Lecture 6 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
- Sá, F. (2015), "Immigration and House Prices in the UK", *Economic Journal*, Vol.125(587), pp.1393-1424
- Background: Wooldridge chapters 13-15

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

- 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 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

- 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

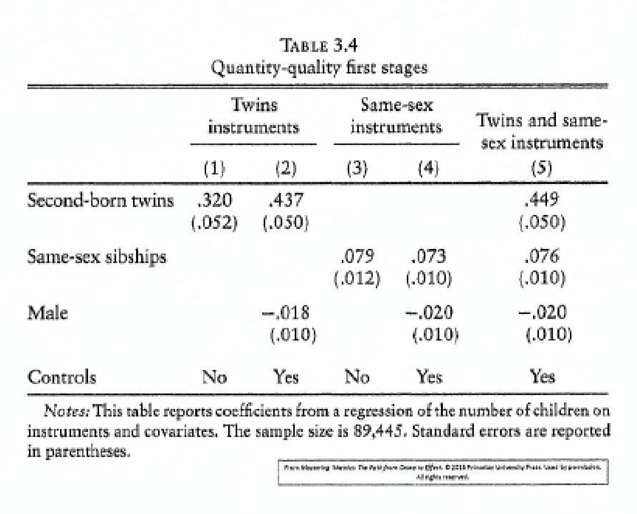

TABLE 3.4
Quantity-quality first stages

	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.

From *Mastering 'Metrics: The Path from Cause to Effect*. © 2015 Princeton University Press. Used by permission. All rights reserved.

The quantity-quality trade-off

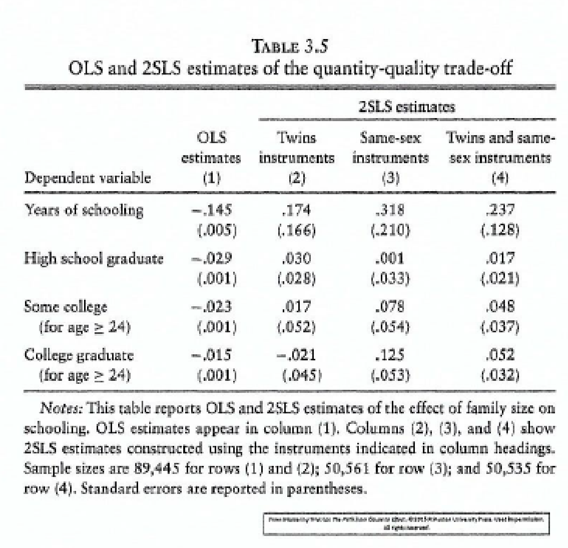

TABLE 3.5
OLS and 2SLS estimates of the quantity-quality trade-off

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

Weak Instruments

- A **weak instrument** is one that is not highly correlated with the regressor being instrumented, so the first-stage coefficient is small and imprecisely estimated
- In this case, the 2SLS estimate is biased
- We can write the 2SLS estimator as the ratio between the reduced-form coefficient and the first-stage coefficient:
- 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:

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

Weak Instruments

- The 2SLS second stage is given by:

$$\lambda_{2SLS} = \frac{\rho}{\phi} = \frac{Cov(Y_i, Z_i)/V(Z_i)}{Cov(D_i, Z_i)/V(Z_i)} = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)}$$

- The model (with selection bias) is:

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

- We can write:

$$\lambda_{2SLS} = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)} = \frac{Cov(\alpha + \lambda D_i + u_i, Z_i)}{Cov(D_i, Z_i)} = \lambda + \frac{Cov(u_i, Z_i)}{Cov(D_i, Z_i)}$$

Weak Instruments

- The **bias of the 2SLS estimator** is the difference between the estimate and the true parameter:

$$bias = \lambda_{2SLS} - \lambda = \frac{Cov(u_i, Z_i)}{Cov(D_i, Z_i)}$$

- If the exclusion restriction is satisfied $Cov(u_i, Z_i) = 0$ and there is no bias
- But we cannot be sure that $Cov(u_i, Z_i)$ is exactly zero in small samples
- If the instrument is weak, $Cov(D_i, Z_i)$ is small. In this case, the bias may be large even if $Cov(u_i, Z_i)$ is small.

Weak Instruments

- When is the problem of weak instruments worth worrying about?
- Rule of thumb:
  - The F-statistic in the first stage should be at least 10
  - The F-statistic is just an extension of the t-statistic to a case where there may be multiple instruments

Immigration and House Prices

- Reading for Problem Set 4: Sá, F. (2015), "Immigration and House Prices in the UK", *Economic Journal*, Vol.125(587), pp.1393-1424
- Model:

$$\Delta \ln P_{it} = \beta \frac{\Delta FB_{it}}{Pop_{it-1}} + \gamma X_{it} + \phi_t + \rho_i + \varepsilon_{it}$$

where i denotes local authority and t denotes year

  - The dependent variable is the change in log of house prices
  - The key independent variable is the share of new immigrants to initial local population
  - X_{it} is a set of **lagged** socioeconomic controls: local unemployment rate, share of the local population claiming state benefits, local crime rate, ratio of number of dwellings to local population and an index of local housing quality.

Immigration and house prices

- The controls (X_{it}) are lagged to reduce potential for endogeneity
- This is a **panel data** model — there is cross-sectional variation (across local authorities i) and time-series variation (across years t)
- The model includes a dummy for each year (ϕ_t) — this is known as **year fixed effects** — and a dummy for each local authority (ρ_i) — **local authority fixed effects**
- The year fixed effects capture macroeconomic conditions that affect all LAs at the same time — for example, the financial crisis
- The LA fixed effects capture observed and unobserved characteristics of the local authorities that we are not controlling for and are fixed over time — for example, a "London effect"

Immigration and house prices

- Problem with the identification of the causal effect of immigration on house prices: locational choices of immigrants are not exogenous
- Sources of bias:
  - **Reverse causality**
    - Immigrants may choose to locate in areas where prices are falling, because they are more affordable
  - **Omitted variable bias**
    - There may be other factors that simultaneously drive house price growth and immigration. For example immigrants may choose to locate in booming areas where more jobs are being created. Those areas will experience both an inflow of immigrants and house price appreciation.
  - **Measurement error**
    - The inflow of immigrants (ΔFB_{it}) may be measured with error, because it is difficult to keep track of all immigrants who come into the UK

Immigration and house prices

- What is the direction of the bias in this case?
  - The bias from reverse causality would be negative
  - The bias from omitted variables would be positive
  - The bias from measurement error can be shown to be always negative — **attenuation bias**
- Solution: use an instrument

Immigration and house prices

- The instrument uses the historical settlement pattern of immigrants to predict the current locational choices of immigrants
- This is based on the idea that immigrants tend to locate in cities where there is already a large number of people from the same country. E.g. Indian immigrants tend to locate in cities where there are already many Indians
- Why? **Immigration networks** are important for job search and assimilation into a new culture

Immigration and house prices

- Construction of the instrument:
  - Let ΔFB_{ct} denote the number of immigrants from source country c that enter the country in year t
  - Let λ_{ci0} denote the fraction of immigrants from country c who are living in region i is some earlier base year
  - The instrument is given by:

$$\frac{\sum_c \lambda_{cit_0} \Delta FB_{ct}}{Pop_{it-1}}$$

Immigration and house prices

- Under what conditions is this a good instrument?
  - The predicted inflow of immigrants based on the historical settlement pattern has a causal effect on the actual inflow of immigrants into the city. This is the **first stage**.
  - The predicted settlement pattern of immigrants is not related with the current economic conditions of different cities. This is the **independence assumption**.
  - The only way through which the predicted settlement pattern of immigrants affects house prices is through its effect on the current inflow of immigrants. This is the **exclusion restriction**.

Immigration and house prices

- Are these conditions likely to be met?
  - We can check how strong the correlation is between the predicted inflow of immigrants based on the historical settlement pattern and the actual inflow of immigrants by looking at the first-stage results.
  - The independence assumption and exclusion restriction are more likely to hold if we can argue that:
    - The unobserved factors determining whether immigrants decide to locate in a city i in the base year are not correlated with changes in the relative economic opportunities offered by different cities in the following years → this requires that economic shocks are not too persistent
    - The total (national) flow of immigrants (ΔFB_{ct}) is exogenous to economic conditions in a given city — for example, it may be determined by national legislation (such as immigrant quotas)