INTERMEDIATE QUANTITATIVE METHODS

Avoiding Pitfalls in Regression Analysis

Per Engzell (Nuffield College)

Hilary term 2018

Sociology Department, University of Oxford

INSTRUMENTAL VARIABLES

CAUSALITY

A common interpretation of regression coefficients is as ”the change in y expected from a one unit change in x”

This is a causal interpretation: change x, and y will follow

However, we do not know this unless we have randomly assigned people to different values of x as in an experiment

In social science, the x’s that we care about will almost always be correlated with other things that are unobserved and might be the true, underlying causes of y

ENDOGENEITY

In general, the regression estimate β_2 is biased as a causal estimate if there is endogeneity: $\text{Cov}(x, u) \neq 0$. This can arise because of:

- Simultaneous or reverse causation
  - x might be caused partly by y
  - Example: recreational drug use and psychiatric disorders
- Omitted variables or unobserved selection
  - An unobserved variable causes both x and y
  - Example: education and labour market earnings
- Measurement error
  - We observe x (or control variables) only with error

Usually, any regression will suffer from all three to some extent

CONDITIONS

Instrumental variables (IV) are a possible solution to these problems but as all estimators they come with assumptions. Good instruments are hard to find.

An instrument is something that predicts the endogenous variable x, but has no impact on y other than through its causing x

In formal terms

- $\mathrm{Cov}(x, u) \neq 0$: x is endogenous
- $\mathrm{Cov}(IV, x) \neq 0$: the instrument is relevant; it predicts x
- $\mathrm{Cov}(IV, u) = 0$: the instrument is exogenous; it does not independently cause, or otherwise correlate with, y

The last assumption cannot be tested, unless there is more than one instrument per endogenous x. Even so, tests can only reject (never confirm) instrument validity

EXAMPLE

One influential study using IV early in economics was by Angrist and Krueger (”Does compulsory school attendance affect schooling and earnings?” Quarterly Journal of Economics, 1991).

They proposed to estimate the causal effect of schooling on earnings using quarter of birth as an instrumental variable

In the US, compulsory schooling starts in the fall of the year you turn 7, but you can drop out of school from the *day* you turn 16

Those born early in the calendar year start school at an older age and are allowed drop out having completed less schooling

EXAMPLE

Does quarter of birth affect schooling? Yes, those born earlier in the year do have lower schooling

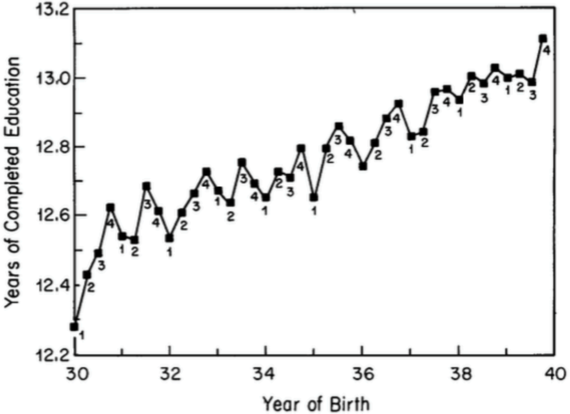
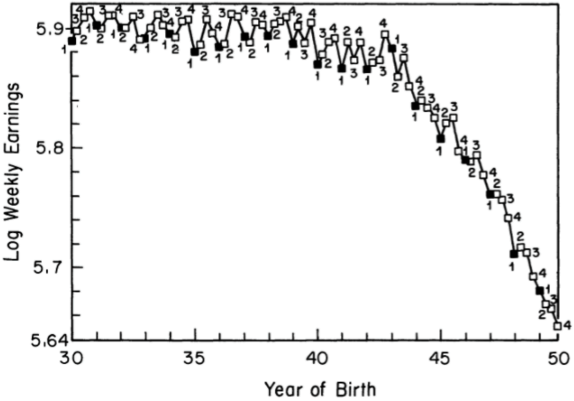


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

EXAMPLE

Do differences in schooling due to different quarter of birth translate into different earnings?



MECHANICS

How (and why) does an IV work? You have seen one description:

$$\widehat{\beta}_{\text{OLS}} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \qquad \widehat{\beta}_{\text{IV}} = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)}$$

If you're like me, you won't find this very intuitive!

MECHANICS

How (and why) does an IV work? You have seen one description:

$$\widehat{\beta}_{\text{OLS}} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \qquad \widehat{\beta}_{\text{IV}} = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)}$$

If you're like me, you won't find this very intuitive! It helps breaking it down into a series of separate regressions:

First stage Regression of the endogenous x on the instrument

$$x = \gamma_1 + \gamma_2 \text{IV} + v$$

Reduced form Regression of the outcome y on the instrument

$$y = \delta_1 + \delta_2 \text{IV} + \varepsilon$$

Second stage Regression of outcome y on predicted x from first stage

$$y = \beta_1^* + \beta_2^* \widehat{x} + e$$

VERSION 1

We want to estimate $y = \beta_1 + \beta_2 x + u$ but can't because $\text{Cov}(x, u) \neq 0$

First stage Regression of the endogenous x on the instrument

$$x = \gamma_1 + \gamma_2 IV + v$$

Reduced form Regression of the outcome y on the instrument

$$y = \delta_1 + \delta_2 IV + \varepsilon$$

In the reduced form, we are using the IV as a proxy variable for x

With classical error, δ_2 is a downward biased estimate of β_2

The weaker proxy, the lower the value of γ_2 in the first stage

Therefore, we adjust δ_2 upward by a factor of $1/\gamma_2$

$$\delta_2 \frac{1}{\gamma_2} = \frac{\delta_2}{\gamma_2} = \frac{\text{Cov}(z, y)/\text{Var}(z)}{\text{Cov}(z, x)/\text{Var}(z)} = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)}$$

VERSION 1

Looking at Angrist and Krueger's data:

	Born early	Born late	Difference	
ln(weekly wage)	5.892	5.905	0.013	(δ_2)
Years of education	12.688	12.839	0.151	(γ_2)

IV estimate of the % expected wage return to one year of schooling:

$$\frac{\text{reduced form}}{\text{first stage}} = \delta_2/\gamma_2 = 0.013/0.151 = 0.086$$

(OLS estimate of the % expected wage return to one additional year of schooling in this sample: 0.070)

VERSION 2

We want to estimate $y = \beta_1 + \beta_2 x + u$ but can't because $\mathrm{Cov}(x, u) \neq 0$

First stage Regression of the endogenous x on the instrument

$$x = \gamma_1 + \gamma_2 IV + v$$

Second stage Regression of outcome y on predicted x from first stage

$$y = \beta_1^* + \beta_2^* \widehat{x} + e$$

The two-stage least squares (2SLS) estimate of β_2 is β_2^* from the second stage, and it is the same as the IV estimator

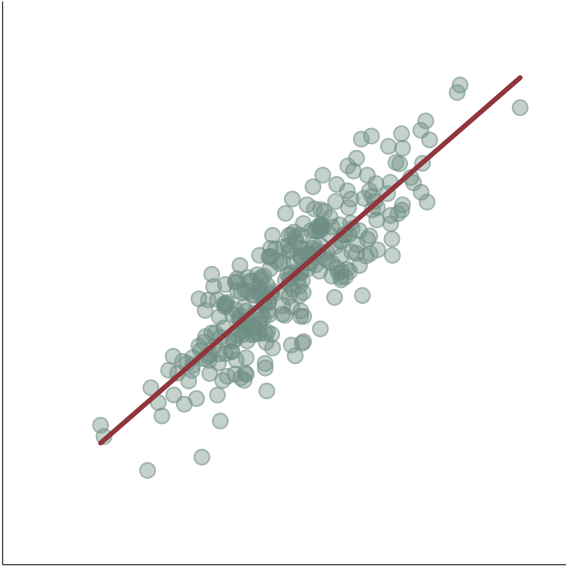
Why no correction? Isn't $\widehat{x}$ a noisy proxy for x just as the IV in the reduced form was?

Yes it is! But the error here is not classical but of a different form

Instead of proxy = true + error, we have: true = proxy + error ($x = \gamma_1 + \gamma_2 IV + v$) and there is no downward bias in this case

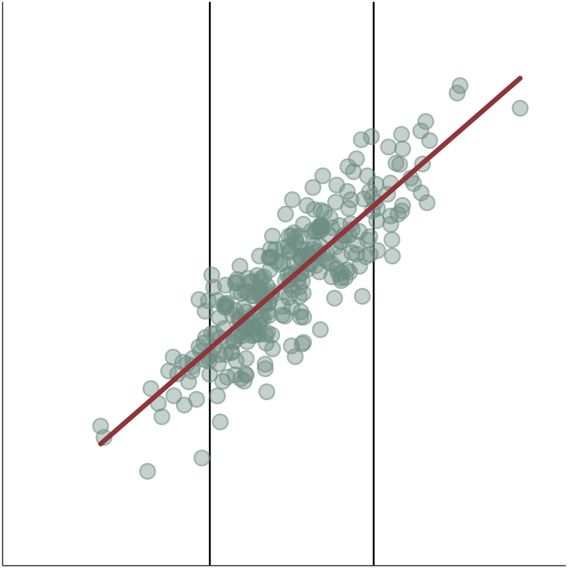
BERKSON ERROR

- Suppose we have grouped data and x is an unbiased average of x* in each group



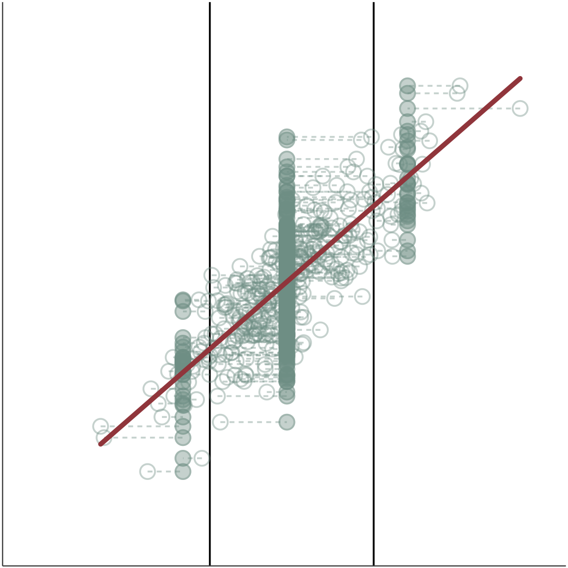
BERKSON ERROR

- Suppose we have grouped data and x is an unbiased average of x* in each group
- In this case, $\text{Cov}(w, x) = 0$



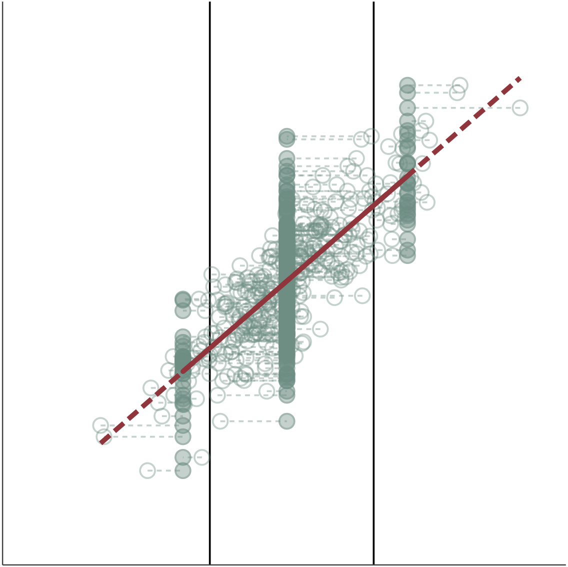
BERKSON ERROR

- Suppose we have grouped data and x is an unbiased average of x* in each group
- In this case, $\text{Cov}(w, x) = 0$
- Different from the classical model which assumed $\text{Cov}(w, x^*) = 0$



BERKSON ERROR

- Suppose we have grouped data and x is an unbiased average of x* in each group
- In this case, $\mathrm{Cov}(w, x) = 0$
- Different from the classical model which assumed $\mathrm{Cov}(w, x^*) = 0$
- The estimate of β_2 is unbiased!



EXAMPLE

We want to know whether attending classes (x) improves student grades (y). Should we trust the OLS estimate

$$\text{final grade} = \beta_1 + \beta_2 \text{attendance} + \text{u}$$

EXAMPLE

We want to know whether attending classes (x) improves student grades (y). Should we trust the OLS estimate

$$\text{final grade} = \beta_1 + \beta_2 \text{attendance} + \text{u}$$

No, classroom attendance is not random. Students that show up are perhaps more motivated, and might have had better grades regardless

Ideally, to answer this question we would randomise who goes to class and not but this is either unethical, unfeasible, or both

EXAMPLE

Can we find an instrument to answer the causal question?

- Prior attendance: % classes attended in previous term
- Foreign student: 1 if i is a foreign student, 0 otherwise
- Strike: Number of lectures cancelled due to strike

One of these is potentially a valid instrument, two are not. Can you tell which?

EXAMPLE

Can we find an instrument to answer the causal question?

- Prior attendance: % classes attended in previous term
- Foreign student: 1 if i is a foreign student, 0 otherwise
- Strike: Number of lectures cancelled due to strike

One of these is potentially a valid instrument, two are not. Can you tell which?

Instruments that are "internal" to the actor are unlikely to be exogenous and therefore fail as valid IVs

Good instruments should be such that they are imposed from outside, and not open to manipulation. The assignment mechanism is know and random

IV CAVEATS

IVs only use a small fraction of total variance in x and therefore have high sampling variance (and low statistical power)

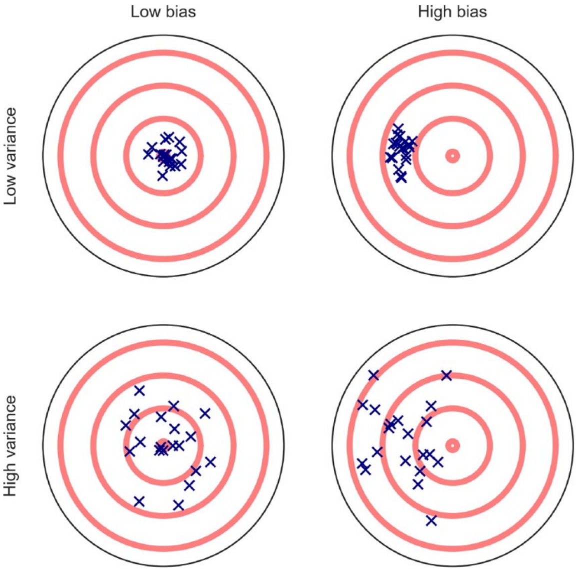
IV always estimates a local average treatment effect, that is, it generalises only to those who are affected by the instrument

Credible instruments are usually policy experiments, but these are ”opportunistic”: we cannot decide which populations to study

Consequences of endogeneity or (nonclassical) measurement error tend to be more damaging in more elaborate causal models, not less

If the IV has a weak relationship to x, the estimate is biased upward (weak instrument problem)

BIAS VS VARIANCE



Source: Yarkoni, Tal, and Jacob Westfall. ”Choosing prediction over explanation in psychology: Lessons from machine learning.” Perspectives on Psychological Science 12, 6 (2017): 1100-1122.

Artwork by Viktor Beekman
viktor.beekman@gmail.com

”When effect size is tiny and measurement error is huge, you’re essentially trying to use a bathroom scale to weigh a feather – and the feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down”

Andrew Gelman

QUESTIONS?