Advanced Applied Econometrics[*]

April 25, 2024

Exercise on Panel Data

In this exercise you are going to investigate the determinants of wages in a panel data setting. The data for this exercise were originally supplied by Jeffrey Wooldridge, contains information on 545 young men in the United States covering the period 1980-1987. The dataset is an extract from the National Longitudinal Survey of the U.S.Department of Labor. The dataset consists of three different parts which contains the following variables:

nr person identifier

year year

lwage log of the wage rate

educ education (years of schooling)

exper work experience (years)

black dummy = 1 if person is black

hisp dummy = 1 if person is hispanic

married dummy = 1 if person is married

Use the dataset to answer the following questions:

(a) The data for this exercise are contained in four different datasets which you need to combine into one dataset: marriage-data.xls, wage-data.dta, experience-data.dta and background-data.dta. Start by importing the Excel data into Stata and saving this as a new data file in Stata format. Then use the commands reshape and merge to combine the information in the three dataset into a consistent panel in the long form.

(b) Is the final dataset a balanced or an unbalanced panel dataset?

[*]

(c) As a first step ignore the fact that the data are a panel and regress the logarithm of the wage on the dummy variable married, which is equal to one if a person is married in a particular year. What could explain your results?

(d) Now include the variables for experience (exper), membership in a union (union), the years of education (educ), and the dummy variables (black) and (hisp). Howdoes the inclusion of these variables in the regression change the marriage premium? What do you conclude from this?

(e) A key omitted variable from the regression in (c) is the unobservable variable “talent”. To capture the effect of this variable include “person fixed effects” in the regression and estimate the regression with the help of the within transformation (you need the STATA commands areg or xtreg for this, or to specify variables using the i. pre-script). What light do the results of this regression shed on the possible explanations for the observed correlation between marriage and wages? Why does STATA now drop the variables educ, black and hisp from the regression?

(f) A common problem in panel data regressions is autocorrelation. Are your results robust to allowing for autocorrelation of the error termacross the observations for each person?

(g) In the lecture you have been told that estimating a regression with fixed effects with the within transformation is completely equivalent to using the dummy variable regression. To check that this is true (and that it takes a lot longer to estimate the dummy variable regression in larger datasets) estimate the model using the dummy variable regression. Why are the results not completely identical with those that you obtained using the within transformation?

(h) Also include time fixed effects in the regression. What might these effects control for? Do your results change as a result of this?

(i) If the variable married changes over time, this could be because a person marries or because a person gets divorced. How many people get divorced in the dataset? Estimate your the regression from the previous question again excluding the people who get divorced. Are your results robust to this change in the sample?