Lecture 6: Instrumental Variables (Part I)

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

Introduction

- In Lecture 4, we studied the potential for bias in OLS estimates
- The ZCM assumption, $E(u|x) = 0$, could break down for different reasons
  - Omitted variables bias
  - Simultaneity
  - Measurement error
- A convenient RCT is not always (or usually) available
- What to do then? This is what the rest of the course will be about!
- This lecture: Introduction to instrumental variables

Plan for Today

1. Introduction
2. Deriving the IV estimator
3. Consistency of the IV estimator
4. 2SLS and the IV requirements
5. Example 1: Angrist and Krueger (1991)
   - Weak instruments
   - Validity
6. Example 2: Acemoglu, Johnson, and Robinson (2001)
   - Identification
   - Criticism
7. Summary

Introducing Instrumental Variables: An Omitted Variables Example

- Consider a general problem where the data generating process (DGP) is:

$$y_i = \beta_0 + \beta_1 x + \beta_2 q + u$$

- But q is unobserved and so you can only estimate the following:

$$y_i = b_0 + b_1 x + e$$

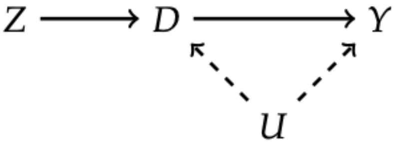
- Classic OVB set-up: $E(b_1) \neq \beta_1$ unless either $Cov(x, q) = 0$ or $\beta_2 = 0$
- Suppose y_i is wages, x is schooling, q is unobserved ability (assume mean 0)
- IV is a potential method for getting consistent estimator for β_1, even if we do not observe q

Introducing Instrumental Variables: An Omitted Variables Example

- Suppose we could find a third variable (z_i) which satisfies the following two conditions:
  - It is correlated with x, i.e., $Cov(x, z) \neq 0$
  - It is not correlated with q or other determinants of y

- One way of thinking about this is that any correlation between z and y only works through x
  - z does not belong in the "structural equation" determining y

- If you can find such a variable, then we can consistently estimate β_1!



Deriving the IV Estimator

- **Problem:** We do not observe q
- Assuming $Cov(x, q) \neq 0$ and $\beta_2 \neq 0$, coefficient(s) when regressing y only on x will be biased
- Now see the following:

$$Cov(z, y) = Cov(z, \beta_1 x + \beta_2 q + u)$$

- Defining $\eta \equiv \beta_2 q + u$, we can write:

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, \eta)$$

Deriving the IV estimator

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, \eta)$$

- Given $Cov(z, \eta) = 0$ (validity) and $Cov(z, x) \neq 0$ (relevance):

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)}$$

$$\Rightarrow \beta_1 = \frac{Cov(z, y)/Var(z)}{Cov(z, x)/Var(z)}$$

- **Numerator:** coefficient on z from the population regression of y on z
- **Denominator:** coefficient on z from the population regression of x on z

Deriving the IV Estimator

- β_1 is identified as the population covariance between z and y divided by the population covariance between z and x

- Given a random sample, we estimate these population quantities by the sample analogs to get the IV estimator:

$$\widehat{\beta}_1 = \frac{1/N \sum_{i=1}^{n} (z_i - \overline{z})(y_i - \overline{y})}{1/N \sum_{i=1}^{n} (z_i - \overline{z})(x_i - \overline{x})}$$

- The IV estimator for β_0 is

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

Consistency of the IV estimator

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\frac{1}{N}\sum_i^n (z_i - \bar{z})(\beta_0 + \beta_1 x_i + \eta_i - \bar{y})}{\frac{1}{N}\sum_i^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\frac{1}{N}\sum_i^n (z_i - \bar{z})(\beta_0 + \beta_1 x_i + \eta_i - (\beta_0 + \beta_1 \bar{x}))}{\frac{1}{N}\sum_i^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\frac{1}{N}\sum_i^n (z_i - \bar{z})(\beta_1 (x_i - \bar{x}) + \eta_i)}{\frac{1}{N}\sum_i^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \beta_1 + \frac{\frac{1}{N}\sum_i^n (z_i - \bar{z})\eta_i}{\frac{1}{N}\sum_i^n (z_i - \bar{z})(x_i - \bar{x})}
\end{aligned}
$$

Consistency of the IV estimator

$$plim(\widehat{\beta}_1) = \beta_1 + \frac{plim\left(\frac{1}{N}\sum_i^n (z_i - \bar{z})\eta_i\right)}{plim\left(\frac{1}{N}\sum_i^n (z_i - \bar{z})(x_i - \bar{x})\right)}$$

$$plim(\widehat{\beta}_1) = \beta_1 + \frac{cov(z, \eta)}{cov(z, x)}$$

- Hence, given validity and relevance:

$$plim(\widehat{\beta}_1) = \beta_1$$

- Note: while IV is consistent, it is still biased in finite samples
  - The bias decreases as the sample size gets larger, and as the instrument is more highly correlated with the endogenous explanatory variable (more on this to follow!)

2SLS

- The IV estimator also has a two stages least squares (2SLS) interpretation: it can be obtained from the regression

$$y = \beta_0 + \beta_1 \widehat{x} + u$$

where $\widehat{x}$ is the fitted value (the exogenous part of x) from the OLS estimation of the first stage

$$x = \pi_0 + \pi_1 z + v$$

- This highlights the cost of doing IV: if OLS was unbiased, then it would have smaller variance than IV estimator
  - Intuition: using only part of the information in x
  - If $R^2_{x,z}$ is small, the IV variance can end up being much larger than the OLS variance

IV Jargon

- Given an outcome variable y_i, an endogenous variable x_i, an instrument z_i and a vector of controls A_i:
  - **First-stage regression:**

$$x_i = b_0 + b_1 z_i + A_i^{'} \gamma + e$$

  - **Second-stage (structural) regression:**

$$y_i = \beta_0 + \beta_1 \hat{x}_i + A_i^{'} \phi + u$$

  - **"Reduced-form" regression:**

$$y_i = a_0 + a_1 z_i + A_i^{'} \lambda + \xi$$

- We are going to ignore covariates for now

Assessing the Two IV Requirements

- The two conditions for IV are conceptually very distinct
- **Relevance:** $Cov(z, x_1) \neq 0$
  - This is testable
  - Regressing x_1 on z should give a coefficient different from 0 on z
  - We will come back to testing for relevance later...
- **Validity:** $Cov(z, \eta) = 0$
  - This requires z to not be a direct determinant of y
  - z must be uncorrelated with any other determinants of y except x_1
  - This restriction is called the **exclusion restriction** for the instrument
  - This is not directly testable, needs justification from theory, knowledge of institutions etc.—oftentimes, a lot of story telling!
  - Even with randomization, exclusion restriction might not hold

IV Estimates of the Economic Returns to Schooling

- Let wages (Y_i) be determined by the following way:

$$Y_i = \alpha + \rho S_i + A_i^{'}\gamma + v_i$$

- We do not observe A_i (let us assume just scalar ability for now)

- So we have:

$$Y_i = \alpha + \rho S_i + \eta_i$$

  where $\eta_i \equiv A_i^{'}\gamma + v_i$

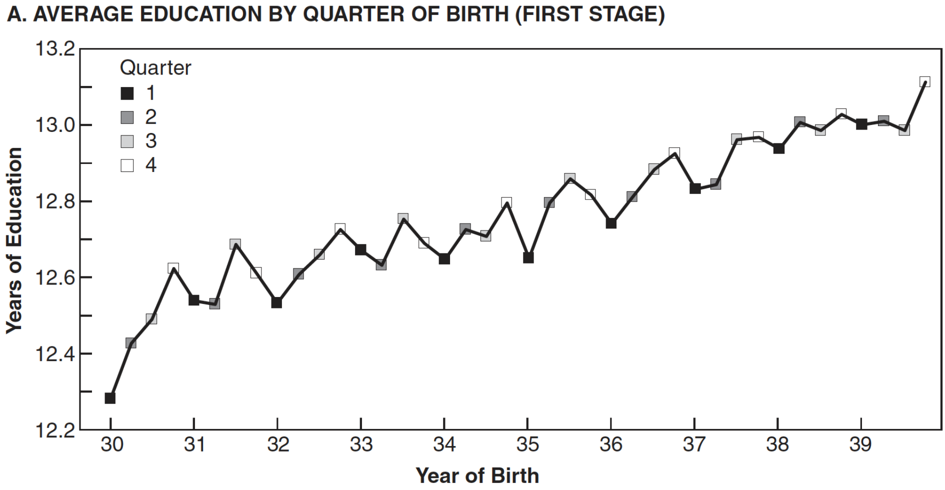
- Requirements for an IV:
  - Relevance: $Cov(z, S_i) \neq 0$
  - Validity: $Cov(z, \eta) = 0$

Angrist and Krueger (1991): Schooling Laws and Quarter-of-Birth

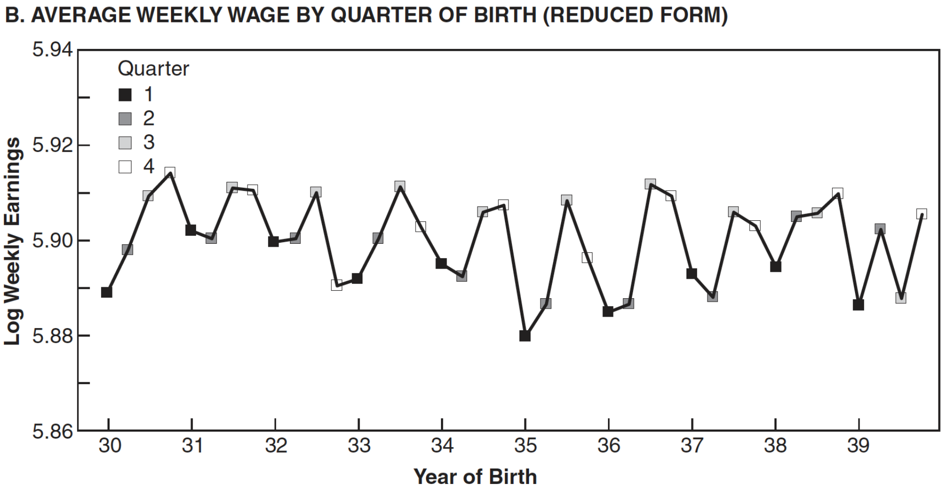
- Angrist and Krueger (1991) find an IV based on compulsory schooling laws in the US:
  - Students were required to join schooling in the calendar year they turned 6
  - Students could drop out of schooling when they turned 16
  - So the compulsory amount of schooling a student had to attend depended on when they were born within the year

- Angrist and Krueger use the student's quarter-of-birth as instrument(s) to isolate the causal effect of schooling
  - Relevance: QOB should affect schooling
  - Validity: QOB should not affect wages in any other way

First-Stage Variation: The Relationship Between QOB and S_i



Reduced-Form Variation: The Relationship Between QOB and *ln*(*wages*)



IV (2SLS) Estimates: The Causal Effect of Schooling on Wages

TABLE 4.1.1
2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	.071	.067	.102	.13	.104	.108	.087	.057
	(.0004)	(.0004)	(.024)	(.020)	(.026)	(.020)	(.016)	(.029)
Exogenous Covariates								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year-of-birth dummies		✓			✓	✓	✓	✓
50 state-of-birth dummies		✓			✓	✓	✓	✓
Instruments								
dummy for QOB = 1			✓	✓	✓	✓	✓	✓
dummy for QOB = 2				✓		✓	✓	✓
dummy for QOB = 3				✓		✓	✓	✓
QOB dummies interacted with year-of-birth dummies (30 instruments total)							✓	✓

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.

Weak Instruments

- A particular issue in the IV case relates to weak instruments—the instrument (z) is only weakly correlated to the endogenous variable (x)

- There are two main problems with such a case:
  - Finite sample bias may be very large with weakly-informative IVs (the bias is inversely proportional to the correlation between x and z)
  - Even a modest violation in the exclusion restriction could lead to severe inconsistency, and this bias may be greater than in the original OLS estimates!

- How to detect a "weak instrument"? Make sure that the first-stage F-statistic is greater than 10 (more on this later)

Violation of Validity Leads to Large Inconsistency

- To see this, write the probability limit of the IV estimator as

$$
\begin{aligned}
plim\widehat{\beta}_1^{IV} &= \beta_1 + \frac{Cov(z,u)}{Cov(z,x)} \\
&= \beta_1 + \frac{Corr(z,u)\,\sigma_u\sigma_z}{Corr(z,x)\,\sigma_z\sigma_x} \\
&= \beta_1 + \frac{Corr(z,u)}{Corr(z,x)}\frac{\sigma_u}{\sigma_x}
\end{aligned}
$$

  where σ_u and σ_x are the standard deviations of u and x in the population, respectively

- Even if $Corr(z,u)$ is small, the inconsistency of the IV estimator can become very large if $Corr(z,x)$ is also small

Violation of Validity Leads to Large Inconsistency

$$plim\widehat{\beta}_1^{IV} = \beta_1 + \frac{Corr\,(z,u)}{Corr\,(z,x)}\frac{\sigma_u}{\sigma_x}$$
$$plim\widehat{\beta}_1^{OLS} = \beta_1 + Corr\,(x,u)\,\frac{\sigma_u}{\sigma_x}$$

- For instance, if $Corr\,(z,u) > 0$ and $Corr\,(x,u) > 0$ and $Corr\,(z,x) > 0$, then the inconsistency in IV is smaller than that in OLS only if:
$$\frac{Corr\,(z,u)}{Corr\,(z,x)} < Corr\,(x,u)$$

Criticism by Bound et al. (1995)

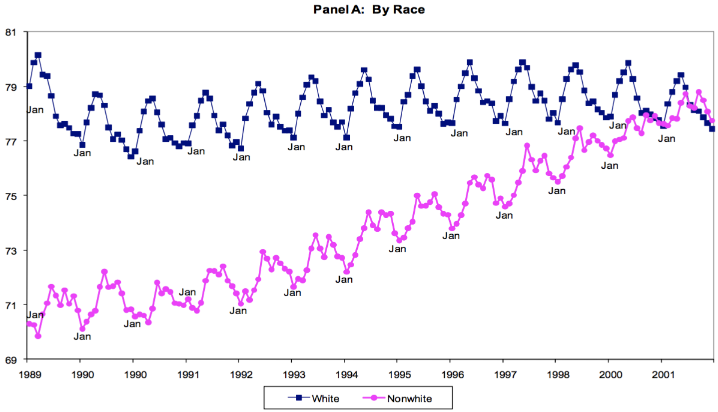
- The quarter of birth instruments explain only a tiny fraction of the variation in education
- Even a small correlation between quarter of birth and earnings for reasons other than education might lead to a large inconsistency
- Bound et al. (1995) argue there is some evidence that quarter of birth is correlated with school performance, physical and mental health, and parental income...

Criticism by Bound et al. (1995)

- These correlations are expected to be small but with weak instruments could wreak havoc
- Bound et al. suggest there is some evidence for this already in the Angrist-Krueger paper
- When Angrist and Krueger add control variables for race, marital status, and location of residence, the rate of return falls by more in the IV model than in the OLS model, suggesting a relationship between these variables and quarter of birth
- Buckles and Hungerman (2013) show much more concerning evidence of this: people who are born in different months, differ in many other relevant characteristics

Systematic Differences in Characteristics by Birth Month (Buckles and Hungerman 2013)

FIGURE 2. PERCENT OF WOMEN GIVING BIRTH EACH MONTH WHO HAVE A HIGH SCHOOL DEGREE, NATALITY FILES, 1989-2001



IVs - and the Act of Persuasion

- Good IVs are very hard to find
  - As you see, even a priori plausible instruments may have issues
  - An IV paper lives (or dies!) by the plausibility of the exclusion restriction...

- Now we will look at one more empirical example so that you get a clear sense of what IV requirements are

Example 2: Acemoglu, Johnson, and Robinson (2001)

- One of the biggest questions in economics: Why are some countries so much richer than others?

- One potential explanation is "institutions" such as property rights
  - Countries with more secure legal and property rights see more investments in physical and human capital
  - This leads on to more growth and higher per cap incomes

- But investigating this link is not easy:
  - Maybe richer countries can afford better institutions?
  - Maybe countries with strong property rights also have other factors that are helpful for growth?

Colonialism, Institutions, and Growth

- Acemoglu et al. study this using cross-country data on gross national income per capita, the current risk of expropriation, and settler mortality in the colonial era

- Their basic argument can be shown as follows:

(potential) settler mortality $\Rightarrow$ settlements $\Rightarrow$ early institutions $\Rightarrow$ current institutions $\Rightarrow$ current performance.

Colonialism, Institutions, and Growth

- Acemoglu et al. use settler mortality as an instrumental variable for property rights (average expropriation risk between 1985-1995)
- The dependent variable of interest is logarithm of GDP per capita in 1995
- What do we require for settler mortality to be a valid instrumental variable in this setting?

First-Stage Relationship

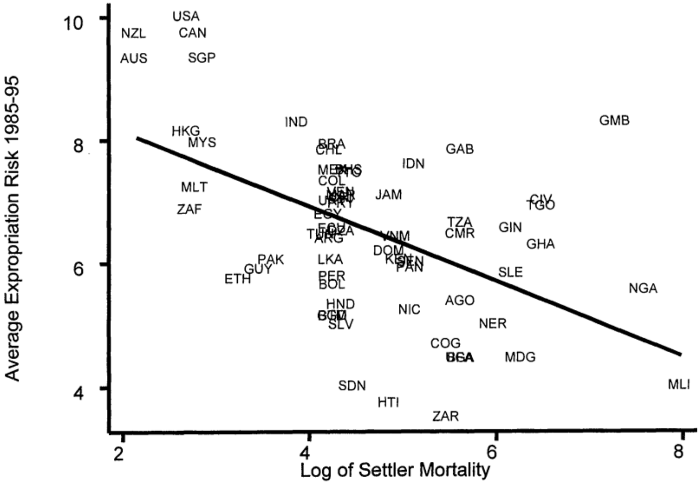


Figure 3. First-Stage Relationship Between Settler Mortality and Expropriation Risk

Reduced-Form Relationship

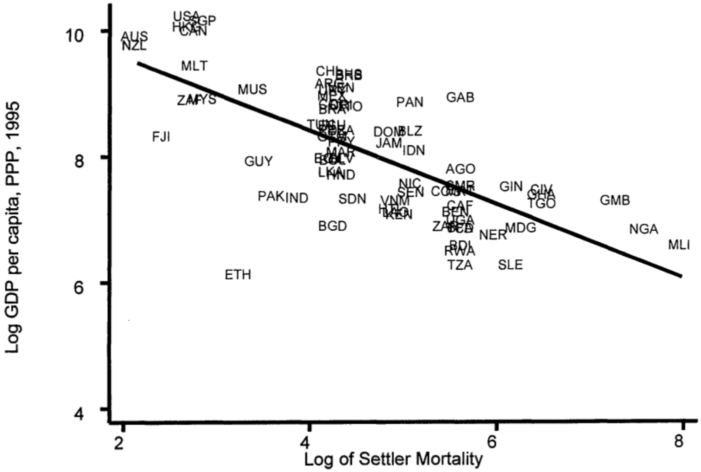


FIGURE 1. REDUCED-FORM RELATIONSHIP BETWEEN INCOME AND SETTLER MORTALITY

What Does Instrument Validity Require?

- Here, validity would need that settler mortality only affect current national income through institutions
- What are possible violations of the exclusion restriction?
  - Things that killed people 100 years ago might relate to disease environments...
  - ...which could also be bad for economic growth today!
  - Acemoglu et al. argue this is unlikely since the diseases that Europeans died from have higher rates of immunity in indigenous populations (malaria, yellow fever)
  - Is this plausible?
- To bolster their case, they show that controlling for a bunch of stuff (incl. geography, current disease environment) does not change results substantively

IV Results: First Stage

Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995

Log European settler mortality	−0.61	−0.51	−0.39	−0.39	−1.20	−1.10	−0.43	−0.34	−0.63
	(0.13)	(0.14)	(0.13)	(0.14)	(0.22)	(0.24)	(0.17)	(0.18)	(0.13)
Latitude		2.00		−0.11		0.99		2.00	
		(1.34)		(1.50)		(1.43)		(1.40)	
Asia dummy							0.33	0.47	
							(0.49)	(0.50)	
Africa dummy							−0.27	−0.26	
							(0.41)	(0.41)	
"Other" continent dummy							1.24	1.1	
							(0.84)	(0.84)	
R^2	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33	0.28

IV Results: Second Stage

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
				Panel A: Two-Stage Least Squares					
Average protection against expropriation risk 1985–1995	0.94 (0.16)	1.00 (0.22)	1.28 (0.36)	1.21 (0.35)	0.58 (0.10)	0.58 (0.12)	0.98 (0.30)	1.10 (0.46)	0.98 (0.17)
Latitude		−0.65 (1.34)		0.94 (1.46)		0.04 (0.84)		−1.20 (1.8)	
Asia dummy							−0.92 (0.40)	−1.10 (0.52)	
Africa dummy							−0.46 (0.36)	−0.44 (0.42)	
"Other" continent dummy							−0.94 (0.85)	−0.99 (1.0)	

Criticisms: Dodgy Data? (Albuoy 2012)

This comment argues that there are several reasons to doubt the reliability and comparability of their European settler mortality rates and the conclusions that depend on them. First, out of 64 countries in the sample, only 28 countries have mortality rates that originate from within their own borders. The other 36 countries in the sample are assigned rates based on conjectures the authors make as to which countries have similar disease environments. These assignments are generally unfounded and potentially contradictory. Six assignments are based on an incorrect interpretation of former colonial names for Mali. Another 16 assignments are extrapolated from thin bishop mortality data in Latin America from Gutierrez (1986), using a "benchmarking" procedure that can produce highly contradictory rates, depending on how the data are benchmarked. At a minimum, the sharing of mortality rates across countries requires that statistics be corrected for clustering (Moulton 1990). This correction alone noticeably reduces the significance of the results. If, in the hope of reducing measurement error, the 36 conjectured mortality rates are dropped from the sample, the point estimates relating mortality rates with expropriation risk become substantially smaller, particularly in the presence of covariates, which often gain significance.

Criticisms: What Else Did Settlers Bring?

- We have discussed two issues:
  - Data issues
  - Underlying omitted variables about the countries settlers went to (e.g., disease environment)
- Even setting these aside, there is a deeper issue: what if the settlers brought more than just their institutions?
  - What if the extent of settler presence also led to greater flow of technology and of human capital from Europe to these countries?
  - This is not implausible given the vast difference in tech in 19th century Europe and colonized countries
  - And much more limited opportunities for non-European descent to travel to Europe and learn
  - This is a criticism raised by Glaeser et al. (2004)

Summary

- IV is a powerful method to deal with various forms of bias in OLS (e.g., OVB)
- The 2SLS estimator gives the ratio of the reduced-form coefficient divided by the first-stage coefficient
- IV relies on two conditions:
  - Relevance (which can be directly tested)
  - Validity (which is not directly testable)
- Under maintained IV assumptions, the IV estimator is consistent
  - But still biased in finite samples
  - This bias declines with N and with the strength of the first-stage
  - When IVs are weak, even small violations of validity can cause large inconsistency

Readings

- Angrist and Pischke (2009) Chapter 4 up till Section 4.1.2 (p. 133)
- Cunningham (2020) Chapter on instrumental variables (pp. 205-232)

Optional:

- Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. Journal of Experimental Criminology, 2(1), 23-44.
- Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. The Journal of Economic Perspectives, 20(4), 111-132.
- Wooldridge (2013) Chapter 15 (except 15.7 on time-series applications)

References

- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical association, 90(430), 443-450.
- Buckles, K. S., & Hungerman, D. M. (2013). Season of birth and later outcomes: Old questions, new answers. Review of Economics and Statistics, 95(3), 711-724.
- Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth?. Journal of Economic Growth, 9(3), 271-303.

- Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. The Quarterly Journal of Economics, 106(4), 979-1014.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. The American Economic Review, 91(5), 1369-1401.
- Albouy, D. Y. (2012). The colonial origins of comparative development: an empirical investigation: comment. The American Economic Review, 102(6), 3059-3076.