Instrumental Variables Regression

Kritee Gujral, PhD

Feb 1st , 2023

I acknowledge the contributions of Dr. Christine Chee and Dr. Lindsey Woodworth for the preparation of this course.

HERC Health Economics Resource Center

VA HSR&D HEALTH SERVICES RESEARCH & DEVELOPMENT

Outline

- Recall Endogeneity
- Introduce Instrumental Variables Regression
  - Intuition
  - Regression
  - Assessing instrument validity
  - Implementation using one example
  - More examples
- Summary

Introduction: Estimating Causal Effects

- A common aim of health services research is the estimation of a causal effect
  - What is the effect of *[treatment]* on *[outcome]*?
- Ideally estimate the effect using a randomized controlled trial
  - Conducting a randomized controlled trial is often not possible
- An alternative is to perform regression analysis using observational data
  - To estimate the causal effect of *[treatment]* on *[outcome]*, unobserved variables must not be driving the outcome, i.e. treatment must be *exogenous*

Recall: Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

- Y: outcome variable of interest
- X: explanatory variable of interest or *treatment*
- e: error term
  - e contains all other factors besides X that determine the value of Y
- β_1: the change in Y associated with a unit change in X
- In order for $\widehat{\beta_1}$ to be an unbiased estimate of the *causal effect* of X on Y, X must be **exogenous**

Recall: Exogeneity

- Assumption: $E\left(e_i \mid X_i\right) = 0$
  - Conditional mean of e_i given X_i is zero
  - Additional info. in e_i does not help us better predict Y_i
  - X is "exogenous"
  - Implies that X_i and e_i **cannot** be correlated

- X_i and e_i are correlated when there is:
  - Omitted variable bias
  - Sample selection
  - Simultaneous causality

- If X_i and e_i are correlated then X is endogenous
  - $\widehat{\beta_1}$ is biased

Introducing Instrumental Variables (IV)

- When X or treatment is not exogenous, another method is necessary for estimating the causal effect of X or treatment on Y.

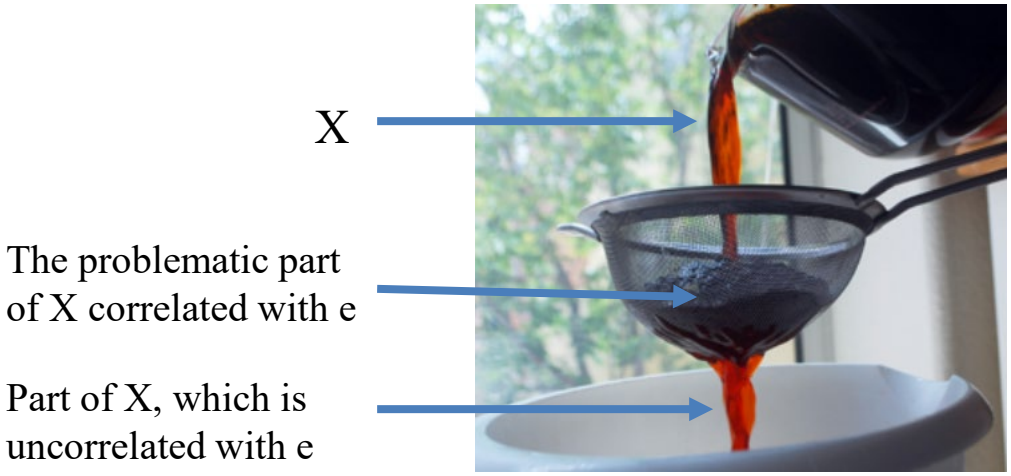
- One possibility: **instrumental variables (IV) regression**

IV Regression: Intuition

- $Y_i = \beta_0 + \beta_1 X_i + e_i$
- X is endogenous
- Think of variation in X having two components
  - One component is correlated with e - Causes endogeneity
  - Other component is uncorrelated with e - "Exogenous" variation
- An instrumental variable is a variable that uses only the exogenous variation in X to estimate β_1

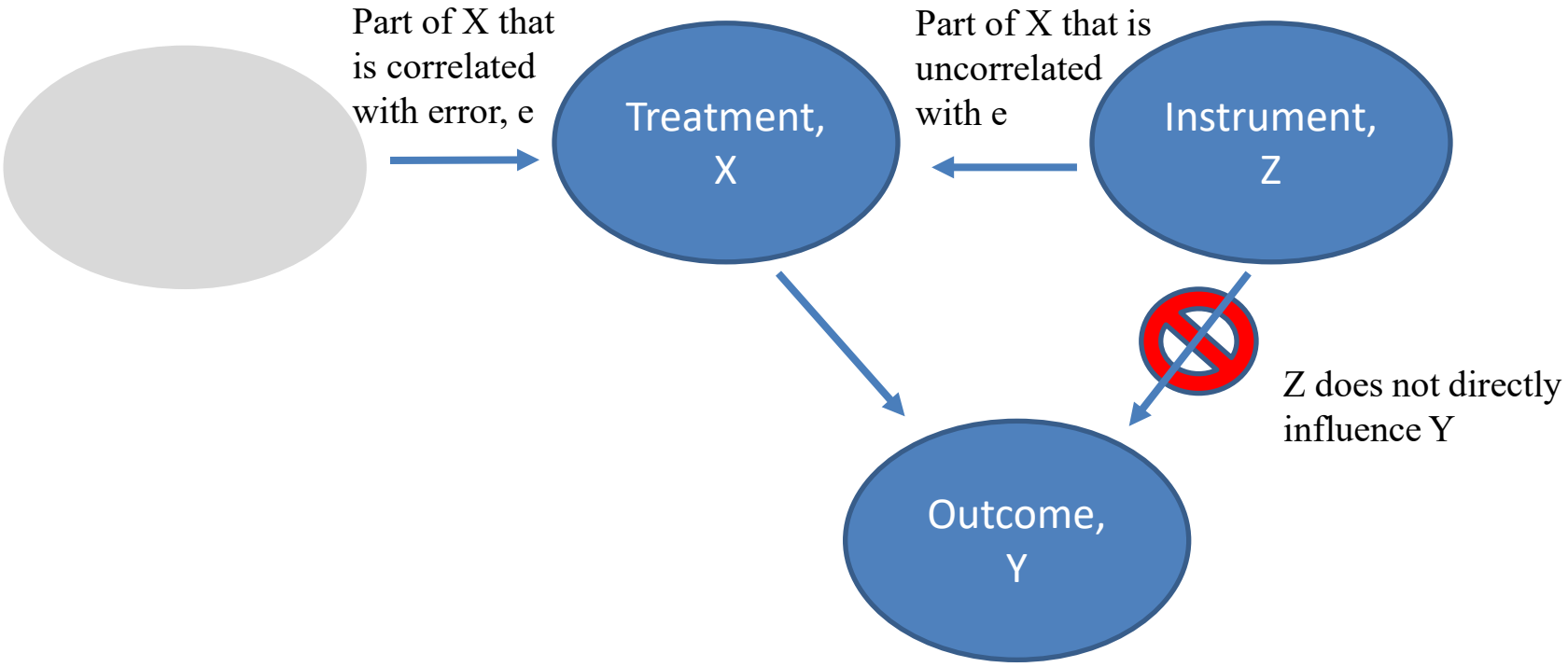
IV Regression: Intuition

- We want to isolate the exogenous variation in X that is uncorrelated with *e*



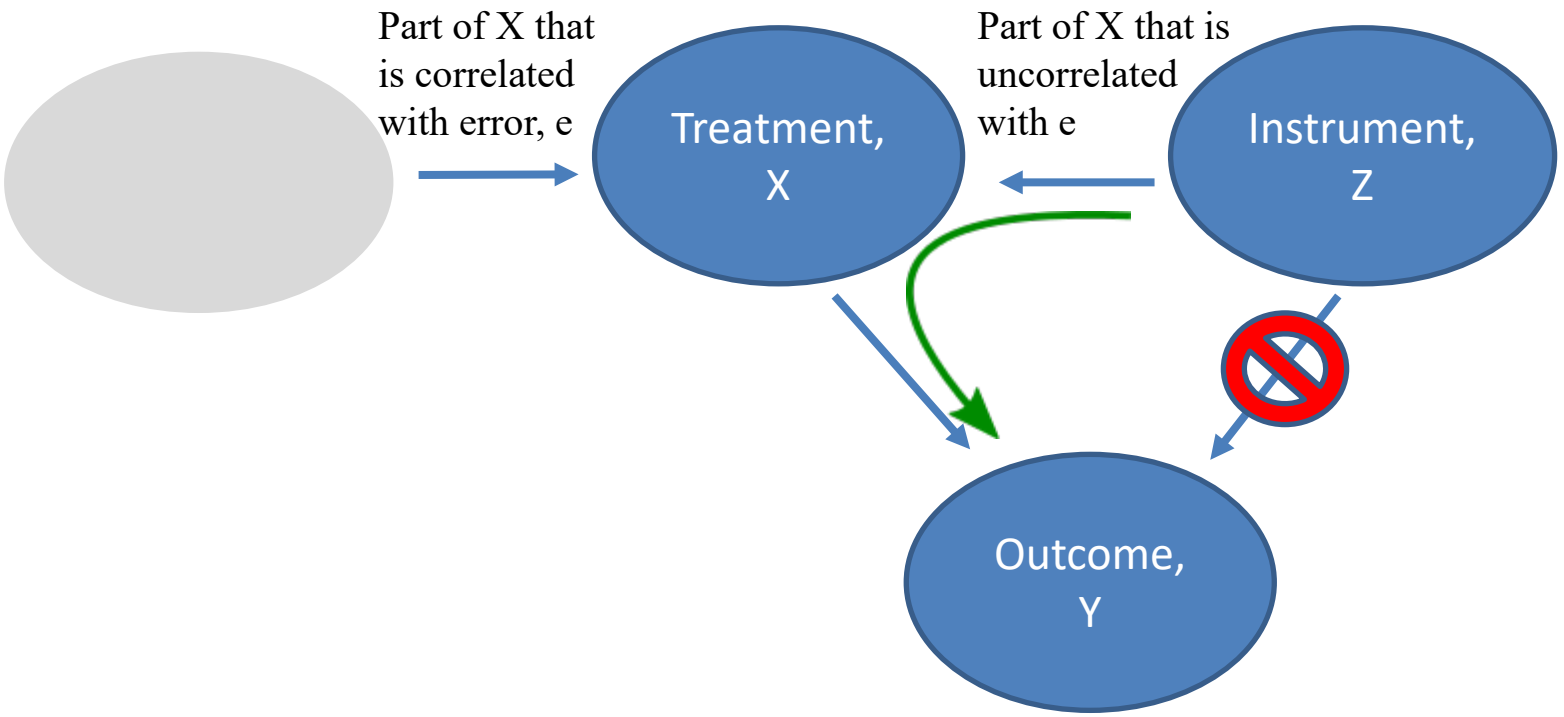
IV Regression: Intuition

- Recall that variation in X has two components
- An instrument, Z, is a variable that can capture only the exogenous variation in X – we need to look for such a variable!



IV Regression: Intuition

- Z can be used to isolate the exogenous variation in X. Since Z is itself exogenous, its correlation with X is exogenous.



IV Regression: Two Stage Least Squares (TSLS)

- Two consecutive OLS regressions
- First stage:
  - Regress X on Z:

$$X_i = \underbrace{\pi_0 + \pi_1 Z_i}_{\text{uncorrelated with } e} + \underbrace{\gamma_i}_{\text{correlated with } e}$$

  - Predict X:

$$\widehat{X_i} = \widehat{\pi_0} + \widehat{\pi_1} Z_i$$

IV Regression: Two Stage Least Squares (TSLS)

- Second stage:
  - Regress Y on $\hat{X}$

$$Y_i = \beta_0^{TSLS} + \beta_1^{TSLS} \hat{X}_i + e_i$$

  - Estimate $\widehat{\beta_1}^{TSLS}$
    - $\hat{X}$ is uncorrelated with e from the original regression model $Y_i = \beta_0 + \beta_1 X_i + e_i$
    - $\widehat{\beta_1}^{TSLS}$ is an unbiased estimate of β_1
    - Note: standard errors in the second stage TSLS regression need to be adjusted

IV Reg.: Generalizes to case of ≥ 1 endogenous regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + e_i$$

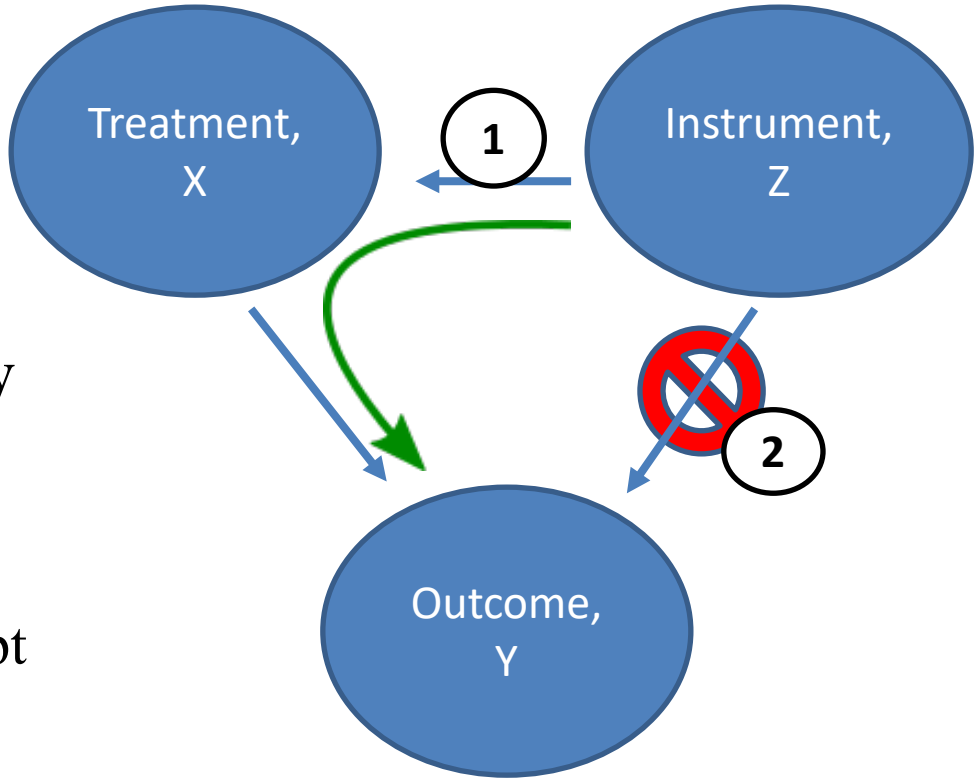
- k endogenous regressors: $X_{1i}, \ldots, X_{ki}$
- r exogenous regressors or control variables: $W_{1i}, \ldots, W_{ri}$
- m instrumental variables: $Z_{1i}, \ldots, Z_{mi}$

- There must be at least as many instruments as there are endogenous variables: $m \geq k$

How to identify a valid instrument?

Two conditions:

1) **Instrument relevance**
   - Z is correlated with X, Corr(Zi, Xi) $\neq 0$
2) **Instrument exogeneity**
   - Z must be uncorrelated with e, Corr(Zi, ei) = 0
   - Z does not affect Y except through Z's correlation with X



Violation of condition 1/ relevance: weak instruments

- Instruments that explain little variation in X are **weak**
- IV regression with weak instruments provide unreliable estimates
- Can test for weak instruments using a rule of thumb :
  - Regress X on Z
  - F-statistic > 10 indicates instruments are not weak
  - Note: this is a rule of thumb; we still need a convincing argument that the instrument is relevant (strong)

Violation of condition 2/ exogeneity: endogenous instruments

- Instruments that are correlated with the error term (other factors that affect the outcome variable) are **endogenous**
- IV regression with endogenous instruments provide unreliable estimates
- Infeasible to formally test for endogenous instruments - need a convincing argument that the instruments are exogenous

IV Regression: Implementation

Wage= $\alpha + \beta_1$Education + β_2Experience+ ε

First, simple OLS without instrumental variables:

```
. reg wage educ exper
```

Source	SS	df	MS		
				Number of obs	= 935
				F(2, 932)	= 73.26
Model	20747023.1	2	10373511.5	Prob > F	= 0.0000
Residual	131969145	932	141597.795	R-squared	= 0.1359
				Adj R-squared	= 0.1340
Total	152716168	934	163507.675	Root MSE	= 376.29

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	76.21639	6.296604	12.10	0.000	63.85922	88.57355
exper	17.63777	3.161775	5.58	0.000	11.43275	23.84279
_cons	-272.5279	107.2627	-2.54	0.011	-483.0323	-62.02344

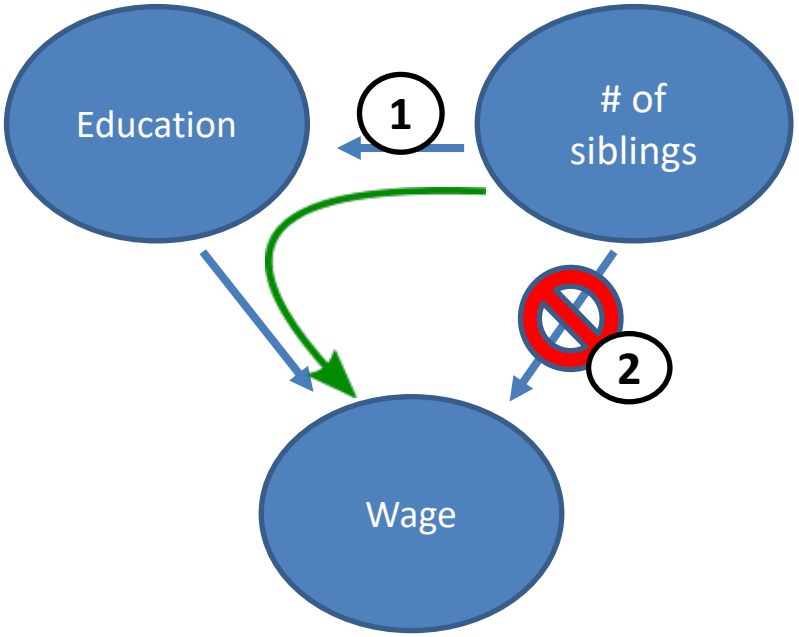
We're concerned that education may be endogenous. A person's innate intellectual ability could be driving both education and wages. $\widehat{\beta_1}$ will be biased.

IV Regression: Recall Intuition

Now consider using an instrumental variable: # of siblings

$$\text{Wage} = \alpha + \beta_1 \text{Education} + \beta_2 \text{Experience} + \varepsilon$$

# of siblings



First Stage of TSLS:

```
. reg educ exper sibs
```

Source	SS	df	MS		
Model	1134.9333	2	567.466652	Number of obs	= 935
Residual	3371.88595	932	3.61790338	F(2, 932)	= 156.85
				Prob > F	= 0.0000
				R-squared	= 0.2518
				Adj R-squared	= 0.2502
Total	4506.81925	934	4.82528828	Root MSE	= 1.9021

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	-.2219521	.0142567	-15.57	0.000	-.2499309	-.1939732
sibs	-.2008413	.0270426	-7.43	0.000	-.2539127	-.1477699
_cons	16.62573	.1889113	88.01	0.000	16.25499	16.99647

```
. predict educHat , xb
```

Second Stage of TSLS (note standard errors are incorrect):

```
. reg wage educHat exper
```

Source	SS	df	MS		
Model	3894404.63	2	1947202.32	Number of obs	= 935
Residual	148821764	932	159680.004	F(2, 932)	= 12.19
				Prob > F	= 0.0000
				R-squared	= 0.0255
				Adj R-squared	= 0.0234
Total	152716168	934	163507.675	Root MSE	= 399.6

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educHat	139.6838	28.28731	4.94	0.000	84.16961	195.198
exper	32.15667	7.127979	4.51	0.000	18.16792	46.14542
_cons	-1295.227	457.3103	-2.83	0.005	-2192.704	-397.7498

IV Regression: Implementation

TSLS in one step (with corrected standard errors):

```
. ivregress 2sls wage exper (educ = sibs)

Instrumental variables (2SLS) regression          Number of obs   =        935
                                                  Wald chi2(2)    =      24.88
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0417
                                                  Root MSE        =     395.64

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   139.6838   28.00689     4.99   0.000     84.79132    194.5763
       exper |   32.15667   7.057316     4.56   0.000     18.32458    45.98875
       _cons |  -1295.227   452.7768    -2.86   0.004    -2182.653   -407.8006
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper sibs
```

IV Regression: Implementation

- Sebastian Wai shows how to run the procedure using two OLS regressions and then using one *ivregress* procedure with corrected standard errors.
  - Also shows manual test of endogeneity using predicted residuals $\widehat{\gamma_i}$ from the first stage as regressors in the original equation $Y_i = \beta_0 + \beta_1 X_i + \beta_2 \widehat{\gamma_i} + e_i$. Endogenous if coefficient on $\widehat{\gamma_i}$ is stat. significant
- Chuck Huber shows how to run built-in tests in Stata: test of endogeneity, first stage statistics, etc.
- Ani Kachova shows how to run IV reg. in SAS

IV Regression: More Examples

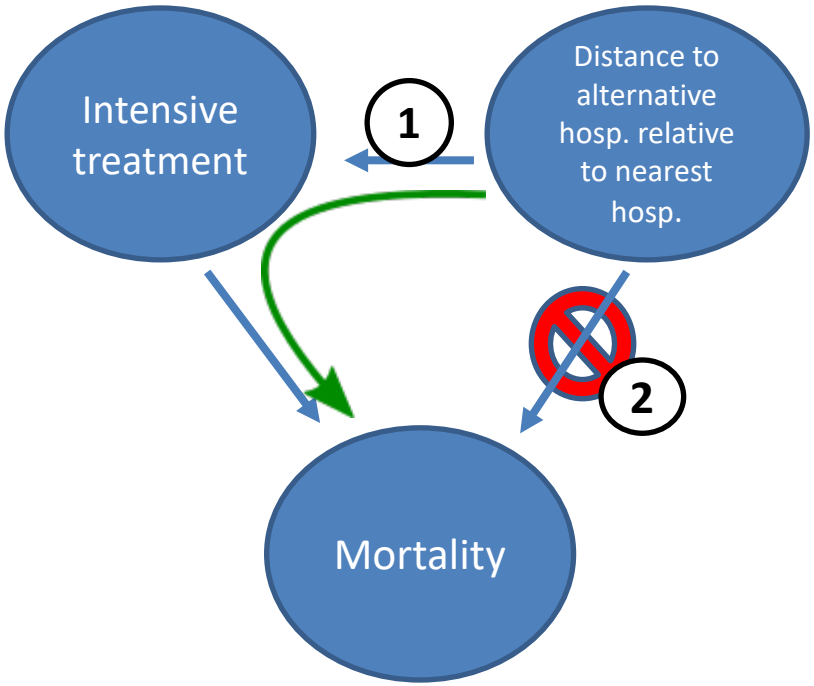
- Will help us understand IVs illustratively
- Will help us better assess the quality of the IV
- For determining IV quality, we should look for/discuss/raise critiques of assumptions being made about the two IV validity conditions:
  - IV relevance
  - IV exogeneity
- I encourage you to revisit these example papers later to look for ways that authors may have addressed some of your critiques

1

McClellan, M., McNeil, B. J., & Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables. *Jama*, *272*(11), 859-866.

Whether AMI patient dies = α + βIntensive treatment (vs. regular) + ε

Patient's differential distance to alternative types of hospitals



Outcome: Death among elderly patients with acute myocardial infarction (AMI)

Treatment: Intensive treatment (vs. regular)

Endogeneity concern: Factors that are difficult to observe such as comorbid diseases, severity of illness, complex details of a patient's health status and patient/physician preferences could be influencing both intensive treatment and mortality.

Instrument: Distance to alternative hospital minus distance to nearest hospital

1) Relevance assumption: Patients with lower differential distance to alternative hospitals are more likely to undergo intensive treatment

2) Exogeneity assumption: Differential distance has no impact on mortality directly

1

McClellan, M., McNeil, B. J., & Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables. *JAMA*, *272*(11), 859-866.

Whether AMI patient dies = α + βIntensive treatment (vs. regular) + ε

Patient's differential distance to alternative types of hospitals

Endogeneity Concerns:

Table 1.—Characteristics of Elderly Patients With Acute Myocardial Infarction in 1987*

Characteristic	All Patients (N=205 021)	No Catheterization Within 90 d (n=158 261)	Catheterization Within 90 d (n=46 760)
Demographic Characteristics			
Female	50.4	53.5	39.7
Black	5.6	6.0	4.3
Mean age, y (SD)	76.1 (7.2)	77.4 (7.3)	71.6 (5.0)
Urban	70.5	69.6	73.8
Comorbid Disease Characteristics			
Cancer	1.9	2.2	0.8
Pulmonary disease, uncomplicated	10.7	11.1	9.3
Dementia	1.0	1.2	0.1
Diabetes	18.0	18.3	17.1
Renal disease, uncomplicated	1.9	2.3	0.7
Cerebrovascular disease	4.8	5.4	2.8

1

McClellan, M., McNeil, B. J., & Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables. *JAMA*, *272*(11), 859-866.

Considering exogeneity and relevance of distance as an IV:

Table 4.—Patient Characteristics by Differential Distance to a Catheterization or Revascularization Hospital*

Characteristic	Differential Distance ≤2.5 Miles (n=102 516)	Differential Distance >2.5 Miles (n=102 505)
Comorbid Disease Characteristics		
Cancer	1.9	1.9
Pulmonary disease, uncomplicated	10.4	10.9
Dementia	0.99	0.94
Diabetes	18.1	18.0
Renal disease, uncomplicated	2.0	1.9
Cerebrovascular disease	4.8	4.8
Treatments		
Initial admit to catheterization hospital†	34.4	5.0
Initial admit to revascularization hospital†	41.7	10.7
Catheterization within 7 d	20.7	11.0
Catheterization within 90 d	26.2	19.5
CABG‡ within 90 d	8.6	6.9
PTCA§ within 90 d	6.4	4.3

1

McClellan, M., McNeil, B. J., & Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables. *JAMA*, *272*(11), 859-866.

Results without accounting for selection bias:

Table 2.—Estimated Cumulative Effect of Catheterization, Not Accounting for Selection Bias

Adjustment for Observable Differences Using ANOVA*	Percentage-Point Changes in Mortality Rates (SE)					
	1 d	7 d	30 d	1 y	2 y	4 y
None (unadjusted differences)	−9.4 (0.2)	−18.7 (0.2)	−19.2 (0.3)	−30.5 (0.3)	−34.0 (0.3)	−36.8 (0.3)
After adjustment for demographic and comorbidity differences	−6.8 (0.2)	−13.5 (0.2)	−17.9 (0.3)	−24.1 (0.3)	−26.6 (0.3)	−28.1 (0.3)

Results with instrumental variables:

Table 7.—Instrumental Variable Estimates of the Effects of Patient Location, High-Volume Hospital, and Catheterization on Mortality at Indicated Time Intervals After Acute Myocardial Infarction

Average Effect	Time After Acute Myocardial Infarction, Percentage-Point Change (SE)						
	1 d	7 d	30 d	1 y	2 y	3 y	4 y
Catheterization within 90 d							
Cumulative	−8.8 (2.0)	−11.5 (2.5)	−7.4 (2.9)	−4.8 (3.2)	−5.4 (3.3)	−5.0 (3.2)	−5.1 (3.2)

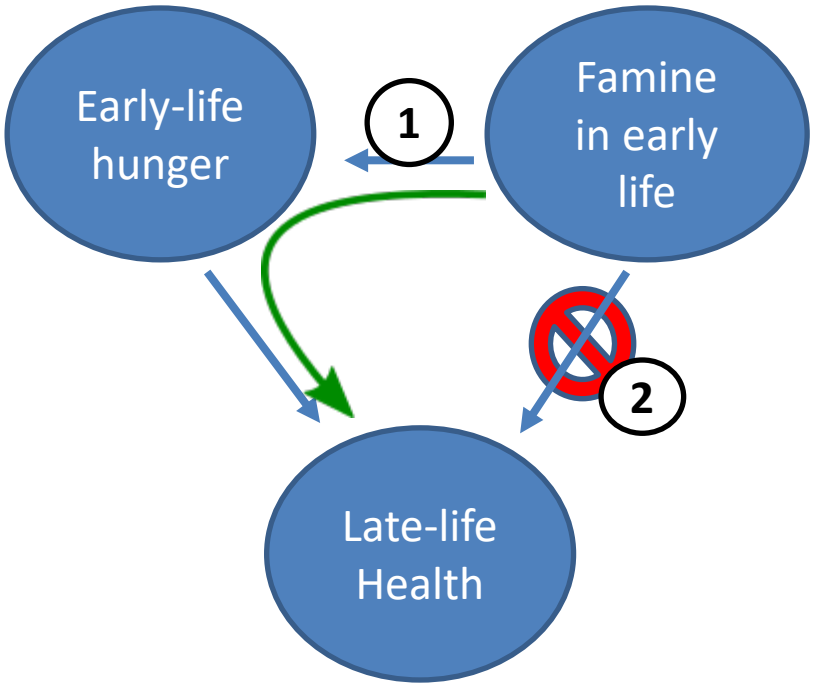
2

Van den Berg, G. J., Pinger, P. R., & Schoch, J. (2016). Instrumental variable estimation of the causal effect of hunger early in life on health later in life. *The Economic Journal*, *126*(591), 465-506.

$$\text{Late-life health} = \alpha + \beta\ \text{Early-life hunger} + \varepsilon$$

↑

Famine in early life



Outcome: Health in later life (measured by adult height)

Treatment: Hunger in early life (measured by self-report)

Endogeneity concern: Later life outcomes and early life conditions in parents' household jointly depend on unobserved confounders.

Instrument: Exposure to a famine early in life

1) Relevance assumption: Famine during early life increases hunger in utero or at ages 0-4

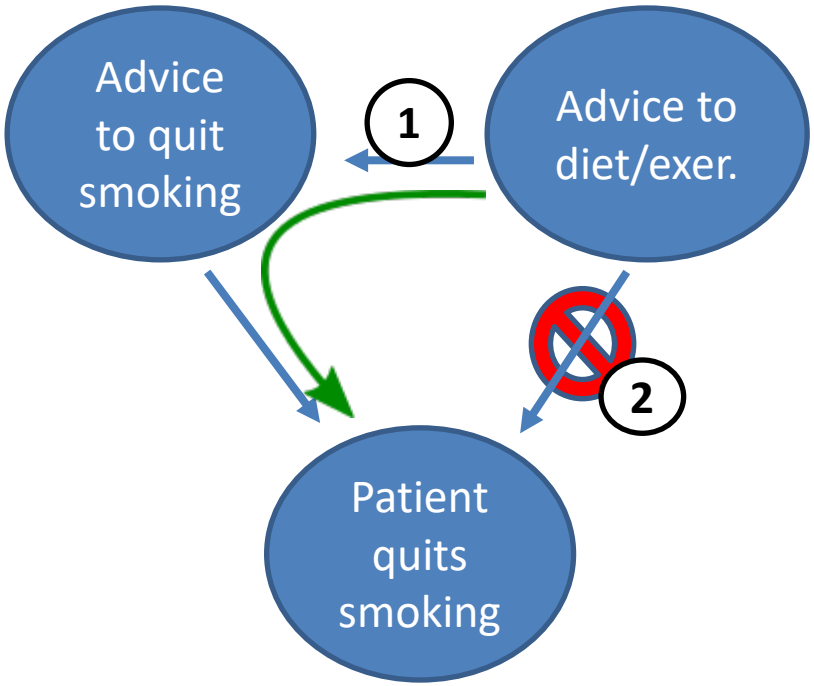
2) Exogeneity assumption: Famines do not impact health in later life except through hunger in early life.

3

Bao, Y., Duan, N., & Fox, S. A. (2006). Is some provider advice on smoking cessation better than no advice? An instrumental variable analysis of the 2001 National Health Interview Survey. *Health services research*, *41*(6), 2114-2135.

Quit smoking = α + β Doc says don't smoke + ε

Doc gave advice on diet/nutrition



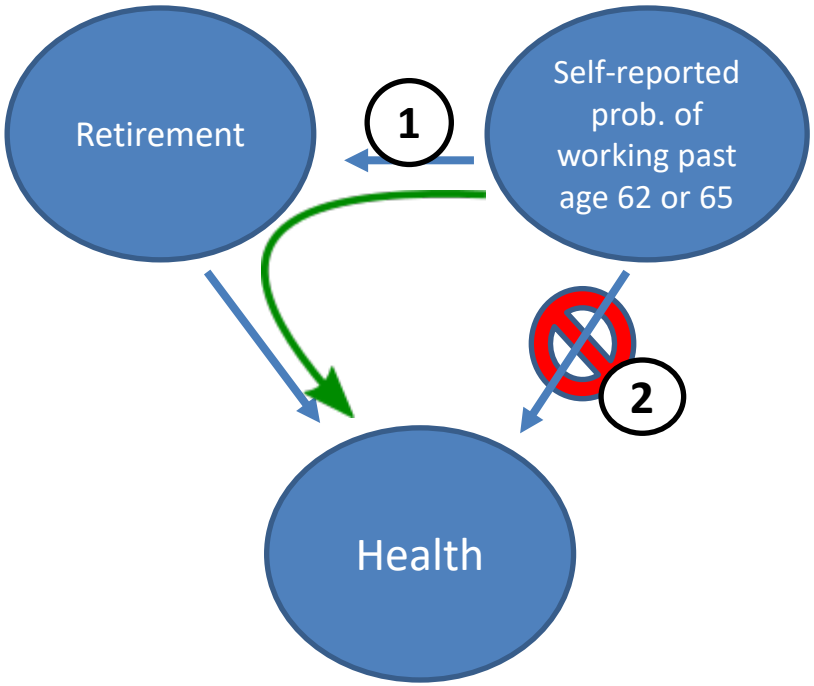
Outcome: Smoking cessation
Treatment: Provider advice to quit smoking
Endogeneity concern: Providers may be more likely to advise heavier smokers and/or those who have already been diagnosed with smoking-related conditions
Instrument: Provider advice to diet or exercise (measure of provider tendency to advise)
1) Relevance assumption: Provider advice to diet or exercise is correlated with advice to quit smoking
2) Exogeneity assumption: Provider advice for diet/nutrition and for physical activity are not directly correlated with the patient's likelihood of success in smoking cessation except through increased likelihood of provider advice for smoking cessation

4

Insler, M. (2014). The health consequences of retirement. *Journal of Human Resources*, *49*(1), 195-233..

Health= α + β Retirement+ ε

self-reported prob. of working past 62 and 65



Outcome: Health

Treatment: Retirement

Endogeneity concern: Declines in health can compel people to retire – difficult to disentangle simultaneous causal effects

Instrument: Self-reported probability of working past ages 62 and 65 when indivs. were employed

1) Relevance assumption: People who indicate high probability of working past these milestone ages are less likely to retire

2) Exogeneity assumption: After controlling for hereditary health trends and past health history, self-reported probability captures the *preference* to retire and not *expectation* to retire (which may be correlated with the error term).

Other IV Examples

- Zulman, Pal Chee, et al. (2015): effect of VA intensive management primary care on VA health care costs; instrument: random assignment to treatment vs. usual care groups

- Bhattacharya, et al. (2011): effect of insurance coverage on body weight; instruments: distribution of firm size and Medicaid coverage for each state and year

- Doyle (2013): effect of foster care on long- and short-term outcomes; instrument: random assignment to investigators

Summary

- IV reg. is a powerful tool for estimating causal effects
- Conditions for a valid instrument:
  - Relevance: the instrument must affect treatment
  - Exogeneity: the instrument must be uncorrelated with all other factors that may affect outcomes
- Using invalid (weak or endogenous) instruments will give meaningless results
- The hardest part is finding good/convincing IVs
- Examples can help us get better at identifying potential instruments and at assessing the validity of IVs
- Some tests available to check instrument validity, but what is absolutely necessary is a good "story" for why an instrument is relevant and exogenous

Thank You

- Questions?
- Please email me if you have any additional questions: Kritee.Gujral@va.gov