Assignment 4: Regression Models with Nonlinear Functions and Binary Dependent Variables

ECO 321

DUE: Thursday, April 19 in class

Instructions: You need to write a Stata do-file to answer these questions. Your do-file must produce a log file that contains your Stata output. You must submit your log file along with your answers to the assignment. Assignments that do not contain log files will be marked down.

1. In this question, we are interested in how education, experience, and race explain differences in wages across people. Use wage2.dta to answer this question.

a) Use the wage2.dta to estimate the following model where the return to education depends upon the amount of work experience (and visa versa):

$$log(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 (educ_i \times exper_i) + u_i$$

Show that the marginal effect of another year of education equals $\hat{\beta}_1 + \hat{\beta}_3 exper$. Plug in the values of $\hat{\beta}_1$ and $\hat{\beta}_3$ to get the general form for the marginal effect of education.

b) Test whether the return to education depends on the level of experience.

c) Now allow education, experience, job tenure, marriage status, race, and geographic location to determine wages by estimating the following model:

$$log(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 married_i + \beta_5 black_i + \beta_6 south_i + \beta_7 urban_i + u_i$$

Report the results in standard form. Holding other factors fixed, what is the approximate difference in monthly salary between married people and nonmarried people? Is this difference statistically significant at the 5% level?

d) Modify the model in part (c) by allowing $log(wage)$ to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married nonblack people and nonmarried nonblack people?

e) On a graph with education on the x-axis and $log(wage)$ on the y-axis, draw the sample regression functions for the four groups of people: (1) married and black, (2) married and nonblack, (3) single and black, and (4) single and nonblack, holding constant experience, tenure, and geographic location (i.e., south and urban). You do not need to create this graph in Stata, but you will use the

results from part (d) to draw the graph on paper.

2. In 1975, the U.S. passed the Home Mortgage Disclosure Act. This law requires all mortgage lenders to release annual data on all mortgage applications. One goal of the act is to identify discriminatory lending—cases where banks do not lend to individuals based on race or gender. The data set LOANAPP.dta contains a random sample of loan applications from HMDA data. The variables are defined as follows:

- *approve* **is a dummy variable equal to one if the loan application is approved**
- *black*, *hispan*, **and** *male* **are dummy variables equal to one if the borrower is black, hispanic, or male, respectively**
- *apr* **is the annual-percent rate on the loan (i.e. the interest rate)**
- *term* **measures the length of time over which the loan will be paid off**
- *bankruptcy* **is a dummy variable equal to one if the individual has ever declared bankruptcy**
- *gdlin* **is a variable equal to one if the borrower's credit history meets the lender's typical guidelines**
- *hh_expenditures* **is a measure of annual household expenditures as a share of annual income**

a) Estimate a linear probability model that relates loan approval rates to the applicant's demographic characteristics (black, hispan, male), the loan characteristics (apr, term), and the applicant's credit worthiness (bankruptcy, gdlin, hh_expenditures). Interpret the coefficients on black, hispan, and male. Although we are interested in the relationship between demographics and loan approval, why is it important to control for loan characteristics and credit worthiness?

b) Find the predicted values based on the regression in part (a). Summarize the distribution of the predicted values. How does the average predicted approval rate compare to the average actual approval rate? Do you notice anything unusual about the predicted values? Explain what you see.

c) Repeat the regression in part (a) using only the regressors that are individually statistically significant at the 10% level using t-tests. Do your results change substantially? Are your results sensitive to dropping these controls? What does might this tell you about OVB?

d) Estimate the relationship in part (c) using a probit model. What is the difference in the probability of approval between black and nonblack applicants who meet the borrowing guidelines ($gdlin = 1$)? How about the difference between hispanic and non-hispanic borrowers? Do the results differ much from the linear probability model?

e) Repeat part (d) using a logit model instead of probit model. Do the results change?

f) Based on your work in parts (a) through (e), comment on the disadvantages and advantages of using the linear probability model vs. probit/logit to estimate models with binary dependent variables.

g) Based on your work in parts (a) through (e), what do you conclude about discrimination by lenders?