Problem Set 4

Labour Economics, Winter Semester 2025/26

*Submit by Sunday, 18 January, 22:45h **on Moodle!***

Learning objectives

- Read a scientific paper and extract key overarching information.
- Study functional forms of earnings profiles in education and over the life-cycle.
- Different ways of computing standard errors and interpretation.
- Fixed effects versus differences estimation.

Tasks

The purpose of this assignment is to estimate the rate of return to education using family fixed effect methods by replicating the results of a classic paper on the topic. In particular, you are asked to reproduce some of the results of the paper *Ashenfelter, O. and A. Krueger, "Estimates of the Economic Return to Schooling from a New Sample of Twins," American Economic Review, Vol. 84 (Dec., 1994): 1157-1173.* The dataset used by Ashenfelter and Krueger is on Moodle.

1. Read Ashenfelter and Krueger's paper. Answer the following:

   a) What is their research question and what are the difficulties in answering this question?

   b) What is the autors' approach to overcome these difficulties and (to what extent) does it convince you?

   c) What are their key results? Are they surprising / convincing in your view?

2. Read in and explore "`twins_long.dta`". Refer to the article and information below for explanations of the variables. Plot workers' education, age, and experience against their (log) wages.

3. Run an OLS regression of log wages on a constant, schooling, age, age-squared, gender and a racial indicator ("`white`" dummy).

   a) What do the coefficients on age and age-squared imply about the life-cycle profile of earnings? Compare also to the plot from 2.

   b) Do your estimates match those in column (1) of Table 3 in Ashenfelter and Krueger? What kinds of biases is the OLS coefficient on `educ` likely to incorporate?

4. Post-estimation adjust the standard errors, once reporting coefficients with HC1 robust errors (e.g., `coeftest(model, vcov = vcovHC, type = "HC1")`) and once by additionally clustering at the family (`famid`)level. Do you know why one might want to do this and does the significance of the coefficients change?

5. Now create dummy variables for each level of schooling and run the above OLS regression with those education dummies instead of the educ variable. Ideally, omit the dummy for high school (`educ=12`) from the regression, so that all other education levels' coefficients are relative to this largest category. Do the estimated coefficients indicate that the effect of education of log wages is linear?

6. Return to the specification that is linear in education. Compare your OLS estimate with one that incorporates a "family fixed effect" (`factor(famid)`). Focus on the coefficient for education (variation in the other variables is mostly removed by the family fixed effect). What are in your view the (economic) reasons for any differences?

7. An alternative to fixed effect estimation is to run a model on first differences. Write out the equation in first differences, estimate that differenced equation and compare the results to column (v) in Table 3. You can first-difference in `twins_long.dta` or compute differences between twins in the "wide" version of the same data `twins.dta`, also provided on Moodle. Should the coefficient of `educ` in the family fixed effect and the first-differenced model be the same? Are they the same?

Notes: You can work in teams of 1–3 students. Please upload your pdf-file with responses. It should be clear which answers in the `.pdf` refer to which question. If you work in a team, each member has to upload the group's solution and note whom they worked with.

... for some further background information, turn to next page ...

1) Background for the study

This classic twins study sought to answer what seems a simple question: By how much will another year of schooling most likely raise one's income? Attempts had been made to estimate the value of a year's education in previous studies, but previous estimates may have been imprecise for two reasons. The first, most obvious reason is the difficulty of extracting the education's effect on income from the effect that other variables related to education have on income. That is, a worker's natural ability, his family background, and his innate intelligence are all possible confounding factors that must be controlled for to estimate the effect of education on income accurately. Thus this study interviewed twins, collecting information about education, income, and background. Because monozygotic twins (twins from a single egg) are genetically identical and have similar family backgrounds, they provide an excellent control for confounding variables.

The second difficulty in measuring the effect of income on education has to do with the false reporting of education levels, and this study is the first to address it. Since people are more likely to report a higher education level than they have actually attained, especially in face to face interviews, the data will contain a number of people with lower education levels in the higher education categories. Thus, since education usually increases income, estimates for the precise amount of this increase will be too low. To correct for this bias the researchers interviewed the twins separately and recorded two entries for each individual's education level: his self-reported education level and the education level reported by his twin. This allowed them to estimate the "measurement error" of reported education levels and correct for it. The result was a much higher estimate of the effect a year of education is likely to have on one's income. In fact, this study's estimates were higher than those of all previous studies, which did not correct for measurement error in education level.

2) Brief description of the data

The data were collected by a team of five interviewers at the 16th Annual Twins Day Festival in Twinsburg, Ohio, in August 1991. A booth was set up at the festival's main entrance, and an ad inviting all adult twins to participate in the survey was placed in the festival program. In addition, the interviews roamed the festival grounds, approaching all adult twins for an interview, and almost every pair of twins accepted.

The key variables are:
famid = family id
age = age of the person
educ1 and educ2 = education attainment of twin 1 and 2 , respectively
lwage1 and lwage2 = the natural log (ln) of the hourly wage of twin 1 and 2 , respectively
male = an indicator variable equal to one if the person is male, zero otherwise
white = an indicator variable equal to one if the person is white, zero otherwise