Problem Set 3

Statistics for International Relations and Political Science 2
IHEID
2021-2022

May 5, 2022

Please complete the following problem set and submit the answers on Moodle. For R-related questions, please submit the R script in addition to the results.

1. You would like to estimate the effect of job training grant (*grant*) on the hours of job training per employee (*hrsemp*), controlling for the number of employees in the firm (*employ*). Please load the data in JTRAIN (Wooldridge database). The dataset includes a panel data from year 1978, 1988, and 1989. There are two time dummies, *d88* and *d89*, as well as an identification code for each firm (*fcode*).

    - First, you would like to see whether job training grant had any effect on *hrsemp* by using the data from year 1987. Write your equation for this cross-sectional model and report OLS results.
    - Now you want to use the data from 1987 and 1988. Write your equation for this panel data, including the term for unobserved heterogeneity.
    - Why would you need panel data? What are the added advantages?
    - Write the first-differenced equation. Use the first difference estimation and report your main results. Does having a grant significantly increase the hours of training per employee?
    - Now write the time-demeaned equation for fixed effects estimation. Do you expect the estimated coefficient to be different from what you get from the first difference method? If so, why, and if not, why not? Report the results from the FE estimation.
    - Now you want to use all three years of data. Report the results from both FE and FD estimation. Do you find any difference in the estimate effect of *grant*?
    - Both FE and FD estimation require strict exogeneity. In this case, what would this imply? Can you conduct any test to see if the assumption holds? If so, please conduct the test and report the result.

- You also want to see if controlling for the presence of a trade union (*union*) makes any difference. Conduct either the FD or the FE estimation while adding *union* as an additional control variable. What do you find? Explain your finding.

2. Carefully read Carneige and Marinov (2017).

   - What is the IV used in this paper, and why do you think the authors used an IV approach? Do you think that it was necessary to use the IV estimation for the research question?
   - What are the assumptions underlying the IV approach here. What are the conditions for a valid IV and what do authors do to show that the conditions are met? Do you think that the authors provide enough justification? If so, how? If not, why not?
   - Now we are going to replicate the results from the paper. Use the provided *Final_Main.dta*. All of the variables are explained in the codebook. Write the equation for the first and second stage estimation.
   - First, conduct 2SLS manually to estimate the effect of aid on the CIRI Human Empowerment Index. Run the first-stage regression. Be cautious to include only the observations that have non-missing values for the dependent variable and from year 1987 and onward. What do you find? Can you conclude that that the IV is a strong instrument? Then, use the estimated values to run the second stage regression and report the results. Can you directly report the standard errors from this analysis?
   - Now use the *ivreg* command to conduct the same 2SLS estimation and report the result. Do you find the same result as from manual estimation?
   - Conduct OLS estimation and compare the OLS estimator and IV estimator. Do you find a difference? Does the result indicate that an IV approach is needed?
   - Alternatively, we can also conduct an endogeneity test. Run the first-stage regression and import the estimated residuals to the second stage regression. What do you find? Can you reject the null hypothesis that the explanatory variable is exogenous?