SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics: Advanced

Instrumental Variables - Applications

Nikiforos Zampetakis based on Florian Oswald's slides
SciencesPo Paris
2024-03-13

Status

What Did we Do Last Week?

- We learned about John Snow's grand experiment in London 1850.
- We used his story to motivate the IV estimator.

Today

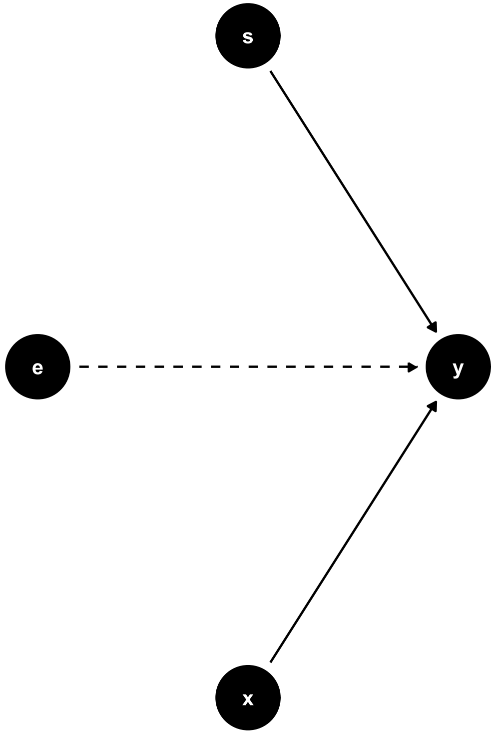
- We'll look at further IV applications.
- We introduce an extension called *Two Stage Least Squares*.
- We will use `R` to compute the estimates.
- Finally we'll talk about *weak* instruments.

Back to school!

Returns To Schooling

- What's the causal impact of schooling on earnings?
- Jacob Mincer was interested in this important question.
- Here's his model:

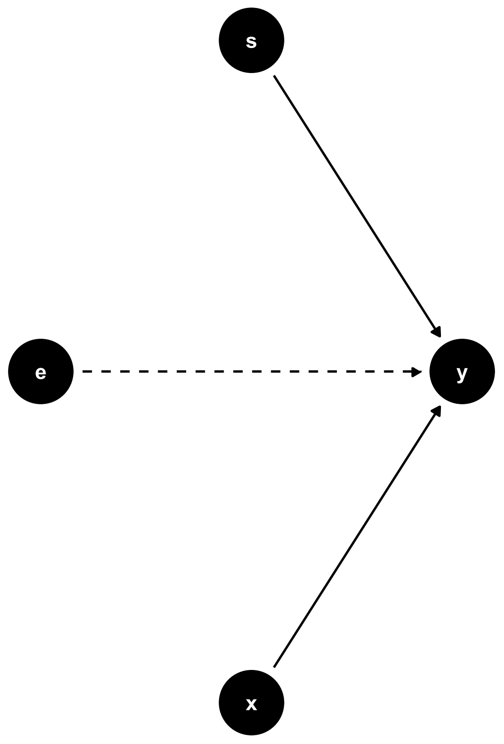
$$\log Y_i = \alpha + \rho S_i + \beta_1 X_i + \beta_2 X_i^2 + e_i$$



Returns To Schooling

$$\log Y_i = \alpha + \rho S_i + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- He found an estimate for ρ of about 0.11,
- 11% earnings advantage for each additional year of education
- Look at the DAG. Is that a good model? Well, why would it not be?



Ability Bias

- We compare earnings of men with certain schooling and work experience
- Is all else equal, after controlling for those?
- Given X,
  - Can we find differently diligent workers out there?
  - Can we find differently able workers?
  - Do family connections of workers vary?
- Yes, of course. So, *all else* is not equal at all.
- That's an issue, because for OLS consistency we require the orthogonality assumption

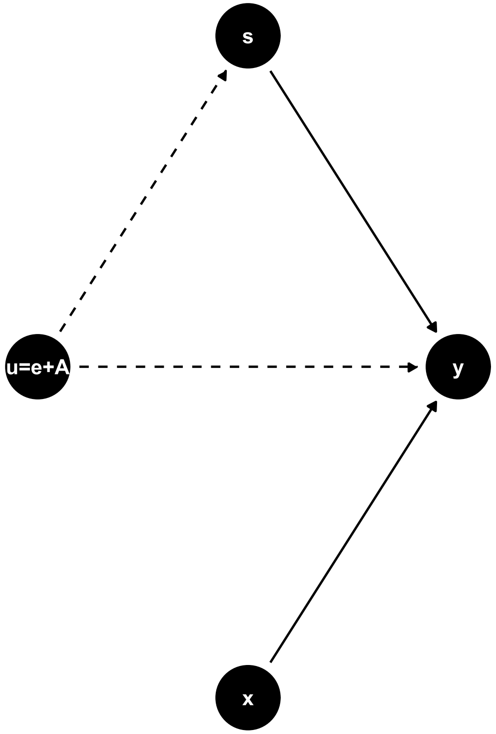
$$E[e_i|S_i, X_i] \neq 0$$

- Let's introduce **ability** A_i explicitly.

Mincer with Unobserved Ability

- In fact we have *two* unobservables: e and A
  .
- Of course we can't tell them apart.
- So we defined a new unobservable factor

$$u_i = e_i + A_i$$

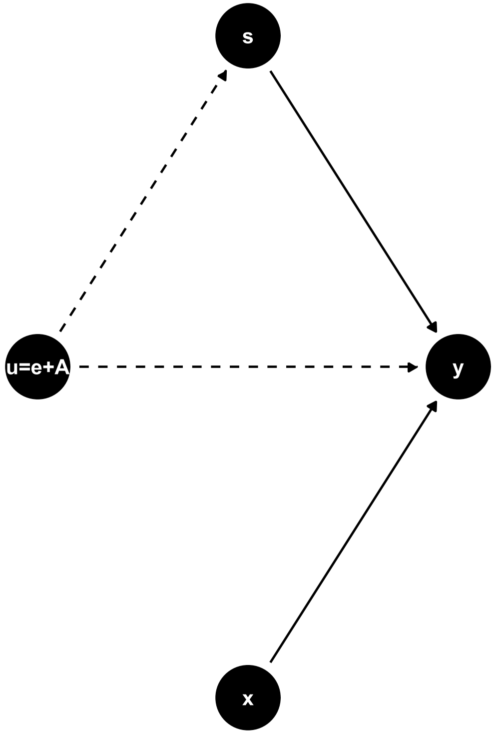


Mincer with Unobserved Ability

- In terms of an equation:

$$\log Y_i = \alpha + \rho S_i + \beta_1 X_i + \beta_2 X_i^2 + \underbrace{u_i}_{A_i + e_i}$$

- Sometimes, this does not matter, and the OLS bias is small.
- But sometimes it does and we get it totally wrong! Example.



Angrist and Krueger (1991): Birthdate is as good as Random

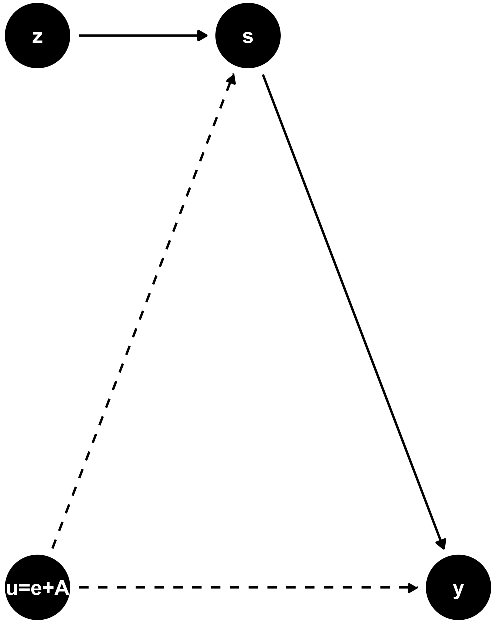
- Angrist and Krueger (AK91) is an influental study addressing ability bias.
- Idea:

  1. construct an IV that encodes *birth date of student*.
  2. Child born just after cutoff date will start school later!

- Suppose all children who reach the age of 6 by 31st of december 2021 are required to enroll in the first grade of school in september 2021.
- If born in September 2015 (i.e. 6 years prior), will be 5 years and 3/4 by the time they start school.
- If born on the 1st of January 2016 will be 6 and 3/4 years when *they* enter school in september 2022.
- However, people can drop out of school legally on their 16-th birthday!
- So, out of people who drop out, some got more schooling than others.
- AK91 construct IV *quarter of birth* dummy: affects schooling, but not related to A!

AK91 IV setup

- *quarter of birth* dummy z: affects schooling, but not related to A!
- In particular: whether born in 4-th quarter or not.



AK91 Estimation: Two Stage Least Squares (2SLS)

AK91 allow us to introduce a widely used variation of our simple IV estimator: **2SLS**

1. We estimate a **first stage model** which uses only exogenous variables (like z) to explain our endgenous regressor s.

2. We then use the first stage model to *predict* values of s in what is called the **second stage** or the **reduced form** model. Performing this procedure is supposed to take out any impact of A in the correlation we observe in our data between s and y.

$$\begin{aligned}
&\text{1. Stage: } s_i = \alpha_0 + \alpha_1 z_i + \eta_i \\
&\text{2. Stage: } y_i = \beta_0 + \beta_1 \hat{s}_i + u_i
\end{aligned}$$

Conditions:

1. Relevance of the IV: $\alpha_1 \neq 0$
2. Independence (IV assignment as good as random): $E[\eta|z] = 0$
3. Exogeneity (our exclusion restriction): $E[u|z] = 0$

Let's do Angrist and Krueger (1991)!

Data on birth quarter and wages

Let's load the data and look at a quick summary

```
#devtools::install_github("jrnold/masteringmetrics", subdir = "masteringmetrics")
data("ak91", package = "masteringmetrics")
# from the modelsummary package
datasummary_skim(data.frame(ak91),histogram = TRUE)
```

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	
lnw	26732	0	5.9	0.7	-2.3	6.0	10.5	
s	21	0	12.8	3.3	0.0	12.0	20.0	
yob	10	0	1934.6	2.9	1930.0	1935.0	1939.0	
qob	4	0	2.5	1.1	1.0	3.0	4.0	
sob	51	0	30.7	14.2	1.0	34.0	56.0	
age	40	0	45.0	2.9	40.2	45.0	50.0	

AK91 Data Transformations

- We want to create the `q4` dummy which is `TRUE` if you are born in the 4th quarter.
- create `factor` versions of quarter and year of birth.

```
ak91 <- mutate(ak91,
               qob_fct = factor(qob),
               q4 = as.integer(qob == "4"),
               yob_fct = factor(yob))
# get mean wage by year/quarter
ak91_age <- ak91 %>%
  group_by(qob, yob) %>%
  summarise(lnw = mean(lnw), s = mean(s)) %>%
  mutate(q4 = (qob == 4))
```

AK91 Figure 1: First Stage!

Let's reproduce AK91's first figure now on education as a function of quarter of birth!

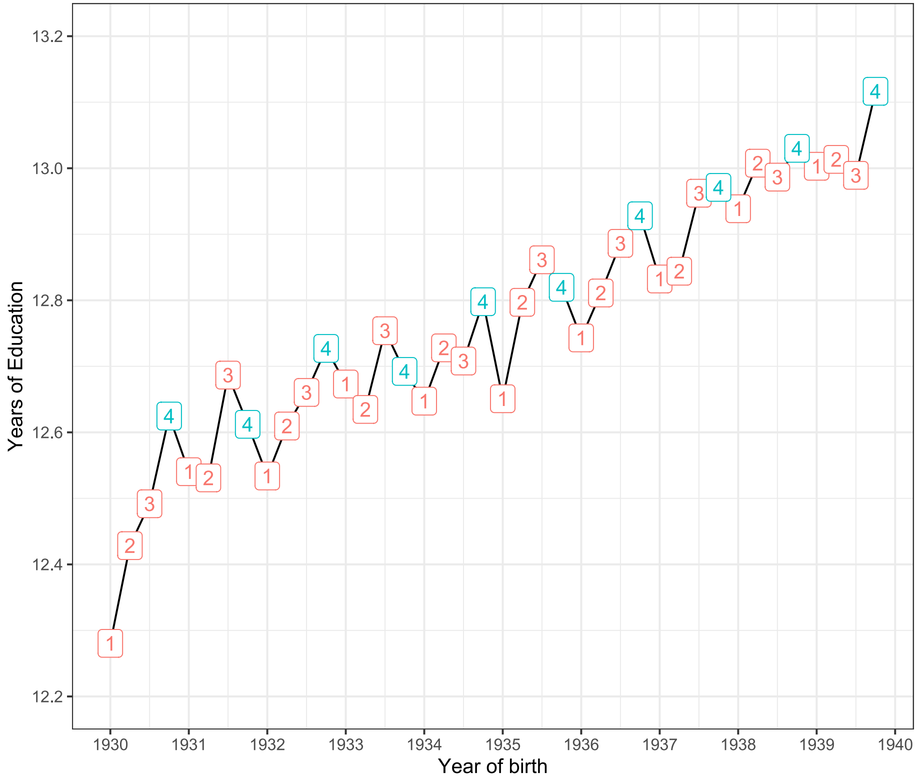
```
ggplot(ak91_age, aes(x = yob + (qob - 1) / 4, y = s )) +
  geom_line() +
  geom_label(mapping = aes(label = qob, color = q4)) +
  guides(label = FALSE, color = FALSE) +
  scale_x_continuous("Year of birth", breaks = 1930:1940) +
  scale_y_continuous("Years of Education", breaks = seq(12.2, 13.2, by = 0.2),
                     limits = c(12.2, 13.2)) +
  theme_bw()
```

AK91 Figure 1: First Stage!

1. The numbers label mean education *by* quarter of birth groups.

2. The 4-th quarters **did** get more education in most years!

3. There is a general trend.



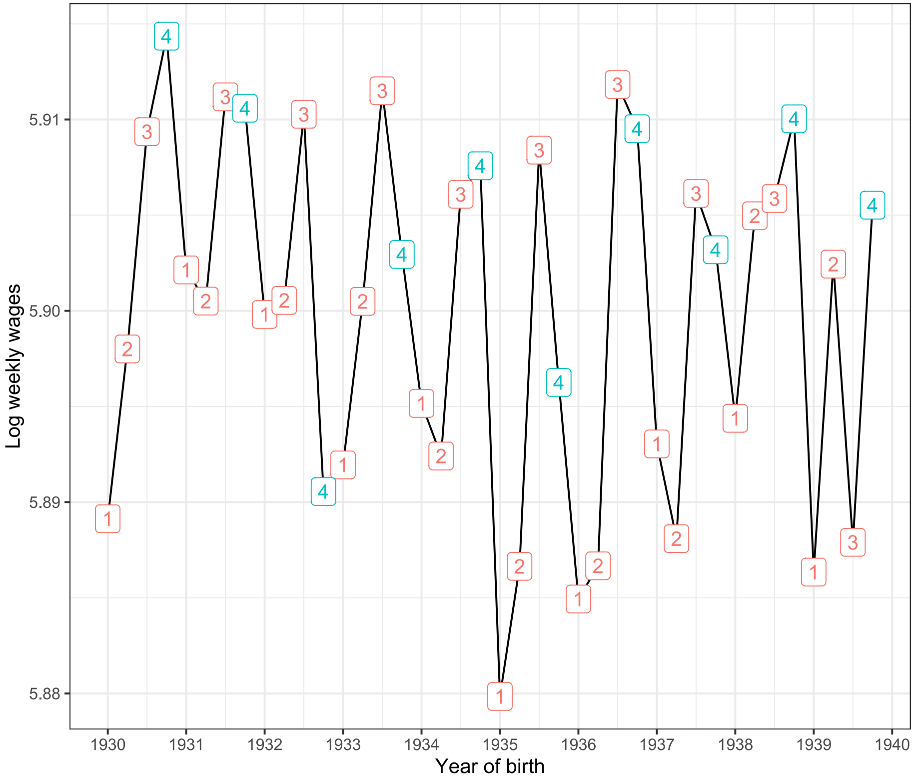
AK91 Figure 2: Impact of IV on outcome

What about earnings for those groups?

```
ggplot(ak91_age, aes(x = yob + (qob - 1) / 4, y = lnw)) +
  geom_line() +
  geom_label(mapping = aes(label = qob, color = q4)) +
  scale_x_continuous("Year of birth", breaks = 1930:1940) +
  scale_y_continuous("Log weekly wages") +
  guides(label = FALSE, color = FALSE) +
  theme_bw()
```

AK91 Figure 2: Impact of IV on outcome

1. The 4-th quarters are among the high-earners by birth year.
2. In general, weekly wages seem to decline somewhat over time.



Running IV estimation in R

- Several options (like always with `R`! 😉)
- Will use the `feols` function from the `fixest` package.
- We will compute *robust* standard errors which are correcting for heteroskedasticity. Details here.
- Notice the `predict` to get $\hat{s}$.

```r
library(fixest)
# create a list of models
mod <- list()

# standard (biased!) OLS
mod$ols <- lm(lnw ~ s, data = ak91)

# IV: born in q4 is TRUE?
# doing IV manually in 2 stages.
mod[["1. stage"]] <- lm(s ~ q4, data = ak91)
ak91$shat          <- predict(mod[["1. stage"]])
mod[["2. stage"]] <- lm(lnw ~ shat, data = ak91)

# run 2SLS
# doing IV all in one go
# notice the formula!
# formula = y ~ 1 | x ~ z
mod$`2SLS`  <- feols(lnw ~ 1 | s ~ q4,
                     data = ak91,
                     vcov = "hetero")
```

AK91 Results Table

	ols	1. stage	2. stage	2SLS
(Intercept)	4.995***	12.747***	4.955***	4.955***
	(0.004)	(0.007)	(0.381)	(0.358)
s	0.071***			
	(0.000)			
q4		0.092***		
		(0.013)		
shat			0.074*	
			(0.030)	
fit_s				0.074**
				(0.028)
R2	0.117	0.000	0.000	0.117
F	43782.556	48.095	6.146	
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001				

1. OLS likely downward biased (measurement error in schooling)
2. First Stage: IV `q4` is statistically significant, but small effect: born in q4 has 0.092 years of educ. R^2 is 0%! But F-stat is large. 😅
3. Second stage has same point estimate as `2SLS` but different std error (2. stage one is wrong)

Remember the F-Statistic?

- We encountered this before: it's useful to test restricted vs unrestricted models against each other.
- Here, we are interested whether our instruments are *jointly* significant. Of course, with only one IV, that's not more informative than the t-stat of that IV.
- This F-Stat compares the predictive power of the first stage with and without the IVs. If they have very similar predictive power, the F-stat will be low, and we will not be able to reject the H0 that our IVs are **jointly insignificant** in the first stage model. 😞

Additional Control Variables

- We saw a clear time trend in education earlier.
- There are also business-cycle fluctuations in earnings
- We should somehow control for different time periods.
- Also, we can use more than one IV! Here is how:

Additional Control Variables

```r
# we keep adding to our `mod` list:
mod$ols_yr  <- update(mod$ols, . ~ . + yob_fct)  #  previous OLS model
# add exogenous vars!
mod[["2SLS_yr"]] <- feols(lnw ~ yob_fct | s ~ q4, data = ak91, vcov = "hetero")
# use all quarters as IVs
mod[["2SLS_all"]] <- feols(lnw ~ yob_fct | s ~ qob_fct, data = ak91, vcov = "hetero")
```

	ols	2SLS	ols_yr	2SLS_yr	2SLS_all
(Intercept)	4.995***	4.955***	5.017***	4.966***	4.592***
	(0.004)	(0.358)	(0.005)	(0.354)	(0.251)
s	0.071***		0.071***		
	(0.000)		(0.000)		
fit_s		0.074**		0.075**	0.105***
		(0.028)		(0.028)	(0.020)
R2	0.117	0.117	0.118	0.117	0.091
Instruments	none	Q4	none	Q4	All Quarters
Year of birth	no	no	yes	yes	yes
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001					

Additional Control Variables

	ols	2SLS	ols_yr	2SLS_yr	2SLS_all
(Intercept)	4.995***	4.955***	5.017***	4.966***	4.592***
	(0.004)	(0.358)	(0.005)	(0.354)	(0.251)
s	0.071***		0.071***		
	(0.000)		(0.000)		
fit_s		0.074**		0.075**	0.105***
		(0.028)		(0.028)	(0.020)
R2	0.117	0.117	0.118	0.117	0.091
Instruments	none	Q4	none	Q4	All Quarters
Year of birth	no	no	yes	yes	yes
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001					

Adding year controls…

- leaves OLS mostly unchanged
- slight increase in 2SLS estimate

Using all quarters as IV…

- Increases precision of 2SLS estimate a lot!
- Point estimate is 10.5% now!

AK91: Taking Stock - The Quarter of Birth (QOB) IV

- This will produce consistent estimates if

    1. The IV predicts the endogenous regressor well.
    2. The IV is as good as random / independent of OVs.
    3. Can only impact outcome through schooling.

- How does the QOB perform along those lines?

1. Plot of first stage and high F-stat offer compelling evidence for **relevance**. ✅

2. Is QOB **independent** of, say, *maternal characteristics*? Birthdays are not really random - there are birth seasons for certain socioeconomic backgrounds. highest maternal schooling give birth in second quarter. (not in 4th! ✅)

3. Exclusion: What if the youngest kids (born in Q4!) are the disadvantaged ones early on, which has long-term negative impacts? That would mean $E[u|z] \neq 0$! Well, with QOB the youngest ones actually do better (more schooling and higher wage)! ✅

Mechanics of IV

Identification and Inference

IV Identification

Let's go back to our simple linear model:

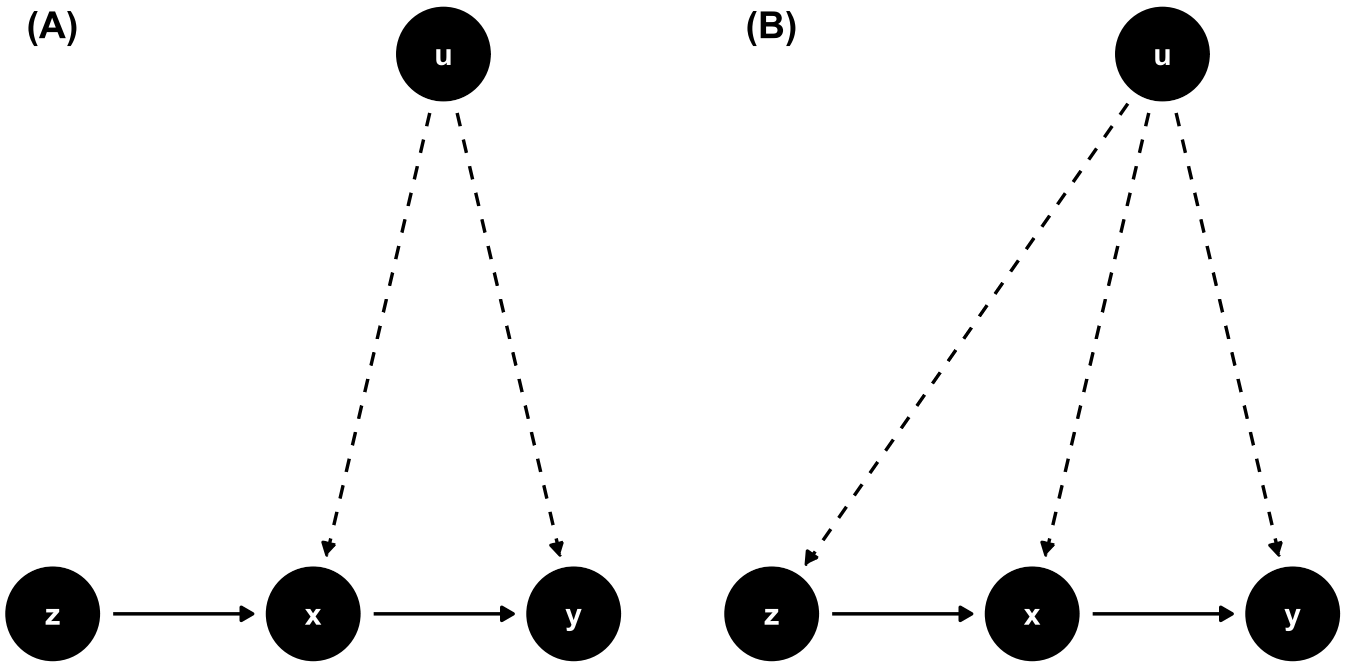
$$y = \beta_0 + \beta_1 x + u$$

where we fear that $Cov(x, u) \neq 0$, x is *endogenous*.

Conditions for IV

1. **first stage** or **relevance**: $Cov(z, x) \neq 0$
2. **IV exogeneity**: $Cov(z, u) = 0$: the IV is exogenous in the outcome equation.

Valid Model (A) vs Invalid Model (B) for IV z



IV Identification

> ## Conditions for IV
>
> 1. **first stage** or **relevance**: $Cov(z,x) \neq 0$
> 2. **IV exogeneity**: $Cov(z,u) = 0$: the IV is exogenous in the outcome equation.

- How does this *identify* β_1?
- (How can we express β_1 in terms of population moments to pin it's value down?)

IV Identification

$$\begin{aligned} Cov(z,y) &= Cov(z, \beta_0 + \beta_1 x + u) \\ &= \beta_1 Cov(z,x) + Cov(z,u) \end{aligned}$$

Under condition 2. above (**IV exogeneity**), we have $Cov(z,u) = 0$, hence

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

and under condition 1. (**relevance**), we have $Cov(z,x) \neq 0$, so that we can divide the equation through to obtain

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

- β_1 is *identified* via population moments $Cov(z,y)$ and $Cov(z,x)$.
- We can *estimate* those moments via their *sample analogs*

IV Estimator

Just plugging in for the population moments:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}$$

- The intercept estimate is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Given both assumptions 1. and 2. are satisfied, we say that *the IV estimator is consistent for* β_1. We write

$$\text{plim}(\hat{\beta}_1) = \beta_1$$

in words: the *probability limit* of $\hat{\beta}_1$ is the true β_1.

- If this is true, we say that this estimator is **consistent**.

IV Inference

Assuming $E(u^2|z) = \sigma^2$ the variance of the IV slope estimator is

$$Var(\hat{\beta}_{1,IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

- σ_x^2 is the population variance of x,
- σ^2 the one of u, and
- $\rho_{x,z}$ is the population correlation between x and z.

You can see 2 important things here:

1. Without the term $\rho_{x,z}^2$, this is **like OLS variance**.
2. As sample size n increases, the **variance decreases**.

IV Variance is Always Larger than OLS Variance

- Replace $\rho^2_{x,z}$ with $R^2_{x,z}$, i.e. the R-squared of a regression of x on z:

$$Var(\hat{\beta}_{1,IV}) = \frac{\sigma^2}{n\sigma^2_x R^2_{x,z}}$$

1. Given $R^2_{x,z} < 1$ in most real life situations, we have that $Var(\hat{\beta}_{1,IV}) > Var(\hat{\beta}_{1,OLS})$ almost certainly.
2. The higher the correlation between z and x, the closer their $R^2_{x,z}$ is to 1. With $R^2_{x,z} = 1$ we get back to the OLS variance. This is no surprise, because that implies that in fact $z = x$.

So, if you have a valid, exogenous regressor x, you should *not* perform IV estimation using z to obtain $\hat{\beta}$, since your variance will be unnecessarily large.

Returns to Education for Married Women

Consider the following model for married women's wages:

$$\log wage = \beta_0 + \beta_1 educ + u$$

Let's run an OLS on this, and then compare it to an IV estimate using *father's education*. Keep in mind that this is a valid IV z if

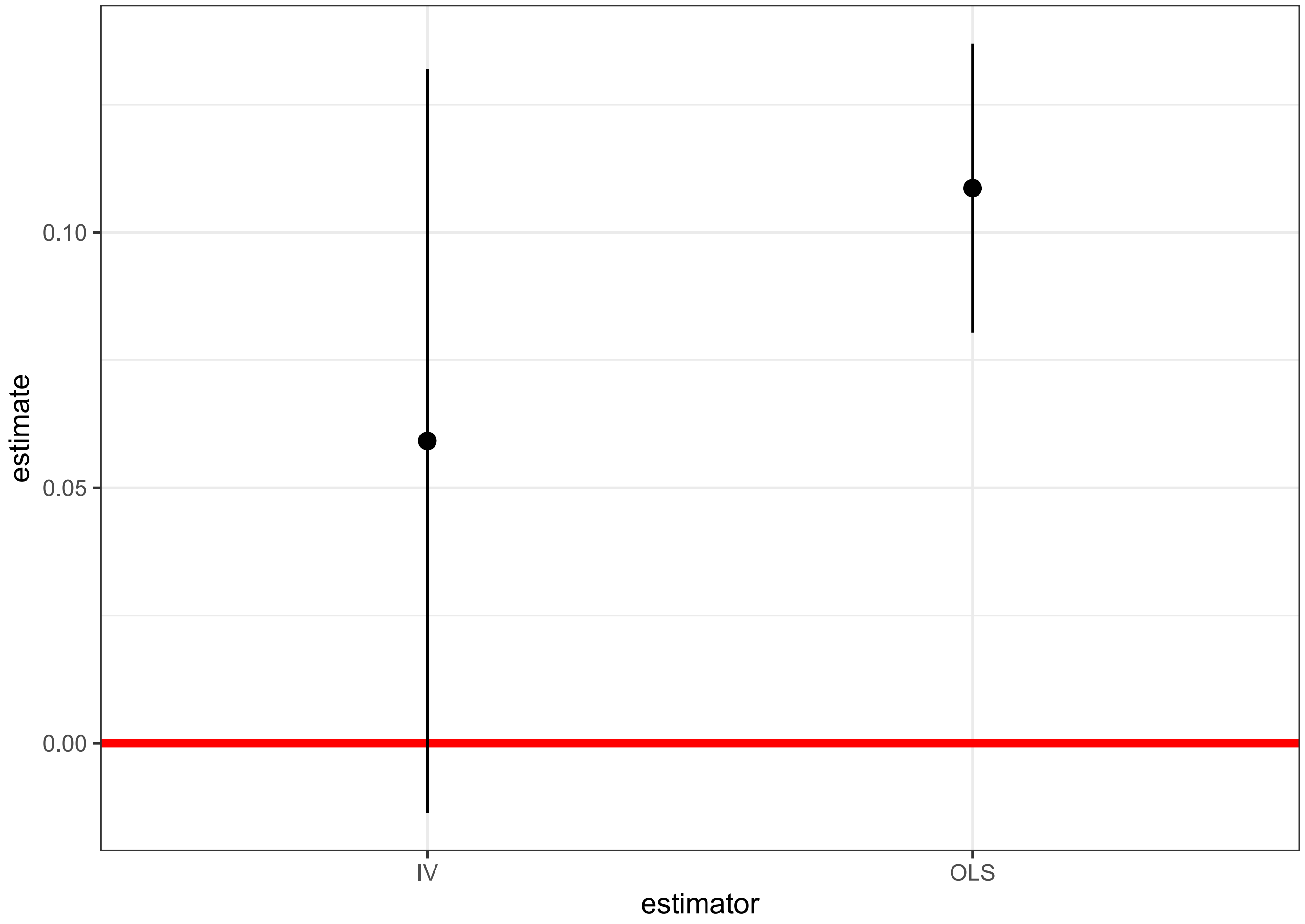
1. *fatheduc* and *educ* are correlated
2. *fatheduc* and u are not correlated.

Returns to Education for Married Women

```r
data(mroz,package = "wooldridge")
mods = list()
mods$OLS <- lm(lwage ~ educ, data = mroz)
mods[['First Stage']] <- lm(educ ~ fatheduc, data = subset(mroz, inlf == 1))
mods$IV  <- feols(lwage ~ 1 | educ ~ fatheduc, data = mroz, vcov = "hetero")
```

	OLS	First Stage	IV
(Intercept)	-0.185	10.237***	0.441
	(0.185)	(0.276)	(0.465)
educ	0.109***		
	(0.014)		
fatheduc		0.269***	
		(0.029)	
fit_educ			0.059
			(0.037)
Num.Obs.	428	428	428
R2	0.118	0.173	0.093

IV Standard Errors



IV with a Weak Instrument

- IV is consistent under given assumptions.
- However, *even if* we have only very small $Cor(z, u)$, we can get wrong-footed
- Small corrleation between x and z can produce **inconsistent** estimates.

$$\text{plim}(\hat{\beta}_{1,IV}) = \beta_1 + \frac{Cor(z,u)}{Cor(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

- Take $Cor(z, u)$ is very small,
- A **weak instrument** is one with only a small absolute value for $Cor(z, x)$
- This will blow up this second term in the probability limit.
- Even with a very big sample size n, our estimator would *not* converge to the true population parameter β_1, because we are using a weak instrument.

Weak Stuff

To illustrate this point, let's assume we want to look at the impact of number of packs of cigarettes smoked per day by pregnant women (*packs*) on the birthweight of their child (*bwght*):

$$\log(bwght) = \beta_0 + \beta_1 packs + u$$

We are worried that smoking behavior is correlated with a range of other health-related variables which are in u and which could impact the birthweight of the child. So we look for an IV. Suppose we use the price of cigarettes (*cigprice*), assuming that the price of cigarettes is uncorrelated with factors in u. Let's run the first stage of *cigprice* on *packs* and then let's show the 2SLS estimates:

Weak Stuf (Different Package)

```r
library(estimatr)
data(bwght, package = "wooldridge")
mods <- list()
mods[["First Stage"]] <- lm(packs ~ cigprice, data = bwght)
mods[["IV"]] <- estimatr::iv_robust(log(bwght) ~  packs | cigprice, data = bwght,
                                    diagnostics = TRUE)
```

	First Stage	IV
(Intercept)	0.067	4.448
	(0.103)	(0.940)
cigprice	0.000	
	(0.001)	
packs		2.989
		(8.996)
R2	0.000	-23.230

Weak Stuff

- The first columns shows: very weak first stage. *cigprice* has zero impact on packs it seems!
- R^2 is zero.
- What is we use this IV nevertheless?

- in the second column: very large, positive(!) impact of packs smoked on birthweight. 🤨
- Huge Standard Error though.
- An R^2 of -23?!
- F-stat of first stage: 0.121. Corresponds to a p-value of 0.728 : we **cannot** reject the H0 of an insignificant first stage here *at all*.
- So: **invalid** approach. ❌

SciencesPo
DEPARTMENT OF ECONOMICS

END

	nikiforos.zampetakis@sciencespo.fr
	Slides
	Book
	@ScPoEcon
	@ScPoEcon