Assignment 4 Solutions: 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.

The regression results give:

$$\hat{log(wage_i)} = \underset{(0.25)}{5.95} + \underset{(0.018)}{0.044} educ_i - \underset{(0.019)}{0.021} exper_i + \underset{(0.0015)}{0.0032} (educ_i \times exper_i) + u_i$$

$$\frac{d\hat{log(wage_i)}}{deduc} = 0.044 + 0.0032 exper$$

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

From the Stata output, we see that $\hat{t} = 2.16 > 1.645$, which is the critical value for a (greater than) one-sided test with a 5% significance level. Therefore, we can reject the null hypothesis, conclude that the interaction term is statistically significant, and that the return to experience is higher for people with more years of education.

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 black people and nonblack people? Is this difference statistically significant at the 5% level?

The regression results give:

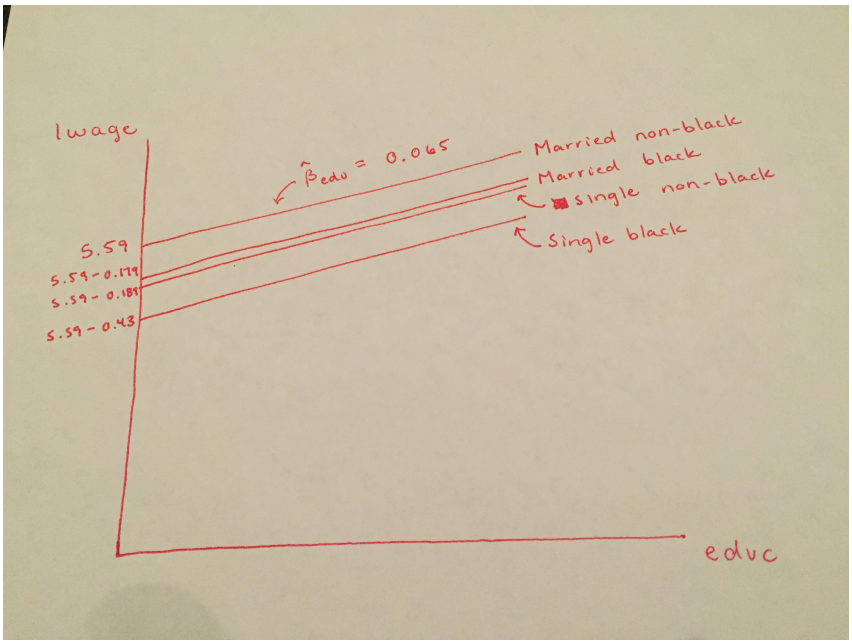
$$\begin{aligned} \widehat{log(wage_i)} &= \underset{(0.11)}{5.40} + \underset{(0.006)}{0.065}educ_i + \underset{(0.003)}{0.014}exper_i + \underset{(0.003)}{0.012}tenure_i + \underset{(0.040)}{0.20}\, married_i - \underset{(0.037)}{0.19}\, black_i \\ &- \underset{(0.027)}{0.09}\, south_i + \underset{(0.027)}{0.184}urban_i + u_i \end{aligned}$$

Black people earn 19% less on average than nonblack people, even after controlling for education, experience, job tenure, marriage status, and geographic location. The result is statistically significant because $\hat{t} = \frac{-0.19}{0.037} = |-5.14| > 1.95$. The result could suggest that there is discrimination in the labor force, but we would probably want to control for more variables to be sure.

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?

Married nonblack people earn 19% more than nonmarried nonblack people, on average, even after controlling for education, experience, job tenure, marriage status, and geographic location.

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 (g) 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?

From the regression we see that both black and Hispanic applicants both have about a 8 percentage point lower probability of approval (significant at 1% and 5% level, respectively). However, there does not appear to be any significant discrimination based on gender.

It is important to control for loan characteristics and credit worthiness—these may correlate with demographic characteristics and may be important determinants of approval (although most of these controls do not appear to be significant).

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.

The predicted approval rate is the average of the predicted values: 87.75%, which is very close to the actual approval rate of 87.84%. However, the predicted values predict a greater than 100% probability of approval for several applicants, which makes no sense. This result is a product of the assumption of linearity—there is nothing restricting the regression line from exceeding one (or falling below 0).

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?

The results do not change substantially—the difference in probability of approval for black and Hispanic applicants vs. non-black or -Hispanic applicants is only slightly larger than above (8.8 and 8.5 percentage points, respectively). Omitting credit history and loan characteristics does not appear to bias the estimates very much, although the influence of these characteristics may be entirely captured by the credit guidelines variable.

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 ($gdlin = 1$)? (Assume that black borrowers cannot be Hispanic and vice versa.) How about the difference between Hispanic and non-Hispanic borrowers? Do the results differ much from the linear probability model?

From the probit results we can compare $\Phi(\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_3(1))$ to $\Phi(\hat{\beta}_0 + \hat{\beta}_3(1))$ to see that black applicants have a 7.6 percentage-point lower probability of approval. This is similar to the results above, although a bit smaller. Doing the same comparison for Hispanic applicants, we see a difference of about 7.8 percentage points.

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

From the logit results we can compare $\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1(1)+\hat{\beta}_3(1))}}$ to $\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_3(1))}}$ to see that black applicants have a 7.2 percentage point disadvantage, while doing the same for Hispanic applicants reveals a 7.6 percentage point disadvantage. These results are not substantially different from part (d).

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.

Ultimately, the results do not change much across the three models, so which model to use will probably not matter if we only want to look at average effects. However, we did see the failure of the OLS for probability predictions for certain applicants — the linear model predicted a nonsensical probability of approval of greater than 100% for some applicants. On the other hand, the logit and probit are a bit more cumbersome to implement and interpret. So which model we choose really depends on the goal.

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

There seems to be evidence of discrimination in these models—even controlling for credit history and loan characteristics, black and Hispanic applicants are significantly less likely to be approved for loans.

However, our control variables may not capture all the important omitted variables. Demographics might correlate with neighborhood of residence, which determines things like house price growth that affect the ability to repay a loan. Demographics might also correlate with the bank that borrowers apply to, and different banks may have different approval rates.