` and `exp_squared`

vi. Dependent variable: `lnwage`; explanatory variables are `female nonwhite union college some_college exper -` Interpret the effect of having at least 4 years of college on wages

b. Why did we leave no_college out of the final model (estimated in L1, parts b, vi)?

c. In L1, parts c, d, and e you used STATA to calculate an F-statistic. Write out the formula for that F-statistic here. What is the null hypothesis for this F-test? If the critical F-value with 2, 1282 degrees of freedom at the 5% level is 3.00, would you reject the null hypothesis? What is your conclusion here?

d. Is the test in L1 part f the same as the test in part e? What is your conclusion based on the result in part f?

e. In L1 part g and h, you created a new variable and ran a regression on that new variable. What is the null hypothesis you are testing? Based on that hypothesis, demonstrate (show) how we transform our original equation (L1 part b, vi) to the one in L1 part h.

f. Based on the results from L1 part h, what is your conclusion about the hypothesis you are testing (in other words, the hypothesis you outlined in Q2, part d)?

g. What does the test in L1 part i tell us? Why couldn't we calculate this particular test using scalars?

h. Is there a problem with multicollinearity in the model in L1 part b iv? How do you know?