EMPIRICAL EXERCISE

Labor and Personnel Economics
2025/26

Remarks:

All work must be your own, with the exception that you are encouraged to discuss your papers with each other but must turn in separate work demonstrating independent thought and investigation.
Answers must not be poems, so try to keep your answers short. But don't forget to explain your results, because a single number without any explanation or the appropriate algebraic passages will be graded lower.
The submission will be by email only (lorenzo.neri@uniroma2.it), no later than one week before the expected exam. Please attach to the email at least one commented text file with the answers, one Stata log file, and one Stata do file, and one text file.

1. On the course web page is a dataset called "acs2007-2008.dta" which has about 96,000 observations on married women aged 18-64 taken from the 2007-2008 American Community Survey. The dataset contains information on hours worked per week, two variables describing weeks worked per year, individual and family income, and other demographic characteristics. From the income, weeks worked and hours worked variables, two measures of hourly wages are constructed. There is no need to submit pages and pages of Stata log files. If possible, edit your results and present them neatly.

a) Describe and summarize the data (no need to report the results). Why does the variable `wkswork1` have so many missing values? Why does the variable `hrwage1` have even more missing values? You will need to do a bit of investigative work, cross-tabulating the variables, browsing the data, etc.

b) There is an indicator variable kidslt5 which takes value 1 if women have children younger than age 5 and an indicator variable inlf which takes value 1 if women are in the labor force. Let's start by examining how labor force participation of these women depends on the age of their children. Write "`tabulate kidlt5 inlf`" in STATA and answer the following questions:

i. What is the overall rate of labor force participation in the sample?

ii. What is the rate of labor force participation among women who have children younger than age 5?

iii. What is the rate of labor force participation among women who have don't have children younger than age 5?

c) Now estimate the elasticity of female labor supply. Regress `lnhours` on `lwage1` for women who are in the labor force. What is your estimate of the female labor supply elasticity?

Re-estimate the model in part (c), but now add log(nonwife's income), the indicator for whether nonwife income is equal to zero, the education variables, age, age squared, the indicator for any children less than 5, the indicator for any children aged 6 and older, the race indicators, and the Hispanic indicator as additional control variables.

d) Does the estimated labor supply elasticity stay the same as in part c)? Explain

e) What is the elasticity of labor supply with respect to non-wife income? Does it have the expected sign?

f) Comment briefly on the coefficients on the education and age variables.

g) How does the elasticity depend on whether women have children younger than six? Test the hypothesis that the elasticity of labor supply is the same for women without children younger than age 5 and with children younger than age 5.

Now we want to repeat the analysis using the largest possible sample, not just the observations with non-missing values of `hrwage1`. For this purpose, create a new variable `hrwage2`, defined just like `hrwage1`, but using the `wkswork2` variable. Note that `wkswork2` is a categorical variable: replace the categorical values of `wkswork2` with the midpoint of the range of values that they represent. So, for example, if `wkswork2 =1` (1-13 weeks), use 7 weeks in calculating `hrwage2`. Take the natural logarithm of this variable, and call it `lwage2`.

h) Rerun the regression in part d) but using lwage2 instead of lwage1. How do the results change? What happens if you estimate the regression with lwage2, but use exactly the same sample as in part d)? [You can mark the observations used in part d) by using the `e(sample)` command. Type `help postest` for further details].

i) Which variable do you think is measured with less error, lwage2 or lwage1? What do your results in part h) say about division bias?

j) Repeat the estimation in part (d), (using lwage1 as your measure of log wages) but now use all the sample, including those that are not in the labor force. Create a new dependent variable `generate lnhours2 = ln(1+hours)` (why is this necessary?). Impute the wage for non-workers as:

   i. $0.01.

ii. The predicted value from a regression of log wages on years of schooling, experience, experience squared, kids less than 6, kids more than 6, and the race indicators.

What is the estimated labor supply elasticity in these two approaches? Are the estimates reasonable? What are the advantages and disadvantages of these two approaches (if any)?