Instrumental Variable Estimation

Data Science and Causal Inference Workshop 2025

Dr Christian Engels

START RECORDING NOW

University of St Andrews | FOUNDED 1413

3 September 2025

What I hope you will get out of today

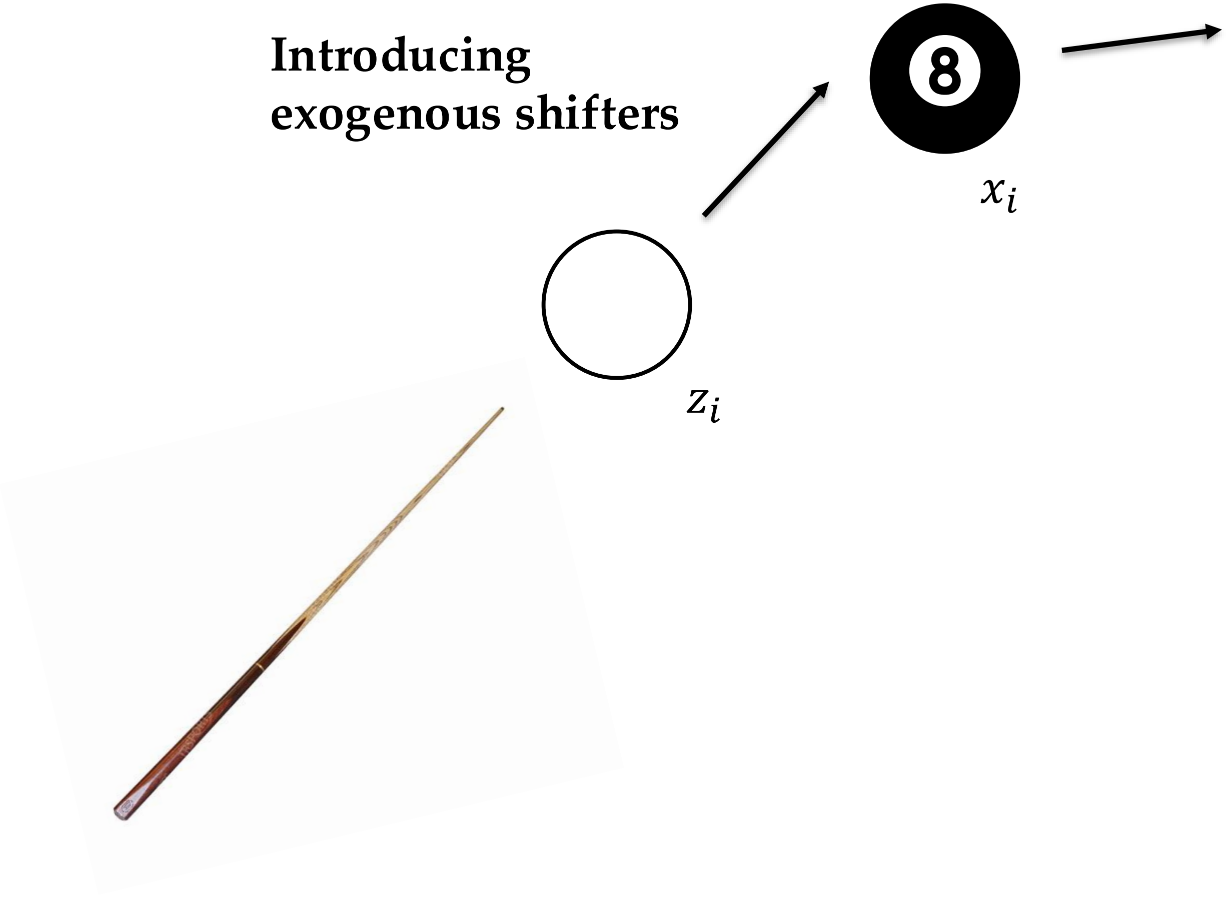
Learning objectives

- Understand intuition of using instrumental variables (IV) for overcoming omitted variable bias
- Understand interpretation and importance of IV relevance and exogeneity restrictions
- Understand IV estimation approaches and weak instrument testing
- Gain first knowledge of Generalized Method of Moments (GMM) estimation
- Appreciate role of creativity and subject knowledge in IV approach for empirical research

Outline of today

Instrumental variable estimation

1. Instrumental variable (IV) intuition
   a) Motivation: omitted variable bias
   b) Key ideas
   c) Key assumptions
   d) Simple IV estimates
2. Estimation techniques
   a) 2SLS / Generalized IV estimator
   b) Popular weak instrument tests
   c) GMM and overidentifying restrictions
3. Stata illustration



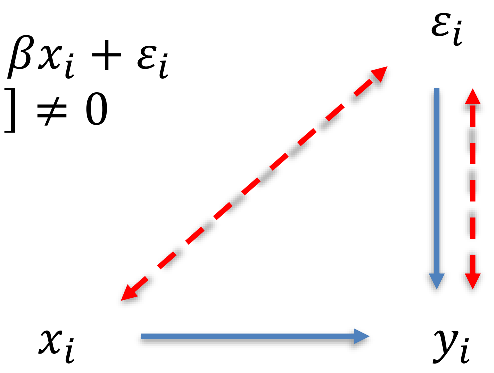
Introducing exogenous shifters

Intuition of using instrumental variables (IV) for overcoming omitted variable bias

Omitted variable bias recap

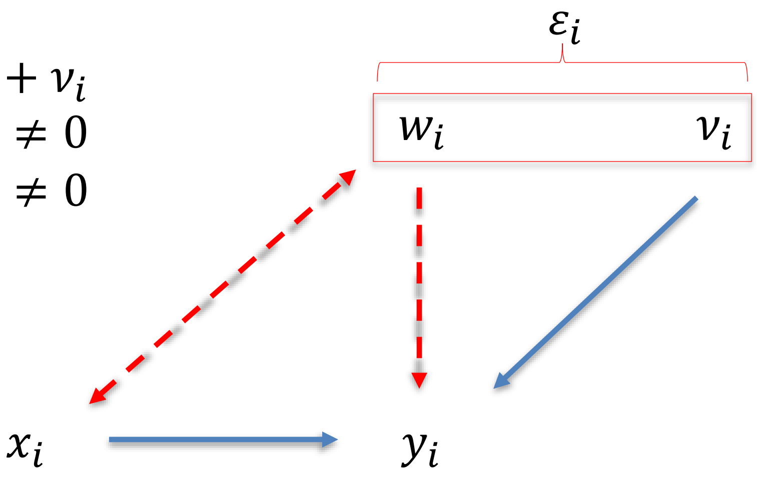
Endogenous regressor x_i

$$y_i = \alpha + \beta x_i + \varepsilon_i$$
$$E[x_i \varepsilon_i] \neq 0$$

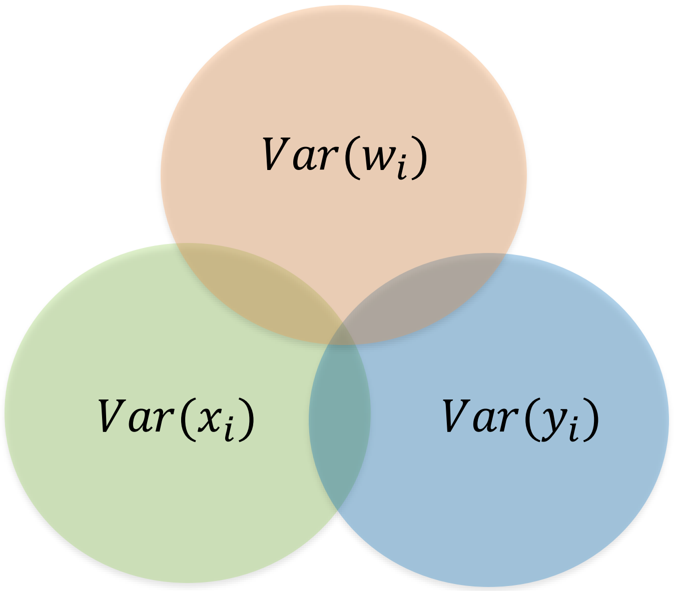


Omitted variable w_i

$$\varepsilon_i = w_i + v_i$$
$$E[x_i w_i] \neq 0$$
$$E[y_i w_i] \neq 0$$



An alternative visualization

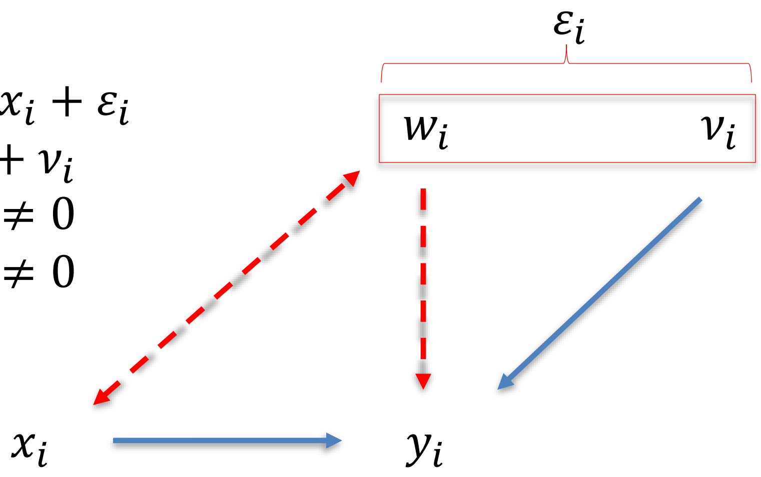


The unaccounted covariances $Cov(x_i, w_i)$ and $Cov(y_i, w_i)$ are the source of bias.

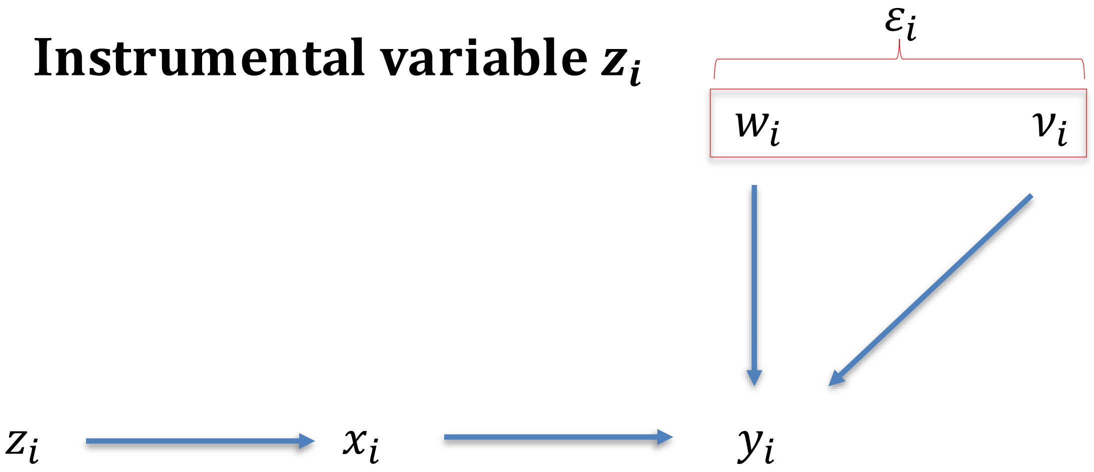
Using instrumental variables (IVs) to overcome bias

Omitted variable w_i

$$\begin{aligned} y_i &= \alpha + \beta x_i + \varepsilon_i \\ \varepsilon_i &= w_i + v_i \\ E[x_i w_i] &\neq 0 \\ E[w_i y_i] &\neq 0 \end{aligned}$$



Instrumental variable z_i



An exogenous shifter

- The variable z_i changes x_i directly and nothing else
- Changes in y_i can therefore be attributed to the change in x_i induced by z_i
- z_i is known as an “instrumental variable”

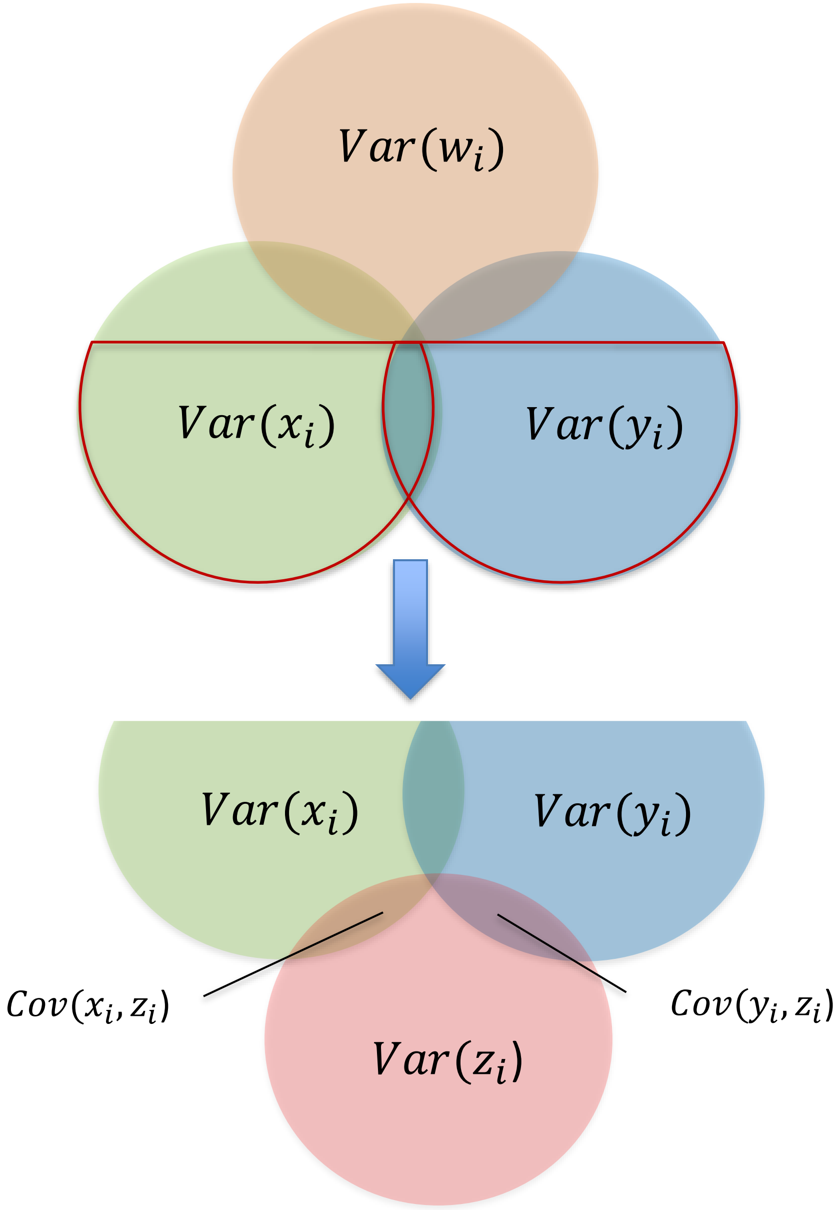
In particular this means that:

- z_i does not affect y_i directly
- z_i does not suffer from omitted variable bias itself

Decoding sample variability

Exploiting variability unrelated to w_i

- Instrumental variables use part of the variability in x_i that is unrelated to w_i
- Sample moments and assumptions on structure yield an estimate of how x_i affects y_i free of the influences of w_i
- Specifically, say we can observe:
  - the *correlation* between y_i and z_i
  - the *correlation* between x_i and z_i
- The instrumental variable assumptions then imply that $Cov(y_i, z_i)$ results from $Cov(x_i, z_i)$



Building IV estimator intuition

Simple IV estimate

$$\begin{tikzcd}
z_i \arrow[r] & x_i \arrow[r] & y_i \\
\Delta z_i = 1 \arrow[r] & \Delta x_i = 0.1 \arrow[r] \arrow[d] & \Delta y_i = -0.2 \\
\textcolor{red}{\leftrightarrow} \ \textcolor{red}{\text{Rescaling}} & \Delta x_i = 1 \arrow[r] & \Delta y_i = -2
\end{tikzcd}$$

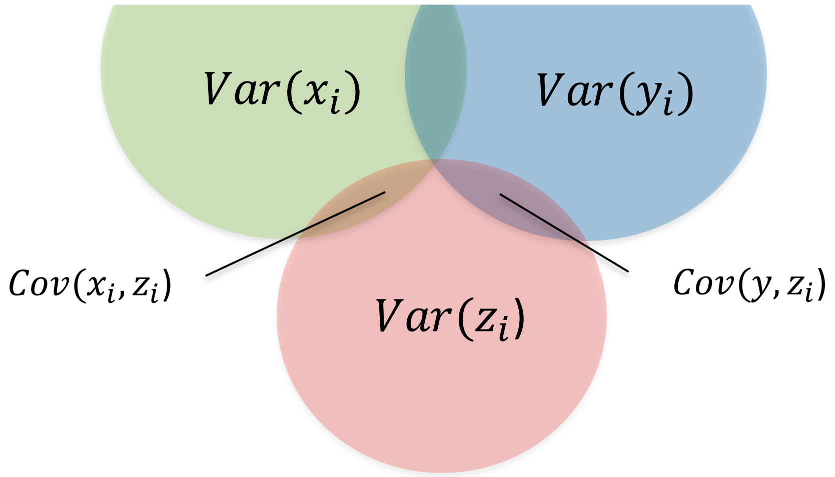
Seeing it algebraically

$$\frac{\Delta y_i}{\Delta z_i} = \frac{-0.2}{1} = -0.2 \qquad \frac{\Delta x}{\Delta z} = \frac{0.1}{1} = 0.1$$

$$\frac{\Delta y_i / \Delta z_i}{\Delta x_i / \Delta z_i} = \frac{\Delta y_i}{\Delta x_i} = \frac{-0.2}{0.1} = -2$$

Relation to sample moments

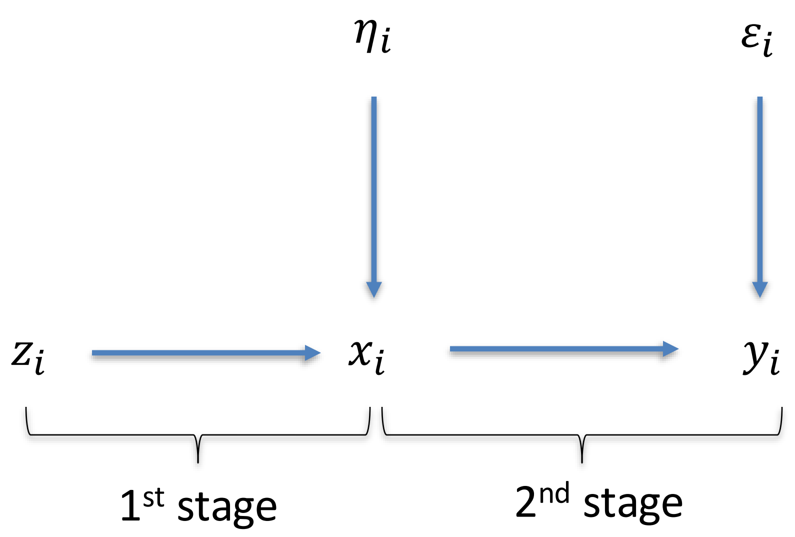
$$Simple\ IV\ estimator = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$$



Only valid given assumptions on structure!

Two stage least squares (2SLS) intuition

Two stage modelling



where:

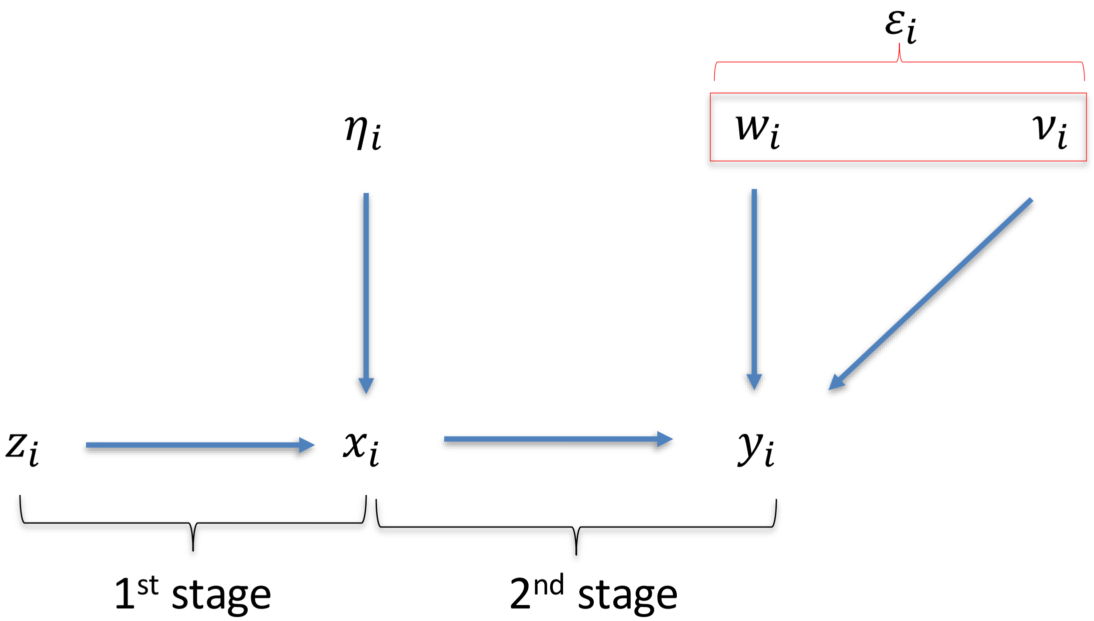
- η_i = first stage error term
- ε_i = first stage error term

Valid instrument requirements

From this we can see two requirements for valid instruments:

1. Instrument relevance. The instrument needs to be (strongly) related to the endogenous regressor, $E[z_i x_i] \neq 0$.
2. Instrument exogeneity. The instrument needs to be unrelated to the (second-stage) error term, $E[z_i \varepsilon_i] = 0$.

Instrument relevance: what does $E[z_i x_i] \neq 0$ mean?



Strong relation between z_i and x_i

This condition is intuitive:

- For the effects of z_i to be noticeable in meaningful ways, it needs to have a strong impact on x_i
- If it does not, attributing the correlation between z_i and y_i to that between z_i and x_i is problematic
- This can lead to misleading inferences!

Exclusion restriction and what does $E[z_i\varepsilon_i] = 0$ mean?

Assuming coefficient is zero

Consider the simple regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

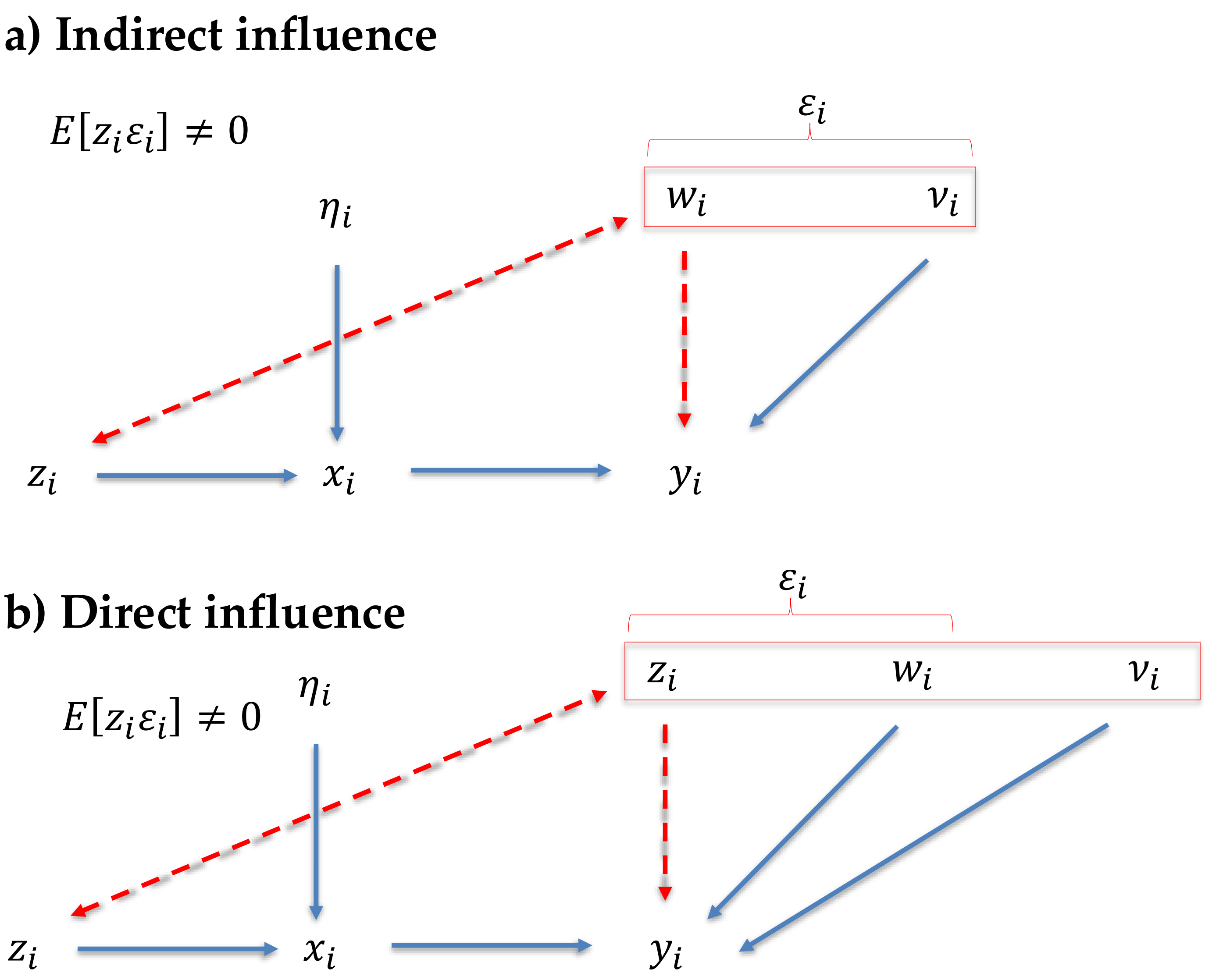
where $\varepsilon_i = \delta z_i + v_i$. Exclusion restriction equals $\delta = 0$.

The resulting model is then

$$y_i = \alpha + \beta x_i + v_i$$

The instrumental variable z_i drops out of this regression equation in line with the assumption that the only way it influences y_i is through the effect it has on x_i.

Violations of the exclusion restriction are one source of why instrument exogeneity may not hold ($E[z_i\varepsilon_i] \neq 0$).



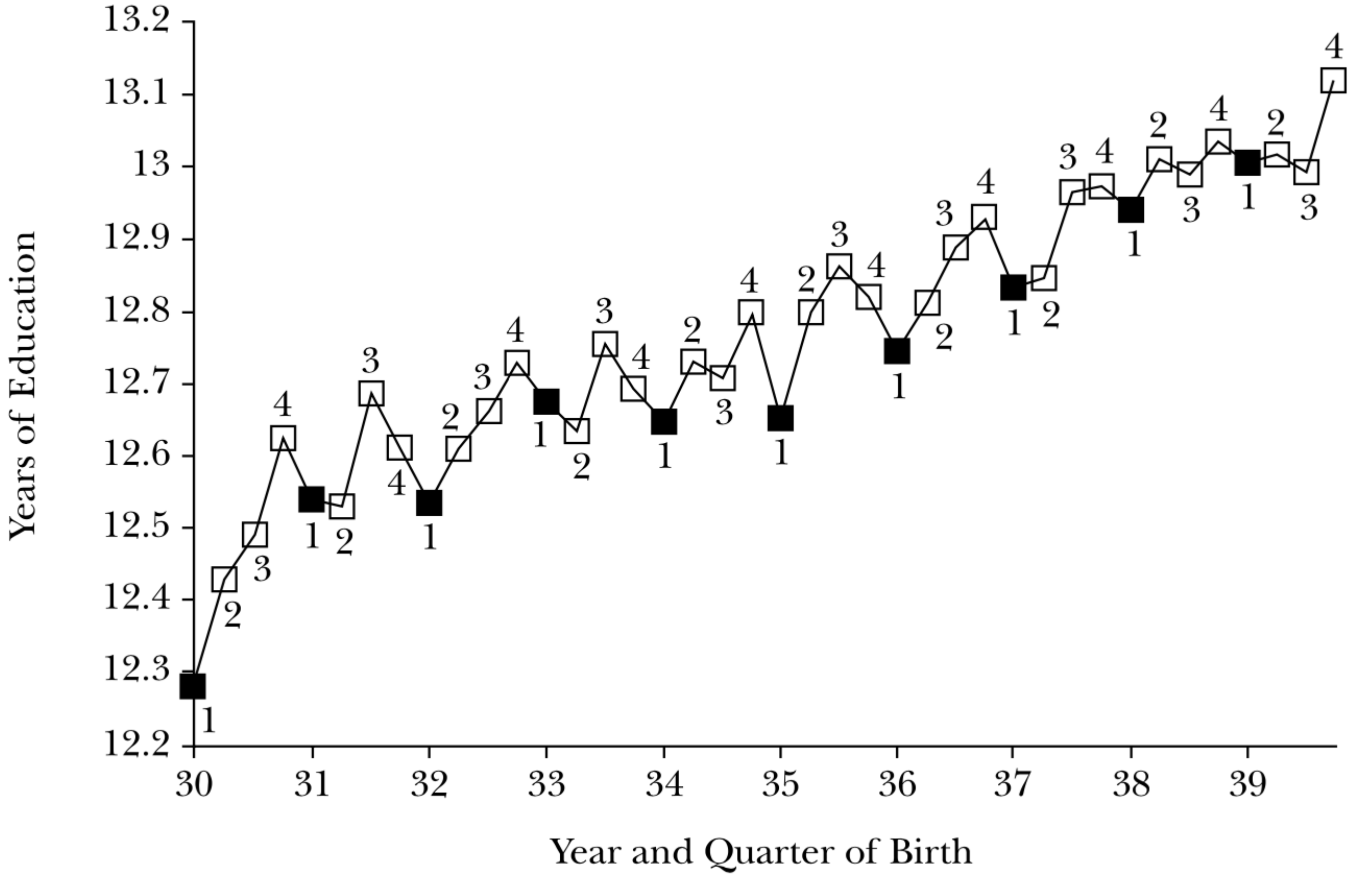
a) Indirect influence

b) Direct influence

Angrist and Krueger (2001, JEP)

Figure 1

Mean Years of Completed Education, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

School start policy

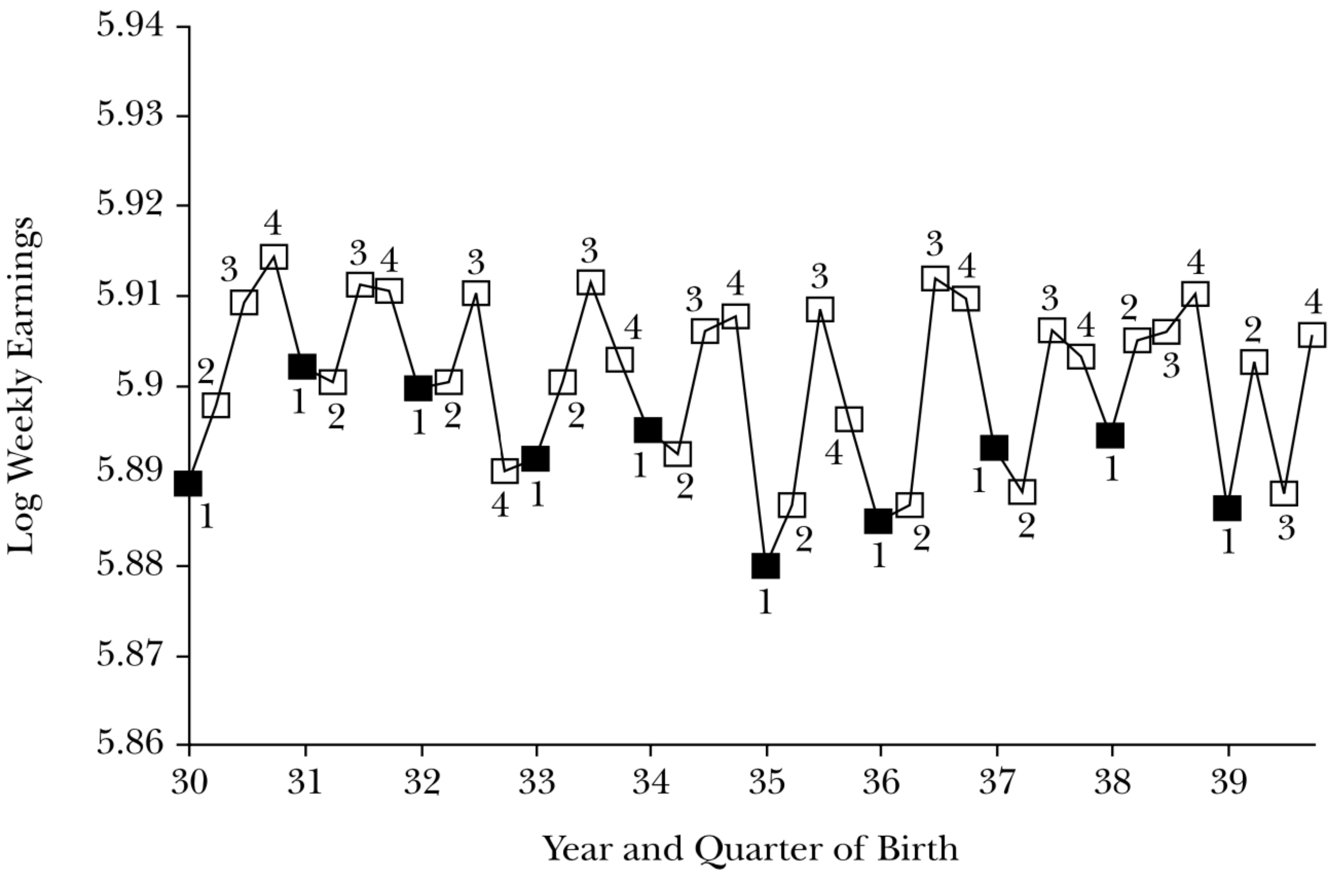
- Start school in April of year when you turn 6
- Those born in first quarter (January – March) will be oldest
- Oldest school starters get less schooling

$$Simple\ IV\ estimator = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$$

Angrist and Krueger (2001, JEP)

Figure 2

Mean Log Weekly Earnings, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

School start policy

- Start school in April of year when you turn 6
- Those born in first quarter (January – March) will be oldest
- <u>Oldest school starters earn less wages</u>

$$Simple\ IV\ estimator = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$$

Local Average Treatment Effect (LATE)

IV estimates are valid for "compliers"

- In IV estimation we exploit variability in the endogenous regressor induced by the instrument
- But not everyone's behaviour will be affected by the instrument
- In previous example: *"the quarter-of-birth instrument is most relevant for those who are at high probability of quitting school as soon as possible, with little or no effect on those who are likely to proceed on to college."* (Angrist & Krueger, 2001)

Person index	Counterfactual outcomes		Difference
	Outcome without IV treatment	Outcome with IV treatment	
1	2	2	0
2	2	3	1
3	3	4	1
4	1	1	0

Generalizing from IV estimates

- Treatment with instrumental variable was only relevant for person 2 and 3
- These are "compliers" – those for which the IV was influential
- Inferences hold for population of compliers

Efficient estimation techniques

2SLS computational illustration

Notation

- y_i = dependent variable
- x'_{1i} = vector of K exogenous regressors
- x_{2i} = endogenous regressor
- z_i = instrumental variable
- v_i = homoskedastic first stage error
- ε_i = homoskedastic second stage error

Key assumptions

- Regressor exogeneity: $E[x'_{1i}\varepsilon_i] = 0$
- Regressor endogeneity: $E[x_{2i}\varepsilon_i] \neq 0$
- Instrument exogeneity: $E[z_i\varepsilon_i] = 0$
- Instrument relevance: $E[z_i x_{2i}] \neq 0$

Possible estimation approach

Assume the following system of equations:

$$\begin{aligned} y_i &= x'_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \; (2nd\ stage) \\ x_{2i} &= x'_{1i}\pi_1 + z_i\pi_2 + v_i \quad (1st\ stage) \end{aligned}$$

We can estimate the first stage and obtain:

$$\hat{x}_{2i} = x'_{1i}\hat{\pi}_1 + z_i\hat{\pi}_2$$

Then we substitute $\hat{x}_{2i}$ for x_{2i} in the second stage and regress y_i on $\hat{x}_{2i}$ and x'_{1i}.

In this way we obtain the 2SLS estimate of β_2, which we denote by $\hat{\beta}_{2,2SLS}$.

Standard error adjustment

Correct standard errors for $\hat{\beta}_{2,2SLS}$ require a small adjustment resulting from sampling error of $\hat{\pi}_1$ and $\hat{\pi}_2$ in the first stage estimation.

2SLS identification

Exploiting moment conditions

In the first half of the lecture, we have built intuition that exogeneity is key for identification. First, note that the coefficients for the K exogenous regressors are identified from their moment conditions:

$$E[x'_{1i}\varepsilon_i] = E[x'_{1i}(y_i - x'_{1i}\beta_1 - x_{2i}\beta_2)] = 0$$

Or, equivalently, for each $k = 1, \dots, K$:

$$E[x_{ik}\varepsilon_i] = E[x_{ik}(y_i - x'_{1i}\beta_1 - x_{2i}\beta_2)] = 0$$

But we cannot stop here; our model is otherwise *under identified*.

Exact identification

Identifies necessitates at least one moment condition for each parameter. We therefore require one further moment condition to identify β_2. We know we cannot use the one associated with the endogenous regressor ($E[x_{2i}\varepsilon_i] \neq 0$), so instead we use that resulting from our instrumental variable:

$$E[z_i\varepsilon_i] = E[x_i(y_i - x'_{1i}\beta_1 - x_{2i}\beta_2)] = 0$$

These $K + 1$ moment conditions *exactly identify* the $K + 1$ coefficients of interest.

Generalized IV estimator

Tweaked, more general model setup

Suppose we have the following model:

$$y_i = x_{1i}'\beta_1 + x_2'\beta_2 + \varepsilon_i = x_i'\beta + \varepsilon_i$$

where:

- $x_i' = (x_{1i}', x_{2i}')$ is a $K \times 1$ vector of explanatory variables
- $\beta =$ conformable the $K \times 1$ coefficient vector
- x_{1i}' collects the exogenous variables
- x_{2i}' collects the endogenous variables

Definition of z_i'

Further we have the $R \times 1$ vector z_i' of <u>*all*</u> exogenous variables, including our instrumental variables.

Expression for $\boldsymbol{\beta_{IV}}$ (not a proof)

Pre-multiply both sides of $y_i = x_i'\beta + \varepsilon_i$ by z_i:

$$z_i y_i = z_i x_i'\beta + z_i \varepsilon_i$$

Then taking expectations leads to:

$$E[z_i y_i] = E[z_i x_i']\beta + E[z_i \varepsilon_i] = E[z_i x_i']\beta$$

given that each instrument is exogenous. Rearranging leads to:

$$\beta_{IV} = E[z_i x_i']^{-1} E[z_i y_i]$$

Single endogenous variable and instrument

In this case β_{IV} it is easy to see that:

$$\beta_{IV} = \frac{E[z_i y_i]}{E[z_i x_i]}$$

Ratio of expectations need not exist: IV *estimator is biased in small samples* and relies on consistency. (Remember simulation!)

Relative bias and weak instruments

Relative bias

From the preceding discussions we know:

- OLS is biased in presence of endogeneity
- 2SLS is biased in small samples

What now?

An important question arises:

> *"Is the finite sample bias of two-stage least squares smaller than that of ordinary least squares?"* (Murray 2006, p. 124)

This is a function of instrument strength. Weak instruments mean smaller relative bias: strong similarity of 2SLS and OLS estimates.

Special overidentified case

In the case where $R > K$, more instruments than endogenous regressors, and no exogenous variables:

$$\frac{Bias(\beta_{2SLS})}{Bias(\beta_{OLS})} \approx \frac{l}{n\tilde{R}^2}$$

where:

- $l =$ number of instrumental variables
- $n =$ number of observations
- $\tilde{R}^2$ $=$ R-squared in first stage regression

given the normalization assumption that error term variances equal unity.

Popular tests for weak instruments

Simple F test in first stage

Recall from earlier the following system of equations with one endogenous regressor (x_{2i}) and one instrumental variable (z_i):

$$y_i = x'_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \; (2nd\ stage)$$

$$x_{2i} = x'_{1i}\pi_1 + z_i\pi_2 + v_i \quad (1st\ stage)$$

A popular test statistic for weak instruments is the F test of the null hypothesis $H_0 : \pi_2 = 0$ (known as the Cragg-Donald Wald F-statistic).

Nonstandard critical values

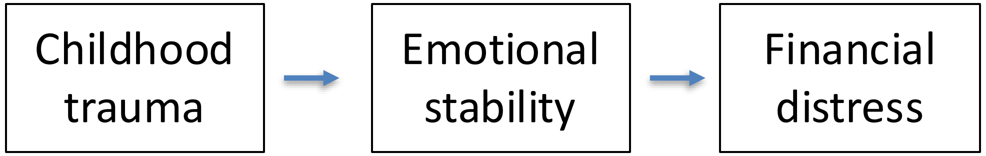
However standard critical values are not appropriate. Instead:

- Stock and Yogo (2005) provide critical values based on the relative bias $Bias(\beta_{IV})/Bias(\beta_{OLS})$ being less than a certain threshold; for example, 5% or 10%
- Staiger and Stock (1997) provide a rule of thumb: if $F > 10$ then no weak instrument concerns

These approaches extend to more general cases!

Parise and Peijnenburg (2019, RFS)

Instrumenting emotional stability with childhood trauma



Noteworthy

- Cragg-Donald Wald F-statistic 618.8!
- Significance of childhood trauma (p<0.01)
- Estimates not sensitive to comprehensive set of additional controls

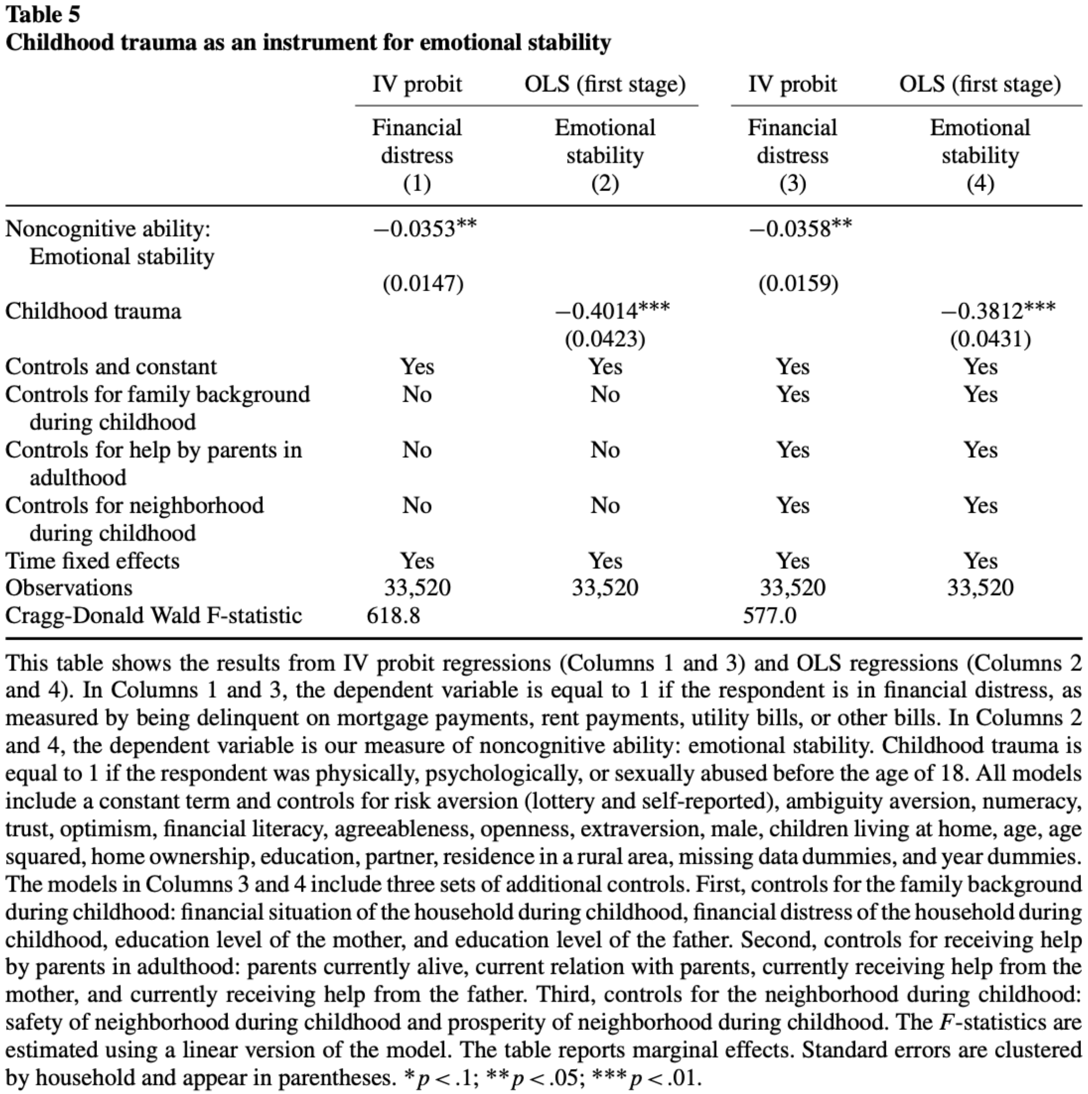


Table 5
Childhood trauma as an instrument for emotional stability

	IV probit	OLS (first stage)	IV probit	OLS (first stage)
	Financial distress (1)	Emotional stability (2)	Financial distress (3)	Emotional stability (4)
Noncognitive ability: Emotional stability	−0.0353**		−0.0358**	
	(0.0147)		(0.0159)	
Childhood trauma		−0.4014***		−0.3812***
		(0.0423)		(0.0431)
Controls and constant	Yes	Yes	Yes	Yes
Controls for family background during childhood	No	No	Yes	Yes
Controls for help by parents in adulthood	No	No	Yes	Yes
Controls for neighborhood during childhood	No	No	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
Observations	33,520	33,520	33,520	33,520
Cragg-Donald Wald F-statistic	618.8		577.0	

This table shows the results from IV probit regressions (Columns 1 and 3) and OLS regressions (Columns 2 and 4). In Columns 1 and 3, the dependent variable is equal to 1 if the respondent is in financial distress, as measured by being delinquent on mortgage payments, rent payments, utility bills, or other bills. In Columns 2 and 4, the dependent variable is our measure of noncognitive ability: emotional stability. Childhood trauma is equal to 1 if the respondent was physically, psychologically, or sexually abused before the age of 18. All models include a constant term and controls for risk aversion (lottery and self-reported), ambiguity aversion, numeracy, trust, optimism, financial literacy, agreeableness, openness, extraversion, male, children living at home, age, age squared, home ownership, education, partner, residence in a rural area, missing data dummies, and year dummies. The models in Columns 3 and 4 include three sets of additional controls. First, controls for the family background during childhood: financial situation of the household during childhood, financial distress of the household during childhood, education level of the mother, and education level of the father. Second, controls for receiving help by parents in adulthood: parents currently alive, current relation with parents, currently receiving help from the mother, and currently receiving help from the father. Third, controls for the neighborhood during childhood: safety of neighborhood during childhood and prosperity of neighborhood during childhood. The *F*-statistics are estimated using a linear version of the model. The table reports marginal effects. Standard errors are clustered by household and appear in parentheses. $*p<.1$; $**p<.05$; $***p<.01$.

Generalized Method of Moments (GMM) estimation

The overidentified case

Assume the simple model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where $E[x_i \varepsilon_i] \neq 0$ and two instruments are available: the model is *over identified*.

To identify β in $E[x_i \varepsilon_i] \neq 0$ we employ the following moment conditions arising from the two valid instruments $z_{1,i}$ and $z_{2,i}$:

$$E\left[z_{1,i}\varepsilon_i\right] = 0$$
$$E\left[z_{1,i}\varepsilon_i\right] = 0$$

Can this extra information from the second instrument yield useful insights? GMM provides the appropriate framework.

A sketch of Feasible Efficient GMM

The corresponding sample moments from our instruments for $l = 1, 2$ and n observations are

$$\frac{1}{n}\sum_{i=1}^{n} z_{l,i}\, \hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n} z_{l,i}\,(y_i - \hat{\beta} x_i)$$

These are collected in the column vector $\bar{g}_n(\hat{\beta})$.

The GMM objective function equals

$$J_n(\hat{\beta}) = n\bar{g}_n(\hat{\beta})'\hat{S}^{-1}\bar{g}_n(\hat{\beta})$$

where $\hat{S}^{-1}$ is the optimal GMM weighting matrix.

The Feasible Efficient GMM estimator $\hat{\beta}_{FEGMM}$ results from minimizing the objective function.

Overidentifying restrictions

Overidentifying restrictions test

The value of the minimized objective function $J_n(\hat{\beta}_{FEGMM})$ provides a test statistic for the null hypothesis that all moment conditions are jointly met.

The critical values come from a χ^2distribution with $L - K$ degrees of freedom (L moment conditions and K coefficients to estimate).

The number of overidentifying restrictions equals $L - K$ and therefore this is an overidentifying restrictions test.

Rejection of the null hypothesis provides evidence *against* instrument validity.

Stata implementation

The command -ivreg2- is the major instrumental variable estimation command in Stata. It does 2SLS, FEGMM and more.

On your own computer, you can install it by executing

ssc install ivreg2, replace

Make sure to have a look at the accompanying papers in the Stata Journal as well as the documentation in -help ivreg2-.

Illustration with R and Stata

Omitted variable bias simulation: OLS vs 2SLS

Further reading

Indicative sections in textbooks

Verbeek (4e) Sections 5.3-5.6

Wooldridge (6e) Sections 15-1, 15-2, 15-3, 15-4, 15-5

Further resources

ANGRIST, J. D. & KRUEGER, A. B. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives,* 15**,** 69-85.

MURRAY, M. P. 2006. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives,* 20**,** 111-132.

PARISE, G. & PEIJNENBURG, K. 2019. Noncognitive Abilities and Financial Distress: Evidence from a Representative Household Panel. *The Review of Financial Studies,* 32**,** 3884-3919.

Further reading

Two key references that develop the Stata command -ivreg2-

BAUM, C. F., SCHAFFER, M. E. & STILLMAN, S. 2003. Instrumental variables and GMM: Estimation and testing. *The Stata Journal,* 3**,** 1-31.

BAUM, C. F., SCHAFFER, M. E. & STILLMAN, S. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal,* 7**,** 465-506.

Even more

CRAGG, J. G. & DONALD, S. G. 1993. Testing identifiability and specification in instrumental variable models. *Econometric Theory***,** 222-240.

JIANG, W. 2017. Have Instrumental Variables Brought Us Closer to the Truth. *The Review of Corporate Finance Studies,* 6**,** 127-140.

STAIGER, D. & STOCK, J. H. 1997. Instrumental variables regression with weak instruments. *Econometrica***,** 557-586.

STOCK, J. H. & YOGO, M. 2005. Testing for weak instruments in linear IV regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg,* 80**,** 1.