ECONOMICS 145 Economic Research Methods

Assignment 4 Regression Analysis

Part 1: Cleaning and Summary of Data

You do not need to turn this part in until part 2 (regression) is completed. Use height.dta (or height.xls) to answer the following questions. The data come from 2005 PSID for California. All individuals are household heads. The data includes

Age: in years

Sex: 1 for male, 2 for female

Weight: in pounds

Height: in inches

Race: see below

Education: highest degree earned, see below

Wage: annual wage in 2004, in dollars

Total income: total family income (not just the household head)

Please answer the following questions using a word processor. Make sure you present your tables clearly.

1. Calculate the summary statistics (sample size, mean, standard deviation, median, min, and max) for the variables wage, height, weight, and age.

2. Tabulate the frequency of male and female individuals. Notice that the individual is assigned the value of 1 if the individual is male and 2 for female.

3. Generate a female variable. That is, assign the value of 1 if the individual is female and 0 otherwise.

4. Generate a series of race dummy variables: Black, Asian, Native American, Pacific Islander, and Others. Tabulate each dummy variable and present its frequency. The coding goes as follows
   1. white
   2. black
   3. native American
4 Asian
5 Pacific Islander
7 Other

5. The education attainment variable measures education by highest degree received. Generate a series of education attainment variables: bachelor and above (be very very careful!!!!), Master, Doctorate, law, and medical. Again, report the frequency of these 5 variables.

0 No college degree
1 Associate of Arts
2 Bachelor
3 Master
4 Doctorate; Ph.D. (except codes 5 and 6)
5 LLB; JD (law degrees)
6 MD; DDS; DVM; DO (medical degrees)

6. Generate the natural log of wage and age squared. Notice that some observations will be missing. Present the summary statistics of sample size, mean, and standard deviation.

Part 2: Regression analysis

7. Do a scatter plot for wages against height. Put wage on the Y-Axis and height on the X-Axis. Before you do the graph, you should first drop the observations with zero wages and the wage above 2,000,000 (there is only one of these). In Stata, you can use the drop down menu under “graphics”, as shown in class.

8. Regress the natural log of wage on age, age squared, height, female, race variables, and the dummy of college graduates. Report the regression output.

9. The height is measured in inches. What is the sign of the coefficient to height? Is it significant at 5% level? Interpret the height coefficient from your equation above.

10. Some race variables are significant at the 5% level. Which ones are they? Interpret the coefficient to those variables

11. Are the wages of Asians and Pacific Islanders significantly different? Test using your results. (use 5% significance level)

12. If you suspect that height may affect earning differently for men and women, how do you test that? State the null hypothesis, and then setup the regression model based on question 8. Run the regression and test the hypothesis at the 5% level.
13. Look at your results in question 7. Do you see any reason to add any more variables? If so, what would you add and how would you generate that variable? Add the variable to your regression in question 8 and present the results. Is the coefficient of the newly added variable significantly different from 0?