Workshop 4 – Sample Selection and Heckman Models

Motivation: This workshop investigates some determinants of hourly wages for 2,108 Italian individuals who were surveyed five years after graduating from university in the academic year 1999/2000.[1] By the time of the survey, not all individuals were employed, which implies that hourly wages are missing for those graduates. This might create a potential sample selection bias problem. We will therefore focus on the role of selection into the labor force when estimating the effects of certain variables, such as gender – based on which we would draw conclusions about gender wage gaps among Italian graduates.

1. **Data:**

Use the data file **ReflexItaly.dta** that you can download from CANVAS. Describe the variables in the file and make sure you understand what they mean (type *describe* in Stata).

2. **Initial descriptive statistics:**

- Describe the key outcome of interest: *lnwage*. Examine the descriptive statistics of this variable and comment on its main characteristics. (*Hint: use the sum and hist commands to get an overview of this variable.)*
- Note the missing values of *lnwage* for those who are not employed by the time of the survey. Do you see any potential problems with this fact?
- Compare employed and not employed individuals in terms of the remaining variables in the dataset? Do you find any significant differences in any of those variables? Which implications could these differences have when investigating wage determinants?

1 This dataset is only a small subset of the original data (REFLEX survey) used by Figueiredo, Rocha, Biscaia, and Teixeira (2015), "Gender pay gaps and the restructuring of graduate labor markets in Southern Europe", *Cambridge Journal of Economics*, 39(2): 565-598, and therefore it does not allow to replicate the analysis of the paper.

3. **OLS regression for hourly wages**

Run an OLS regression for *lnwage* using the following variables: *female age stem goodgrades unemp6more*.

- Which variables are significantly determining hourly wage according to this regression? How would you interpret the coefficients?
- Would you expect this regression to produce consistent estimates of the gender wage gap and the returns to a STEM degree among Italian graduates? Why/why not?

4. **Probit model of labour force participation (Heckman 1st step)**

Run a probit model for the probability of being employed by the time of the survey, using the variables *female, age, stem, goodgrades, unemp6more, child, livewpartner, livewparents, lookjobbefore*.

- What main conclusions do you derive based on the regression output? How would you interpret the coefficient of *female*?
- Which variables are significantly determining labor force participation? What theoretical hypotheses would you be considering in this part of the analysis?
- Test whether the variables *child, livewpartner, livewparents* and *lookjobbefore* jointly affect labor force participation. What is the main assumption you would be considering here when including these extra variables in the selection equation, and not in the outcome (wage) equation?
- Construct an "inverse Mills ratio" term based on the results of this probit model. This term corresponds to the ratio between the standard normal pdf and the standard normal cdf, evaluated at **x*beta** where **x** are the explanatory variables of the probit, including the constant term, and **beta** is the vector of parameters. (*Hint: use predict xbeta, xb after running the probit. The Stata functions normalden() and normal() give you the pdf and cdf, respectively.*)

>>>>>>>>>>>>>>>>>> Follow-up **<<<<<<<<<<<<<<<<<<**

5. **OLS with control for selection (Heckman 2nd step)**

Repeat the same regression as in exercise 3, but now include the “inverse Mills ratio” term as an additional regressor.

- Would you expect this regression to produce consistent estimates of the returns to STEM education and of the gender wage gaps? Why/why not?
- What is your estimate of the gender wage gap in this case? Why do you think your conclusions changed compared to Exercise 3?
- How do you interpret the remaining coefficients? Compare with the previous results obtained with the baseline OLS model and discuss why some results might have changed and others did not. *(Hint: check the correlation between each of the regressors and the IMR in the sample of employed individuals)*
- How do you interpret the coefficient of the inverse Mill’s ratio? Is it significant? What are the implications of your conclusion?

6. **Heckman two-step procedure in Stata**

- Now run the same analysis using the Heckman procedure in Stata with the twostep option. Compare this with your results from question 5. Are there any differences? Why (not)?
- Run the above analysis using the Heckman procedure in Stata with maximum likelihood (the default option). Compare this with your results from the Stata two-step procedure. Are there any differences? Why?
- Which conclusions do you derive from the sign and significance of “rho” in this output? What would you conclude regarding the selection effects in this example?

7. **Exclusion restrictions**

Our main wage equation leaves out the four variables *child, livewpartner, livewparents* and *lookjobbefore*.

- Test whether these variables are significantly related to hourly wages. What do you conclude regarding the strength of these variables as exclusion restrictions?
- Run the Heckman procedure without any exclusion restrictions. Compare to your results in question 6. Comment on your results.

>>>>>>>>>>>>>>>>>> Follow-up **<<<<<<<<<<<<<<<<<<<**