MOEC0021
Empirical Methods
University of Zurich HS 2020
PROF: Emanuele Dicarlo / TA: A. Jenni, M.R. Greco, S. Bagagli, L. Schubert

---

Problem Set 3 – Instrumental Variables

This problem set is due on the **30th of November** at **23:59**. Solutions should be turned in on OLAT in a single PDF file. Please name your file as GroupName_PS3.pdf. Include any code you wrote to answer the questions in the file.
One of your goals is to communicate efficiently. Please keep your answers succinct. Lengthy answers will be marked down.

---

1. Theory – Endogenous Instruments

Consider the simple bivariate model:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

Recall that in the bivariate case:

$$plim\hat{\beta}_{ols} = \beta + \frac{cov(x, \varepsilon)}{var(x)}.$$

Suppose x suffers from endogeneity, i.e. it is likely that $cov(x, \varepsilon) \neq 0$. You aim to address this problem by proposing an instrument z.

(a) What conditions must z satisfy to be a good instrument?

(b) Can these condition be tested in this context? How?

(c) When would z be a *weak* instrument?

Consider now the following ratio:

$$\frac{plim\hat{\beta}_{IV} - \beta}{plim\hat{\beta}_{ols} - \beta}$$

(d) What is the numerator of this ratio? What is the denominator?

(e) Derive an expression for this ratio in terms of variance and covariance terms.
Hint: Recall that in the bivariate case with a single instrument:

$$plim\hat{\beta}_{IV} = \beta + \frac{cov(z, \varepsilon)}{cov(z, x)}$$

(f) Discuss *intuitively* what it means if this ratio equals 0. What does it mean if it is larger than 1?

(g) Recall our *only* assumption so far: $cov(x, \varepsilon) \neq 0$. Under which additional condition(s) is the ratio equal to 0?

(h) Discuss under which conditions it would be preferable to rely on the OLS estimator instead of the IV estimator. You can base your argumentation on an example.

2. Empirical Application: IV Regression

This question is based on the paper by Bonjour, Cherkas, Haskel, Hawkes and Spector ("Returns to Education: Evidence from UK Twins," The American Economic Review, 2003), which we will refer to in what follows as BCHHS. Start by reading the paper - it is available on OLAT and not very long. The dataset of BCHHS is available on to download here: `http://bit.ly/1YATkWe`.
The same data set is also available on the website of the American Economic Association (http://www.aeaweb.org/articles.php?doi=10.1257/000282803322655554), i.e. it is exactly the same dataset that BCHHS have submitted to the journal alongside their paper.
The data set contains the following variables: $family$ (family number), *twinno* (twin number within the family: 1 or 2), *earnings* (hourly wage), *highqua* (each twin's reported years of schooling), age, and some other variables that we will not be using here. It also includes *twihigh*, which is each twin's estimate of the years of schooling *of the other twin.* Open the data set in your preferred statistical software and familiarize yourself with the variables listed above.
Generate the variables *lnearn* (log earnings) and *agesq* (the square of age).

(a) Use the data set to reproduce the results in columns (2) and (3) of Table 2 in BCHHS. That is, perform the following regressions:

- Regress log earnings on years of schooling, age and age squared using OLS.
- Estimate the same model, but use twins estimated years of schooling as an instrument for years of schooling.

i. Do you find any discrepancy between your results and those reported in BCHHS? Is the discrepancy "serious"? Why or why not?

ii. The authors do not report the coefficient on the constant term. Is there any important information contained in that coefficient?

iii. The main coefficient of interest here is the coefficient on education (= years of schooling). Assume for the moment that your IV results are consistent and explain the interpretation of this coefficient in both regressions.

(b) Let's think harder about the IV regression we just ran.

i. Try to provide at least *two* reasons why years of education (*highqua*) might be endogenous. Your answer should not be longer than 10 lines.

ii. For *each* reason, conjecture about the likely sign of the bias the source of endogeneity would have on your estimate of the returns to schooling.

iii. For each reason, evaluate the relevance and exogeneity of using a twin's sibling's report of their years of schooling (*twihigh*) as an instrument for their own report of their years of schooling (*highqua*).

iv. Examine the difference between the OLS and IV results. Did the IV results move in the direction you expected? Why or why not?

v. Report the First Stage regression for your IV estimation and test whether the instrument is weak.

vi. Do you "believe" these results? Why or why not?

Next, reshape the data set such that the unit of observation becomes the family (i.e a twin pair) instead of a twin. This reduces the number of observations by a factor of two. Furthermore, generate new variables that contain the difference of log earnings between the twins (*dlnearn*), the difference in schooling (*dhigh*) and the difference of twins estimated years of schooling between twins (*dtwihigh*).

(c) Now reproduce the results in columns (4) and (5) of Table 2 in BCHHS. That is, perform the following regressions:

- Regress the difference in log earnings (*dlnearn*) on the difference in schooling (*dhigh*) using OLS without a constant.
- Estimate the same model, but use twins estimated years of schooling as an instrument for years of schooling.

i. Why don't we include differences in age and age-squared in the regression?

ii. Why did we exclude the constant regressor from our regressions? What do you expect for the estimated coefficient on the constant if it were to be included in your estimation? Redo the estimations above including the constant. What do you find?

iii. Compare the **OLS** results here to those you obtained and analyzed in Questions (a) and (b). What is the advantage of taking differences between the observations of identical twins before estimating the returns to schooling? Does this address any of the endogeneity problems you mentioned in your answer to Question (b).i.?

iv. Now compare the **IV** results here to those obtained and analyzed in Questions (a) and (b). What is the advantage of taking differences *and* instrumenting with the difference in twins' sibling-estimated years of schooling (*dtwihigh*)?

v. Do you "believe" in these results? Why or why not?

We now follow another paper (V Amin "Returns to education: Evidence from UK twins: A Comment", American Economic Review, 2011). The paper is again on OLAT, and I recommend you read it. The idea of the paper is to check whether the BCHHS result is robust to outliers.[1] For this purpose create a new variable that contains the absolute value of the difference in earnings between individual twins (*absearn*). Any observation for which this absolute value is greater than 60 (British Pounds per hour) is considered an outlier (There should be 4 such instances).

(d) Repeat the estimation as described in Question (c) but drop the four outlier families.

i. Report your findings. What happens to the magnitude and statistical significance of the education coefficient in column (5) of Table 2 in BCHHS?

ii. How do you now interpret your IV-on-the-differenced data results?

iii. Does it make sense to drop the outliers from the sample or are these regular observations that contain important information?

[1] An outlier is defined as "a data point that is very much bigger or smaller than the next nearest data point."