# STATA Exercise 4: Instrumental variable estimation and the returns to education

May 2, 2019

Understanding the causal effect of education on (productivity and) earnings is of key interest in policy-making. We will consider this question by using a data set from Angrist and Krueger (1991).[1] The data set, *AngristKrueger1991.dta*, contains individual observations for birth cohorts 1920-1949 from the US 1980 Census, 5 percent Public Use Sample. In this exercise, we will only consider the birth cohorts 1930-1939 (the year of birth variable is *yob*) and you may already when you have read the data set exclude all other observations. This will reduce the sample from $1,063,634$ to $329,509$ observations.

1. Regress the log weekly earnings (the variable is called *lnwklywge*) on explanatory variables, $X_i$, and years of schooling $S_i$ (the variable is called *educ*)
$$\ln w_i = X_i\beta + \gamma S_i + u_i$$
Estimate this regression twice using the following two sets of explanatory variables:

   a) Only include year of birth dummies, which can be constructed from the *yob* variable.[2]

   b) Include years of birth dummies, a dummy for race (*race* equals 1 if black), a dummy for living in center city (*smsa*), a dummy for being married (*married*), and the following 8 regional dummies: *neweng*, *midatl*, *enocent*, *wnocent*, *soatl*, *esocent*, *wsocent*, and *mt*.

2. What is the (potential) problem with this regression? In which direction do you expect the bias to go?

Angrist and Krueger (1991) suggest to use quarter of birth as instrument for schooling. In the US, most states require students to enter school in the calendar year they turn 6 years. Therefore, school starting age is a function of date of birth:

[1] Angrist, J.D. and A.B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics*, Vol. 106, No. 4, pp. 979-1014.

[2] You can include year of birth dummies by writing i.yob in the list of variables when using the command regress.

- A child born on December 31: Will enter school at age 5 years and 8 months.
- A Child born on January 1: Will enter school at age 6 years and 8 months.

Compulsory schooling laws typically require that students remain in school until their 16th birthday:

- A child born on December 31: Will have to be in school for at least 10 years and 4 months.
- A Child born on January 1: Will have to be in school for at least 9 years and 4 months.

However, there are exemptions so school starting age is not a deterministic function of date of birth. There are two reasons why compulsory school laws presumably should be effective. First, in most US states, children are prohibited to work during school hours before they have reached the compulsory schooling age in the state. Second, young workers need a work permit, which is issued by the schools.

3. Construct two graphs: A) A graph which shows the first-stage (i.e. as primary axis use quarter of birth, i.e. 1930q1, 1930q2, 1930q3,..., 1939q4[3], and as secondary axis use average years of schooling) and B) A graph which shows the reduced-form relationship (i.e. still use quarter of birth as primary axis, but now average log weekly earnings (*lnwklywge*) as secondary axis). Hint: Use Stata's `graph twoway connected` with the `mlabel` option to get Stata to write each quarter (*qob*) next to each point in the figure. The resulting figures should look like those in A&P (2009) p. 119.

4. We want to understand which level of education the quarter of birth instrument affects. To examine this define the following three dummies: i) A dummy for being high school graduate (12 years or more), ii) a dummy for being college graduate (16 years or more), and iii) a dummy for having a doctoral degree (20 years or more). Then, examine at which education level quarters of birth have most effect.

5. Let $Z = 1$ if $qob = 1$ and $Z = 0$ otherwise. Compute each of the components of the Wald estimator, i.e. estimates of $E\left(Y|Z=1\right)$, $E\left(Y|Z=0\right)$, $E\left(S|Z=1\right)$, and $E\left(S|Z=0\right)$, where $S$ is year of schooling and $Y$ is log weekly earnings. Test if the estimate of $E\left(Y|Z=1\right)$ is statistically different from the estimate of $E\left(Y|Z=0\right)$ and similarly for $E\left(S|Z=1\right)$ and $E\left(S|Z=0\right)$. Compute the Wald estimate from the components listed above.

[3] You can construct this variable by using the variables *qob* and *yob*. For example, you can let the first quarter in 1930 be 1930.1 and the second quarter in 1930 be 1930.2 etc.

6. Use the Stata command `ivregress` to estimate the Wald estimator using heteroscedasticity-robust standard errors. Compare the estimate to the OLS estimate. Are the relative magnitudes as expected?

7. Characterize the compliers by computing: i) The share of compliers, ii) The share of compliers among treated, iii) The share of compliers among non-treated, iv) Share of compliers among blacks ($race = 1$). In order to do this, we will redefine the endogenous variable to a dummy for being high school graduate (i.e. $D = 1$), and let the instrument be defined a dummy for being born in the 3rd an 4th quarter of births (i.e. $Z = (qob \geq 3)$).

8. Estimate 2SLS and include respectively explanatory variables a) and b) from question 1. Besides this, you should examine if it matters whether you only use a dummy for first quarter of birth as instrument compared to using dummies for the first three quarters of birth (i.e. three instruments). Why would you prefer the 2SLS estimate over the Wald estimate?

9. To obtain a more efficient estimate, Angrist and Krueger (1991) construct a full set of interaction dummies between the three first quarters of birth dummies and year of birth dummies (29 dummies in total[4]), which they use as instruments. Estimate 2SLS using these instruments with the two sets of control variables a)-b) in question 1. Compare your estimate of the returns to schooling to the OLS estimates from question 1. What variation in the data identifies the effect of education on the log weekly wages?

10. To explain the lower estimates of the returns to schooling in the previous question, repeat the estimations and use the option first to Stata's `ivregress` command to examine the first stage results. Furthermore, use the post estimation command `estat firststage`. What is the problem?

11. Use LIML and JIVE to estimate the models with i) the three instruments, ii) the full set of interaction dummies between the three first quarters of birth dummies and year of birth dummies as instruments. Do LIML and JIVE give noticeable different estimates compared to 2SLS?

[4] You can include these interaction variables by writing i.yob#i.qob in the (relevant) variable list when using `ivregress`.