Assignment 3: Multiple and Nonlinear Regression Models

ECO 321

DUE: Thursday, April 12 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 the determinants of years of education. Use the data set HTV.dta for this question. The data set includes information on education, ability, parents' education, and several other variables for men in 1991.

a) Load the data set in Stata (via the `use` command). First, use the `sum, detail` command to show the range of values that the education variable takes in the sample. Second, use the `gen` command to create a dummy variable that equals 1 if men completed exactly 12 years of education and equals 0 otherwise. Third, use the `tab` command to show what percentage of men completed exactly 12 years of education. Fourth, use the `sum` command to determine whether the men or their parents have higher levels of education.

The *educ* variable ranges from values 8 to 20, which makes sense because it measures years of education for the men in the sample. 42% of men in the sample have exactly 12 years of education, which makes sense because 12 years of education is the equivalent of a high school degree. On average, men in the sample have 13.03 years of education whereas their mothers have only 12.17 years of education and their fathers have only 12.44 years of education on average.

b) Use the `reg` command to estimate the regression model:

$$educ_i = \beta_0 + \beta_1 motheduc_i + \beta_2 fatheduc_i + u_i$$

Make sure to use heteroskedasticity-robust standard errors. How much of the variation in *educ* is explained by parents' education? Interpret the coefficient on *motheduc*.

The R^2 in the regression shows that parents' education explains 25.28% of the variation in men's education. The coefficient on mother's education shows that for each additional year of education the mother has, her son has 0.31 more years of education, on average. Last, the regression F-statistic equals 166.79 with a p-value$<$ 0.0000, so we can conclude that mothers' and fathers' education are jointly statistically significant.

c) Add the variable *abil* (a measure of cognitive ability) to the model in part (b). Does ability help to explain variation in education, even after controlling for parents' education? Use at least one test statistic to justify your answer.

Ability helps to explain the variation in education, even after controlling for parents' education. The easiest way to see this is to look at the statistical significance for the coefficient on ability. Its t-statistic= 17.63 and its p-value= 0.000, so we can reject the null hypothesis – that ability does not affect years of education – with a very high degree of confidence (i.e., a very small significance level). We can also see that the R^2 increased from 0.25 in model (1) to 0.42 in model (2), which is a large increase in R^2 for only adding one explanatory variable.

d) Compare the coefficient estimates on *motheduc* and $fatheduc$ from the model in part (b) to the model in part (c). What do the differences tell us about omitted variable bias in the model in part (b)? Be specific.

The effect of mother's education on men's education changed from 0.31 in model (b) to 0.19 in model (c) after we controlled for men's ability. Similarly, the effect of father's education on men's education changed from 0.18 in model (b) to 0.11 in model (c) after we controlled for men's ability. Since both estimates became smaller positive numbers after controlling for ability, the results suggest that the estimates in model (b) had positive bias. For the estimates to exhibit positive bias as a result of omitted variable bias, that means $Corr(ability, mothedu) > 0$, $Corr(ability, fathedu) > 0$, and $Corr(ability, wages) > 0$. In words, it means that men with more ability have mothers and fathers with more education, and men with more ability have higher wages.

e) Use the results from part (c) to test the hypothesis that the effect of mother's education on men's education equals two times the effect of father's education on men's education. First, write down the null hypothesis and the alternative hypothesis. Second, evaluate the hypothesis using the F-statistic and the `test` command. Third, rearrange the regression model and evaluate the hypothesis using a t-statistic. Do you arrive at the same conclusion using the F- and t-statistics?

$H_0 : \beta_1 = 2\beta_2$ vs. $H_1 : \beta_1 \neq 2\beta_2$

Using the F-statistic method:

The F-statistic for the joint hypothesis equals 0.12 and has a p-value= 0.73, so we cannot reject the null hypothesis that the effect of mother's education on men's education equals two times the effect of father's education on men's education.

Using the t-statistic method:

First, we can rewrite the hypothesis test as: $H_0 : \beta_1 - 2\beta_2 = 0$ vs. $H_1 : \beta_1 - 2\beta_2 \neq 0$

Next, to rearrange the regression, we will add and subtract $2\beta_2 motheduc$ to the model from part (c):

$$educ_i = \beta_0 + \beta_1 motheduc_i - 2\beta_2 motheduc_i + 2\beta_2 motheduc_i + \beta_2 fatheduc_i + \beta_3 abil_i + u_i$$

Combine like terms for *mothedu* to get the null hypothesis as a new coefficient in the model:

$$educ_i = \beta_0 + (\beta_1 - 2\beta_2) motheduc_i + 2\beta_2 motheduc_i + \beta_2 fatheduc_i + \beta_3 abil_i + u_i$$

Combine remaining like terms involving β_2 to see what new variable needs to be created:

$$educ_i = \beta_0 + (\beta_1 - 2\beta_2)motheduc_i + \beta_2(2motheduc_i + fatheduc_i) + \beta_3 abil_i + u_i$$

Then we will generate a new variable called "*newvar*" and set it equal to $2motheduc_i + fatheduc_i$

Then we regress *educ* on *motheduc*, *newvar*, and *abil*

Finally, we will use the t-statistic on *motheduc* in the rearranged model to test the null hypothesis.

The t-statistic= -0.34 and yields a p-value= 0.73, which is the same p-value that we obtained from the F-statistic above. The p-value from the F-statistic in the original model and the p-value from the t-statistic on the rearranged (i.e., transformed) regression should be the same because the tests are identical. Therefore, we still fail to reject the null hypothesis. We cannot reject the hypothesis that the effect of mother's education on men's education equals two times the effect of father's education on men's education.

2. Next we are interested in the determinants of wages. Use the data set wage1.dta to answer this question.

a) Use OLS to estimate the equation:

$$log(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i$$

Write the regression results in standard form (i.e., plug in the correct values for the parameters estimates and their standard errors):

$$\hat{log(wage_i)} = \underset{(SE(\hat{\beta}_0))}{\hat{\beta}_0} + \underset{(SE(\hat{\beta}_1))}{\hat{\beta}_1} educ_i + \underset{(SE(\hat{\beta}_2))}{\hat{\beta}_2} exper_i + \underset{(SE(\hat{\beta}_3))}{\hat{\beta}_3} exper_i^2$$

Writing the regression results in standard form gives:

$$\hat{log(wage_i)} = \underset{(0.106)}{0.128} + \underset{(0.007)}{0.090} educ_i + \underset{(0.005)}{0.04}\ exper_i - \underset{(0.0001)}{0.0007} exper_i^2$$

b) Interpret β_1 and test its statistical significance at the 1% level. What can we say about the relationship between education and $log(wage)$?

An additional year of education increases wages by 9% on average. It is statistically significant at the 1% level because the p-value$<0.000<0.01$. We can say that there is a positive and statistically significant relationship between education and log(wages).

c) Test whether experience has a statistically significant effect on $log(wage)$.

Experience and experience-squared are jointly statistically significant above the 1% level because the p-value on the F-statistic is less than 0.0000<0.01. We conclude that experience has a statistically significant effect on log(wages).

d) Is $exper^2$ statistically significant at the 5% level? What does this hypothesis test say about the relationship between $log(wage)$ and experience?

Experience-squared is statistically significant at the 5% level because the p-value on $exper^2$ from part (a) is less than 0.000<0.01. This says that the relationship between experience and log(wages) is nonlinear.

e) Use the approximation of the marginal effect of experience:

$$\%\Delta\hat{wage}_i = \frac{dlog(wage_i)}{dexper_i} = (\hat{\beta}_2 + 2\hat{\beta}_3 exper_i) \times 100\%$$

to find the approximate return to going from 5 to 6 years of experience. What is the approximate return to going from 20 to 21 years of experience? Do you think experience exhibits diminishing marginal returns?

The return to the 5th year of experience is 3.33%. The return to the 20th year of experience is 1.2%. Experience exhibits diminishing marginal returns because the return to the 20th year of experience is less than the return to the 5th year of experience. We could also see this in part (a) where the coefficient on $exper$ is positive, but the coefficient on $exper^2$ is negative.

f) Find the predicted difference in $log(wage)$ between someone with 10 years of experience and someone with 20 years of experience. Is the difference statistically significant? (Note: Do not use the approximation from part (e). We only use the approximation in part (e) when $\Delta X = 1$. In this case, $\Delta X = 10$.)

The predicted difference in log(wages) between someone with 20 years of experience and someone with 10 years of experience is 0.20. To interpret this result, we say that someone with 20 years of experience earns 20% more than someone with 10 years of experience, holding all else constant.

g) At what value of $exper$ does an additional year of experience reduce predicted $log(wage)$? How many people have experience beyond that turning point in the sample? Do you think the turning point is a problem or do you think the results make sense?

The turning point where the return to experience becomes negative is at 28.74 years. About 23% of the sample has more experience than 28.74 years.

On one hand, this could be a problem if the functional form of the regression is incorrect and/or if omitted variable bias is present. It could also be the case that we drew an unusual sample of adults. $n = 526$ is not very large, and it's possible that this is not a random sample.

On the other hand, it's possible that age discrimination exists in the workforce and older people earn less with each additional year of experience compared to "middle-aged" people. There could also be practical reasons why older people earn less – they may switch jobs, switch to part time jobs, or stop learning new skills – all of which would result in lower wages. The only way to understand

what is going on is to either control for other variables that might explain the variance in wages or draw another sample to confirm the results.