**Part A: Quarter of birth, returns to schooling, instrumental variables and more**

This part is based on an extract from the data used by Angrist and Krueger in their seminal study “Does Compulsory School Attendance Affect Schooling and Earnings?” ̧ *Quarterly Journal of Economics*, November 1991 (https://www.jstor.org/stable/2937954). The data can be found on the course website in Stata format, AngristKrueger\_3039Cohort.dta.

As usual, make sure to present your results in a neat format.

1. a)  Angrist and Krueger are interested in estimating the causal effect of schooling on log earnings.
   1. Estimate an OLS regression of log weekly wages (lwklywage) on years of schooling (educ), without any additional controls.
   2. Repeat part (i), but now add controls for race, SMSA, marital status, regional dummies, and year of birth dummies.

Compare the coefficients in the two specifications, and comment on the results. Is the coefficient on years of schooling likely to be a consistent estimator of the causal effect of interest? Why or why not? What are the possible sources of bias?

1. b)  Angrist and Krueger propose using quarter of birth as an instrument for years of schooling. Explain why this can be a valid instrument. Specifically, why does it satisfy the first stage (or “relevance” condition, Cov(Z,X) ≠ 0)? Does it satisfy the exclusion restriction?

c) Define a new variable QOB1, equal to 1 if the individual is born in the first quarter of the year, and equal to zero otherwise.

1. Estimate the *first stage regression*: the regression of years of schooling on QOB1 (without any additional explanatory variables), and report the coefficient and standard error.
2. Estimate the *reduced form regression:* the regression of lon weekly wage on QOB1 (without any additional explanatory variables), and report the coefficient and standard error.
3. Estimate a 2SLS regression of log weekly wage on years of schooling, using QOB1 as an instrument. Use the ivregress 2sls command in Stata. Report the coefficient and standard error.
4. What is the relationship between the estimates you obtained in parts i, ii and iii of this question?
5. d)  Suppose that you have a faulty version of Stata, in which the ivregress command doesn’t work. Explain how you could obtain the 2SLS coefficient (of part c.iii) using only the regress command. Implement this estimate. Is the coefficient and the standard error the same as in part c.iii?
6. e)  Suppose that you have an even faultier version of Stata, in which neither the ivregress nor the regress commands work. Explain how you could obtain the 2SLS coefficient of part c.iii using only the summarize command (and if statements). Implement this estimate.
7. f)  Repeat parts c.i-c.iii but now use three dummies for quarter of birth as your instruments (QOB1, QOB2 and QOB3), and include all the explanatory variables (the race dummy, the SMSA dummy, the marital status dummy, the region dummies, and year of birth dummies). What is now the relationship between the first stage, the reduced form, and the 2SLS coefficient?
8. g)  Compare the 2SLS estimate you obtained in part f) with the OLS coefficient you obtained in part a.ii. Comment on the results. Is the sign of the bias of the OLS coefficient what you expected?
9. h)  Regress each one of the control variables you used in part f) on the three quarter of birth dummies. Report the F-statistic and the p-value for the test of the null hypothesis that all the coefficients are equal to zero. Is there any systematic pattern between the quarter of birth and the control variables? Why would one be interested in carrying out this test?

i) Define a new set of variables: 1 educ ≥ 𝑗  
educ\_gt\_𝑗 = 0 otherwise , for 𝑗 = {8,9, ... ,16}

Regress each one of these variables on the three quarter of birth dummies, and all of the control variables that you used in part f). Calculate the F-statistic for the test that all three coefficients on the quarter of birth dummies are equal to zero. For which values of 𝑗𝑗 is the test statistic greater than 10? Why does this result support Angrist and Krueger’s identification strategy?

**Part B: The earnings-experience profile**

For this question, you will download the data set ps4\_CPS-ASEC\_1970-2020.dta. The data set contains information on earnings, wages, and demographic characteristics taken from the Annual Social and Economic Supplement of the CPS (the “March CPS”) for 1970, 1980, 1990, 2000, 2010 and 2020.

Use the describe and summarize commands in Stata to familiarize yourself with the data set. The key variable, lnearn, is the log of yearly earnings in the previous calendar year. The data includes only individuals aged 20 to 64, who were salaried workers in the previous year, and worked full-time (i.e., more than 35 hours a week) and year-round (more than 40 weeks a year). Values of earnings, weekly wages and hourly wages below the bottom 1st percentile and above the top 99th percentile were set to missing. Answer the following questions.

a) It is well known that the earnings-experience profile is concave: earnings rise with labor market experience, but at a declining rate.

1. How does the theory of human capital rationalize this finding? In your answer, make sure to explain the role of on-the-job training, and how it evolves over the life-cycle.
2. Do you expect the earnings-experience profile steeper for college graduates or for high-school graduates? Why?
3. Do you expect the earnings-experience profile to be steeper for men or for women? Why?
4. b)  Theoretical models typically predict that earnings rise with *actual* labor market experience, which is not observed in the data. Instead, the data set contains a variable, exper, equal to *potential* experience. This is defined as: age – 16 for high school dropouts (educ\_hsdrop==1); age – 18 for high school graduates (educ\_hsgrad==1); age – 20 for people with some college (educ\_somecoll==1); age – 22 for college graduates (educ\_collgrad==1); and age – 24 for people with more than a college degree (educ\_collplus==1).

Discuss the advantages and disadvantages of using potential experience as a proxy for actual labor market experience. Do you think that this proxy is equally good for men and women?

1. c)  Using only males (sex==1) for the year 2020, estimate the following three regressions (no need to report the results), and calculate and store the predicted values.
   1. Log real earnings experience and experience squared. (Quadratic specification)
   2. Log real earnings on experience, experience squared, experience cubed, and experience to the fourth. (Quartic specification)
   3. Log real earnings on a full set of experience dummies.

On the same graph, plot the predicted values from the three regressions against experience. Which parametric specification does a better job of capturing the earnings-experience profile? The quadratic or quartic specification?

1. d)  Repeat part a) using only female (sex=2) and the year 2020.
2. e)  Plot on the same graph the experience-earnings profile for males and females (i.e., plot the predicted values from the regressions with a full set of experience dummies). Are there visible differences in the returns to experience by gender?
3. f)  Pool the male and female samples together (just for the year 2020). Estimate a regression of log earnings on a quartic in experience, a gender dummy, and the interaction between the gender dummy and the quartic in experience.
   1. Calculate the difference in predicted ln earnings between a worker with 0 and 10 years of experience, and the difference in predicted ln earnings between a worker with 10 and 20 years of experience, separately for men and women.
   2. Test whether the difference in predicted ln earnings between a worker with 0 and 10 years of experience is the same for men and women. How about the difference in predicted ln earnings between a worker with 10 and 20 years of experience?

For this part, use the lincom command in Stata to calculate linear predictions of the regression coefficients and perform statistical tests).

1. g)  Do the results in parts e)-f) support the notion that women invest less in on-the-job training?
2. h)  Now, use the model in part f) to complete the table on the following page.
3. i)  Comment on the results in the table. In particular, have the returns to experience changed over time? Are there differences between men and women? Between men with and without a college degree? Between women with and without a college degree?