Homework #4
***** Due Tuesday, March 12, 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 `arclab.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.

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 (so, STATA log files and typed answers first, then written problems).

Lab Problems

L1. The dataset `arclab.dta` contains data on 557 U.S. Counties in thirteen states in the eastern U.S. Of these, some counties are in the federally-designated Appalachian Regional Commission (ARC) region which has historically been an economically-disadvantaged region. The other counties are those which are immediately adjacent to the ARC counties. This question looks at the determinants of employment growth in the ARC region and the surrounding counties. Look at the labels for the data to see what each data element represents. Create a do file `arclab.do` and a log file `arclab.log` for this problem. Submit the log file with your answers.

- **a.** Create a dummy variable, `arc` that is equal to one if the `arc_county` variable indicates it is an ARC county. Otherwise, non-ARC counties should have a value of zero. This variable will be used to see if there is a difference in employment growth between the ARC and non-ARC counties.
- **b.** Create a set of dummy variables for each of the 13 states in the ARC region. Use `tabulate` with the prefix `state` to create these variables. Hint: the `state` variable has the unique names of the 13 states for each county in the dataset.
- **c.** Create a new variable `empgrowth_9006` that is the percent change in employment from 1990 to 2006. Use the `label` command to label this "`Percent change in employment from 1990 to 2006`". Make sure you save the data as `arcdata1.dta` so you can use it for the next problem.
- **d.** Use the `summarize` command to look at the description of the variables in this dataset.
- **e.** Use the `regression` command to estimate the following linear regression model, under the assumption that college educated people contribute to employment growth:

$$empgrowth\_9006_i = \beta_0 + \beta_1 percoll90_i + u_i.$$

f. What is the estimated slope coefficient for *percoll90*? What is the interpretation of this slope coefficient?

g. Use the `regression` command to estimate the following linear regression model, where we now test whether college educated people and having a higher percentage of self-employed individuals are important to employment growth in this region:

$$empgrowth\_9006_i = \beta_0 + \beta_1 percoll90_i + \beta_2 perse90_i + u_i.$$

h. What is the estimated slope coefficient for *percoll90* in the model from part g? What is the interpretation of this slope coefficient?

i. Using your results from parts e and g, does the regression model in part e suffer from omitted variable bias? Explain.

j. Use the `regression` command to estimate the following linear regression model:

$$percoll90_i = \beta_0 + \beta_1 perse90_i + u_i.$$

k. Use the `predict` command to capture the residuals from the regression in part j in a new variable named `res`.

l. Use the `covariance` option of the `correlate` command to examine the covariance between `res` and `empgrowth_9006`.

L2. Use your data from L1, `arcdata1.dta` for this section. We will continue to look at employment growth from 1990 to 2006 in the ARC counties and the surrounding region. Again, create a do file and a log file. Save the log file as `arclab1.log` and submit with your assignment.

a. Now, we want to test if additional variables would add to the explanatory value of employment growth in the region. Add the following additional variables to the model from L1 part g: `pci90_thousands` and `pci90.`

b. Does the STATA output from part a) include both the new variables? Explain what happened.

c. Now, use the `regression` command to estimate a linear regression where y=`empgrowth_9006` and x includes `percoll90, perse90`, and the dummy variable `arc.`

d. What is the coefficient on `arc?` What is the interpretation of this coefficient in our model? (Hint: Look at the t-statistic!)

e. Now, use the `regression` command to estimate a linear regression where y=`empgrowth_9006` and x includes `percoll90, perse90`, `and` dummy variables for each of the states in the model (`state1,` … `state13`). Exclude the dummy variable for `state2` (Georgia) from the model.

f. Why did we exclude Georgia from the model in part e? What is the interpretation of the coefficient on `state1` (Alabama)?

g. Now, add the following additional variables to the model from part e: `popsqmi_60` and `rural.`

h. Test for multicollinearity in your data by running the `vif` command.

i. Are there any problems with multicollinearity in your model? Explain.

j. Compare the final model in part g to the model in L1, part g in terms of how much they explain the variance in employment growth. Explain.

Questions

Q1. Suppose you are interested in whether there is a gender bias in setting wages.

a. You get data from the Current Population Survey. You then use STATA to estimate a regression function as follows:

$$wage_i = \beta_0 + \beta_1 Female_i + \beta_2 Nonwhite_i + \beta_3 Union\ Member_i + \beta_4 Education_i + \beta_5 Experience_i + u_i.$$

Some of the STATA output is as follows:

Source	SS	df	MS	Number of obs	1289
				F(5, 1283)	122.61
Model	25967.3	5	5193.46	Prob > F	0
Residual	54342.5	1283	42.3558	R-squared	
				Adj R-squared	
Total	80309.8	1288	62.3523	Root MSE	6.5081

Wage	Coef.	SE	t	P>t
Female	-3.0749	0.36462		
Nw	-1.5653	0.50919		
Un	1.09598	0.50608		
education	1.3703	0.0659		
experience	0.16661	0.01605		
_cons	-7.1833	1.01579		

What is the coefficient on female? What is the interpretation of this coefficient? Calculate the t-statistic and test whether it is statistically significant at the 5% level. Based on the regression, do you think that women earn less than men?

b. What are the R^2 and Adjusted R^2 of this regression model?

c. As an alternative to the regression in part a, you collect data about gender and salaries from people who stop by a table at the local mall. You then use a "difference in means" test to see if the average salary for women is less than the average salary for men. You find that there is a statistical difference in the means with women's average salary statistically lower than men's.

d. Even though both approaches give you the same answer, explain which method is a better way to test if women earn less than men.

e. What if you also want to know whether women and men get the same additional wages for each additional year of school? To test this, you generate a new variable (female*education) which interacts the female dummy variable and education. The results of the regression are below:

```
      Source |       SS       df       MS              Number of obs =    1289
-------------+------------------------------           F(  6,  1282) =  103.19
       Model |  26154.5202     6   4359.0867           Prob > F      =  0.0000
    Residual |  54155.3045  1282  42.2428273           R-squared     =  0.3257
-------------+------------------------------           Adj R-squared =  0.3225
       Total |  80309.8247  1288  62.3523484           Root MSE      =  6.4994

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .5205381   1.746147     0.30   0.766    -2.905081    3.946157
          nw |  -1.597437   .5087366    -3.14   0.002    -2.595484   -.5993888
          un |   1.171601   .5066774     2.31   0.021     .1775931    2.165609
   education |   1.492911   .0878829    16.99   0.000     1.320501    1.665321
  experience |   .1643591   .0160616    10.23   0.000     .1328491    .1958691
  women_educ |  -.2733575     .12984    -2.11   0.035    -.5280798   -.0186352
       _cons |  -8.773084   1.264614    -6.94   0.000    -11.25402   -6.292143
```

What is the coefficient on women_educ? Interpret the meaning of this coefficient in this regression model.

Q2. Use the results from L1 to do the following. Follow the directions carefully in terms of what to calculate. Even if other parts of L1 include the answer, show how these elements are calculated, assuming you only had certain results. Please show the formula you use and the steps you take to get to the final answer. You can always use the output to see if you did it right!

a) Use the results from L1 parts j and l to calculate the coefficient of percoll90 in the following regression model:

$$empgrowth\_9006_i = \beta_0 + \beta_1 percoll90_i + \beta_2 perse90_i + u_i.$$

b) Calculate the t-statistic to test whether the coefficient on percoll90 is different than zero. Note: In this part it is OK to use the standard error calculated by the regression output rather than having to calculate it.