PUBPOL 639: ASSIGNMENT 5
Winter 2011

Due: Monday, April 11th at the start of class

NOTE: This assignment explores various transformations that let you estimate non-linear relationships using OLS. Be sure to copy your do and log files into the back of your solutions. Courier 8 or 9 pt font works well. Your log file may be long, so don't worry about cleaning/formatting it.

In 1995, the newly elected government (headed by Nelson Mandela) has identified racial differences in earnings as a major concern. The administration believes that education is the key to reducing observed differences in pay. Prior to implementing new policy, the South African Labor and Development Unit (SALDRU) is hired to asses the current value of education in the labor market. Your task as one of the primary researchers in the unit is to complete the preliminary analysis below using the 1994 October Household Survey Data (ohs94.dta).

Final Notes:

- After completing the analysis, be sure to create two tables to present your regression output (one for each part)
- In running the models below, use the following omitted categories throughout:
  - Race => blacks
  - Gender => females
  - Union => non-union member
  - Location => rural

PART I

1. Estimate a model relating monthly income to education (continuous), age, race, and gender.

   a. Interpret the coefficient on education
   b. Interpret the coefficients on the race indicator variables

2. Now run the same regression, but including a quadratic age variable.

   a. Holding race, gender, and education constant, what is the predicted change in income associated with a one-year increase in age for a 34 year old respondent?
   b. Holding race, gender, and education constant, what is the predicted change in income associated with a one-year increase in age for a 54 year old respondent?
   c. Can we reject the linear model from question 1 in favor of this quadratic model? Explain.

3. To simplify the task of interpreting quadratic terms, write a program (called "coef2") that will provide you with the information found in question 2, parts (a) and (b). At the very least, the program should work for the model estimated in question 2, but if possible, it should work with any model estimated with any number of independent variables. Once the program is completed, confirm your answers to question 2 parts (a) and (b) using the program.

4. Now estimate a model relating income to education (continuous), race, and interactions between education and race.

   a. Which racial group receives the largest return for their educational attainment?

b. Calculate the change in predicted income associated with a one year increase in education for each racial group, holding race constant.
c. Check your calculations by plotting the predicted (fitted) relationship between income and education for each racial group. To do this, you should first generate a variable called pinc (predicted income) by typing "predict pinc" after your regression. You can then plot this predicted value by education separately by racial group by typing:

```
twoway line pinc educ if race == 2, lc(red) ||
       line pinc educ if race == 3, lc(blue) ||
       line pinc educ if race == 4, lc(green) ||,
       legend(lab(1 "coloured") lab(2 "indian") lab(3 "white")) ;
```

Note: You may need to sort the data by education first depending on version of Stata, otherwise can get crazy graph.

PART II

Prior to extending your analysis in Part I, the head researcher at SALDRU suggests you use a log transform on income given the distribution is positively skewed. Using this suggestion, complete the items below.

5. Now regress log(income) on education (continuous), race, and gender.

   a. Interpret the coefficient on education
   b. Interpret the coefficients on the race indicator variables
   c. Does the transformation on income seem appropriate given the data and your findings? Explain.

6. Now regress log(income) on education (continuous), age, race, and gender.

   a. Interpret the coefficient on education
   b. Why has the coefficient on education changed from the regression in question 5? Explain in terms of the framework used in class.

7. Now regress log(income) on education (continuous), age, age-squared, race, gender, union membership, location (urban/rural), and interactions between education and race.

   a. Using the model specified, in addition to previous results, summarize the general overall findings.
   b. Discuss how the findings inform the administration's assumption (listed below). Additionally, provide the administration with your recommendation on how to best reduce racial differences in earnings.

   Assumption: The administration assumes that increasing educational attainment will reduce racial differences in earnings.

8. Using the existing data, are there improvements that could be made to the "final" model specified in question 7? If so, describe any changes you believe would improve the model.

9. Describe how any data limitations impact the analysis in questions 1-7. Be as specific as possible. [Note: I would like you to briefly discuss possible variables omitted from the data file, but a majority of the discussion should focus on measurement and operationalization of variables in the data file]

10. You are reminded that the OHS data is not a simple random sample and that weights should be utilized. Use the weights provided to re-estimate the "final model" requested in question 7. How, if at all, does this change your findings? Note: To use the weights insert the following code at the end of your regress command -> [aw=weight]

OPTIONAL CHALLENGE QUESTIONS

Note: The questions below are completely optional and are not required.

I. Write a program that finds the difference in estimated standard errors for a given regression model when the robust option is and is not utilized.

II. Provide evidence for the improvements suggested in question 8. More specifically, actually run the analysis suggested to provide evidence of your claim.