Instrumental Variables

EC 421, Set 11

Edward Rubin
05 March 2019

Prologue

Schedule

Last Time

Causality

Today

- Econ. Masters program
- Review: Causality
- New: Instrumental variables

Upcoming

Assignment soon.

Master's Program

Applied Economics

You could be a master of (applied) economics...

- 1-year program including courses on applied econometrics, data science, and "big data".
- Awesome opportunity to focus on applying economic methods to real-world questions/scenarios.
- Applications are due by **May 1 (2019)**.

More information: https://economics.uoregon.edu/masters/

Causality

Review

Causality

Review

In our last lecture, we returned to the concept of **causality**.

Causality

Review

In our last lecture, we returned 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

Causality

Review

In our last lecture, we returned 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

Causality

Review

In our last lecture, we returned 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*.

Review

In our last lecture, we returned 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.*,

Causality

Review

In our last lecture, we returned 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}.

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.*,

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

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.*,

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

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?

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),

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

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**?

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.

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.

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.

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.*,

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

Review

Practical problem: Selection bias is also difficult to observe

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

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

Review

Practical problem: Selection bias is also difficult to observe

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

(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.

Review

Practical problem: Selection bias is also difficult to observe

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

(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?

Causality

Review

Practical problem: Selection bias is also difficult to observe

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

(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 makese 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.

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

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

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.

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

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

They both have the same treatment effect (return to college)

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.

$$\text{Earn}_{1,\text{Al}} = \$60\text{K} \qquad \text{Earn}_{1,\text{Bri}} = \$140\text{K}$$
$$\text{Earn}_{0,\text{Al}} = \$30\text{K} \qquad \text{Earn}_{0,\text{Bri}} = \$110\text{K}$$

They both have the same treatment effect (return to college)

$$\tau_{\text{Al}} = \text{Earn}_{1,\text{Al}} - \text{Earn}_{0,\text{Al}} = \$60\text{K} - \$30\text{K} = \$30\text{K}$$

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.

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

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

They both have the same treatment effect (return to college)

$$\tau_{\text{Al}} = \text{Earn}_{1,\text{Al}} - \text{Earn}_{0,\text{Al}} = \$60\text{K} - \$30\text{K} = \$30\text{K}$$

$$\tau_{\text{Bri}} = \text{Earn}_{1,\text{Bri}} - \text{Earn}_{0,\text{Bri}} = \$140\text{K} - \$100\text{K} = \$30\text{K}$$

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.

$$\text{Earn}_{1,\text{Al}} = \$60\text{K} \qquad \text{Earn}_{1,\text{Bri}} = \$140\text{K}$$
$$\text{Earn}_{0,\text{Al}} = \$30\text{K} \qquad \text{Earn}_{0,\text{Bri}} = \$110\text{K}$$

They both have the same treatment effect (return to college)

$$\tau_{\text{Al}} = \text{Earn}_{1,\text{Al}} - \text{Earn}_{0,\text{Al}} = \$60\text{K} - \$30\text{K} = \$30\text{K}$$
$$\tau_{\text{Bri}} = \text{Earn}_{1,\text{Bri}} - \text{Earn}_{0,\text{Bri}} = \$140\text{K} - \$100\text{K} = \$30\text{K}$$

but any real-world estimate would have serious selection issues since $\text{Earn}_{0,\text{Al}} \neq \text{Earn}_{0,\text{Bri}}$.

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...

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.

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

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

The selection bias...

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

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

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):

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

Causality

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.

Causality

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.

Causality

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

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).

Causality

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

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.

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

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

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.

Introductory example

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

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

Instrumental variables

Introductory example

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

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

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

Instrumental variables

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…

Instrumental variables

Introductory example

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

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

Introductory example

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

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

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

Introductory example

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

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

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

Q: What would exogenous variation in veteran status mean?

Instrumental variables

Introductory example

Q: What would exogenous variation in veteran status mean?

Instrumental variables

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.

Instrumental variables

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?

Instruments

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

A: Find an instrument (an instrumental variable).

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?

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**).

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.

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

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

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

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

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

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.**

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

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

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

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

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

Instrumental variables

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**.

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.

Instrumental variables

Back to our example

For veteran status we considered three potential instruments:

Instrumental variables

Back to our example

For veteran status we considered three potential instruments:

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

   Exogenous

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

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**

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?

Estimation

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

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

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:

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.*,

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

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.*,

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

Instrumental variables

Estimation

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

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

and we know that the draft affected veteran status.

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**)

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**.

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!

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!

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

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.*,

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

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:

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**) =

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

Instrumental variables

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**) = $\dfrac{\pi_1}{\gamma_1}$

Let's work an example in `R`.

Instrumental variables

Example in R

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

```
#> # A tibble: 722 x 4
#>     wage  educ educ_dad educ_mom
#>    <int> <int>    <int>    <int>
#>  1   769    12        8        8
#>  2   808    18       14       14
#>  3   825    14       14       14
#>  4   650    12       12       12
#>  5   562    11       11        6
#>  6   600    10        8        8
#>  7  1154    15        5       14
#>  8  1000    12       11       12
#>  9   930    18       14       13
#> 10   900    15       12       12
#> # … with 712 more rows
```

Instrumental variables

Example in R

OLS for the returns to education is likely (definitely) biased...

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    177.      89.2       1.98 4.81e- 2
#> 2 educ            58.6      6.44      9.10 8.76e-19
```

but what if father's or mother's education provide a valid instrument?

Instrumental variables

Example in R

OLS for the returns to education is likely (definitely) biased...

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    177.      89.2       1.98 4.81e- 2
#> 2 educ            58.6      6.44      9.10 8.76e-19
```

but what if father's or mother's education provide a valid instrument?

Q: Why/why not?

Instrumental variables

Example in R

Live coding here.

Table of contents