Instrumental Variables

Prof. Jacob M. Montgomery

Quantitative Political Methodology (L32 363)

November 20, 2017

Road map

Where we have been:

- What is regression?
- How to interpret coefficients?
- Interactions/Dummies
- Regression assumptions
- Using regression for causal inference
- Using difference-in-differences to make causal claims
- Regression discontinuity

Today:

- Instrumental variables

Instrumental variables (IV) analysis: Two frameworks

- How to handle non-compliance in experiments

Instrumental variables (IV) analysis: Two frameworks

- How to handle non-compliance in experiments
  - What do you do if some people in an experiment don't do what they are told?

Instrumental variables (IV) analysis: Two frameworks

- How to handle non-compliance in experiments
  - What do you do if some people in an experiment don't do what they are told?
- How to make causal inference in the presence of endogenous regressors

Instrumental variables (IV) analysis: Two frameworks

- How to handle non-compliance in experiments
  - What do you do if some people in an experiment don't do what they are told?
- How to make causal inference in the presence of endogenous regressors
  - An approach that can sometimes work when you can't do anything else

Framework 1: Who get's the milk?

A non-hypothetical example: The setup

Let's imagine a nutrition intervention:

- Randomly assign schools to get extra provisions of school milk at lunch
- At all schools teachers allocate milk and keep track of who gets it
- After one year, follow up and measure weights of all children

A non-hypothetical example: The hitch

Two-sided noncompliance

- People in the treatment condition failed to receive the treatment

A non-hypothetical example: The hitch

Two-sided noncompliance

- People in the treatment condition failed to receive the treatment
  - Some kids don't like milk

A non-hypothetical example: The hitch

Two-sided noncompliance

- People in the treatment condition failed to receive the treatment
  - Some kids don't like milk
  - More importantly, who would you give milk to?

A non-hypothetical example: The hitch

Two-sided noncompliance

- People in the treatment condition failed to receive the treatment
  - Some kids don't like milk
  - More importantly, who would you give milk to?
- People in the control condition receive the treatment

A non-hypothetical example: The hitch

Two-sided noncompliance

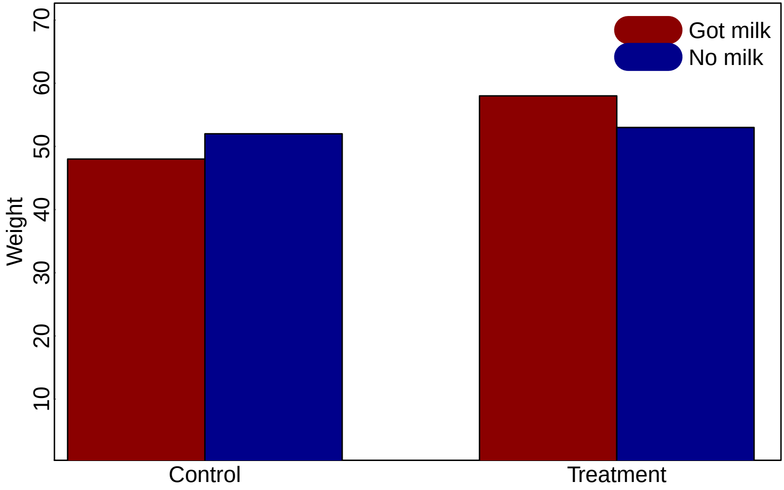
- People in the treatment condition failed to receive the treatment
  - Some kids don't like milk
  - More importantly, who would you give milk to?
- People in the control condition receive the treatment
  - Some people have cows

A non-hypothetical example: The hitch

Two-sided noncompliance

- People in the treatment condition failed to receive the treatment
  - Some kids don't like milk
  - More importantly, who would you give milk to?
- People in the control condition receive the treatment
  - Some people have cows
  - Rich people

Hypothetical experimental results



Instrumental variables

We are going to solve two equations at the same time:

$$x_i = \tau + T_i\gamma + \epsilon_{i1} \tag{1}$$

$$y_i = \alpha + x_i\beta + \epsilon_{i2} \tag{2}$$

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N(\mathbf{0}, \Sigma)$$

How does this work? In the abstract

1. We have an endogenous regressor x and we want to know how it affect y.
2. We have a randomly assigned variable T that has a strong effect on x
3. T only affects y through x (exclusion restriction)
4. No one does the opposite on purpose (no defiers)

If we have all of these, we can correctly estimate β ...

How does this work? In the abstract

1. We have an endogenous regressor x and we want to know how it affect y.
2. We have a randomly assigned variable T that has a strong effect on x
3. T only affects y through x (exclusion restriction)
4. No one does the opposite on purpose (no defiers)

If we have all of these, we can correctly estimate β ...
... but only for the subset of observations that are compliers.

How does this work? In our example

1. We have an endogenous regressor x (milk consumption) and we want to know how it affect y (weight).
2. We have a randomly assigned variable T (which schools get lunch) that has a strong effect on x (milk consumption)
3. T only affects y through x (assignment not related to weight except through milk)
4. No one does the opposite on purpose (no one drinks extra milk because their school was added to the control)

If we have all of these, we can correctly estimate β.

How does this work? In our example

1. We have an endogenous regressor x (milk consumption) and we want to know how it affect y (weight).
2. We have a randomly assigned variable T (which schools get lunch) that has a strong effect on x (milk consumption)
3. T only affects y through x (assignment not related to weight except through milk)
4. No one does the opposite on purpose (no one drinks extra milk because their school was added to the control)

If we have all of these, we can correctly estimate β.
. . . but only for the subset of observations that are compliers.

Framework 2: Guns and money

Some things can't be randomized

- Bad economy $\rightarrow$ more war

Some things can't be randomized

- Bad economy $\rightarrow$ more war
- More war $\rightarrow$ bad economy

Some things can't be randomized

- Bad economy $\rightarrow$ more war
- More war $\rightarrow$ bad economy
- Bad government $\rightarrow$ more war, bad economy

Instrumental variables

We are going to solve two equations at the same time:

$$x_i = \tau + T_i\gamma + \epsilon_{i1} \tag{3}$$

$$y_i = \alpha + x_i\beta + \epsilon_{i2} \tag{4}$$

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N(\mathbf{0}, \Sigma)$$

How does this work? In the abstract

1. We have an endogenous regressor x and we want to know how it affect y.
2. We have a randomly assigned variable T that has a strong effect on x
3. T only affects y through x (exclusion restriction)
4. No one does the opposite on purpose (no defiers)

If we have all of these, we can correctly estimate $\beta \ldots$

How does this work? In the abstract

1. We have an endogenous regressor x and we want to know how it affect y.
2. We have a randomly assigned variable T that has a strong effect on x
3. T only affects y through x (exclusion restriction)
4. No one does the opposite on purpose (no defiers)

If we have all of these, we can correctly estimate β . . .
. . . but only for the subset of observations that are compliers.

Using rainfall as an instrument

Using rainfall as an instrument

How does this work? In our example

1. We have an endogenous regressor x (economy) and we want to know how it affect y (civil war).
2. We have a randomly assigned variable T (rainfall in the previous year) that has a strong effect on x (economic growth)
3. T only affects y through x (rain does not directly affect civil war)
4. No one does the opposite on purpose (countries do not have negative economic growth as a consequence of lot's of rain)

If we have all of these, we can correctly estimate β.

How does this work? In our example

1. We have an endogenous regressor x (economy) and we want to know how it affect y (civil war).
2. We have a randomly assigned variable T (rainfall in the previous year) that has a strong effect on x (economic growth)
3. T only affects y through x (rain does not directly affect civil war)
4. No one does the opposite on purpose (countries do not have negative economic growth as a consequence of lot's of rain)

If we have all of these, we can correctly estimate β.
... but only for the subset of observations that are compliers.

R code for two staged least squares

```
library(AER)
ivreg(y ~ x| T, data = myData)
```

Negative ads and turnout

Ansolabehere, Iyengar, and Simon (1999), "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout," *American Political Science Review*.

The basic approach

Using survey data from the 1992 election

- y: Self-reported probability of voting

The basic approach

Using survey data from the 1992 election

- y: Self-reported probability of voting
- x: Self-reported ad tone

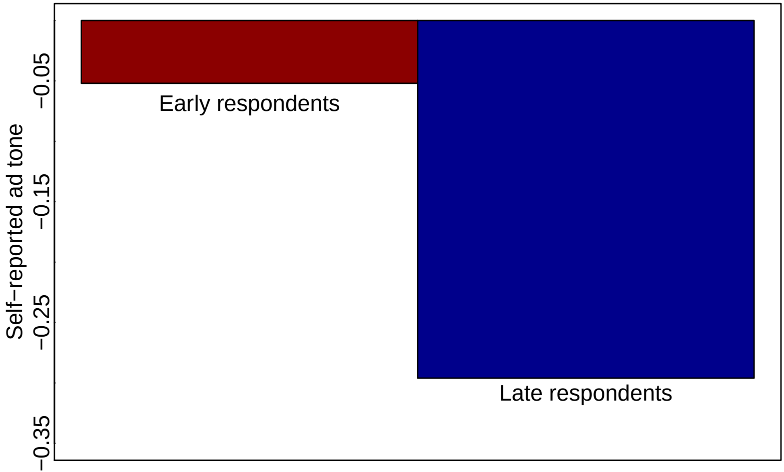
The basic approach

Using survey data from the 1992 election

- y: Self-reported probability of voting
- x: Self-reported ad tone
- T: Date of the interview

Check our assumptions

Assumption 1: We have a randomly assigned variable T that has a strong effect on x



Check our assumptions

- T only affects y through x
- No one does the opposite on purpose

2SLS Results

	2SLS Instrumental Variables for Pos & Neg	2SLS Instrumental Variables for Neg Only
Recall of positive ad	.527	.017
	(.496)	(.022)
	−.184	−.090
Recall of negative ad	(.093)	(.049)

2SLS Results

	2SLS Instrumental Variables for Pos & Neg	2SLS Instrumental Variables for Neg Only
Recall of positive ad	.527	.017
	(.496)	(.022)
	−.184	−.090
Recall of negative ad	(.093)	(.049)

Remember, this is the effect of the treatment on compliers.