Module 4: Instrumental Variables

Econometrics II

Max Heinze (mheinze@wu.ac.at)
Department of Economics, WU Vienna

Sannah Tijani (stijani@wu.ac.at)
Department of Economics, WU Vienna

December 4, 2025

What are Instrumental Variables

Two Stage Least Squares

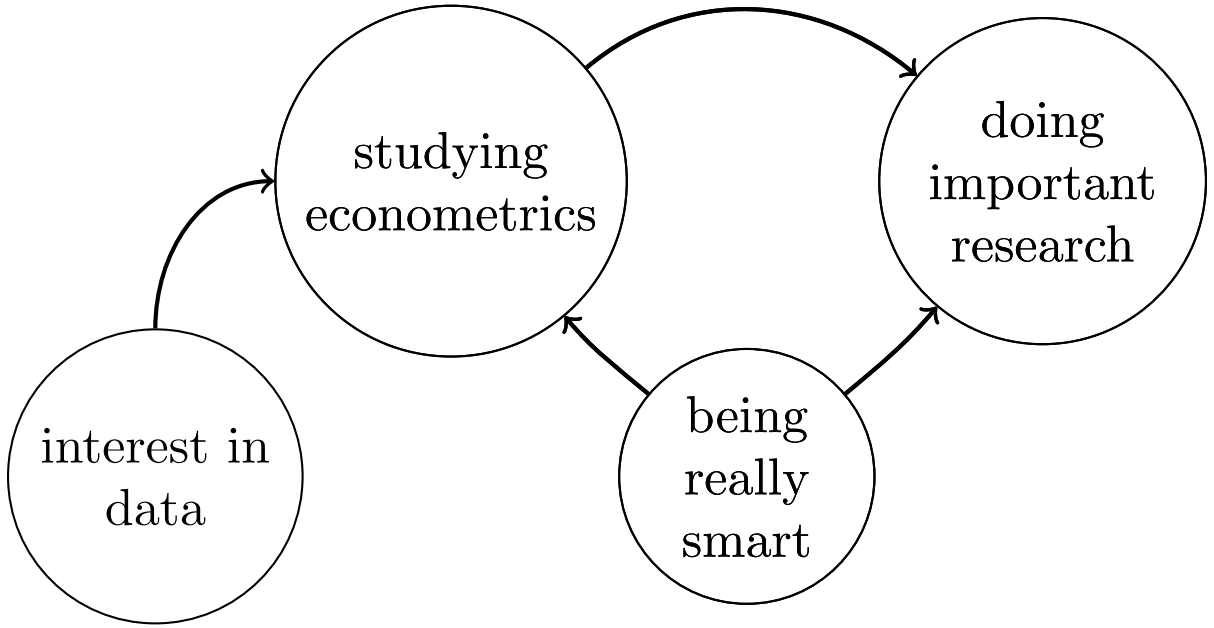
An IV Example

Weak Instruments

Endogeneity: Is all Hope Lost?

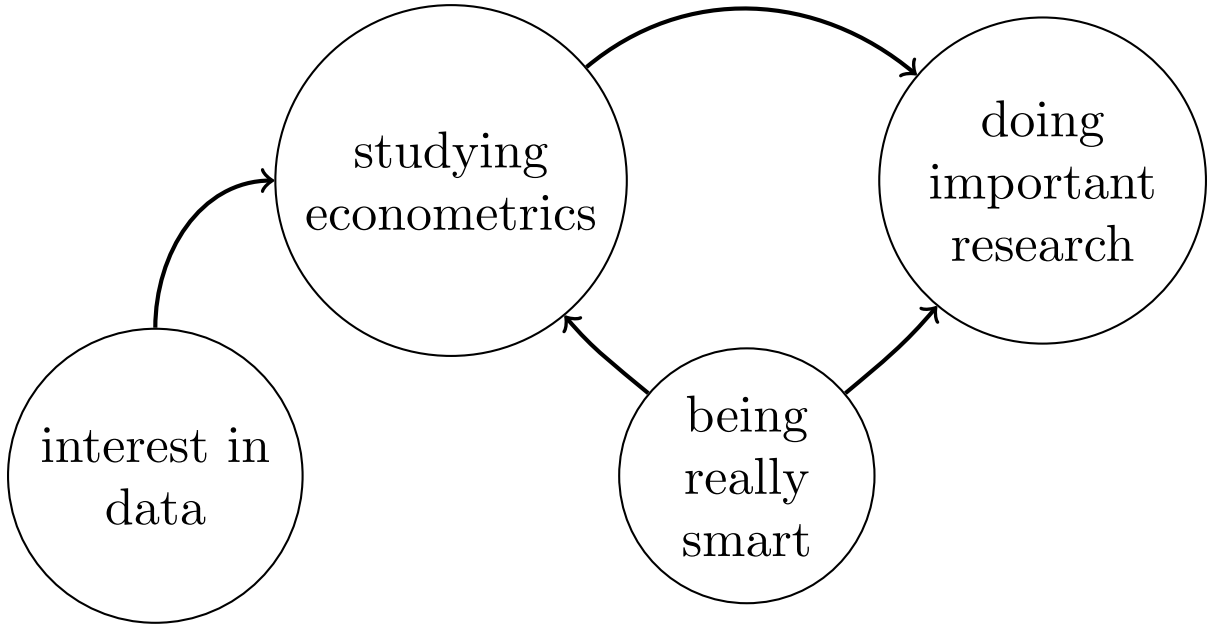
We have discussed why **confounders** are a source of **endogeneity**.

Imagine the confounder is unobserved. How can we fend off this **threat to identification**?



The **basic idea** is: If we can find a so-called **instrumental variable** that explains the endogenous regressor, we can use this variable to **circumvent the issue**. We can also use this technique to deal with **other sources of endogeneity**.

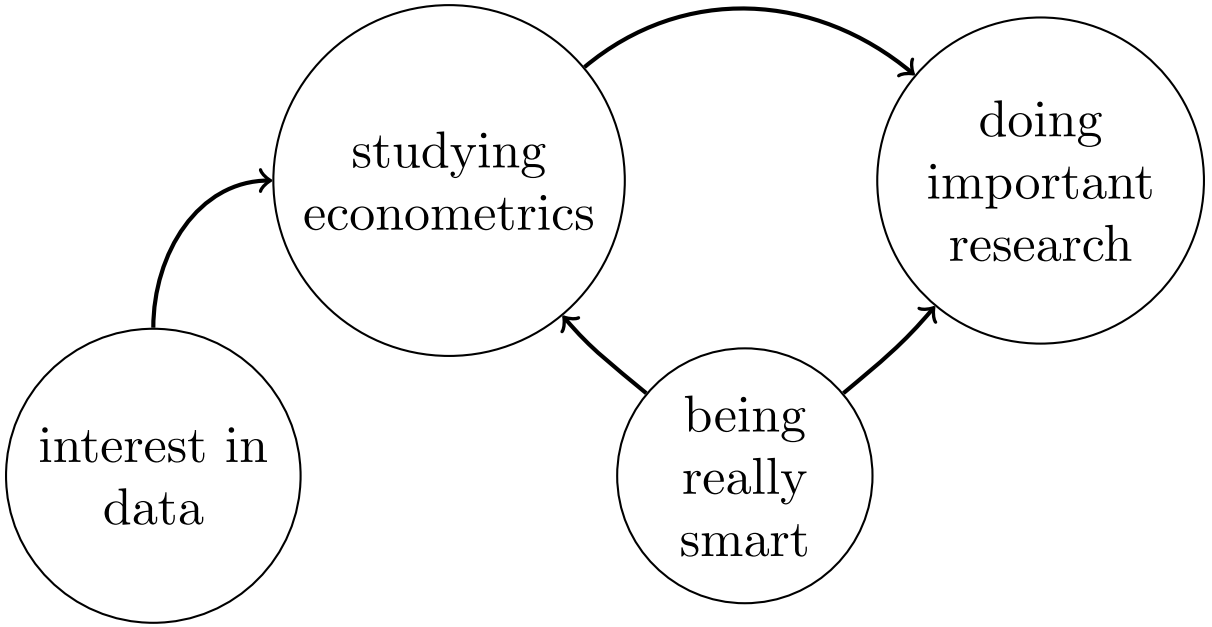
How Does This Work?



Intuitively, we can think of the depicted situation like this:

- We cannot identify the causal effect from studying econometrics on being really smart since they are confounded.
- But if we **find an instrumental variable** that **explains only the endogenous regressor**, we can **isolate** the part of the co-variation that is our **causal effect**.

When Does This Work?



There are **two conditions** our **instrumental variable** must fulfill.

(1) **Relevance Condition**: The instrument must be **correlated with the endogenous regressor**, i.e., it must actually explain this endogenous regressor.

(2) **Exclusion Restriction**: The instrument must **affect the outcome only through the endogenous regressor**.

Instruments, More Formal

Consider the following case where omitting a confounder is the source of **endogeneity**:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}, \\ \boldsymbol{u} &= \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \end{aligned}$$

where $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{S}) \neq 0$, and thus $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{u}) \neq 0$.

- If we can observe $\boldsymbol{S}$, we can easily include it in our regression, estimating $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$.
- If we **do not observe** $\boldsymbol{S}$, we can use an **instrument** $\boldsymbol{Z}$. This instrument has to satisfy
  - the **Relevance Condition**, i.e. $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Z}) \neq 0$, and
  - the **Exclusion Restriction**, i.e. $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{u}) = 0$. Alternatively, this is called the **Exogeneity Condition**.

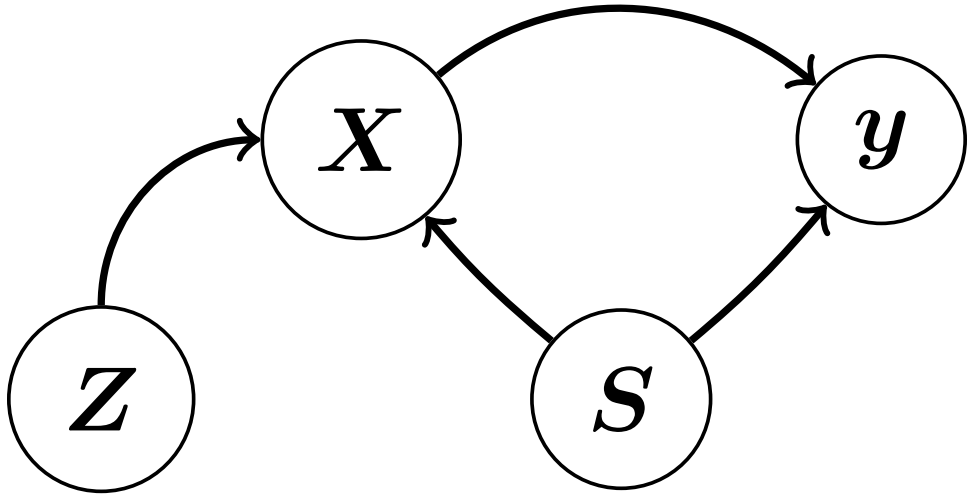
What does "Using an Instrument" Mean?

We can think of the process as containing **two steps**:

(1) In the **First Stage**, we regress the **endogenous regressor** X on the **instrument** Z.

(2) In the **Second Stage**, we take the **predictions** $\hat{X}$ from the first stage and regress the **outcome** y on the **predictions** $\hat{X}$.

This is the intuition behind what we will call the **Two Stage Least Squares** (**2SLS**) estimator.



What are Instrumental Variables

Two Stage Least Squares

An IV Example

Weak Instruments

More Examples

Setup

Consider the following general model:

$$\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{Q} = [\boldsymbol{S}\ \boldsymbol{X}]$, with $\mathrm{Cov}(\boldsymbol{S}, \boldsymbol{u}) = 0$ and $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{u}) \neq 0$, that is,

- $\boldsymbol{S}$ contains L **exogenous regressors**, and
- $\boldsymbol{X}$ contains K **endogenous regressors**.

Assume in addition to that that $\boldsymbol{Z}$ contains M instrumental variables.

If $M \geq K$, we can **identify** the effect of the **endogenous regressors**. In that case, there is at least **one instrument per endogenous regressor**.

- If $M = K$, we call the coefficients **just identified**.
- If $M > K$, we call them **overidentified**.

Estimating 2SLS – First Stage

The concept behind the 2SLS estimator is similar to before. In the **First Stage**, we regress the **endogenous regressors** X on the **exogenous variables** S and the **instruments** Z.

Assume (for simplicity) that there are no exogenous regressors:

$$X = Z\delta + v, \qquad \hat{\delta} = (Z'Z)^{-1}Z'X.$$

Using $\hat{\delta}$, we can now obtain a prediction $\hat{X} = Z\hat{\delta}$ for the next stage.

We can express this in a very simple way using a **projection matrix**:

$$P_Z = Z(Z'Z)^{-1}Z'.$$

The math behind projection matrixes is out of scope for this class, so we just accept that **pre-multiplying** a matrix P_Z of this form **yields** a variable's **predictions**:

$$\hat{X} = P_Z X$$

Two nice **features of projection matrices**, which we need for the following derivations, are:

- **symmetry**, i.e., $P_Z' = P_Z$, and
- **idempotency**, i.e. $P_Z P_Z = P_Z$.

Estimating 2SLS – Second Stage

In the **Second Stage**, we replace the endogenous variables with their prediction $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{P_Z}\boldsymbol{X}$. This allows us to obtain the **2SLS estimator**:

$$
\begin{aligned}
\boldsymbol{y} &= \hat{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{u}, \\
\hat{\boldsymbol{\beta}} &= (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{y} \\
\beta_{2SLS} &= (\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{y}.
\end{aligned}
$$

The covariance matrix of the 2SLS estimator is $\mathrm{Cov}(\beta_{2SLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}$.

The IV Estimator

When the coefficients are **just identified** ($M = K$), the dimensions of $(\boldsymbol{Z}'\boldsymbol{X})^{-1}$ and $\boldsymbol{Z}'\boldsymbol{y}$ match and we can use the **IV estimator**[1].

$$\boldsymbol{\beta}_{IV} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

We can derive it by pre-multiplying $\boldsymbol{Z}'$ in the standard model.

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{Z}'\boldsymbol{y} &= \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{u}\end{aligned}$$

Now, we can impose the moment condition $\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{IV}) = \boldsymbol{0}$, the sample analog of the exogeneity assumption $\mathrm{E}(\boldsymbol{Z}'\boldsymbol{u}) = 0$,

$$\begin{aligned}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}_{IV} &= \boldsymbol{Z}'\boldsymbol{y} \\ \boldsymbol{\beta}_{IV} &= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}.\end{aligned}$$

1. Some people use the term “IV estimator” to describe *any* kind of estimator that uses instrumental variables, we don’t.

The IV Estimator is Consistent, ...

We can easily sketch a proof for **consistency** of the **IV estimator**:

$$\begin{aligned}\boldsymbol{\beta}_{IV} &= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u} \\ &= \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u} \\ &= \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X}N^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{u}N^{-1}\end{aligned}$$

From the **exogeneity** and **relevance** conditions we get

- $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{u}) = 0$, which implies that $\boldsymbol{Z}'\boldsymbol{u}N^{-1} \xrightarrow{p} 0$,
- $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{X}) \neq 0$, which implies that $\boldsymbol{Z}'\boldsymbol{X}N^{-1} \xrightarrow{p} \mathbb{E}[\boldsymbol{Z}'\boldsymbol{X}]$.

Thus[1], $\boldsymbol{\beta}_{IV} \xrightarrow{p} \boldsymbol{\beta} + \frac{0}{c} = \boldsymbol{\beta}$ as $N \to \infty$.

1. This proof relies on the fact that $\operatorname{plim} \frac{a}{b} = \frac{\operatorname{plim} a}{\operatorname{plim} b}$ Note that the analogous statement does not hold for expectations.

... But the IV Estimator Is Also Biased

The **IV estimator** is consistent, but almost certainly **biased** in small samples.

$$\begin{aligned}\boldsymbol{\beta}_{IV} &= \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u}, \\ \mathbb{E}[\boldsymbol{\beta}_{IV}] &= \boldsymbol{\beta} + \mathbb{E}[(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u}].\end{aligned}$$

We cannot separate the second term:

(1) If we conditioned on $\boldsymbol{Z}$, would be stuck with $(\boldsymbol{Z}'\boldsymbol{X})^{-1}$.

(2) If we conditioned on $\boldsymbol{X}$ and $\boldsymbol{Z}$, would have a problem with $\mathbb{E}[\boldsymbol{u}|\boldsymbol{Z}, \boldsymbol{X}]$:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\beta}_{IV}] &= \boldsymbol{\beta} + \mathbb{E}\left[\mathbb{E}\left[(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u} \mid \boldsymbol{Z}, \boldsymbol{X}\right]\right] \\ &= \mathbb{E}\left[(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\mathbb{E}[\boldsymbol{u} \mid \boldsymbol{Z}, \boldsymbol{X}]\right].\end{aligned}$$

What are Instrumental Variables

Two Stage Least Squares

An IV Example

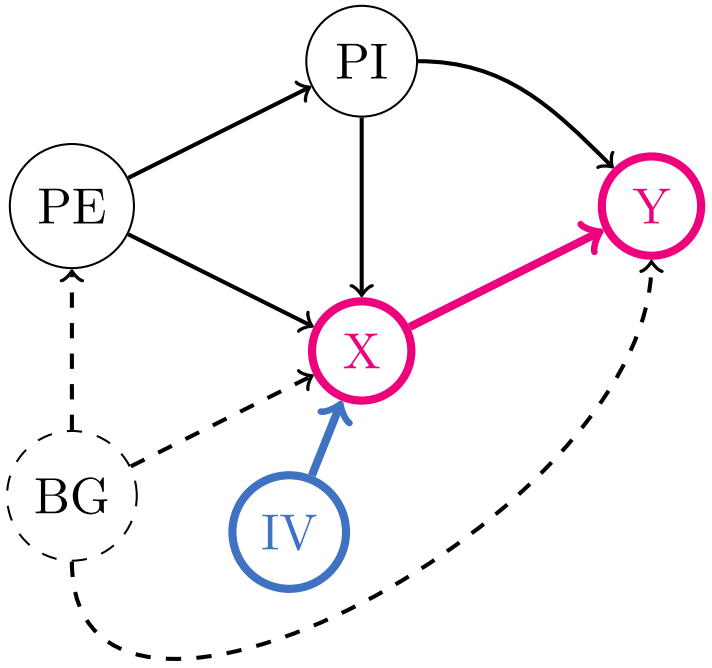
Weak Instruments

More Examples

Shift-Share Instruments

How Do We Use This?

We now know what **instrumental variables** are, and how we can **estimate** β in an instrumental variables setting. We know that instruments must be **exogenous** and **relevant**, and that we need at least one instrument per endogenous variable. Now consider the following example:



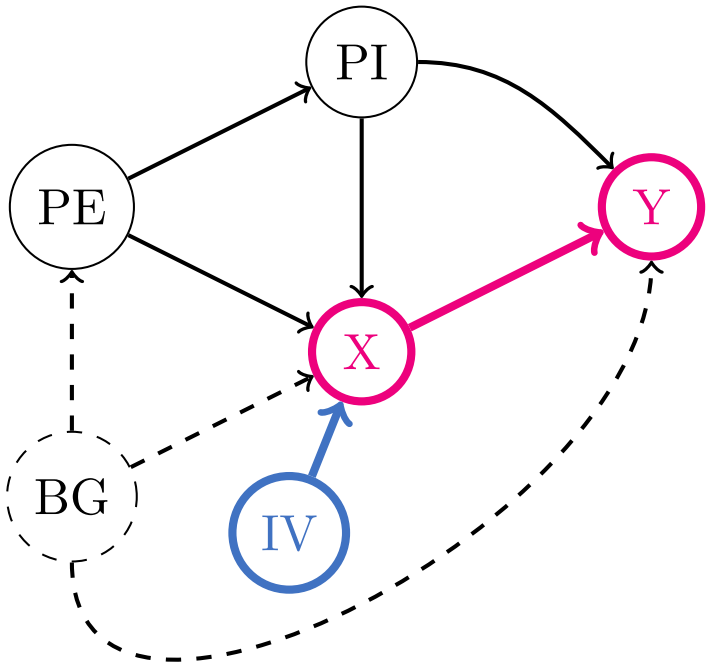
Say we want to find out the effect of **education** X on **income** Y.

But we know that both **parental education** PE and **parental income** PI influence the level of education. We could control for these since we can observe them.

However, there are likely other **background factors** BG that influence parental education, education and income. These background factors are unobserved, meaning we **cannot identify a causal effect**.

Only if we find an **instrument** IV, we can bypass this restriction.

An IV for Education



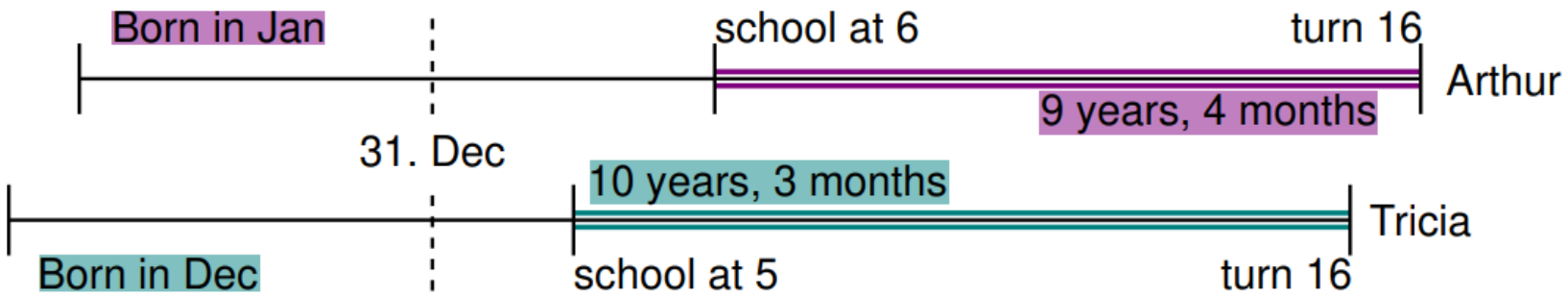
This example is from Angrist & Krueger (1991)[1]. They came up with a novel instrument for education: the **quarter of birth** of a given individual.

How does this work?

In the United States, students must attend school from **the calendar year in which they turn six** until their 16th birthday. School entry is once per year, so the **length of schooling at age 16 differs**, and students who drop out at 16 create variation in education.

1. Also see Angrist & Krueger (2001).

Quarter of Birth as an Instrument



Is this instrument both **exogenous** and **relevant**?

- **Exogeneity** implies that the quarter of birth does not directly affect income. There is no statistical test for this, and so whether you believe this depends on how much you trust the authors' argument.
- **Relevance** is easier to investigate quantitatively. Angrist & Krueger (1991) show that men born earlier in the year tend to have lower education on average.

There is a rule of thumb that a **good instrument** must seem **ridiculous**, since it is then likely fulfilling the exclusion restriction.

Figures from Angrist & Krueger (1991)

The length of **completed education** shows a clear cyclical pattern when plotted against the quarter and year of birth.

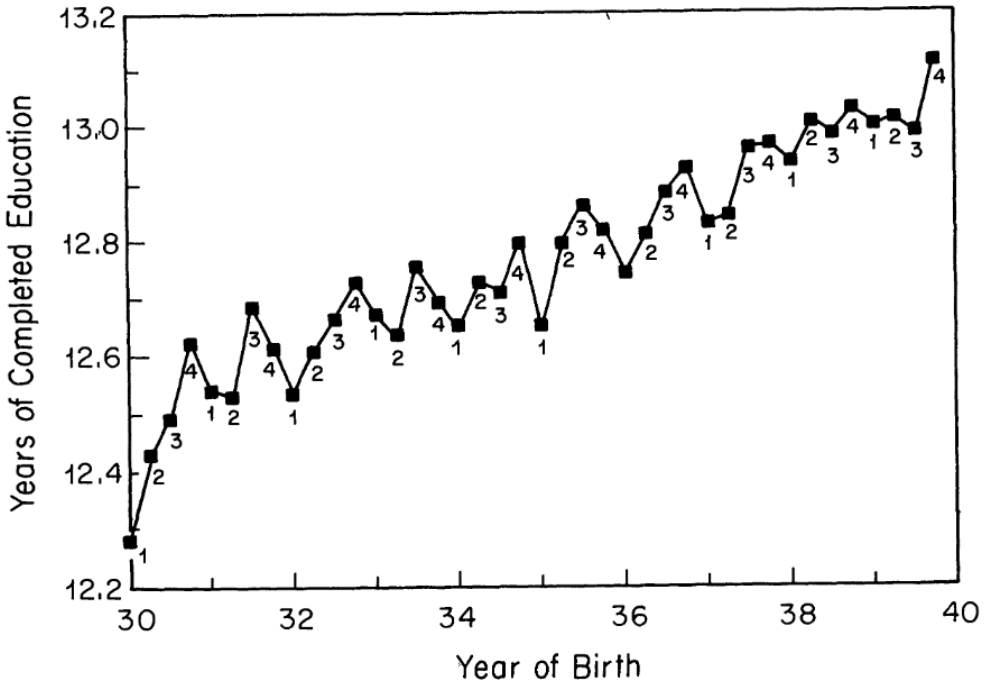


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Log weekly earnings also show a cyclical pattern.

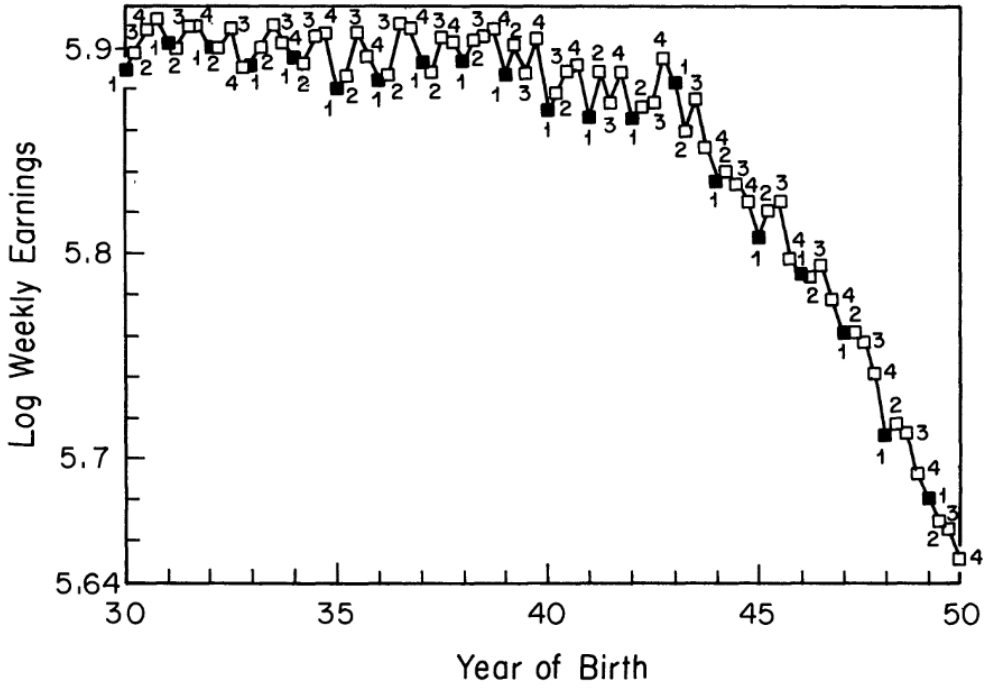


FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

Let's Replicate This

R Code | Start Over | Run Code

```r
library(AER)
library(dplyr)
library(readr)
library(stargazer)

df <- read_csv("https://maxheinze.eu/assets/angrist1991.csv", show_col_types = FALSE)

yr_regs <- paste(paste0("YR", 1930:1938), collapse = " + ")
insts   <- c(unlist(lapply(1:3, function(q) paste0("QTR", q, "_", 1930:1939))),
             paste0("YR", 1930:1938))
inst_formula <- paste(insts, collapse = " + ")

f1 <- as.formula(paste("LWKLYWGE ~ EDUC +", yr_regs))
f2 <- as.formula(paste("LWKLYWGE ~ EDUC +", yr_regs, "|", yr_regs, "+", inst_formula))

ols <- lm(f1, data = df)
iv  <- ivreg(f2, data = df) # ivreg() from the AER package
```

Which Formula Do We Actually Use?

R Code | Start Over | Run Code

```
1 print("f1")
2 f1
3 print("f2")
4 f2
```

Hope We Get the Same Results

R Code | Start Over | Run Code

```
stargazer(ols, iv, type = "text", keep = "EDUC", dep.var.labels = "log(weekly wage)",
          digits = 4, omit.stat = c("f","ser","adj.rsq"),
          title = "Angrist & Krueger (1991) – Table V – Columns (1) and (2)"
)
```

Choosing Between IV and OLS

Angrist & Krueger (1991) — Table V — Columns (1) and (2)

	Dependent variable:	
	log(weekly wage)	
	OLS	*instrumental variable*
	(1)	(2)
EDUC	0.0711***	0.0891***
	(0.0003)	(0.0161)
Observations	329,509	329,509
R^2	0.1177	0.1102
Note:	*p<0.1; **p<0.05; ***p<0.01	

- The estimates of the **OLS** and **IV** specifications are similar.
- If the instrument works as intended, we find that **omitted variable bias** is **limited** and **reduces** the effect.
- If we hypothesize that the only omitted variable is **ability**, we would expect positive bias instead.
- But is the use of IV regression actually **justified** in this case?
  - **2SLS** is **consistent**, if instruments are exogenous and relevant.
  - **OLS** is more **efficient**, but only consistent in the absence of endogeneity.

Durbin-Wu-Hausman Test

In the **absence of endogeneity**, we **prefer OLS**. So it makes sense to have a **test for endogeneity**.

The **Durbin-Wu-Hausman Test** compares an **consistent estimator** to a more **efficient, potentially inconsistent estimator** by following these three steps:

(1) Use the **residuals of the first stage** as explanatories in the **regular model**.

(2) **Test** whether this variable is **relevant** ($H_0 : \beta_j = 0$).

(3) **Rejecting** the null hypothesis means **rejecting exogeneity** of that explanatory variable and thus **rejecting consistency of OLS**.

Using this test, we can justify **using IV regression**, but we **cannot assess** the **quality** of our **instruments**.

What are Instrumental Variables

Two Stage Least Squares

An IV Example

Weak Instruments

More Examples

Shift-Share Instruments

Appendix

Relevance of Instruments

How do we know whether our **instruments** are **good**? Recall that

$$\hat{\boldsymbol{\beta}}_{IV} = \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u},$$

where $(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u}$ should disappear as $N \to \infty$.

- If our **instruments** have **little relevance**, then $\boldsymbol{Z}'\boldsymbol{X}$ will be small. That means that the term will disappear more slowly.
- If our **instruments** have **no relevance** at all, then $\boldsymbol{Z}'\boldsymbol{X}$ will be zero, which is bad because we cannot invert zero.

In such a case of **weak instruments**, we run into multiple problems:

- **Inconsistency** from small violations of the exogeneity condition is magnified,
- The small-sample **bias** of the 2SLS estimator is large, and
- **Confidence intervals** will be inaccurate.

Checking for Weak Instruments

One approach to find out whether instruments are weak is to check their explanatory power using an **F-Test**.

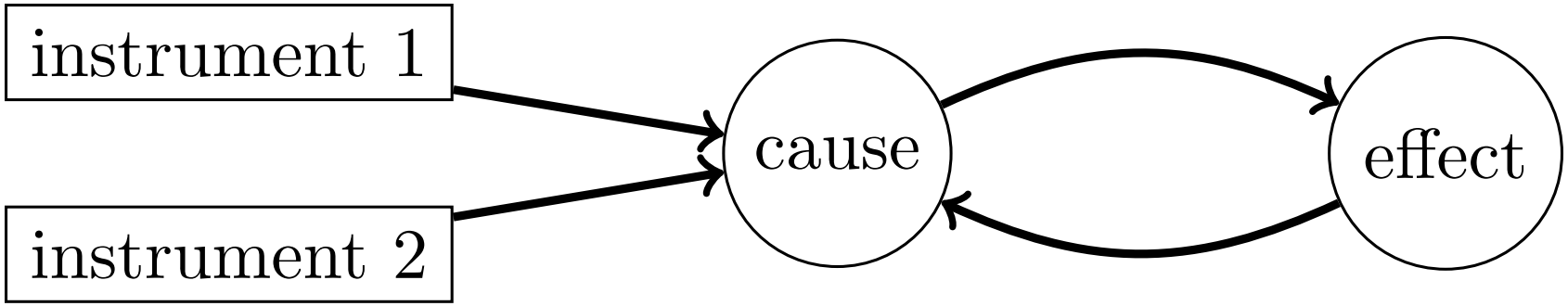
- A frequently used **rule-of-thumb cutoff** in settings with a single endogenous regressor and a usual number of instruments is a first-stage F-statistic of $F = 10$. If the F-statistic is below that, instruments are considered weak.
- Settings with multiple endogenous regressors, or with heteroskedastic errors, require different tests and different critical values.

If instruments are weak, we can e.g. use Anderson-Rubin Confidence Sets (Anderson & Rubin, 1949), which are robust to weak instruments.

Andrews et al. (2019) provide a good review of weak instruments and how to respond to them.

Overidentification

If we have more instruments than endogenous regressors, we have **overidentification**.



In a setting of overidentification, we can use Sargan's J-test to assess exogeneity of our instruments. The idea is to *compare estimates* using different instruments:

- If instruments are exogenous, **estimates** should be **the same**.
- The test's **null hypothesis** is that all instruments are exogenous.

Unfortunately, we do not learn which instrument is not valid, and estimates could always be similar or different by chance.

Recap: Quarter of Birth as Instrument

Angrist & Krueger (1991) — Table V — Columns (1) and (2)

	Dependent variable:	
	log(weekly wage)	
	OLS	*instrumental variable*
	(1)	(2)
EDUC	0.0711***	0.0891***
	(0.0003)	(0.0161)
Observations	329,509	329,509
R^2	0.1177	0.1102
Note:	*p<0.1; **p<0.05; ***p<0.01	

- Last week, we discussed the paper by Angrist & Krueger (1991), in which the authors use **quarter of birth** to instrument for **education**.
- We discussed, and replicated, Columns 1 and 2, the simplest specifications from the output table.
- These columns use no controls, they do use fixed effects, and they use quarter of birth × birth year interactions as instruments.
- This yields a total of 30 instruments, and they include other specifications in the paper that contain up to 180 instruments.
- Do we run into a **weak instruments** problem here?

Let's Run Last Week's Code Again

R Code | Start Over | Run Code

```r
library(AER)
library(dplyr)
library(readr)
library(stargazer)

df <- read_csv("https://maxheinze.eu/assets/angrist1991.csv", show_col_types = FALSE)

yr_regs <- paste(paste0("YR", 1930:1938), collapse = " + ")
insts   <- c(unlist(lapply(1:3, function(q) paste0("QTR", q, "_", 1930:1939))),
             paste0("YR", 1930:1938))
inst_formula <- paste(insts, collapse = " + ")

f1 <- as.formula(paste("LWKLYWGE ~ EDUC +", yr_regs))
f2 <- as.formula(paste("LWKLYWGE ~ EDUC +", yr_regs, "|", yr_regs, "+", inst_formula))

ols <- lm(f1, data = df)
iv  <- ivreg(f2, data = df) # ivreg() from the AER package
```

Is Quarter of Birth a Weak Instrument?

R Code | Start Over | Run Code

```
cat("\nFirst-stage F-test (weak instruments test):\n")
print(summary(iv, diagnostics = TRUE)$diagnostics["Weak instruments", , drop = FALSE])
```

We get an F-statistic of about 4.9 for the case with 30 instruments, which is **much lower** than the rule-of-thumb cutoff of 10, pointing to that **instruments are weak**.

Bound et al. (1995) concur with the result of this assessment and go a step further: They randomly generate an **irrelevant instrument** and show that it leads to similar results.

Is Quarter of Birth Exogenous?

R Code | Start Over | Run Code

```
1 cat("\nSargan's J test (overidentification test):\n")
2 diag_iv <- summary(iv, diagnostics = TRUE)$diagnostics
3 print(diag_iv[grep("Sargan", rownames(diag_iv)), , drop = FALSE])
```

We get a J-statistic of about 25.4, which **does not indicate** a violation of exogeneity.

Even so, Buckles & Hungerman (2013) argue that **exogeneity may be violated** because there is seasonality in mothers' characteristics. On average, women that give birh in winter are younger, less educated, and less likely to be married; which may affect the income of their children.

Two Stage Least Squares

An IV Example

Weak Instruments

More Examples

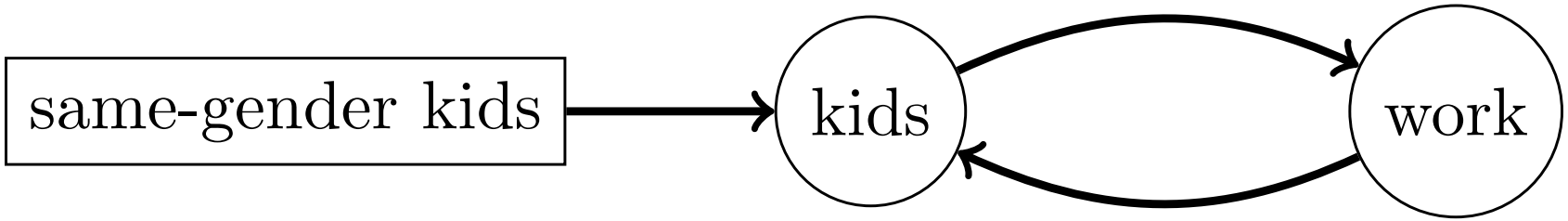
Shift-Share Instruments

Appendix

Family Size and Female Labor

Say we want to learn about the way **family size** affects the **labour supply** of women — e.g. to better understand discrimination, or to design policies for more equality.

- Women with **more children** tend to **work less** (outside the home).
- This is **unlikely to be exogenous** since children are not randomly assigned.



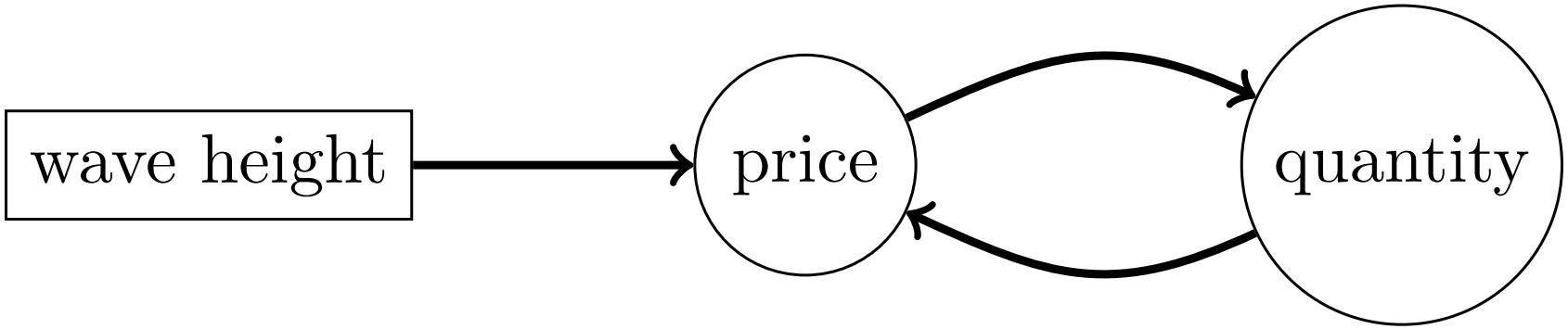
Now consider the fact that mothers whose **first two children are of the same gender** work fewer hours than others. **How is this related to labour supply**?

It **probably is not** related to labor supply. But it may be *related to family size* since parents may have a preference for mixed genders and choose to have a third kid.

Fish Market

Suppose we want to understand how the **price of fish** affects the **quantity sold** at a fish market.

- This is a **simultaneity** issue: Price and quantity are determined simultaneously by supply and demand.



However, on days after a period with especially **high waves**, **prices** on the fish market are usually **higher**. **How and why**?

When waves are high, it is more **difficult to fish**, which means that the quantity sold at the fish market will be lower. Note, however, that we need to rely on the assumption that the kind of fish caught is not affected by waves.

An IV Example

Weak Instruments

More Examples

Shift-Share Instruments

Appendix

Shift-Share Instruments (Bartik Instruments)

Shift-Share Instruments, or **Bartik Instruments** after Bartik (1991), are instruments that use a **national-level shock** (the **shift**) in combination with local **shares** to instrument for a local shock.

Say we are interested in how **immigration** im in some municipality m affects **wages** y in that place (with t being a time index and $\boldsymbol{x}$ being a vector of controls):

$$y_{mt} = im_{mt}\beta + \boldsymbol{x}'\boldsymbol{\gamma} + u_{mt}.$$

The problem with this is that while immigration affects wages, wages likely affect immigration as well. However, **national immigration** changes are credibly exogenous to local wage changes. We can thus use **national-level immigration figures from different countries of origin** as the **shifts**, and initial (at $t = 0$) **shares** of different immigrant nationalities $q = 1, \ldots, Q$ in the place to construct the **Bartik Instrument**:

$$B_{mt} = \sum_{q=1}^{Q} \text{share}_{mq,t=0} \times \text{shift}_{qt}$$

Shifts or Shares?

$$B_{mt} = \sum_{q=1}^{Q} \text{share}_{mq,t=0} \times \text{shift}_{qt}$$

Once we have constructed this instrument, we can use it like any other instrument.

There are **two perspectives** about what is needed for identification:

- **The Shares Perspective**: Following Goldsmith-Pinkham et al. (2020), the **initial shares** provide the **exogenous variation**. Having exogenous **shares** is sufficient for identification. In the previous example, this would mean that the researcher would need to argue for that initial shares of migrants are unrelated to local incomes.
- **The Shifts Perspective**: Borusyak et al. (2021) offer the alternative framework that even if **shares** are endogenous, exogenous **shifts** can identify causal effects, as long as they are **uncorrelated with the bias** of the shares. In the example, this would mean that national immigration shocks need to be unrelated to local incomes.

Examples

Autor et al. (2013) want to find out how competition from **Chinese imports** affects local **labor markets** in the U.S. They use an instrument like this:

$$B_{it} = \sum_{j=1}^{J} l_{ijt} \times g_{jt},$$

where i are regions, t is a time index, and j are industries; l_{ijt} is the share of people working in (manufacturing) industry j in region j at time t and g_{jt} is the growth of Chinese imports in industry j in a group of countries that are comparable to the U.S.

Nunn & Qian (2014) investigate the effect of **U.S. food aid** on **conflict** in non-OECD countries. To circumvent the endogeneity issue, they use the following instrument (simplified):

$$B_{it} = \overline{D}_i \times P_{t-1},$$

where $t = 1, \ldots, T$ are years and $i = 1, \ldots, N$ are countries; $\overline{D}_i$ is the share of years in which the country received aid, $\overline{D}_i = T^{-1} \sum_{t=1}^{T} D_{it}$, and P_{t-1} is U.S. wheat production the previous year.

References

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, *20*(1), 46–63. https://doi.org/10.1214/aoms/1177730090

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

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*(4), 69–85. https://doi.org/10.1257/jep.15.4.69

Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, *103*(6), 2121–2168. https://doi.org/10.1257/aer.103.6.2121

Bartik, T. J. (1991). *Who benefits from state and local economic development policies?* W.E. Upjohn Institute. https://doi.org/10.17848/9780585223940

Borusyak, K., Hull, P., & Jaravel, X. (2021). Quasi-experimental shift-share research designs. *The Review of Economic Studies*, *89*(1), 181–213. https://doi.org/10.1093/restud/rdab030

Borusyak, K., Hull, P., & Jaravel, X. (2025). A practical guide to shift-share instruments. *Journal of Economic Perspectives*, *39*(1), 181–204. https://doi.org/10.1257/jep.20231370

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. https://doi.org/10.1080/01621459.1995.10476536

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. https://doi.org/10.1162/REST_a_00314

Cunningham, S. (2021). *Causal inference*. Yale University Press. https://doi.org/10.12987/9780300255881

Goldsmith-Pinkham, P., Sorkin, I., & Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, *110*(8),

Weak Instruments

More Examples

Shift-Share Instruments

Appendix

Extracting the Angrist & Krueger (1991) Code File

```r
library(haven)
library(dplyr)

if (!file.exists("NEW7080.dta")) {
  if (!file.exists("NEW7080_1.rar"))
    download.file("https://economics.mit.edu/sites/default/files/inline-files/NEW7080_1.rar",
                  "NEW7080_1.rar", mode = "wb")
  system("unrar x -y NEW7080_1.rar", ignore.stdout = TRUE)
}

df <- read_dta("NEW7080.dta")

nm <- c("v4"="EDUC","v9"="LWKLYWGE","v16"="CENSUS","v18"="QOB","v27"="YOB")
for (k in names(nm)) if (k %in% names(df)) names(df)[names(df)==k] <- nm[[k]]

df <- df %>%
  mutate(AGEQ = ifelse(CENSUS == 80, NA, NA),                # placeholder, dropped later
         COHORT = ifelse(YOB >= 30 & YOB <= 39, 30, NA)) %>%
  filter(COHORT == 30)

# Year-of-birth dummies (YR1930–YR1939)
```