Lecture 5 Instrumental Variables

Filipa Sá

King's College London

Semester 1, 2019/20

Readings

- Angrist and Pischke chapter 3
- Angrist, Lavy, Schlosser (2010), "Multiple Experiments for the Causal Link between the Quantity and Quality of Children", *Journal of Labor Economics*, Vol. 28, No. 4, pages 773-824
- Background: Wooldridge chapters 13-15

The quantity-quality trade-off

- World population increased from 3 billion in 1960 to 6 billion in 1999. The question of how population growth affects living standards has a macro and a micro aspect:
  - Macro: Thomas Malthus argued that, as population size increases when food output increases, productivity gains do not improve living standards. Instead, most people live at subsistence level.
  - Micro: "Quantity-Quality trade-off" — a reduction in family size increases parental investment in children. For example, parents with fewer children invest more in their education.
    - This is the hypothesis tested in Angrist, Lavy, Schlosser (2010) — ALS, using data for Israel

The quantity-quality trade-off

- **Selection bias** — parents in larger families are different from parents in small families. For example, parents in larger families tend to be less educated. And children of less educated parents tend to be less educated themselves.
- The ideal experiment — random assignment
  - We would like to randomly select some families to have additional children and compare the outcomes of children later in life (educational level, earnings, etc.)
  - This is not possible

The quantity-quality trade-off

- But we can take advantage of some situations that cause **exogenous variation** in the number of children:
  - Having twins
  - Sibling sex compositions — families whose first two children are both boys or both girls are more likely to have a third child
- Multiple births are more frequent among mothers who are older and in some racial and ethnic groups, so it is not unrelated to individual characteristics
- But the sex composition of the siblings is largely determined by chance

The quantity-quality trade-off

- These factors that generate exogenous variation in the number of children can be used as **instrumental variables**
- An instrument needs to meet three conditions:
  - The instrument has a causal effect on the variable whose effect we are trying to capture. In this case, having twins or having two children of the same sex increases the number of children that a family has. This causal effect is called the **first stage** (we will see why soon).
  - The instrument is randomly assigned, in the sense of being unrelated to the omitted variables we might like to control for. In this case, having twins or two children of the same sex is not related to family background. This is more likely to hold for sex composition than for multiple births. This is known as the **independence assumption**.

The quantity-quality trade-off

- An instrument needs to meet three conditions:
  - The only way through which the instrument affects the outcome variable is through its effect on the variable of interest. In this case, having twins or having two children of the same sex only affects the level of education of the child through the effect on family size. This is known as the **exclusion restriction**.

The quantity-quality trade-off

- Suppose we are interested in studying the causal effect of number of children in the family (D_i) on the level of education of individual i (Y_i). We would like to estimate:

$$Y_i = \alpha + \lambda D_i + u_i$$

- This regression suffers from selection bias. **Instrumental variables (IV)** estimation proceeds in two steps:
  - First, it estimates the effect of instrument (Z_i) on family size (D_i). For example, Z_i can be a dummy variable for multiple second births. This is the first-stage regression:

$$D_i = \alpha_1 + \phi Z_i + e_{1i}$$

  - Then, it saves the fitted values of the first-stage regression ($\widehat{D}_i = \alpha_1 + \phi Z_i$) and uses them in the second stage. The second-stage regression is:

$$Y_i = \alpha_2 + \lambda_{2SLS} \widehat{D}_i + e_{2i}$$

  - Because we proceed in two stages, IV estimation is also called **two-stage least squares (2SLS)**

The quantity-quality trade-off

- What is the intuition behind IV estimation?
  - Family size (D_i) is endogenous and depends on family characteristics (for example, parental education).
  - We find a variable that is strongly related to family size, but is exogenous (for example, multiple births or sibling sex composition). This variable is the instrument.
  - We look at the link between the instrument and family size (first stage) and retain only the variation in family size that can be explained by the exogenous instrument. This is given by the fitted values $\widehat{D}_i$.
  - We look at the link between earnings and the exogenous variation in family size generated by the instrument. This is the second-stage regression.

The quantity-quality trade-off

- A direct regression of education on family size has no causal interpretation because of the selection bias.
- But we could regress education on the instrument (a dummy for multiple births). This would have a causal interpretation, because the instrument is exogenous.

$$Y_i = \alpha_0 + \rho Z_i + e_{0i}$$

- This regression is the **reduced form**.
- It can be shown that the IV coefficient is the ratio of the reduced form and the first-stage coefficients:

$$\lambda_{2SLS} = \frac{\rho}{\phi}$$

The quantity-quality trade-off

- We can extend this two-step regression framework to include control variables like maternal age (A_i)
- This can be important because the multiple births instrument is unlikely to satisfy the independence assumption — older women are more likely to have twins
- But if we control for maternal age, this instrument is as good as randomly assigned
- Reduced form:

$$Y_i = \alpha_0 + \rho Z_i + \gamma_0 A_i + e_{0i}$$

- First stage:

$$D_i = \alpha_1 + \phi Z_i + \gamma_1 A_i + e_{1i}$$

- The fitted values from the first stage are:

$$\widehat{D}_i = \alpha_1 + \phi Z_i + \gamma_1 A_i$$

The quantity-quality trade-off

- 2SLS estimates are constructed by regressing Y_i on both $\widehat{D}_i$ and A_i:

$$Y_i = \alpha_2 + \lambda_{2SLS}\widehat{D}_i + \gamma_2 A_i + e_{2i}$$

- Notice that the control variable (A_i) is included in both the first and second stages.
- In practice, we do not do 2SLS manually one stage at a time.
- Instead, we use the command *ivregress 2sls* in Stata
- This ensures that all control variables are included in both the first and second stages and also calculates the correct standard errors

The quantity-quality trade-off

- We can extend 2SLS estimation to cases when we have more than one instrument:
  - A dummy for multiple births Z_i
  - A dummy for same sex siblings W_i ($W_i = 1$ if the family has two boys or two girls and 0 otherwise)
  - The first stage is given by:

$$D_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma_1 A_i + \delta_1 B_i + e_{1i}$$

  - We have included an additional control variable (B_i) indicating whether the first child is a boy. This is because boys are born with a slightly higher probability than girls, so the probability of same sex siblings is slightly higher when the first child is a boy.

The quantity-quality trade-off

- The reduced form with two instruments is:

$$Y_i = \alpha_0 + \rho_t Z_i + \rho_s W_i + \gamma_0 A_i + \delta_0 B_i + e_{0i}$$

- The second-stage regression is given by:

$$Y_i = \alpha_2 + \lambda_{2SLS} \widehat{D}_i + \gamma_2 A_i + \delta_2 B_i + e_{2i}$$

where the fitted values come from the first-stage regression:
$\widehat{D}_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma_1 A_i + \delta_1 B_i$

The quantity-quality trade-off

- 2SLS with two instruments produces a weighted average of the estimates we would obtain if we just use the instruments one at a time
- 2SLS is flexible: we can include control variables and more than one instrument. Also, instruments do not need to be dummy variables.

The quantity-quality trade-off

- Table 3.4 reports the first stage estimates
- Twins instrument:
  - Column (2) shows that first-born Israeli adults whose second born siblings were twins were born in families that had about 0.44 more children than those raised in families where the second birth was a singleton
  - The first-stage estimate is larger than the estimate without controls reported in column (1)
  - The OVB formula tells us that twins are associated with characteristics that reduce family size, like maternal age. Controlling for maternal age boosts the twins first stage

The quantity-quality trade-off

- Twins instrument
  - Long first-stage regression

$$D_i = \alpha_1^l + \phi^l Z_i + \gamma_1^l A_i + e_{1i}^l$$

    - where D_i is family size, Z_i is a dummy for twin second births and A_i is maternal age
  - Short first-stage regression

$$D_i = \alpha_1^s + \phi^s Z_i + e_{1i}^s$$

  - OVB = Relation between A_i and $Z_i \times$ Effect of A_i in long
    - Older women are more likely to have twins — Relation between A_i and Z_i is positive
    - Older women tend to have smaller families — Effect of A_i in long is negative ($\gamma_1^l < 0$)
    - The OVB is negative

The quantity-quality trade-off

- Table 3.4 reports the first-stage estimates
- Same-sex instrument:
  - Having two boys or two girls increases family size
  - The first-stage estimates are similar with and without controls, suggesting that this instrument is closer to meeting the independence assumption than the twins instrument

The quantity-quality trade-off

- Table 3.5 reports the 2SLS estimates along with the estimates given by an OLS regression of the form:

$$Y_i = \alpha_3 + \beta D_i + \gamma_3 A_i + \delta_3 B_i + e_{3i}$$

- The OLS estimate is negative and significant — adults who were raised in larger families have a lower level of schooling
- The 2SLS estimates are positive, but insignificant — once we correct for selection bias using IV, there is little evidence of a quantity-quality trade-off

The quantity-quality trade-off

TABLE 3.4
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320	.437			.449
	(.052)	(.050)			(.050)
Same-sex sibships			.079	.073	.076
			(.012)	(.010)	(.010)
Male		−.018		−.020	−.020
		(.010)		(.010)	(.010)
Controls	No	Yes	No	Yes	Yes

Notes: This table reports coefficients from a regression of the number of children on instruments and covariates. The sample size is 89,445. Standard errors are reported in parentheses.

The quantity-quality trade-off

TABLE 3.5
OLS and 2SLS estimates of the quantity-quality trade-off

Dependent variable	OLS estimates (1)	2SLS estimates: Twins instruments (2)	2SLS estimates: Same-sex instruments (3)	2SLS estimates: Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in column (1). Columns (2), (3), and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Sample sizes are 89,445 for rows (1) and (2); 50,561 for row (3); and 50,535 for row (4). Standard errors are reported in parentheses.

Side note: 2SLS estimator

- We can write the 2SLS estimator as the ratio between the reduced-form coefficient and the first-stage coefficient:

$$\lambda_{2SLS} = \frac{\rho}{\phi}$$

- The first stage links instrument and treatment:

$$D_i = \alpha_1 + \phi Z_i + e_{1i}$$

- The reduced form links instrument and outcomes:

$$Y_i = \alpha_0 + \rho Z_i + e_{0i}$$

- The 2SLS second stage is the regression of outcomes on first-stage fitted values:

$$Y_i = \alpha_2 + \lambda \widehat{D}_i + e_{2i}$$

Side note: 2SLS estimator

- The 2SLS second stage is given by:

$$\begin{aligned}\lambda_{2SLS} &= \frac{C(Y_i, \widehat{D}_i)}{V(\widehat{D}_i)} = \frac{C(Y_i, \alpha_1 + \phi Z_i)}{V(\alpha_1 + \phi Z_i)} \\ &= \frac{\phi C(Y_i, Z_i)}{\phi^2 V(Z_i)} = \frac{\rho}{\phi}\end{aligned}$$