Instrumental Variables

EC 421, Set 12

Connor Lennon

03 June 2021

Prologue

Schedule

Last Time

Causality

Today

- Review: Causality
- New: Instrumental variables

Upcoming

The end?

Causality

Review

Causality

Review

In our last few weeks, we returned more robustly to the concept of **causality**.

We worked through the *Rubin causal model*, in which we defined

- y_{1i} : the outcome for individual i if she had received treatment
- y_{0i} : the outcome for individual i if she had not received treatment

and we referred to individuals who did not receive treatment as *control*.

If we were able to know both y_{1i} ***and*** y_{0i}, we could calculate the causal effect of treatment for individual i, *i.e.*,

$$\tau_i = y_{1i} - y_{0i}$$

Causality

Review

Fundamental problem of causal inference:

We cannot simultaneously know y_{1i} and y_{0i}.

Either we observe individual i in the treatment group, *i.e.*,

$$\tau_i = y_{1i} - ?$$

or we observe i in the control group, *i.e.*,

$$\tau_i = ? - y_{0i}$$

but never both at the same time.

Causality

Review

If we want to know τ_i (or at least $\overline{\tau}$), what can we do?

Idea: Estimate the **average treatment effect** as the difference between the average outcomes in the treatment group and the control group, *i.e.*,

$$Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

where $D_i = 1$ if i received treatment, and $D_i = 0$ if i is in the control group.

Review

Result: We showed that even when the treatment effect is constant (meaning $\tau_i = \tau$ for all i),

$$\begin{aligned} Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0) \\ = \tau + \underbrace{Avg(y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)}_{\text{Selection bias}} \end{aligned}$$

which says that the difference in the groups' means will give us a **biased estimate** for the causal effect of treatment **if we have selection bias.**

Causality

Review

Q: What is this **selection bias**?

A: (Informal) We have selection bias when our control group doesn't offer a good comparison for our treatment group.

Specifically, the control group doesn't give us a good **counterfactual** for what our treatment group would have looked like if the members had not received treatment. Basically, the groups are different.

A: (Formal) The average *untreated* outcome for a member of our **treatment group** (which we cannot observe) differs from the average *untreated* outcome for a member of our **control group**, *i.e.*,

$$Avg(y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

Causality

Review

What about using SCM or DAG?

We learned that a SCM is a DAG with a collection of random variable nodes and causal effect edges.

We also learned that an 'experiment' in a SCM is a **mutilation** of the graph...

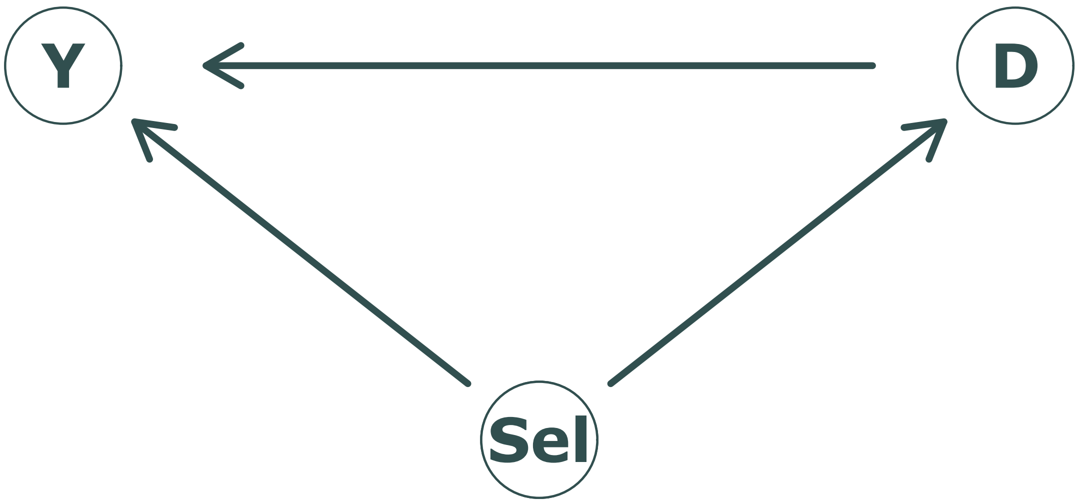
- delete all incoming causal edges to our 'intervened on' node.

We can think of this like our 'omitted variable bias' dag.

Causality

Review

Example Selection Bias



Selection Bias in a DAG

Causality

Review

Example If the Experiment went well...

Causality

Review

Practical problem: Selection bias is unfortunately, difficult to observe, so often we cannot simply condition on `Sel`. In Rubic-Neyman...

$$\underbrace{Avg(y_{0,i} \mid D_i = 1)}_{\text{Unobservable}} - Avg(y_{0,i} \mid D_i = 0)$$

(back to the *fundamental problem of causal inference*)

Bigger problem: If selection bias is present, our estimate for τ is biased, preventing us from understanding the causal effect of treatment.

Sounds a bit like omitted-variable bias, right? Our treatment variable is correlated with something that makes the two groups different.

Causality

Review

Example: Imagine we have two people—Al and Bri—and a single binary treatment, college. We interested in the effect of college on earnings.

$$\text{Earn}_{1,\text{Al}} = \$60\text{K} \qquad \text{Earn}_{1,\text{Bri}} = \$140\text{K}$$

$$\text{Earn}_{0,\text{Al}} = \$30\text{K} \qquad \text{Earn}_{0,\text{Bri}} = \$110\text{K}$$

The selection bias...

If Bri attended college (D_{Bri}=1) and Al did not (D_{Al}=0):

$$\hat{\tau} = \text{Earn}_{1,\text{Bri}} - \text{Earn}_{0,\text{Al}} = \$140\text{K} - \$30\text{K} = \$110\text{K}$$

If Al attended college (D_{Al}=1) and Bri did not (D_{Bri}=0):

$$\hat{\tau} = \text{Earn}_{1,\text{Al}} - \text{Earn}_{0,\text{Bri}} = \$60\text{K} - \$110\text{K} = -\$50\text{K}$$

Review

We have (at least) two problems...

1. Selection bias is difficult to observe
2. If selection bias is present, our estimate for τ is biased, preventing us from understanding the causal effect of treatment.

Solution: Eliminate/minimize selection bias.

- **Option 1: Distribute treatment** in a way such that the treatment and control groups are essentially identical (experiments).
- **Option 2: Build a control** group that *matches* the treatment group (life with observational data).

Instrumental variables

Instrumental variables

Intro

Instrumental variables (IV) is one route econometricians often take toward estimating the causal effect of a treatment/program.

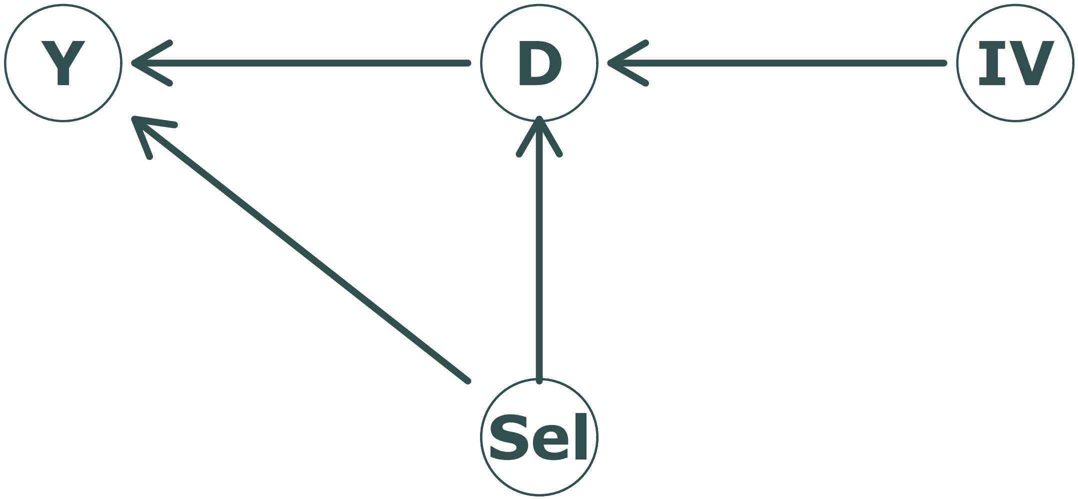
Recall: **Selection bias** means our treatment and control groups differ on some unobserved/omitted dimension. (**Endogeneity**)

Instrumental variables attempts to separate out

- the **exogenous** part of x, which gives us unbiased estimates
- the **endogenous** part of x, which biases our results

If we use only the exogenous (*good*) variation in x, then we can avoid selection bias/omitted-variable bias.

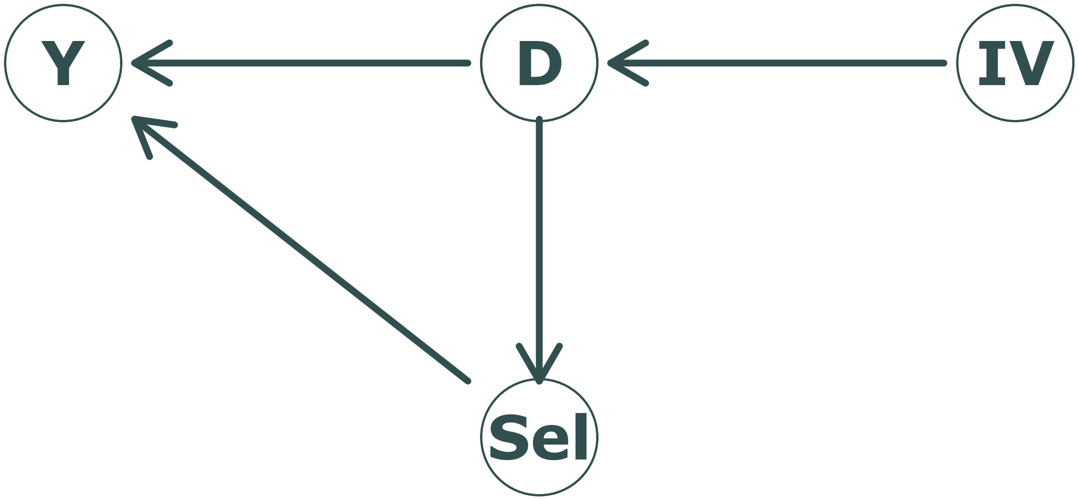
In a **DAG**



Works.

What if selection bias goes the other way?

In a **DAG**



Can Work... IF we make pretty specific assumptions about `IV`

Let's lay those out, shall we?

Introductory example

Example: If we want to estimate the effect of veteran status on earnings,

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Veteran}_i + u_i \tag{1}$$

We would love to calculate $\text{Earnings}_{1i} - \text{Earnings}_{0i}$, but we can't.

And OLS will likely be biased for (1) due to selection/omitted-variable bias.

Instrumental variables

Introductory example

Imagine that we can split veteran status into an exogenous part and an endogenous part...

$$
\begin{aligned}
\textbf{Earnings}_i &= \beta_0 + \beta_1 \textbf{Veteran}_i + u_i \\
&= \beta_0 + \beta_1 \left(\textbf{Veteran}_i^{\text{Exog.}} + \textbf{Veteran}_i^{\text{Endog.}} \right) + u_i \\
&= \beta_0 + \beta_1 \textbf{Veteran}_i^{\text{Exog.}} + \underbrace{\beta_1 \textbf{Veteran}_i^{\text{Endog.}} + u_i}_{w_i} \\
&= \beta_0 + \beta_1 \textbf{Veteran}_i^{\text{Exog.}} + w_i
\end{aligned} \tag{1}
$$

We could use this exogenous variation in veteran status to consistently estimate β_1.

Q: What would exogenous variation in veteran status mean?

Introductory example

Q: What would exogenous variation in veteran status mean?

A_1: Choices to enlist in the military that are essentially random—or at least uncorrelated with omitted variables and the disturbance.

A_2: No selection bias:

$$Avg(\text{Earnings}_{0i} \mid \text{Veteran}_i = 1) - Avg(\text{Earnings}_{0i} \mid \text{Veteran}_i = 0) = 0$$

Instrumental variables

Instruments

Q: How do we isolate this *exogenous variation* in our explanatory variable?

A: Find an instrument (an instrumental variable).

Q: What's an instrument?

A: An **instrument** is a variable that is

1. **correlated** with the **explanatory variable** of interest (**relevant**),
2. **uncorrelated** with the **disturbance** (**exogenous**).

Instrumental variables

Instruments

> **Q:** What's an instrument?
>
> **A:** An **instrument** is a variable that is
>
> 1. **correlated** with the **explanatory variable** of interest (**relevant**),
> 2. **uncorrelated** with the **disturbance** (**exogenous**). So if we want an instrument z_i for endogenous veteran status in
>
> $$\text{Earnings}_i = \beta_0 + \beta_1 \text{Veteran}_i + u_i$$

1. **Relevant:** $\text{Cov}(\text{Veteran}_i,\, z_i) \neq 0$
2. **Exogenous:** $\text{Cov}(z_i,\, u_i) = 0$

Instrumental variables

Instruments: Relevance

Relevance: We need the instrument to cause a change in (correlate with) our endogenous explanatory variable.

We can actually test this requirement using regression and a t test.

Example: For the veteran status, consider three potential instruments:

1. Social security number — **Probably not relevant**
   uncorrelated with military service
2. Physical fitness — **Potentially relevant**
   service may correlate with fitness
3. Vietnam War draft — **Relevant**
   being draw led to service

Instruments: Exogeneity

Exogeneity: The instrument to be independent of omitted factors that affect our outcome variable—as good as randomly assigned.

z_i must be uncorrelated with our disturbance u_i. **Not testable.**

Example: For the veteran status, consider three potential instruments:

1. Social security number — **Exogenous**
   Indep. of other factors of service
2. Physical fitness — **Not exogenous**
   fitness correlates with many things
3. Vietnam War draft — **Exogenous**
   the lottery was random

Instrumental variables

Instrumental review

Let's recap...

- Our instrument must be **correlated with our endogenous variable**.
- Our instrument must be **uncorrelated with any other variable that affects the outcome**.

In other words:

The instrument only affects our outcome through the endogenous variable.

Back to our example

For veteran status we considered three potential instruments:

1. Social security number — **Not relevant** **Exogenous**

2. Physical fitness — **Probably relevant** **Not exogenous**

3. Vietnam War draft — **Relevant** **Exogenous**

Thus, only the Vietnam War's draft lottery appears to be a ***valid* instrument**.

If we have a *valid* instrument (*e.g.*, the draft lottery), how do we use it?

Instrumental variables

Estimation

Recall: We want to estimate the effect of veteran status on earnings.

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Veteran}_i + u_i$$

Let's consider two related effects:

1. The effect of the **instrument** on the **endogenous variable**, *e.g.*,

$$\text{Veteran}_i = \gamma_0 + \gamma_1 \text{Draft}_i + v_i$$

2. The effect of the **instrument** on the **outcome variable**, *e.g.*,

$$\text{Earnings}_i = \pi_0 + \pi_1 \text{Draft}_i + w_i$$

Estimation

Recall: We want to estimate the effect of veteran status on earnings.

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Veteran}_i + u_i$$

and we know that the draft affected veteran status.

Draft $\longrightarrow$ **Veteran status** $\longrightarrow$ **Earnings**

Using our assumptions on independence and exogeneity:

(Effect of **the draft** on **earnings**) =

(Effect of **the draft** on **veteran status**)×

(Effect of **veteran status** on **earnings**)

Instrumental variables

Estimation

We just wrote out an expression for the effect of **the draft** on **earnings**, *i.e.*,

(Effect of **the draft** on **earnings**) =
(Effect of **the draft** on **veteran status**)×
(Effect of **veteran status** on **earnings**)

but we want to know the effect of **veteran status** on **earnings**. Rearrange!

$$\text{(Effect of veteran status on earnings)} = \frac{\text{(Effect of the draft on earnings)}}{\text{(Effect of the draft on veteran status)}}$$

Our **instrument** consistently estimates both parts of this fraction!

Estimation: Bring it all together

By estimating two regressions involving our **instrument**,

1. The effect of the **instrument** on the **endogenous variable**, *e.g.*,

$$\text{Veteran}_i = \gamma_0 + \gamma_1 \text{Draft}_i + v_i$$

2. The effect of the **instrument** on the **outcome variable**, *e.g.*,

$$\text{Earnings}_i = \pi_0 + \pi_1 \text{Draft}_i + w_i$$

we can estimate our desired effect:

(Effect of **veteran status** on **earnings**) = $\dfrac{\pi_1}{\gamma_1}$

Estimation: Bring it all together

So with instrumental variables, we estimate β_1 using

$$\hat{\beta}_1^{\text{IV}} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$$

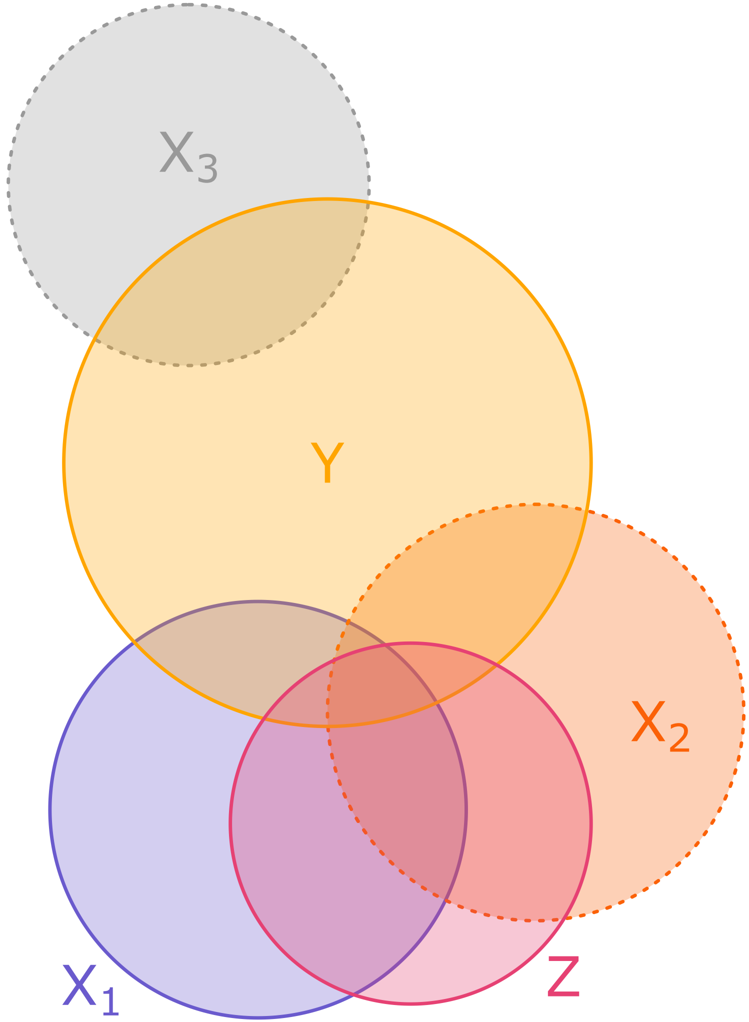
where $\hat{\pi}_1$ and $\hat{\gamma}_1$ come from the two equations we just discussed.

Q: Can we trust $\hat{\beta}_1^{\text{IV}}$?
A: Yes... **if we have a valid instrument.**

$$\text{plim}\left(\hat{\beta}_1^{\text{IV}}\right) = \beta_1 + \frac{\text{Cov}(\text{Instrument},\, u)}{\text{Cov}(\text{Instrument},\, \text{Endog. variable})}$$

which equals β_1 as long as our instrument is **exogenous** (numerator) and **relevant** (denominator).

Figure 4



Venn diagram explanation

In these figures (Venn diagrams)

- Each circle illustrates a variable.
- Overlap gives the share of correlatation between two variables.
- Dotted borders denote *omitted* variables.

Take-aways

- Figure 1: **Valid instrument** (relevant; exogenous)
- Figure 2: **Invalid instrument** (relevant; not exogenous)
- Figure 3: **Invalid instrument** (not relevant; not exogenous)
- Figure 4: **Invalid instrument** (relevant; not exogenous)

Let's work an example in `R`.

Instrumental variables

Example in R

Back to our age-old battle to estimate the returns to education.

Show 4 entries Search:

	wage	education	education_dad	education_mom
1	769	12	8	8
2	808	18	14	14
3	825	14	14	14
4	650	12	12	12

Showing 1 to 4 of 722 entries

Previous 1 2 3 4 5 … 181 Next

Instrumental variables

Example in R

OLS for the returns to education with will likely (definitely) be biased...

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + u_i$$

(Likely biased) OLS results:

term	estimate	std.error	statistic	p.value
(Intercept)	177	89.2	1.98	0.0481
education	58.6	6.44	9.1	8.76e-19

but what if mother's education provides a valid instrument?

Instrumental variables

Example in R

We can check/test the *relevance* of **mother's education** for **education**.

This regression is known as the ***first stage*:**

The effect of the instrument on our endogenous explanatory variable.

$$\text{Education}_i = \gamma_0 + \gamma_1 (\text{Mother's Education})_i + v_i$$

Example in R

We can check/test the *relevance* of **mother's education** for **education**.

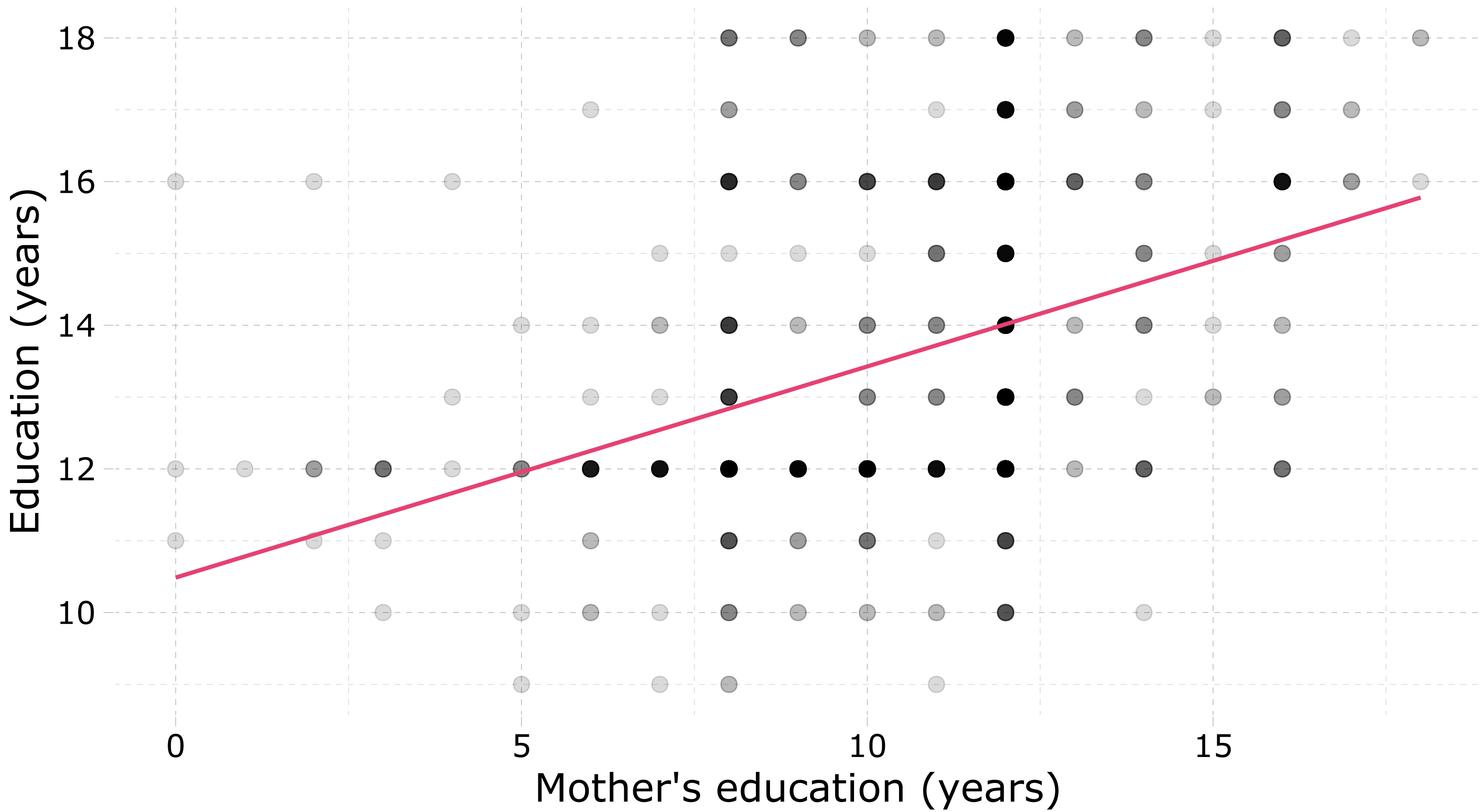
First-stage results:

term	estimate	std.error	statistic	p.value
(Intercept)	10.5	0.306	34.3	1.13e-153
education_mom	0.294	0.0274	10.7	4.41e-25

The *p*-value suggests a very strong relationship (very *relevant*).

Instrumental variables

Visualizing the first stage



Instrumental variables

Exogeneity

Q: What does **exogeneity** mean in this case?

A: We need two things

1. Mother's education (our instrument) must only affect earnings through education (our endogenous explanatory variable).
2. Mother's education must be uncorrelated with other factors that affect wages (our outcome variable).

We want to be able to compare two people (A and B) whose mothers have different levels of education and say that the only differences between the two people (A and B) are due to their mothers' educational levels.

Q: Does *mother's education* seem likely to satisfy exogeneity?

Instrumental variables

Example in R

Now, let's estimate the ***reduced form***:

The effect of the instrument on our outcome variable.

$$\text{Wage}_i = \pi_0 + \pi_1 (\text{Mother's Education})_i + w_i$$

Reduced-form results:

term	estimate	std.error	statistic	p.value
(Intercept)	633	58.6	10.8	2.37e-25
education_mom	31.8	5.24	6.07	2.12e-09

Example in R

So what is our IV-based estimate for the returns to education?

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + u_i$$

We know that the IV estimate for β_1 is

$$\hat{\beta}_1^{\text{IV}} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$$

1. In the **reduced-form equation**, we estimated $\hat{\pi}_1 \approx 31.81$.
2. In the **first-stage equation**, we estimated $\hat{\gamma}_1 \approx 0.294$.

$$\implies \hat{\beta}_1^{\text{IV}} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{31.81}{0.294} \approx 108.2$$

Instrumental variables

Example in R

Alternative: Use the function `iv_robust()` from the `estimatr` package.

This new function `iv_robust` works very similar to our good friend `lm`:

```
iv_robust(y ~ x | z, data = dataset)
```

- `formula` Specify the regression followed by `|` and your instrument (`z`).
- `data` You still need a dataset.

Note: As you might guess by its name, `iv_robust` calculates heteroskedasticity-robust standard errors by default.

Instrumental variables

Example in R

In practice...

```
# Estimate our IV regression
iv_est ← iv_robust(wage ~ education | education_mom, data = wage_df)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-501	226	-2.21	0.0271
education	108	16.8	6.44	2.22e-10

More

So now we know how to "do" instrumental variables *when we have one endogenous variable and one exogenous variable.*

1. Estimate the reduced form (regress outcome var. on instrument).
2. Estimate the first stage (regress expl. var. on instrument).
3. Calculate the IV estimate using the estimates from (1) and (2).

Our magical instrument isolates the exogenous variation in our endogenous variable.

Q: What if we want more? (*E.g.*, more instruments or endog. variables)

A: ~~Too bad.~~ Extend IV to **two-stage least squares (2SLS)**.

Two-stage least squares

Two-stage least squares

Intro

The intuition and insights of IV carry over into two-stage least squares.

Plus: The *first stage* that we've been discussing is actually the *first* of the *two stages* in two-stage least squares.

$$
\begin{aligned}
\text{Endogenous model} \qquad \text{Outcome}_i &= \beta_0 + \beta_1 \left(\text{Endog. var.}\right)_i + u_i \\
\textbf{First stage} \qquad \left(\text{Endog. var.}\right)_i &= \pi_0 + \pi_1 \text{Instrument}_i + v_i \\
\textbf{Second stage} \qquad \text{Outcome}_i &= \delta_0 + \delta_1 \widehat{\left(\text{Endog. var.}\right)}_i + \varepsilon_i \\
\text{Reduced form} \qquad \text{Outcome}_i &= \pi_0 + \pi_1 \text{Instrument}_i + w_i
\end{aligned}
$$

where $\widehat{\left(\text{Endog. var.}\right)}_i$ denotes the predicted values (*fitted values*) from the first-stage regression.

Two-stage least squares

Intro

Two-stage least squares is very flexible—we include other controls, additional endogenous variables, *and* have multiple instruments.

But don't get too distracted by this fancy flexiblity.

We still need **valid** instruments.

Two-stage least squares

In R

Back to our *returns to education* example.

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + u_i$$

Imagine that mother's *and* father's education are both valid instruments.

Then our **first-stage regression** is

$$\text{Education}_i = \gamma_0 + \gamma_1 (\text{Mother's education})_i + \gamma_2 (\text{Father's education})_i + v_i$$

which we can estimate via OLS.

Q: Why?

Two-stage least squares

In R

$$\color{purple}{\text{Education}_i} = \gamma_0 + \gamma_1 \color{red}{(\text{Mother's education})_i} + \gamma_2 \color{red}{(\text{Father's education})_i} + v_i$$

```
stage1 ← lm(education ~ education_mom + education_dad, wage_df)
```

First-stage results:

term	estimate	std.error	statistic	p.value
(Intercept)	9.85	0.305	32.3	3.59e-142
education_mom	0.149	0.0322	4.62	4.47e-06
education_dad	0.216	0.0275	7.84	1.67e-14

Our instruments each appear to be *relevant*. Formally, we want an F test.

Two-stage least squares

In R

Using our estimated first stage, we grab the *fitted* endogenous variable

$$\widehat{\text{Education}}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 (\text{Mother's education})_i + \widehat{\gamma}_2 (\text{Father's education})_i$$

```r
# Add fitted values from first stage
wage_df$education_hat <- stage1$fitted.values
```

Now we use OLS (again) to estimate the **second-stage regression**

$$\text{Wage}_i = \delta_0 + \delta_1 \widehat{\text{Education}}_i + \varepsilon_i$$

Two-stage least squares

In R

$$\color{orange}{\text{Wage}_i} = \delta_0 + \delta_1 \color{purple}{\widehat{\text{Education}}_i} + \varepsilon_i$$

```
stage2 ← lm(wage ~ education_hat, wage_df)
```

Second-stage results:

term	estimate	std.error	statistic	p.value
(Intercept)	-455	198	-2.29	0.022
education_hat	105	14.5	7.25	1.11e-12

Ordinary least squares

term	estimate	std.error	statistic	p.value
(Intercept)	177	89.2	1.98	0.0481
education	58.6	6.44	9.1	8.76e-19

Instrumental variables

term	estimate	std.error	statistic	p.value
(Intercept)	-501	226	-2.21	0.0271
education	108	16.8	6.44	2.22e-10

Two-stage least squares

In R

As you probably guessed, R will do both of the stages for you.

```
iv_robust(y ~ x1 + x2 + … | z1 + z2 + …, data)
```

In our case, we have

- one explanatory variable (`x`) (education)
- two instruments (`z`) (parents' educations)

Let's see how to do this in R

```
iv_robust(wage ~ education | education_mom + education_dad, data = wage_df)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-455	200	-2.27	0.0233
education	105	14.9	7.06	4.05e-12

Two-stage least squares

There's more!

Because 2SLS **isolates exogenous variation in an endogenous variable**, we apply it in other settings that are biased from *endogenous* relationships.

Common applications

- **General causal inference** for observational data (as we've seen).
- **Experiments:** Randomize a treatment that affects an endog. variable.
- **Measurement error:** Regress noisy x_1 on noisy x_2 to capture signal.
- **Simultaneous relationships** (*e.g.*, p and q from supply and demand).

However, in any 2SLS/IV setting, you need to mind the requirements for **valid instruments**—exogeneity and relevance.

Table of contents

Admin

Instrumental variables