STA 235 - Causal Inference: Instrumental Variables

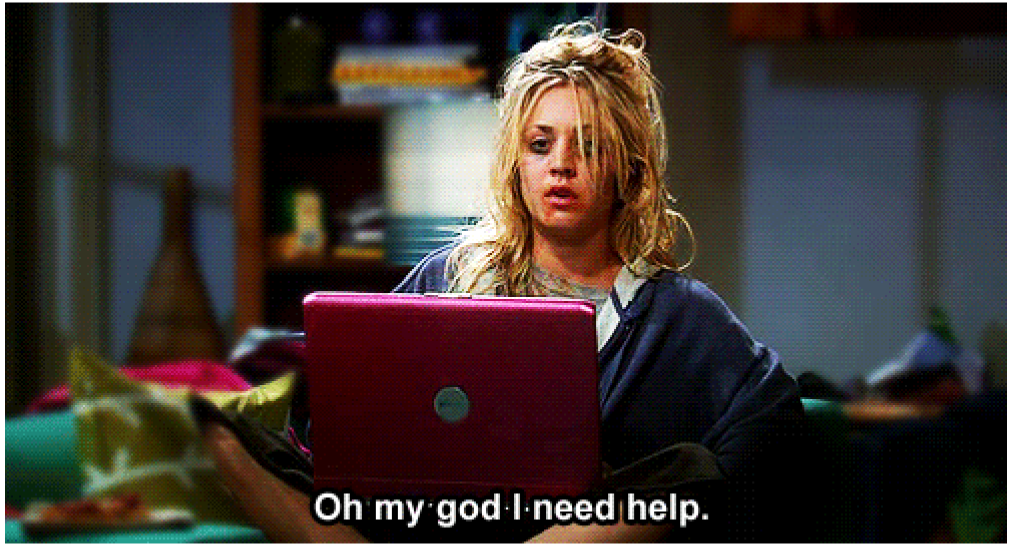
Spring 2021

McCombs School of Business, UT Austin

Introduction to instrumental variables

- We have seen that controlling for covariates is usually **not enough**.
- We might not have **randomization** or **a nice RD**.

What to do?



Instrumental variables could help!

... but first let's review some concepts.

Endogeneity vs Exogeneity

Does education cause higher earnings?

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i$$

Would β_1 give us the causal effect of Education on Earnings?

Endogeneity vs Exogeneity

Endogenous variable

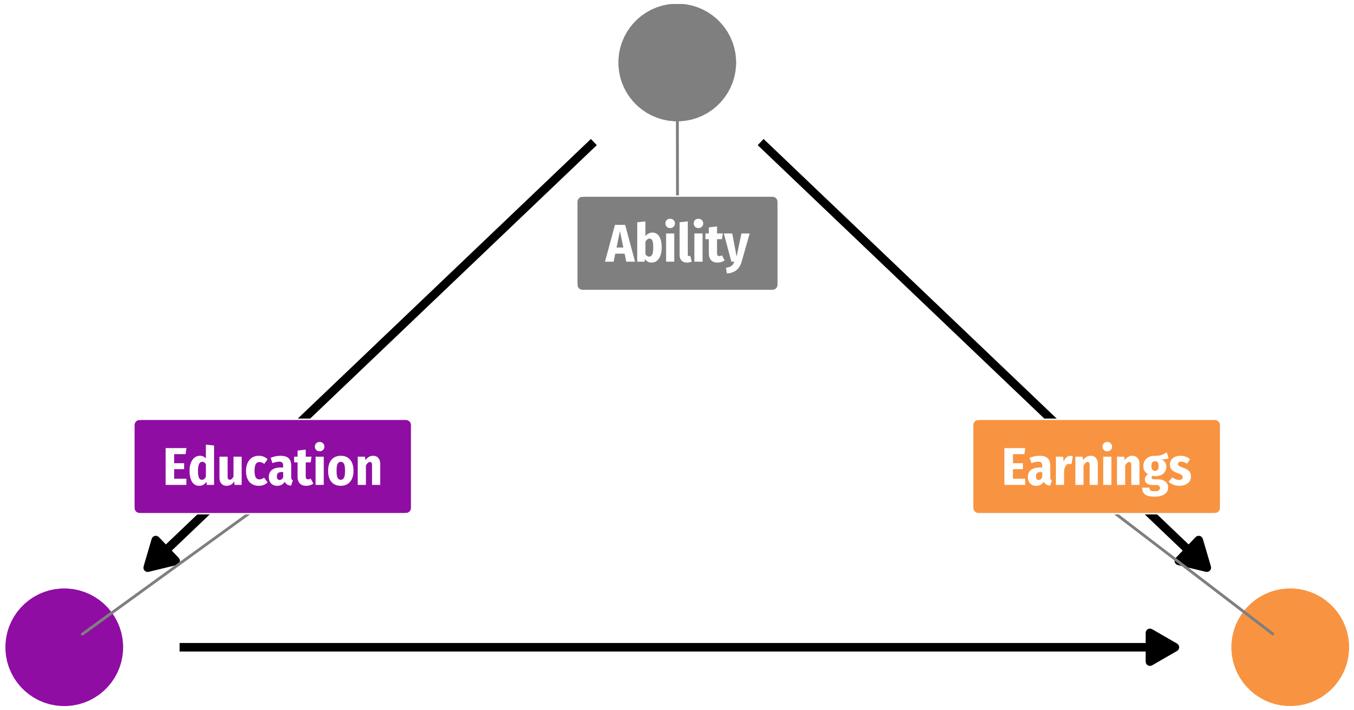
"A variable is said to be endogenous within the causal model M if its value is determined or influenced by one or more of the independent variables (excluding itself)." (Little, 2011)

Exogenous variable

"An exogenous variable is a variable that is not affected by other variables in the system"

Back to earnings and education

- **Education** could be considered and *endogeneous* variable.
- **Ability** could be considered an *exogenous* variable.



Can we do something about this?

- We want some **exogeneous variation** in education:
    - E.g. choices to get more or less education are essentially random (or unless uncorrelated with omitted variables)
- We would like education to be **exogeneous**, but it's not! Part of it is caused by **ability**

... but part of it is not. Can we separate both parts?

Separate an edogenous variable

$$\begin{aligned}
\text{Earnings}_i =& \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i \\
& \beta_0 + \beta_1 (\text{Education}_i^{\text{exog.}} + \text{Education}_i^{\text{endog.}}) + \varepsilon_i \\
& \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \underbrace{\beta_1 \text{Education}_i^{\text{endog.}} + \varepsilon_i}_{\omega_i} \\
& \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \omega_i
\end{aligned}$$

- How do we find $\text{Education}_i^{\text{exog}}$?

Instrumental variables to the rescue?

Instrumental variables (IV) can help.

- What is an IV?
    - Something that is correlated with the treatment: **Relevance**
    - Something that does not directly cause the outcome: **Exclusion**
    - Something that is not correlated with the ommited variables: **Exogeneity**

Assumptions behind IVs

Relevance
Correlated with treatment

$Z \longrightarrow D \qquad Cor(Z, D) \neq 0$

testable with stats

Assumptions behind IVs

Exclusion
Correlated with outcome *only through* treatment

$Z \longrightarrow D \longrightarrow Y$ $\quad Z \not\longrightarrow Y$ $\quad Cor(Z, Y \mid D) = 0$

testable with stats + story

Assumptions behind IVs

Exogeneity
***Not* correlated with omitted variables**

$U \not\rightarrow Z \qquad Cor(Z, U) = 0$

Not testable with stats (only story!)

Who do instruments work for?

- When doing an IV analysis, we are only estimating an effect **for those who are moved by our instrument**

Compliers

- We are not identifying and effect for "always-takers" or "never-takers" (also, we assume no defiers).

LATE

Finding instruments is hard!

- Usually the **exclusion restriction fails**.
- In the previous example of education, researchers have used **distance to college** as an instrument.
  - **Is this valid?** Why yes or why not?
- However, good examples for an instrument could be treatment assignment in:

Fuzzy Regression Discontinuity Design

Noncompliance in RCTs

Two-stage least squares (2SLS)

- **First stage**: Regress endogenous variable (e.g. education) on instrument (e.g. distance to college), and get fitted values.

$$\widehat{\text{Education}}_i = \gamma_0 + \gamma_1 \text{Distance}_i + \eta_i$$

- **Second stage**: Regress outcome (e.g. income) on predicted values of endogenous variable (e.g. $\widehat{\text{Education}}_i$).

$$\text{income}_i = \beta_0 + \beta_1 \widehat{\text{Education}}_i + \varepsilon_i$$

Let's go back to GOTV example

- RCT were households were randomized into GOTV calls.
- We had random treatment assignment, but high noncompliance (e.g. people did not pick up their phone).

What was the outcome of interest?

What is the endogenous variable?

What could be an instrument?

GOTV compliance

```
d <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/

# Drop variables with unlisted phone numbers
d_s1 <- d[!is.na(d$treat_real),]

# Treatment assignment vs Actual treatment
table(d_s1$treat_real, d_s1$contact)

# % of treated by assignment
d_s1 %>% group_by(treat_real) %>% summarise(contact = mean(contact))
```

GOTV compliance

```
d <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/

# Drop variables with unlisted phone numbers
d_s1 <- d[!is.na(d$treat_real),]

# Treatment assignment vs Actual treatment
table(d_s1$treat_real, d_s1$contact)

# % of treated by assignment
d_s1 %>% group_by(treat_real) %>% summarise(contact = mean(contact))
```

```
## 
##           0       1
##   0 1845348       0
##   1   34929   25043
```

GOTV compliance

```r
d <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/

# Drop variables with unlisted phone numbers
d_s1 <- d[!is.na(d$treat_real),]

# Treatment assignment vs Actual treatment
table(d_s1$treat_real, d_s1$contact)

# % of treated by assignment
d_s1 %>% group_by(treat_real) %>% summarise(contact = mean(contact))
```

```
## # A tibble: 2 x 2
##   treat_real contact
##        <int>   <dbl>
## 1          0   0
## 2          1   0.418
```

GOTV: First stage

```r
library(estimatr)

lm1 <- estimatr::lm_robust(contact ~ treat_real, data = d_s1)

summary(lm1)
```

```
##
## Call:
## estimatr::lm_robust(formula = contact ~ treat_real, data = d_s1)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)  CI Lower  CI Upper      DF
## (Intercept) 5.176e-13  8.601e-16   601.8        0 5.160e-13 5.193e-13 1905318
## treat_real  4.176e-01  2.014e-03   207.4        0 4.136e-01 4.215e-01 1905318
##
## Multiple R-squared:  0.4098 ,	Adjusted R-squared:  0.4098
## F-statistic: 4.3e+04 on 1 and 1905318 DF,  p-value: < 2.2e-16
```

```r
d_s1$contact_fitted = lm1$fitted.values
```

GOTV: First stage

```r
library(estimatr)

lm1 <- estimatr::lm_robust(contact ~ treat_real, data = d_s1)

summary(lm1)
```

```
##
## Call:
## estimatr::lm_robust(formula = contact ~ treat_real, data = d_s1)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)  CI Lower  CI Upper      DF
## (Intercept) 5.176e-13  8.601e-16   601.8        0 5.160e-13 5.193e-13 1905318
## treat_real  4.176e-01  2.014e-03   207.4        0 4.136e-01 4.215e-01 1905318
##
## Multiple R-squared:  0.4098 ,    Adjusted R-squared:  0.4098
## F-statistic: 4.3e+04 on 1 and 1905318 DF,  p-value: < 2.2e-16
```

```r
d_s1$contact_fitted = lm1$fitted.values
```

GOTV: Second stage

```
estimatr::lm_robust(vote02 ~ contact_fitted, data = d_s1)
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)   CI Lower
## (Intercept)    0.54528902 0.0003665575 1487.59488 0.000000e+00 0.54457058
## contact_fitted 0.08728695 0.0049029128   17.80308 6.778072e-71 0.07767741
##                  CI Upper      DF
## (Intercept)    0.54600746 1905318
## contact_fitted 0.09689648 1905318
```

GOTV: Intention to Treat

```
lm2 <- estimatr::lm_robust(vote02 ~ treat_real, data = d_s1)

summary(lm2)
```

```
##
## Call:
## estimatr::lm_robust(formula = vote02 ~ treat_real, data = d_s1)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper      DF
## (Intercept)   0.54529  0.0003666  1487.6 0.000e+00  0.54457  0.54601 1905318
## treat_real    0.03645  0.0020473    17.8 6.778e-71  0.03244  0.04046 1905318
##
## Multiple R-squared:  0.0001634 ,    Adjusted R-squared:  0.0001629
## F-statistic: 316.9 on 1 and 1905318 DF,  p-value: < 2.2e-16
```

```
lm2$coefficients[2]/lm1$coefficients[2]
```

```
## treat_real
## 0.08728695
```

GOTV: Intention to Treat

```
lm2 <- estimatr::lm_robust(vote02 ~ treat_real, data = d_s1)

summary(lm2)
```

```
##
## Call:
## estimatr::lm_robust(formula = vote02 ~ treat_real, data = d_s1)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper      DF
## (Intercept)   0.54529  0.0003666  1487.6 0.000e+00  0.54457  0.54601 1905318
## treat_real    0.03645  0.0020473    17.8 6.778e-71  0.03244  0.04046 1905318
##
## Multiple R-squared:  0.0001634 ,    Adjusted R-squared:  0.0001629
## F-statistic: 316.9 on 1 and 1905318 DF,  p-value: < 2.2e-16
```

```
lm2$coefficients[2]/lm1$coefficients[2]
```

```
## treat_real
## 0.08728695
```

GOTV: 2SLS

- You can recover point estimates with the previous methods, but **standard errors will be wrong** (unless you adjust them).
- You can use packages designed for this, e.g. `ivreg` or `iv_robust()` from `estimatr`

```
summary(iv_robust(vote02 ~ contact | treat_real, data = d_s1))
```

```
##
## Call:
## iv_robust(formula = vote02 ~ contact | treat_real, data = d_s1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper      DF
## (Intercept)  0.54529  0.0003666  1487.6 0.000e+00  0.54457  0.54601 1905318
## contact      0.08729  0.0048760    17.9 1.166e-71  0.07773  0.09684 1905318
##
## Multiple R-squared:  0.0005131 ,    Adjusted R-squared:  0.0005126
## F-statistic: 320.5 on 1 and 1905318 DF,  p-value: < 2.2e-16
```

Fuzzy Regression Discontinuity

- The same principal applies when we **don't have full compliance** in an RDD
- **Fuzzy regression discontinuity**
  - If $Z = I(R_i > c)$, then $\Pr(D = 1|Z = 1) < 1$ and/or $\Pr(D = 1|Z = 0) > 0$

```
rdrobust(y = y, x = x, c = c, fuzzy = treat)
```

Example: Entrance exam and tutoring

Use above/below cutoff as instrument: A parametric approach

```r
tutoring <- tutoring %>% mutate(distance = entrance_exam - 70,
                                below_cutoff = entrance_exam <= 70)


summary(iv_robust(exit_exam ~ distance + tutoring | distance + below_cutoff,
  data = filter(tutoring, distance >= -10 & distance <= 10)))
```

```
##
## Call:
## iv_robust(formula = exit_exam ~ distance + tutoring | distance +
##     below_cutoff, data = filter(tutoring, distance >= -10 & distance <=
##     10))
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   60.1414     1.0177  59.098 9.747e-200  58.1407  62.1420 400
## distance       0.4366     0.0993   4.397  1.407e-05   0.2414   0.6318 400
## tutoringTRUE   9.7410     1.9118   5.095  5.384e-07   5.9825  13.4996 400
##
## Multiple R-squared:  0.3646 ,    Adjusted R-squared:  0.3615
## F-statistic: 13.06 on 2 and 400 DF,  p-value: 3.19e-06
```

Use above/below cutoff as instrument: A nonparametric approach

```r
library(rdrobust)

summary(rdrobust(y = tutoring$exit_exam, x = tutoring$distance, c = 0, fuzzy = tutoring$tutoring))
```

```
## Call: rdrobust
##
## Number of Obs.                 1000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 238          762
## Eff. Number of Obs.            170          347
## Order est. (p)                   1            1
## Order bias  (q)                  2            2
## BW est. (h)                 12.985       12.985
## BW bias (b)                 19.733       19.733
## rho (h/b)                    0.658        0.658
## Unique Obs.                    238          762
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     9.683     1.893     5.116     0.000     [5.973 , 13.393]
##         Robust         -         -     4.258     0.000     [5.210 , 14.095]
## =============================================================================
```

Takeaways

- Instruments can be **useful** for recovering treatment effects, even under no random assignment.
- Finding good instruments is **hard**.
- We can easily use them in RCTs or RD designs to go **from an ITT to a LATE**.

References

- Angrist, J. and S. Pischke. (2015). "Mastering Metrics". *Chapter 3*.
- Heiss, A. (2020). "Program Evaluation for Public Policy". *Class 11: Instrumental Variables, Course at BYU*.