Causal Inference
3 - Instrumental Variables

Benjamin Elsner
benjamin.elsner@ucd.ie

Instrumental Variables

We learn:

- What are **instrumental variables (IV)**?
- How can they be used to **identify causal effects**?
- Interpretation of IV estimators is **different from OLS**. We learn why.

This lecture is based on

- Angrist & Pischke (2009), Ch. 4.1, 4.2, 4.4

IV: Starting Point

$$y_i = \alpha + \beta D_i + u_i$$

CIA $cov(D_i, u_i) = 0$ often doesn't hold $\Rightarrow$ OLS estimates of β are biased

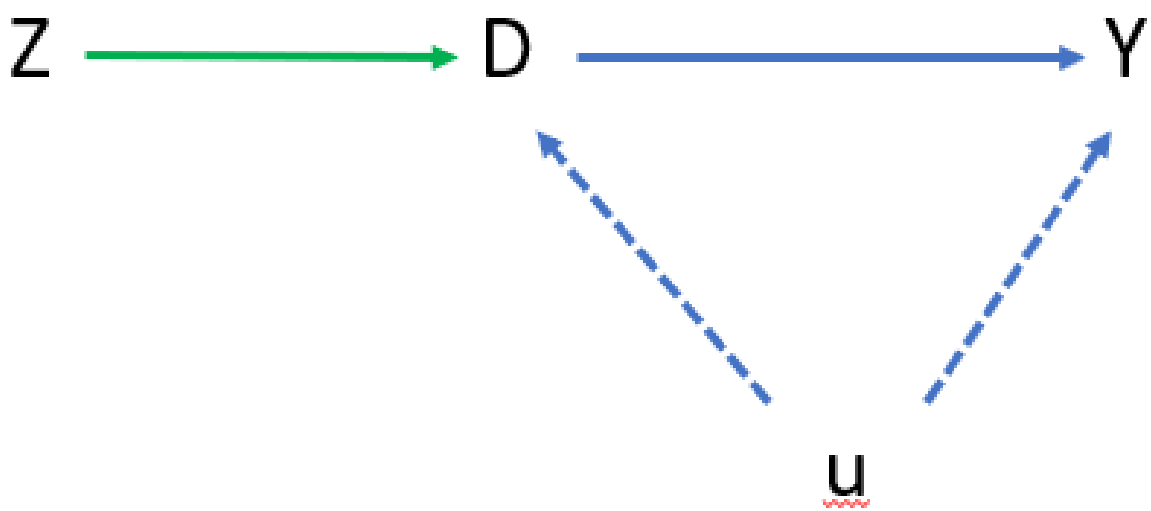
- **Unobserved heterogeneity**: we may not observe all confounding variables
- D_i may be **measured with error**
- Simultaneity or **reverse causality**

Instrumental Variables

In theory, **instrumental variables** offer a way to

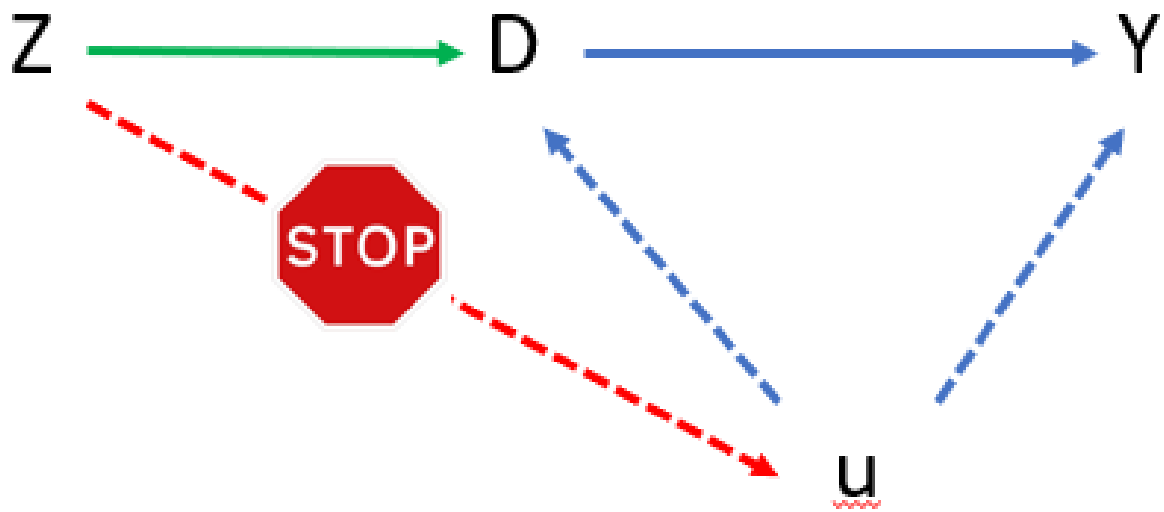
- break the correlation $cov(D_i, u_i)$
- and obtain a **consistent causal estimate of the treatment on y_i**

Idea behind an instrumental variable (Z):



Instrumental Variables

1) An IV **affects Y only through its effect on D**



2) It **must not be correlated with unobservable characteristics**

Instrumental Variables

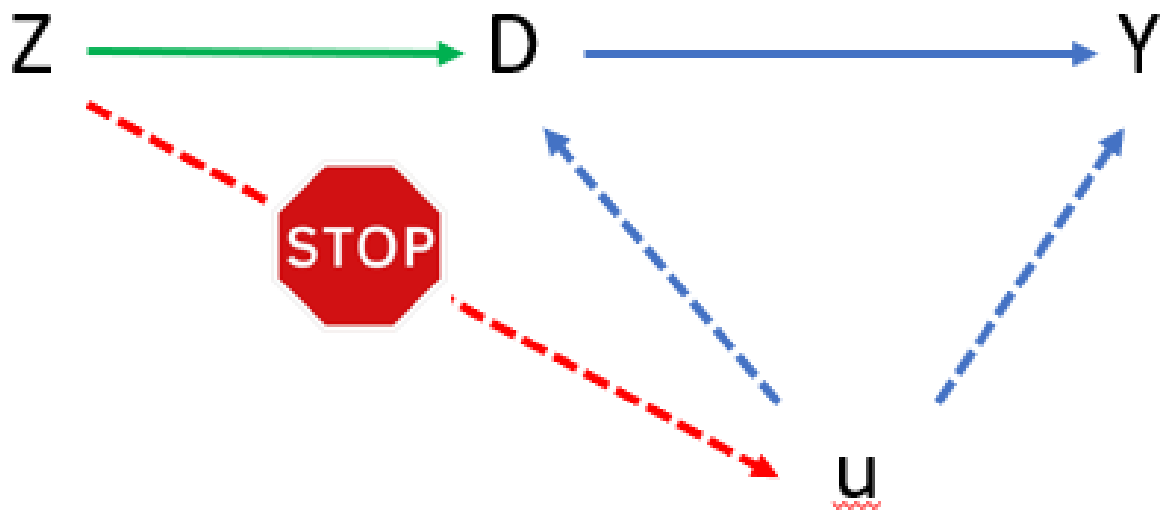
One way to think about an IV:

- **people/firms make optimal choices** that affect their **treatment status**
- Z is a **shock that changes the behavior** of at least some people/firms
- Z has to be **unrelated to people's characteristics**
- i.e. it should be assigned as good as randomly

And another:

- The instrument Z is a **treatment/incentive that is offered** to units/people
- D measures if the unit **actually takes up the treatment**
- The instrument Z should be **as good as randomly assigned**
- Example: randomly assigned **school vouchers**

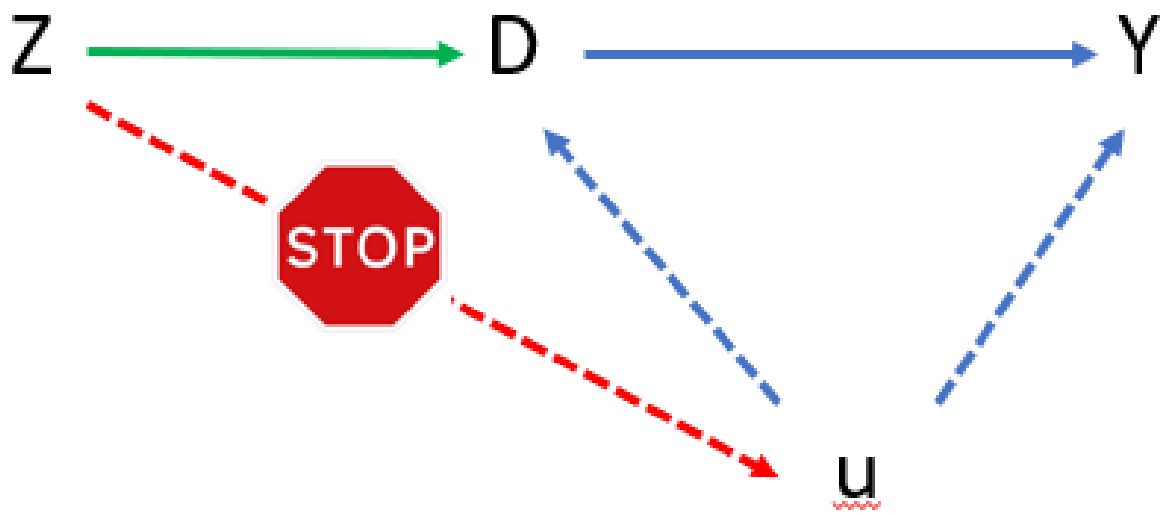
Instrumental Variables



And another:

- OLS uses all the variation in D to explain y
- IV uses **only the variation in D that is related to Z**
- So this means less variation is used, but at least Z is unrelated to u

Instrumental Variables Lingo



IV requires two ingredients:

- **First stage:** $cov(Z, D) \neq 0$
- **Exclusion restriction:** $cov(Z, u) = 0$

First Stage and Exclusion Restriction

The **first-stage relationship is testable**

- we can run a regression of D on Z
- it is also possible to include covariates

The **exclusion restriction is not testable**

- it is an **identification assumption**
- we **need to make a convincing argument** in favor of it
- this is difficult and the reason for heated debates in seminars

Some say: **friends tell their friends not to use IV...**

IV Equations: Two-Stage Least Quares (2SLS)

Relationship of interest

$$y_i = \alpha + \beta D_i + X_i'\gamma + u_i$$

First stage

$$D_i = \delta_0 + \delta_1 Z_i + X_i'\rho + e_i$$

Second stage ($\widehat{D}_i$ from first stage)

$$y_i = \tilde{\alpha} + \tilde{\beta}\widehat{D}_i + X_i'\kappa + \varepsilon_i$$

Reduced form

$$y_i = \lambda_0 + \lambda_1 Z_i + X_i'\sigma + \eta_i$$

IV in Theory

It can be shown that

$$\widehat{\beta^{IV}} = \frac{cov(Y,Z)}{cov(D,Z)} = \frac{\widehat{\lambda_1}}{\widehat{\delta_1}}$$

is a **consistent estimator of** β under the **exclusion restriction** $cov(Z,u) = 0$

This estimator is nothing but the **reduced-form coefficient** $\widehat{\lambda_1} = \frac{cov(y,Z)}{var(Z)}$...

divided by the first stage $\widehat{\delta_1} = \frac{cov(D,Z)}{var(Z)}$

Later we will see that this interpretation is useful

Classic IV Example: Moving to Opportunity

Research question: does **moving to a better neighborhood** affect adults and children?

The **Moving to Opportunity Program (MTO)**

- **Large-scale experiment** with people in public housing in several US cities in 1996
- **Treatment group 1**: voucher for private rental housing in low-poverty neighborhood
- **Treatment group 2**: voucher for private rental housing (no strings attached)
- **Control group**: no voucher

This experiment has been evaluated by Kling *et al.* (2007).

Classic IV Example: Moving to Opportunity

50% of voucher recipients actually moved; most to better neighborhoods

DESCRIPTIVE STATISTICS OF NEIGHBORHOOD CHARACTERISTICS

	Experimental (i)	Section 8 (ii)	Control (iii)
Average census tract poverty rate	.33	.35	.45
Average census tract poverty rate above 30%	.52	.62	.87
Respondent saw illicit drugs being sold or used in neighborhood during past 30 days	.33	.34	.46
Streets are safe or very safe at night	.70	.65	.56
Member of household victimized by crime during past 6 months	.17	.16	.21
Average census tract share on public assistance	.16	.17	.23
Average census tract share of adults employed	.83	.83	.78
Average census tract share workers in professional and managerial occupations	.26	.23	.21
Average census tract share minority	.82	.87	.90

Moving to Opportunity: Empirical Challenge

MTO was a **randomized experiment**

- $Z \in \{0, 1\}$ is the instrument, $D \in \{0, 1\}$ is the treatment
- but **not everyone** who received a voucher **actually moved**

We can estimate an **Intention-to-Treat (ITT)** effect by using the **reduced form**

$$y_i = \gamma_0 + \gamma_1 Z_i + \varepsilon_i$$

ITT is useful for policy evaluation

- But it does not tell us much about the **causal effect of moving**

Moving to Opportunity

Suppose we are interested in the **treatment effect on the treated**, in this case the **causal effect of moving**

- but we cannot force voucher recipients to move...

IV allows us to estimate this treatment effect under three conditions

1. Z is as good as **randomly assigned**
2. Z has **no direct effect** on the outcome
3. Z has a **sufficiently strong effect** on D

Moving to Opportunity: The Wald Estimator

We can estimate **three causal effects**

1. **First stage:** the causal effect of Z on D:
   $P(D = 1|Z = 1) - P(D = 1|Z = 0)$
2. **Reduced form (ITT):** the causal effect of Z on Y:
   $E(Y|Z = 1) - E(Y|Z = 0)$
3. **Treatment effect of interest:** the causal effect of D on Y:
   $Y(1) - Y(0) = E(Y|D = 1) - E(Y|D = 0)$

The **Wald Estimator** relates all three effects

$$E(Y|D = 1) - E(Y|D = 0) = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{P(D = 1|Z = 1) - P(D = 1|Z = 0)} \quad (1)$$

Moving to Opportunity: The Wald Estimator

$$\widehat{\beta}^{IV} = E(Y|D=1) - E(Y|D=0) = \frac{E(Y|Z=1) - E(Y|Z=0)}{P(D=1|Z=1) - P(D=1|Z=0)}$$

- **difference in outcomes** by groups intended and not intended for treatment
- divided by **difference in the actual treatment**

Example:

- Suppose the difference in outcomes $E(Y|Z=1) - E(Y|Z=0)$ is 10
- and we know that 50% of voucher recipients moved (but noone else)
- In this case, $\widehat{\beta}^{IV} = \frac{10}{0.5} = 20$

Moving to Opportunity

	E/S (i)	CM (ii)	ITT (iii)	TOT (iv)	CCM (v)
A. Adult outcomes					
Obese, BMI $\geq$ 30	E – C	0.468	−0.048	−0.103	0.502
			(0.022)	(0.047)	
Calm and peaceful	E – C	0.466	0.061	0.131	0.443
			(0.022)	(0.047)	
Psychological distress, K6 *z*-score	E – C	0.050	−0.092	−0.196	0.150
			(0.046)	(0.099)	
B. Youth (female and male) outcomes					
Ever had generalized anxiety symptoms	E – C	0.089	−0.044	−0.099	0.164
			(0.019)	(0.042)	
	S – C	0.089	−0.063	−0.114	0.147
			(0.019)	(0.035)	
Ever had depression symptoms	S – C	0.121	−0.039	−0.069	0.134
			(0.019)	(0.035)	
C. Female youth outcomes					
Psychological distress, K6 scale *z*-score	E – C	0.268	−0.289	−0.586	0.634
			(0.094)	(0.197)	
Ever had generalized anxiety symptoms	E – C	0.121	−0.069	−0.138	0.207
			(0.027)	(0.055)	
	S – C	0.121	−0.075	−0.131	0.168
			(0.029)	(0.051)	
Used marijuana in the past 30 days	E – C	0.131	−0.065	−0.130	0.202
			(0.029)	(0.059)	
	S – C	0.131	−0.072	−0.124	0.209
			(0.032)	(0.056)	
Used alcohol in past 30 days	S – C	0.206	−0.091	−0.155	0.306
			(0.038)	(0.056)	

Wald estimator: TOT; denominator: CM

Classic IV Example: Angrist & Evans (1998)

Angrist & Evans (1998) study the effect of **children on female labor supply**

Their **most basic regression** is

$$hours_i = \alpha + \beta kids_i + u_i$$

The **number of children** is almost certainly **endogenous**:

- fertility is a choice, and so is labor supply
- richer families can afford more children and lower labor supply
- couples differ in their preferences over fertility and labor supply

Classic IV Example: Angrist & Evans (1998)

Ideal experiment: **randomly assign children to families**

IV in Angrist & Evans (1998): **sex of the first two children**

- the sex of a child is as good as random
- couples tend to have a preferences for mixed-sex offspring
- couples with two boys or two girls are more likely to have a third child

Analysis is purely based on families with two or more children

Classic IV Example: Angrist & Evans (1998)

The **components of the IV estimator**

First stage: effect of same-sex children on the likelihood of having a third child

$$kids_i = \delta_0 + \delta_1 samesex_i + e_i$$

Reduced form:

$$hours_i = \lambda_0 + \lambda_1 samesex_i + \eta_i$$

Exclusion restriction: same-sex children unrelated with personal characteristics $\Rightarrow cov(samesex_i, u_i) = 0$

Classic IV Example: Angrist & Evans (1998)

The following analysis is based on a small sub-sample of Angrist & Evans (1998)

```
. sum hours kids samesex

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       hours |     31857    21.22011    19.49892           0         99
        kids |     31857    2.752237    .9771916           2         12
     samesex |     31857     .502778    .5000001           0          1
```

Descriptive statistics indicate that in **50% of all families** the **first two children had the same sex**

This is **what we would expect**. Any different result would be a red flag

Classic IV Example: Angrist & Evans (1998)

Now let's look at the **simple OLS regression**

```
. reg hours kids, robust

Linear regression                               Number of obs =   31857
                                                F(  1, 31855) =  585.25
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0178
                                                Root MSE      =  19.325
------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.664309   .1101318   -24.19   0.000    -2.880171   -2.448446
       _cons |   28.55292   .3200455    89.22   0.000     27.92562    29.18022
------------------------------------------------------------------------------
```

Each additional child (above two) decreases weekly work hours on average by 2.66

Classic IV Example: Angrist & Evans (1998)

The first stage: is the **instrument relevant to explain the number of kids?**

```
. reg kids samesex, robust
Linear regression                               Number of obs =   31857
                                                F(  1, 31855) =   40.90
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0013
                                                Root MSE      =  .97658
------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0699933   .0109439     6.40   0.000     .0485429    .0914437
       _cons |   2.717045    .007806   348.07   0.000     2.701745    2.732346
------------------------------------------------------------------------------
```

Important things to discuss in an IV paper

- Does the first-stage coefficient make sense (sign, magnitude)?
- Is the first-stage correlation strong enough (is the F-Statistic of the instrument >10)

Classic IV Example: Angrist & Evans (1998)

```
. reg kids samesex, robust
Linear regression                               Number of obs =   31857
                                                F(  1, 31855) =   40.90
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0013
                                                Root MSE      =  .97658
------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0699933   .0109439     6.40   0.000     .0485429    .0914437
       _cons |   2.717045    .007806   348.07   0.000     2.701745    2.732346
------------------------------------------------------------------------------
```

In this case...

- families with same-sex children have more children
- the coefficient is small: out of 14 families with same-sex children, one has an additional child
- the t-statistic of the instrument is strong enough (implied F-Statistic: $F = 40.96$)

Classic IV Example: Angrist & Evans (1998)

2SLS estimate

```
. ivreg hours (kids = samesex), robust

Instrumental variables (2SLS) regression          Number of obs =   31857
                                                  F(  1, 31855) =    3.19
                                                  Prob > F      =  0.0743
                                                  R-squared     =       .
                                                  Root MSE      =  19.534
------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |   -5.58186   3.127136    -1.78   0.074    -11.71117    .5474471
       _cons |   36.58271   8.606509     4.25   0.000     19.71362    53.45179
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   samesex
```

This table reports the **second-stage estimates**

- the regressor is the **number of children predicted by the same-sex instrument**
- the effect is stronger than the OLS estimate (-2.66)
- it is statistically significant at the 10%-level

Classic IV Example: Angrist & Evans (1998)

To **develop a better intuition of how IV works**, it is useful to look at the reduced form and first stage

The IV estimator is the **reduced-form divided by the first stage**

$$\widehat{\beta^{IV}} = \frac{\widehat{\lambda_1}}{\widehat{\delta_1}}$$

```
. reg hours samesex

      Source |       SS           df       MS      Number of obs   =    31,857
-------------+----------------------------------   F(1, 31855)     =      3.20
       Model |  1215.63289         1  1215.63289   Prob > F        =    0.0738
    Residual |    12110681    31,855  380.181477   R-squared       =    0.0001
-------------+----------------------------------   Adj R-squared   =    0.0001
       Total |  12111896.6    31,856  380.207703   Root MSE        =    19.498

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |  -.3906929   .2184891    -1.79   0.074    -.8189399     .0375541
       _cons |   21.41654   .1549237   138.24   0.000     21.11288     21.7202
------------------------------------------------------------------------------
```

Intuition behind the IV

Start with the **reduced form**:

- having same-sex children decreases weekly work hours by 0.39

Now **consider the first stage**

- having same-sex children increases the number of children by 0.07

If an increase in the number of children by 0.07 reduces the number of work hours by 0.39...

- then an **increase by one child** reduces work hours by $0.39/0.07 = 5.56$

For this reason, we often say that we **"scale up" the reduced form by the first stage**

Classic IV Example: Angrist & Evans (1998)

So we have that $\widehat{\beta^{IV}} < \widehat{\beta^{OLS}}$. Does this make sense?

Explanation 1: OLS estimator is upward biased (i.e. closer to zero)

- there could be an omitted variable (for example family wealth)
- both the correlation with kids and the direct effect on hours need to have the same sign
- e.g. $cov(wealth, kids) > 0$ and $cov(wealth, hours|kids) > 0$ or both negative

Explanation 2: IV effect measures the **effect for a specific population**

- only 1 in 14 families "respond" to the instrument
- families who respond may not be the average family...

Local Average Treatment Effects (LATE)

So far, we implicitly assumed that the **potential outcomes are constant across units**. But what if potential outcomes are heterogeneous?

Consider a case with a binary instrument $Z_i \in \{0, 1\}$ the the treatment statuses

- $D_{1i} =$ i's treatment status when $Z_i = 1$
- $D_{0i} =$ i's treatment status when $Z_i = 0$

The **observed treatment status** is

$$D_i = D_{0i} + (D_{1i} - D_{0i})Z_i = \delta_0 + \delta_{1i}Z_i + \eta_i$$

Note that the effect of the IV on treatment may differ between individuals

Local Average Treatment Effects (LATE)

We **divide the population into four groups** depending on their reaction to the instrument

1. **Compliers**: people who react to the instrument as expected, $D_{1i} = 1$ and $D_{0i} = 0$
2. **Always-takers**: people who always take the treatment regardles of Z, $D_{1i} = D_{0i} = 1$
3. **Never-takers**: people who never take the treatment regardless of Z, $D_{1i} = D_{0i} = 0$
4. **Defiers**: people who react to the instrument in the wrong direction, $D_{1i} = 0$ and $D_{0i} = 1$

From any dataset, it is impossible to see who belongs to what group

The Angrist-Imbens-Rubin Causal Model

Angrist *et al.* (1996) define the **minimum set of assumptions** for the **identification of a causal effect** for the relevant subgroup of the population

As an example, consider Angrist (1990): the impact of **being a Vietnam veteran on earnings**

The Vietnam Draft Lottery (Angrist, 1990)

Context:

- In the 1960s and 70s young men in the US were at **risk of being drafted for military service** in Vietnam.
- Fairness concerns led to the institution of a **draft lottery** in 1970 that was used to determine **priority for conscription**

In each year from 1970 to 1972, **random sequence numbers were randomly assigned** to each birth date in cohorts of 19-year-olds.

- Men with lottery numbers below a cutoff were eligible for the draft.
- Men with lottery numbers above the cutoff were not.

But **compliance was not perfect**

- Many eligible men were exempted from service for health or other reasons.
- Others, who were not eligible, nevertheless volunteered for service.

The Vietnam Draft Lottery (Angrist, 1990)

Idea: use **lottery outcome as an instrument** for veteran status

Is there a first stage? the lottery did not completely determine veteran status, but it certainly mattered

What about the exclusion restriction?

- the lottery was random
- it seems reasonable to assume that its only effect was on veteran status

The Vietnam Draft Lottery (Angrist, 1990)

The **instrument is thus defined** as follows:

- $Z_i = 1$ if lottery implied individual i would be draft eligible,
- $Z_i = 0$ if lottery implied individual i would not be draft eligible.

The instrument affects **treatment**, which in this application amounts to **entering military service**.

The econometrician observes **treatment status** as follows:

- $D_i = 1$ if individual i served in the Vietnam war (veteran),
- $D_i = 0$ if individual i did not serve in the Vietnam war (not veteran).

The Angrist-Imbens-Rubin Causal Model

In Angrist (1990), the **main research question** is whether veteran status has a causal effect on earnings

The **causal effect of veteran status**, conditional on draft eligibility status, is defined as

$$Y_i(1, Z_i) - Y_i(0, Z_i)$$

We are **unable to identify individual treatment effects**, because we **do not observe all potential outcomes**

The Angrist-Imbens-Rubin Causal Model: Assumptions

Assumption 1: Random Assignment (ignorability)

All units have the **same probability of assignment to treatment**

$$Pr(Z_i = 1) = Pr(Z_j = 1). \tag{2}$$

Given random assignment we can **identify and estimate the two intention to treat** causal effects:

$$E(D_i|Z_i = 1) - E(D_i|Z_i = 0) = \frac{cov(D_i, Z_i)}{var(Z_i)} \tag{3}$$

$$E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) = \frac{cov(Y_i, Z_i)}{var(Z_i)}. \tag{4}$$

The Angrist-Imbens-Rubin Causal Model: Assumptions

Assumption 2: Non-zero average causal effect of Z on D

The **probability of treatment must be different** in the two assignment groups:

$$Pr(D_{i1} = 1) \neq Pr(D_{i0} = 1) \tag{5}$$

This is the equivalent of the **first stage in the conventional IV** approach.

The Angrist-Imbens-Rubin Causal Model: Assumptions

Assumption 3: Exclusion Restriction

The **instrument** affects the **outcome only through the treatment**

$$Y_i(D_i, 0) = Y_i(D_i, 1) = Y_i(D_i) \tag{6}$$

Given treatment, assignment does not affect the outcome. So we can define the causal effect of D_i on Y_i as

$$Y_{i1} - Y_{i0}. \tag{7}$$

This difference is not observed in the data. We **need to assume that assumption 3 holds** and bring good arguments in favour of it.

The Angrist-Imbens-Rubin Causal Model: Assumptions

Assumption 4: Monotonicity

- The instrument affects the **treatment status of all units** in the **same direction**
- Binary case: **no one does the opposite** of his/her assignment
- I.e. there are **no defiers**

$$D_{i1} \geq D_{i0} \; \forall i \tag{8}$$

Assumptions 2 and 4 together give **Strong Monotonicity** and ensure that:

- there is **no defier** and
- there exists **at least one complier**

Compliance types

		D_{i0}	
		0	1
D_{i1}	0	never-taker	defier
	1	complier	always-taker

Compliance types by treatment status and instrument

		Z_i	
		0	1
D_i	0	complier OR never-taker	never-taker OR defier
	1	always-taker or defier	complier OR always-taker

Compliance types

Compliance types by treatment status and instrument given monotonicity

		Z_i	
		0	1
D_i	0	complier OR never-taker	never-taker
	1	always-taker	complier OR always-taker

Back to the example (Angrist, 1990)

A1: instrument is as good as **randomly assigned**

- draft eligibility was assigned by a lottery...

A2: can have **no direct effect on the outcome** variable (earnings)

- this is debatable. Angrist argues that it holds

A3: **instrument affects the treatment**

- this can be checked

A4: **monotonicity**: a man who serves if not draft eligible, would also serve if draft eligible

- this seems plausible

Local Average Treatment Effect (Angrist, 1990)

Under the assumptions A1-A4, the IV approach in Angrist (1990) identifies a **local average treatment effect (LATE)**

The **effect is "local"** because

- it identifies the **effect on the compliers**
- ... the **causal effect of the draft on earnings** for men whose treatment status is changed by the instrument
- i.e. on men who are **drafted if eligible** but who **wouldn't volunteer if not eligible**

The **LATE is different from the ATE** because it excludes men who

- would be exempt from the draft regardless of their eligibility (never-takers)
- would volunteer regardless of their eligibility (always-takers)

The LATE theorem

Given assumptions 1-4,

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} = E(Y_{i1} - Y_{i0}|D_{i1} > D_{i0}) \quad (9)$$

$$= E(Y_{i1} - Y_{i0}|\text{complier}). \quad (10)$$

It shows that the **Wald estimator** equals the **average treatment effect for compliers**

The LATE theorem

Proof: Let π_c, π_n, π_a be the population proportions of compliers, never-takers and always-takers respectively.
Consider the least squares regression of Y on a constant and Z.
The slope coefficient in that regression estimates

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]. \tag{11}$$

Consider the first term:

$$\begin{aligned} E[Y_i|Z_i = 1] = & E[Y_i|Z_i = 1, \text{complier}]\ Pr(\text{complier}|Z_i = 1) \\ & + E[Y_i|Z_i = 1, \text{never-taker}]\ Pr(\text{never-taker}|Z_i = 1) \\ & + E[Y_i|Z_i = 1, \text{always-taker}]\ Pr(\text{always-taker}|Z_i = 1). \end{aligned}$$

This equals

$$E[Y_{i1}|\text{complier}]\pi_c + E[Y_{i1}|\text{never-taker}]\pi_n + E[Y_{i1}|\text{always-taker}]\pi_a.$$

The LATE theorem

Similarly,

$$\begin{aligned} E[Y_i|Z_i = 0] =& E[Y_i|Z_i = 0, \text{complier}]Pr(\text{complier}|Z_i = 0) \\ &+ E[Y_i|Z_i = 0, \text{never-taker}]Pr(\text{never-taker}|Z_i = 0) \\ &+ E[Y_i|Z_i = 0, \text{always-taker}]Pr(\text{always-taker}|Z_i = 0). \end{aligned}$$

This equals

$$E[Y_{i0}|\text{complier}]\pi_c + E[Y_{i0}|\text{never-taker}]\pi_n + E[Y_{i1}|\text{always-taker}]\pi_a \tag{12}$$

Hence the difference is

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[Y_{i1} - Y_{i0}|\text{complier}]\pi_c. \tag{13}$$

The LATE theorem

The same argument can be used to show that the slope coefficient in the regression of D on Z is

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = \pi_c. \tag{14}$$

The **instrumental variables estimand**, the ratio of the two reduced form estimands, equals the **local average treatment effect**:

$$E(Y_{i1} - Y_{i0}|\text{complier}) \tag{15}$$

The Angrist-Imbens-Rubin approach concludes that the **only causal effect that IV can identify** with a minimum set of assumptions is the causal effect for compliers, i.e. the **Local Average Treatment Effect (LATE)**.

The LATE theorem

Intuitively this makes sense because **compliers are the only group on which the data can be informative**:

- Compliers are the only group with units observed in both treatments (given that defiers have been ruled out).
- Always takers and never-takers are observed only in one treatment.

The LATE is a controversial parameter,

- it is defined for an unobservable sub-population
- it is instrument dependent

Therefore, it is no longer clear **which interesting policy question it can answer**.

Better LATE than never?

Under A1-A4, IV ensures **internal validity of the LATE**

But LATE has **no (or little) external validity**. Without further assumptions

- we cannot generalize to the population
- we cannot generalize to different contexts

Despite these shortcomings, **LATE is often the best we can do**

- similar estimates from different contexts increase external validity

There are many relevant positive and normative questions for which the LATE seems to be an interesting parameter in addition to being the only one we can identify without making unreasonable assumptions.

Extrapolating LATE to the Full Population

We cannot consistently estimate the **average treatment effect** for **always-takers** and **never-takers**

- but have some information about these subpopulations...

We can estimate $E[Y_{i0}|\text{never-taker}]$ and $E[Y_{i1}|\text{always-taker}]$.

We can look for **evidence of heterogeneity in outcomes** by compliance status, by comparing these to $E[Y_{i0}|\text{complier}]$ and $E[Y_{i1}|\text{complier}]$.

If the **outcomes differ substantially** between groups, it is **difficult to extrapolate from LATE**

Extrapolating LATE to the Full Population

We can estimate π_c, π_n, π_a from the population distribution of treatment and instrument status:

$$E[D_i|Z_i = 0] = \pi_a, \; E[D_i|Z_i = 1] = \pi_a + \pi_c.$$

which we can invert to infer the population shares of the different types:

$$\begin{aligned}
\pi_a &= E[D_i|Z_i = 0], \\
\pi_c &= E[D_i|Z_i = 1] - E[D_i|Z_i = 0], \\
\pi_n &= 1 - E[D_i|Z_i = 1].
\end{aligned}$$

Extrapolating LATE to the Full Population

Now consider **average outcomes by instrument and treatment status**:

$$E[Y_i|D_i = 0, Z_i = 0] =$$
$$\frac{\pi_c}{\pi_c + \pi_n} E[Y_{i0}|\text{complier}] + \frac{\pi_n}{\pi_c + \pi_n} E[Y_{i0}|\text{never-taker}]$$

$$E[Y_i|D_i = 0, Z_i = 1] = E[Y_{i0}|\text{never-taker}]$$

$$E[Y_i|D_i = 1, Z_i = 0] = E[Y_{i1}|\text{always-taker}]$$

$$E[Y_i|D_i = 1, Z_i = 1] =$$
$$\frac{\pi_c}{\pi_c + \pi_a} E[Y_{i1}|\text{complier}] + \frac{\pi_a}{\pi_c + \pi_a} E[Y_{i1}|\text{always-taker}]$$

From these relationships we can calculate $E[Y_{i0}|\text{complier}]$ and $E[Y_{i1}|\text{complier}]$.

LATE in Angrist (1990)

The **simple OLS regression** leads to:

$$log(\text{earnings})_i = 5.44 - 0.021 \times \text{veteran}_i$$

The following table contains population sizes of the four treatment/instrument samples.

Treatment status by assignment (Angrist, 1990)

		Z_i	
		0	1
D_i	0	5,948	1,915
	1	1,372	865

LATE in Angrist (1990)

For example, with a **high lottery number** ($Z_i = 0$), 1,372 individuals serve and 5,948 do not.

Using these data we can calculate the **proportions of the various compliance types** under the no-defiers assumption.

For example, the proportion of never-takers is estimated as the conditional probability that $D_i = 0$ given $Z_i = 1$.

$$\Pr(\text{never-taker}) = \frac{1915}{1915 + 865}$$

The **estimated proportions** are never-taker (0.6888), defier (0), complier (0.1237) always-taker (0.187).

LATE in Angrist (1990)

Estimated average log earnings (Angrist, 1990)

		Z_i	
		0	1
D_i	0	5.45	5.40
	1	5.41	5.43

Estimated average log earnings by type (Angrist, 1990)

Type	Y_{i0}	Y_{i1}
Never-taker	5.40	
Complier	**5.69**	**5.46**
Always-taker		5.41

LATE in Angrist (1990)

The second table on the previous slide gives the **estimated averages for the four compliance types**, under the **exclusion restriction**

The **exclusion restriction** is the **key assumption** here. There are a number of reasons why it may be violated, e.g., never-takers taking actions to avoid military service if draft eligible.

The **local average treatment effect is -0.23**, a 23% drop in earnings as a result of serving in the military.

LATE in Angrist (1990)

Is the LATE informative about the total population?

Average log earnings for never-takers are 5.40, 29% lower than average earnings for compliers who do not serve in the military.

- never-takers are substantially different to compliers
- average effect of 23% for compliers need not be informative about never-takers

In contrast, average **log earnings for always-takers are only 6% lower than those for compliers** who serve,

- differences between always-takers and compliers are considerably smaller.

Recent work **formalizes the extrapolation problem**: Mogstad & Torgovitsky (2018), Mogstad *et al.* (2018)

IV with Covariates

It is possible to **condition on pre-treatment covariates**

- the covariates have to be included in the first and second stage (and reduced form)

The **IV assumptions are different**

- First stage: $cov(Z, D|X) \neq 0$ has to be sufficiently strong
- Exclusion restriction: $cov(Z, u|X) = 0$ has to hold conditional on X
- Monotonicity also has to hold conditional on X

IV with Multiple Instruments

It is possible to have **multiple instrumental variables**

2SLS combines instruments to get a single (more precise) estimate

Assumptions needed:

- Each instrument needs to be **as good as randomly assigned**
- Each instrument needs to **satisfy the exclusion restriction**
- The joint **first stage** has to be **strong enough**

IV with Multiple Instruments

We can write the **model in matrix form**

$$y = X\beta + \varepsilon \tag{16}$$

$$W = Z\gamma + \eta \tag{17}$$

- X is a $(N \times K)$ matrix of endogenous and exogenous regressors
- Z is a $(N \times Q)$ matrix of instruments and exogenous regressors, $Q \geq K$

We say

- a model is *just identified* if $Q = K$
- a model is *overidentified* if $Q > K$

Instrument Strength

For any meaningful analysis, the instrument has have a **sufficiently strong effect on the regressor(s)**

If the instrument doesn't shift the treatment a lot, we divide by a very small first stage coefficient

There are three **problems with weak instruments**

- high variance of the estimator $\Rightarrow$ unreliable inference
- inconsistency of the IV estimator
- small sample bias

General Model in Matrix Notation

Remember: the **OLS estimator** is

$$\hat{\beta}^{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

The **TSLS estimator** includes a projection of X on Z and, therefore, is

$$\widehat{\beta}^{2SLS} = \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'Y$$

$$\widehat{\beta}^{2SLS} = (\boldsymbol{X}'\boldsymbol{P}_z\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_z Y$$

Where $\boldsymbol{P}_z = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ is the projection matrix and

$$\hat{\mathbf{X}} = \boldsymbol{P}_z\boldsymbol{X} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$$

Variance of OLS and TSLS

Asymptotic **variance of OLS estimator** with K regressors

$$\widehat{Avar}\left(\hat{\boldsymbol{\beta}}^{OLS}\right) = \hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

with

$$\hat{\sigma}^2 = (N-K)^{-1}\hat{u}'\hat{u}$$

TSLS equivalent:

$$\begin{aligned}\widehat{Avar}\left(\hat{\boldsymbol{\beta}}^{2SLS}\right) &= \hat{\sigma}^2 \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1} \\ &= \hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\end{aligned}$$

$\Rightarrow$ a **weak first-stage** correlation will **increase the variance**

Variance of TSLS, $K = Q = 1$

It can be shown that

$$\widehat{Avar}\left(\hat{\beta}^{2SLS}\right) = \hat{\sigma}^2 \frac{1}{N\rho_{xz}^2\sigma_x^2},$$

where $\rho_{xz} = \text{cov}\left(z_i, x_i\right) / \left(\sigma_z \sigma_x\right)$.

This equation offers **several important insights**:

- An increase in the sample size decreases the standard errors
- The standard error is higher the higher the variance of the residuals $\hat{\sigma}^2$ and the lower the variation in x_i
- The standard error decreases with the strength of the first stage
- Also: $\widehat{Avar}\left(\hat{\beta}^{2SLS}\right) > \widehat{Avar}\left(\hat{\beta}^{OLS}\right)$ because $\rho_{xx} = 1$

Note: we assumed here **homoskedasticity of the error terms**

Weak Instruments: Inconsistency

Consider the simultaneous equations model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$
$$x_i = \mu + \pi z_i + v_i.$$

The OLS and IV estimators are given by

$$\widehat{\beta}_{OLS} = \frac{cov(y_i, x_i)}{var(x_i)}$$
$$\widehat{\beta}_{2SLS} = \frac{cov(y_i, \widehat{x}_i)}{var(\widehat{x}_i)}$$

and the plims of the estimators are

$$\text{plim}\widehat{\beta}_{OLS} = \beta + \frac{\sigma_{x\varepsilon}}{\sigma_x^2}$$
$$\text{plim}\widehat{\beta}_{2SLS} = \beta + \frac{\sigma_{\widehat{x}\varepsilon}}{\sigma_{\widehat{x}}^2}.$$

Weak Instruments: Inconsistency

This yields

$$\frac{\text{plim}\widehat{\beta}_{2SLS} - \beta}{\text{plim}\widehat{\beta}_{OLS} - \beta} = \frac{\sigma_{\widehat{x}\varepsilon}/\sigma_{x\varepsilon}}{\sigma^2_{\widehat{x}}/\sigma^2_x} = \frac{\sigma_{\widehat{x}\varepsilon}/\sigma_{x\varepsilon}}{R^2_{xz}}.$$

The **inconsistency** of the 2SLS estimator relative to the OLS estimator is related to the **relative *endogeneity*** of z and x.

Notice that R^2_{xz}, is the R^2 from the first stage regression.

- The instrument z may be *almost* as good as randomly assigned but not quite.
- Hence, $\sigma_{\widehat{x}\varepsilon}$ may be small but not quite zero.

However, even if $\sigma_{\widehat{x}\varepsilon}$ is small compared to $\sigma_{x\varepsilon}$, the **relative inconsistency of the 2SLS** estimator may still be important as long as R^2_{xz} is also small, i.e. as long as the **correlation of z and x is low**.

Small Sample Bias

We know that **OLS is consistent and is unbiased** (under standard assumptions).

In finite samples, **2SLS is consistent but biased** (see appendix)

Problem is worst when

1. **Instruments are weak** AND
2. The model is heavily **overidentified**

Intuition for Small Sample Bias

Suppose the **instrument is completely uncorrelated** with the endogenous regressor.

- With an **infinite amount of data**, $cov(x_i, z_i)$ is going to be exactly zero, and the IV estimator cannot be computed anymore.
- In a **small sample**, $cov(x_i, z_i)$ is not going to be literally zero, just small.
- So the observed correlation is only due to **sampling variation**, i.e. **noise**

Intuition for Small Sample Bias

In **any particular sample**, z and x will have some **slight random correlation**

- Variation in the sample just comes from ε
- This means $\widehat{x}$ **picks up some variation** that is just like the original variation in x,
- so it is not in any way **purged of the endogenous variation in** x

Since x is correlated with 2nd-stage error, $\widehat{x}$ will also be **correlated with 2nd-stage error**. Consequence:

- OLS and 2SLS estimate the same quantity (on average)
- If the true $cov(x, z) \neq 0$ but small 2SLS is biased towards OLS

Weak Instruments - What to Do?

If you consider using more than one instrument (hint: DON'T!), show first the **results of the just-identified model**, using your best instrument.

- Monte Carlo simulations show that **just-identified 2SLS is approximately unbiased**
- But just-identified estimates are also **unstable and imprecise**

Show the **F-Statistic of the first stage**

- Stock *et al.* (2002) suggest that an F-Statistic > 10 indicates that the instruments are sufficiently strong
- But this is a rule of thumb, nothing more.

Are Weak Instruments a Problem in Practice?

No! say most researchers $\Rightarrow$ after all we can test for weak instruments

Yes! says Alwyn Young (2019): *'Consistency without Inference"*

- Looks at >1300 IV estimates from the AER and AEJs
- Uses bootstrap and jackknife procedures to assess inference

Result: **IVs are weaker than we think**

- In most studies, the results are driven by very few observations
- The F-Statistic does not reflect this $\Rightarrow$ uninformative
- IV estimator has very low power

What about $F > 10$?

Consider the model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$
$$x_i = \mu + \pi z_i + v_i.$$

Lee *et al.* (2020) show that $F > 10$ does **not permit valid inference** about β based on the t-statistic $\widehat{\beta}/se(\widehat{\beta})$.

- The critical values of $t = \pm 1.96$ are too small of inference with a type-I error $\alpha = 0.05$
- They show that reliable inference at the 5% level is possible with $F > 143$

All hope is lost?

- No! Because we can use **larger critical values**
- But we are less likely to detect a statistically significant effect

Weak Instruments without Homoskedasticity

So far we derived the standard case with homoskedastic errors

But this **assumption is often violated** due to **heteroskedasticity, serial or spatial auto-correlation, or clustering**

- The "conventional" (non-robust) F-Statistic results in standard errors that are too small

Alternatives

- Robust F-tests (Kleibergen & Paap, 2006)
- Effective F-statistic (Montiel Olea & Pflueger, 2013) (scales up the non-robust F)

Recommendations: Instrument Strength

Weak instruments are a problem for inference

- inflated variance makes inference less reliable
- it can exacerbate the inconsistency of the estimator as well as the small sample bias

So **what to do?**

- Report the non-robust F and adjust the standard errors using the **critical values from Lee *et al.* (2020)**
- Report the **effective F-Stat** by Montiel Olea & Pflueger (2013) and use their critical values to determine whether IVs are weak
- If the IVs are weak, don't discard the paper but use **robust methods** based on Anderson & Rubin (1949) (Andrews *et al.*, 2006; Moreira, 2009)

For more information, see Andrews *et al.* (2019)

Continuous Instruments

It is also possible to obtain an **IV estimate with a continuous instrument and/or treatment**

The assumptions (first stage, exclusion restriction, monotonicity) remain the same

The **LATE is more difficult to interpret**

- units differ in their *compliance intensity*
- i.e. some react to the instrument more than others
- LATE is the weighted average of unit causal effects over the support of D
- weights are determined by the share of compliers in each bin of D

Often useful to use the **binary case as reference** (high/low intensity of treatment and instrument)

IV Example: Elsner & Wozny (2018)

We revisit the paper on the **effect of radiation on cognitive skills**

$$y_{ims} = \alpha + \beta Cs137_{ms}^{86} + \boldsymbol{X}'_{\boldsymbol{ims}}\boldsymbol{\gamma} + \delta_s + \varepsilon_{ims}$$

Challenge: the initial level of radiation may be endogenous

- balancing tests don't show differences conditional on $\boldsymbol{X}_{\boldsymbol{ims}}$ and δ_s
- but there could be unobserved differences
- especially because the raw differences suggest that the results could be driven by unobservables

IV Example: Elsner & Wozny (2018)

We use an **instrument that predicts the level of radiation** based on two factors

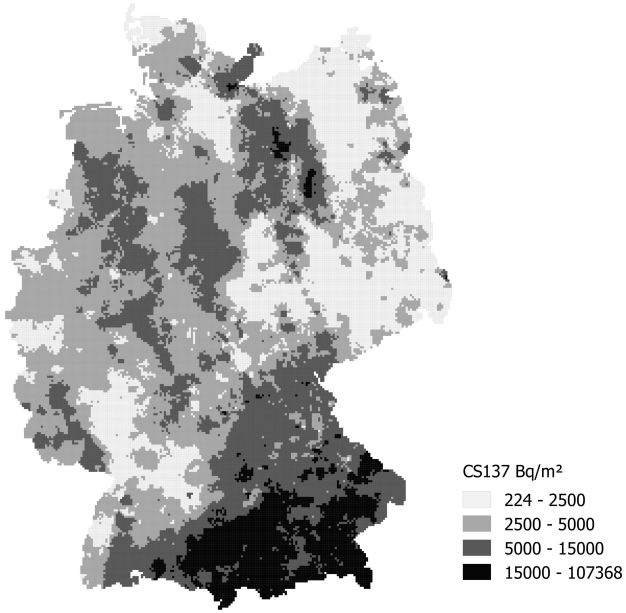
1) The **amount of rainfall** in the critical ten days after the desaster (May 1-10, 1986)

- radioactive plume was transported by (East) winds
- areas where contaminated when it rained while the plume was hanging over it

2) The **amount of radioactive matter in the plume**

- the plume moved from south-east to north-west Germany
- lots of radioactive matter was “rained off” in the South
- a lot less radioactive matter was left in the North

IV Example: Elsner & Wozny (2018)



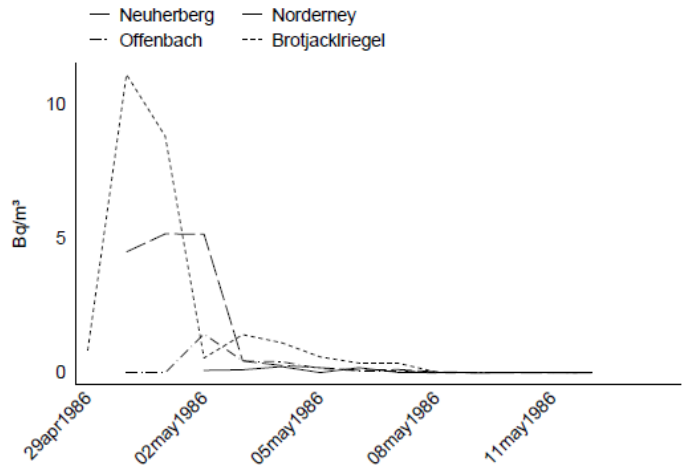
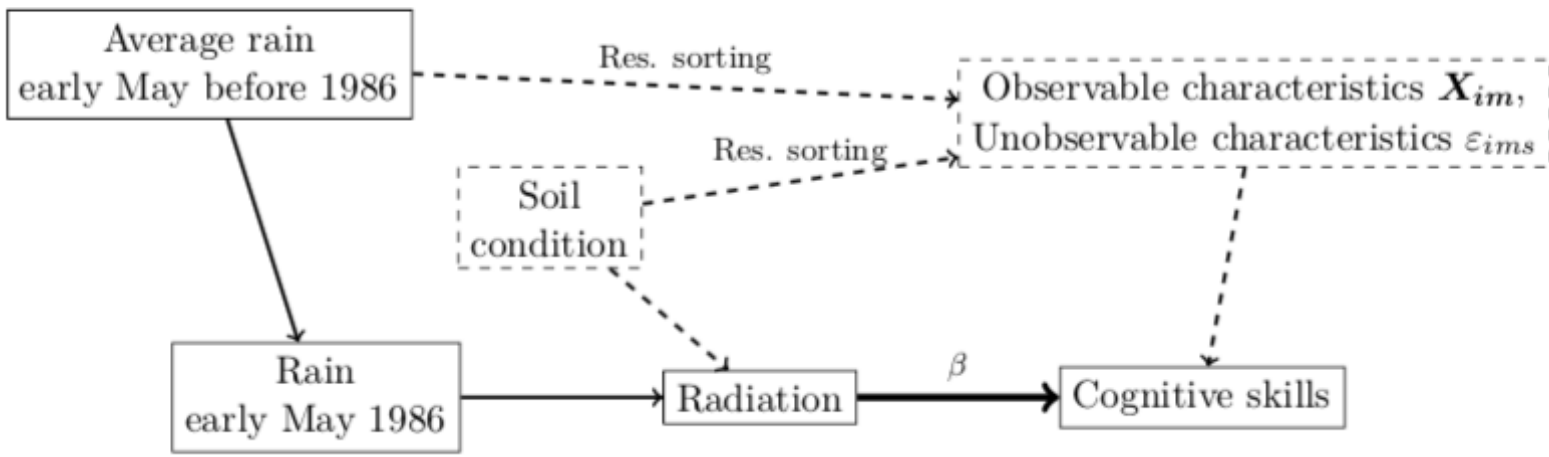


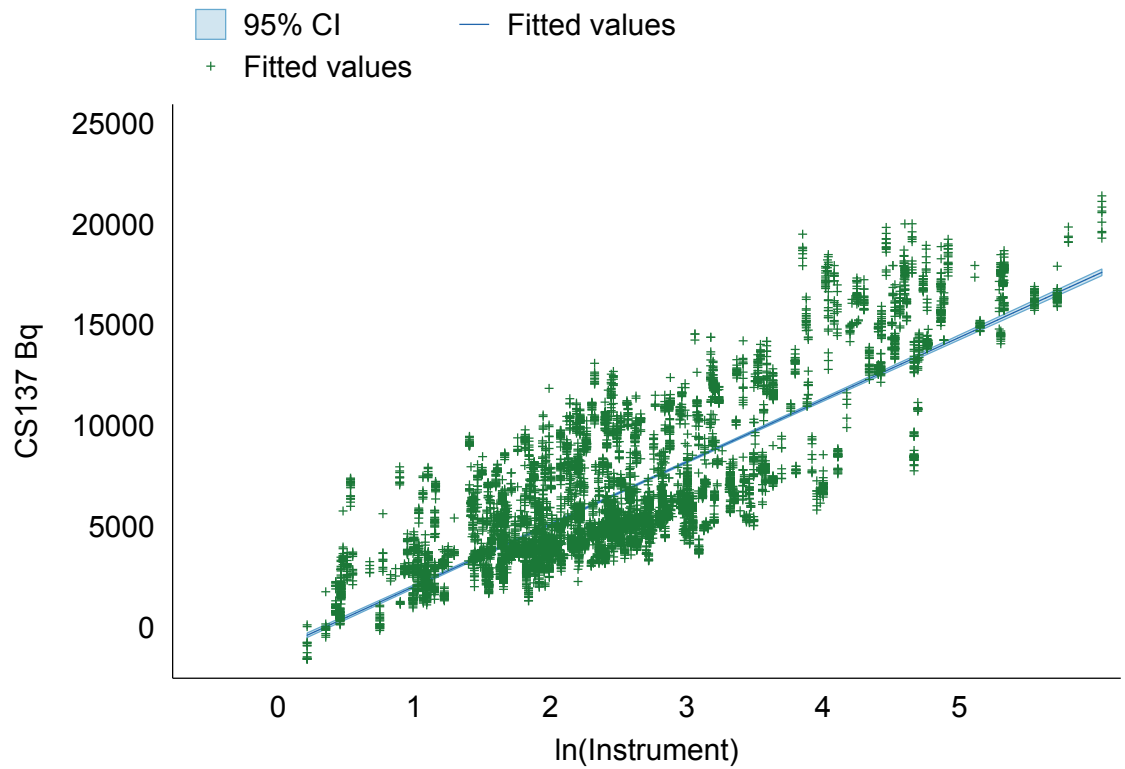
Figure 9: Air concentration of radioactive particles in 1986

IV Example: Elsner & Wozny (2018)

A DAG helps...



IV Example: Elsner & Wozny (2018)

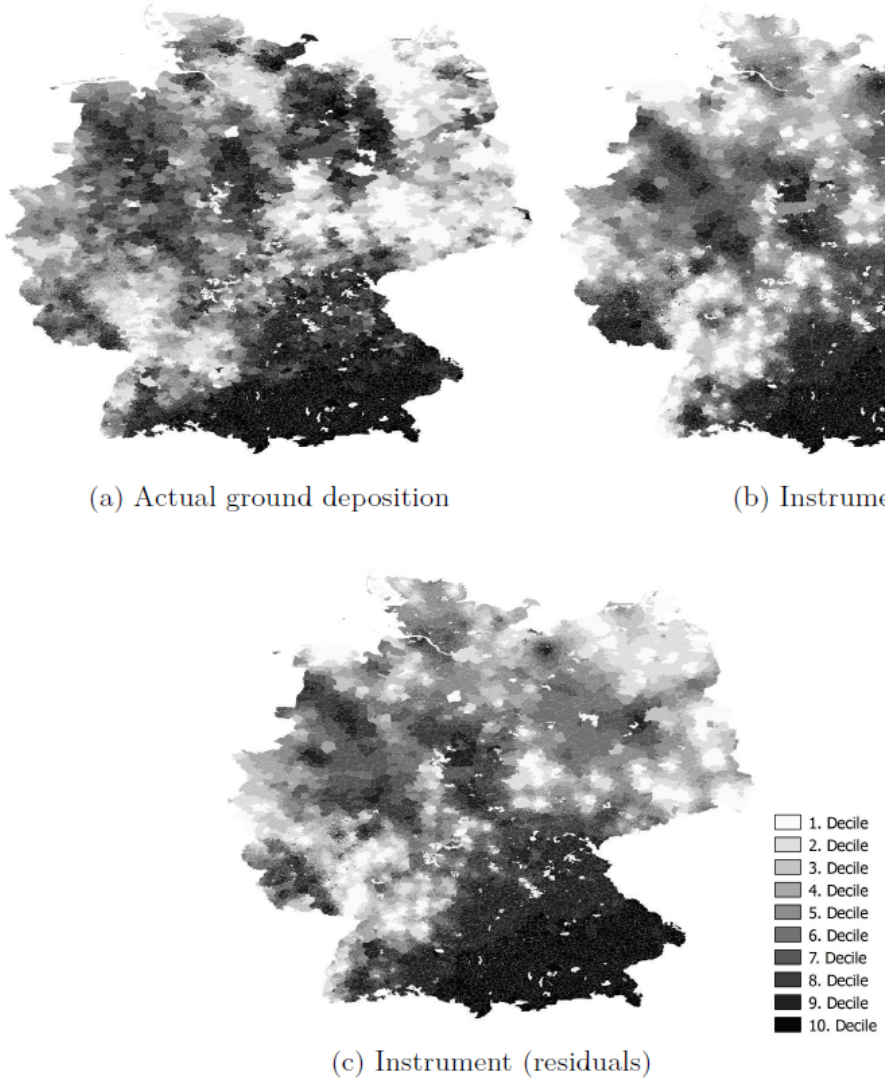


First stage

$$Cs137_{ms}^{86} = \delta_0 + \delta_1 \ln(particles \times rain)_m s + \boldsymbol{X}'_{\boldsymbol{ims}}\boldsymbol{\gamma} + \delta_s + e_i$$

IV Example: Elsner & Wozny (2018)

Graphic first stage



(a) Actual ground deposition

(b) Instrument

(c) Instrument (residuals)

Also show that first stage only exists in 1986

Exclusion Restriction

Assumption: $cov(\ln(particles \times rain)_{ms}, u_i | \boldsymbol{X_{ims}}, \delta_s)$

Important: $\boldsymbol{X_{ims}}$ includes **average rainfall on May 1-10 (1981-1985)**

- hence, the instrument represents **deviations from the average rainfall**
- these deviations are **weighted by the amount of radioactive matter** in the plume

Assumption: **abnormal rainfall on 10 days in 1986** only affects **cognitive skills in 2010** through its effect on radiation

Exclusion Restriction

This **identification assumption is not testable**. But we can bring supportive evidence.

Here we **regress individual characteristics on the instrument**

	(1)	(2)	(3)	(4)
A. Individual characteristics				
Age in 1986	-0.108	-0.037	-0.216	-0.172
	(0.237)	(0.308)	(0.343)	(0.373)
Female	0.004	0.008	-0.020	-0.010
	(0.011)	(0.015)	(0.013)	(0.016)
Native speaker	0.010***	0.018***	0.014**	0.013
	(0.003)	(0.006)	(0.006)	(0.008)
Employed in April 1986	-0.002	0.024	0.005	0.009
	(0.014)	(0.019)	(0.017)	(0.021)
Unemployed in April 1986	0.000	-0.001	-0.004	-0.002
	(0.003)	(0.004)	(0.004)	(0.005)
If employed : Qualified or highly qualified	0.009	-0.030	-0.016	-0.017
	(0.021)	(0.030)	(0.031)	(0.031)
Children before 1986	-0.018**	0.012	-0.002	0.002
	(0.009)	(0.012)	(0.010)	(0.013)
Older siblings	0.010	0.028	0.019	0.042*
	(0.015)	(0.021)	(0.019)	(0.022)
Smoke before 1986	-0.013	0.013	-0.018	0.000
	(0.014)	(0.018)	(0.018)	(0.020)

Column 4: conditional on State FE and municipality characteristics

Exclusion Restriction

Educational attainment in April 1986				
Lower secondary and secondary	0.001	-0.004	-0.003	-0.001
	(0.005)	(0.008)	(0.008)	(0.008)
Upper secondary	0.010	0.023	0.014	0.007
	(0.011)	(0.017)	(0.015)	(0.019)
Tertiary	-0.005	0.006	-0.003	0.004
	(0.010)	(0.011)	(0.013)	(0.013)
In school or college education	-0.008	-0.035*	-0.020	-0.025
	(0.013)	(0.019)	(0.018)	(0.021)
In education, already attained lower secondary and secondary	0.000	-0.003	-0.000	0.003
	(0.009)	(0.012)	(0.012)	(0.013)
In education, already attained upper secondary	-0.004**	-0.006*	-0.006*	-0.004
	(0.002)	(0.003)	(0.004)	(0.004)
In education, already attained tertiary	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Highest parental education				
Lower secondary education	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Secondary education	0.002	0.010	0.012	0.015
	(0.009)	(0.011)	(0.013)	(0.014)
Upper seconary	0.047***	0.006	0.018	0.017
	(0.017)	(0.020)	(0.021)	(0.022)

Column 4: conditional on State FE and municipality characteristics

$\Rightarrow$ instrument **unrelated to observable characteristics**

2SLS Results

	OLS (1)	OLS (2)	OLS (3)	OLS (4)	IV (5)
A. Individual test scores					
Math	0.003	0.002	-0.011***	-0.011***	-0.023**
	(0.004)	(0.003)	(0.003)	(0.003)	(0.009)
Reading	-0.001	0.001	-0.013***	-0.013***	-0.026**
	(0.006)	(0.004)	(0.005)	(0.005)	(0.007)
Listening comprehension	-0.003	-0.003	-0.008**	-0.009*	-0.014**
	(0.004)	(0.003)	(0.004)	(0.004)	(0.007)
ICT	0.000	0.001	-0.003	-0.005	-0.005
	(0.002)	(0.002)	(0.003)	(0.004)	(0.006)
Scientific literacy	0.001	0.002	-0.002	-0.003	-0.007
	(0.003)	(0.002)	(0.003)	(0.003)	(0.006)
Reasoning	0.002	0.001	-0.001	-0.001	-0.007
	(0.003)	(0.003)	(0.004)	(0.004)	(0.007)
Reading speed	-0.001	0.001	-0.010***	-0.008**	-0.018**
	(0.003)	(0.003)	(0.004)	(0.004)	(0.007)
Perceptual speed	0.003	0.003	-0.003	-0.004	-0.000
	(0.003)	(0.002)	(0.003)	(0.003)	(0.007)
B. Indices					
Cognitive skill index	0.001	0.002	-0.007***	-0.008***	-0.014**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)
Crystallized intelligence index	0.000	0.002	-0.007**	-0.008**	-0.016**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)
Fluid intelligence index	0.002	0.003	-0.006**	-0.006*	-0.009
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)
First-stage: dep. var. Cs137 kBq/m^2					
ln(Precipitation (mm/m^3) × Air contamination (mm/m^3))					6.480***
					(1.029)
F statistic					39.652

IV result: 1SD increase in CS137 $\Rightarrow$ 8.4% of a SD reduction in test scores

2SLS Results incl. Placebos

Outcomes: log nr of diagnoses

	(1)	(2)	(3)	(4)
Dementia	-0.017	0.041	0.077**	0.120***
	(0.025)	(0.031)	(0.033)	(0.046)
Hypertonia	-0.021	0.006	0.001	0.101
	(0.026)	(0.032)	(0.031)	(0.098)
Diabetes	0.008	-0.001	-0.003	0.106
	(0.027)	(0.024)	(0.024)	(0.098)
Asthma	0.025	0.050	0.055	0.042
	(0.062)	(0.064)	(0.061)	(0.062)
Injuries	0.048*	0.024	0.008	0.012
	(0.029)	(0.029)	(0.031)	(0.056)
First-stage:				
ln(Precipitation (mm/m^3) $\times$ Air contamination (kBq/m^3))				1.082***
				(0.113)
F statistic				91.169
Controls:				
Municipality characteristics	No	Yes	No	Yes
Municipality characteristics	No	Yes	Yes	Yes
State FE	No	No	Yes	Yes

Reduced Form

	Reduced forms			
	(1)	(2)	(3)	(4)
A. Individual test scores				
Math	0.012	0.012	-0.152***	-0.151***
	(0.038)	(0.034)	(0.049)	(0.055)
Reading	-0.025	0.004	-0.181***	-0.168***
	(0.063)	(0.049)	(0.038)	(0.037)
Listening comprehension	-0.039	-0.034	-0.084**	-0.086*
	(0.040)	(0.035)	(0.042)	(0.045)
ICT	0.010	0.009	-0.042	-0.033
	(0.028)	(0.027)	(0.034)	(0.038)
Scientific literacy	-0.000	0.005	-0.068*	-0.049
	(0.030)	(0.027)	(0.035)	(0.036)
Reasoning	0.007	-0.007	-0.051	-0.046
	(0.032)	(0.032)	(0.038)	(0.041)
Reading speed	-0.026	-0.018	-0.169***	-0.119**
	(0.039)	(0.035)	(0.046)	(0.048)
Perceptual speed	0.063**	0.042	-0.003	-0.002
	(0.029)	(0.026)	(0.036)	(0.042)
B. Indices				
Cognitive skill index	0.000	0.006	-0.107***	-0.093***
	(0.035)	(0.031)	(0.031)	(0.035)
Crystallized intelligence index	-0.006	0.002	-0.106***	-0.101***
	(0.038)	(0.033)	(0.029)	(0.033)
Fluid intelligence index	0.017	0.018	-0.082**	-0.058
	(0.032)	(0.036)	(0.032)	(0.038)
Controls:				
Individual characteristics	No	Yes	Yes	Yes
County characteristics	No	No	Yes	Yes
Municipality characteristics	No	No	Yes	Yes
State FE	No	No	No	Yes

Where do good IVs come from?

Theory combined with clever datas collection. Examples

- Distance from job training centers
- College openings

Variation in policies. This requires a **deep institutional knowledge**. Examples

- assignment to judges with different severity
- differences in budgets across job training centers
- ...

Nature. Examples

- different levels of pollution in different places
- sex of the first two children
- ...

IV: Cookbook

1) Explain your identification strategy very clearly

- start with the **ideal experiment**; why is your setting different? Why is your **regressor endogenous**?
- Explain theoretically **why there should be a first stage** and what coefficient we should expect
- Explain why the instrument is **as good as randomly assigned**
- Explain theoretically **why the exclusion restriction holds** in your setting

2) Show and discuss the first stage

- Best to start with a **raw correlation**
- Do the **sign and magnitude make sense**?
- Assess the **strength of the instrument** using state-of-the-art techniques

IV: Cookbook

3) Bring supportive evidence for instrument validity

- Show that the **instrument does not predict pre-treatment characteristics**
- Can you provide evidence in support of the exclusion restriction?
- Use auxiliary tests, for example Kitagawa (2015) and Huber & Mellace (2015)
- Consider using the *plausibly exogenous* bounding procedure by Conley *et al.* (2012)

4) Discuss the results in detail

- Show the **OLS and 2SLS results**, both with **varying sets of controls**
- Comment on the differences between both (bias, LATE, etc)
- Show the **reduced form**
- If the reduced form isn't there, the effect isn't there (MHE)

Final Comments on IV

Friends tell their friends not to use IV...

This may be an extreme view, but **IVs have become less popular in recent years**

- It is very **difficult to find an IV** that fulfills the exclusion restriction
- The LATE is often not the desired policy parameter
- IV has unfavourable small sample properties

Many **"classic" IVs have been shown to be invalid**

- Quarter of birth is correlated with SES
- Twin births (IV for family size) are related to IVF and, thus, to SES
- Rainfall affects many things...

Final Comments on IV

In what settings are **IVs used these days?**

- In **randomized experiments** with imperfect compliance
- As a **complementary identification strategy**, along with FE estimation, diff-in-diff
- In **fuzzy regression discontinuity designs** (next topic)

References I

Anderson, T. W., & Rubin, Herman. 1949. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics*, **20**(1), 46–63.

Andrews, Donald W. K., Moreira, Marcelo J., & Stock, James H. 2006. Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, **74**(3), 715–752.

Andrews, Isaiah, Stock, James H., & Sun, Liyang. 2019. Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, **11**(1), 727–753.

Angrist, Joshua, & Evans, William. 1998. Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review*, **88**(3), 450–77.

Angrist, Joshua, & Pischke, Jörn-Steffen. 2009. *Mostly Harmless Econometrics - An Empiricist's Companion*. Princeton University Press.

Angrist, Joshua D. 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review*, **80**(3), 313–336.

Angrist, Joshua D., Imbens, Guido W., & Rublin, Donald B. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91**(434), 444–455.

Conley, Timothy G., Hansen, Christian B., & Rossi, Peter E. 2012. Plausibly Exogenous. *Review of Economics & Statistics*, **94**(1), 260–272.

Elsner, Benjamin, & Wozny, Florian. 2018 (Mar.). *The Human Capital Cost of Radiation: Long-Run Evidence from Exposure Outside the Womb*. IZA Discussion Papers 11408. Institute for the Study of Labor (IZA).

Huber, Martin, & Mellace, Giovanni. 2015. Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints. *The Review of Economics and Statistics*, **97**(2), 398–411.

Kitagawa, Toru. 2015. A Test for Instrument Validity. *Econometrica*, **83**(5), 2043–2063.

Kleibergen, Frank, & Paap, Richard. 2006. Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, **133**(1), 97–126.

Kling, Jeffrey R, Liebman, Jeffrey B, & Katz, Lawrence F. 2007. Experimental Analysis of Neighborhood Effects. *Econometrica*, **75**(1), 83–119.

Lee, David S., McCrary, Justin, Moreira, Marcelo J., & Porter, Jack. 2020 (Oct.). *Valid t-ratio Inference for IV*. Papers 2010.05058. arXiv.org.

References II

Mogstad, Magne, & Torgovitsky, Alexander. 2018. Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics*, **10**(1), 577–613.

Mogstad, Magne, Santos, Andres, & Torgovitsky, Alexander. 2018. Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica*, **86**(5), 1589–1619.

Montiel Olea, José Luis, & Pflueger, Carolin. 2013. A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics*, **31**(3), 358–369.

Moreira, Marcelo J. 2009. Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, **152**(2), 131 – 140. Nonparametric and Robust Methods in Econometrics.

Stock, James H., Wright, Jonathan H., & Yogo, Motohiro. 2002. A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics*, **20**(4), 518–529.

Young, Alwyn. 2019. Consistency without Inference. *LSE, mimeo*.

Appendix

Formal Analysis of the small sample bias with multiple instruments

Return to the simultaneous equation model, but with multiple instruments as in

$$y = \beta x + \eta$$
$$x = Z\pi + \xi.$$

where Z is an $N \times Q$ matrix. OLS is biased because η_i is correlated with ξ_i.

Z_i uncorrelated with ξ_i by construction.

Z_i uncorrelated with η_i by assumption.

The 2SLS estimator is

$$\widehat{\beta}_{2SLS} = (x'P_Z x)^{-1} x'P_Z y = \beta + (x'P_Z x)^{-1} x'P_Z \eta$$

$$\widehat{l\beta}_{2SLS} - \beta$$
$$= (x'P_Z x)^{-1}(\pi' Z' + \xi')P_Z \eta = (x'P_Z x)^{-1}\pi' Z'\eta + (x'P_Z x)^{-1}\xi' P_Z \eta$$

Difficult to evaluate bias as expectation doesn't go through $(x'P_Z x)^{-1}$.
However,

$$E(\widehat{\beta}_{2SLS} - \beta) \approx (E[x'P_Z x])^{-1} E[\pi' Z'\eta] + (E[x'P_Z x])^{-1} E[\xi' P_Z \eta]$$

This approximation uses asymptotics as the number of instruments goes to infinity at the same rate as the number of observations.

Can show that,

$$E(\widehat{\beta}_{2SLS} - \beta) \approx \frac{\sigma_{\eta\xi}}{\sigma_\xi^2}\left[\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\xi^2} + 1\right]^{-1}$$

The term $\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\xi^2}$ is the population equivalent of the first-stage F statistic. Then,

$$E(\widehat{\beta}_{2SLS} - \beta) \approx \frac{\sigma_{\eta\xi}}{\sigma_\xi^2}\frac{1}{F+1}$$

The lower the F-statistic, the worse the bias will be.
As $F \to 0, 2SLS\ bias \to \frac{\sigma_{\eta\xi}}{\sigma_\xi^2}$.
$OLS\ bias = \frac{\sigma_{\eta\xi}}{\sigma_x^2}$ which equals $\frac{\sigma_{\eta\xi}}{\sigma_\xi^2}$ if $\pi = 0$.
I.e. if first stage is zero, 2SLS bias equals OLS bias.

In general, 2SLS estimates are biased towards OLS when 1st-stage is weak.

1. Adding useless instruments increases bias because $E[\pi' Z' Z \pi]$ and σ_ξ^2 stay the same, but Q increases.
2. Intuitively, problem arises because 1st stage is estimated. If we knew $Z\pi$, we could use these as the fitted values and they are uncorrelated with the 1st-stage error.
3. Just-identified 2SLS is approximately unbiased (in Monte Carlo simulations). Difficult to show this formally because it has no moments.
4. However with weak instruments, just-identified estimates are unstable and imprecise.
5. Some suggest want to have 1st-stage F-statistic of at least 10 to be safe using 2SLS.