STA 235 - Causal Inference: Instrumental Variables (Cont.)

Spring 2021

McCombs School of Business, UT Austin

Some reminders

Homework 3 is due next Wed.

- This homework has a bit less guidance than previous homework. **Make assumptions and use Google!**

- Remember to post questions on **Piazza** or **come to OH**.

No questions will be answered after 7pm 4/13

Some reminders (Cont.)

Highlight sessions on Thur. 5:00-5:30pm

- Review session to cover the highlights of that week's class.
- Objective: Get all students on the **same page** regarding that week's material.
- **Not an exercise class.**

Not a replacement for OH

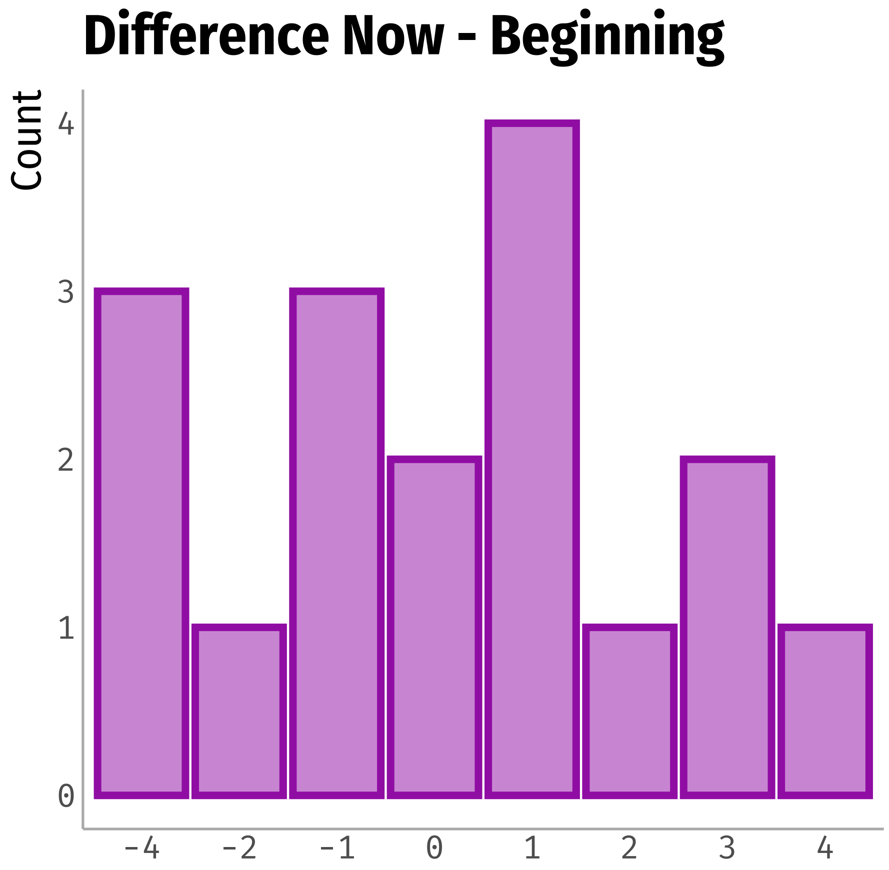
Before we start: About JITTs

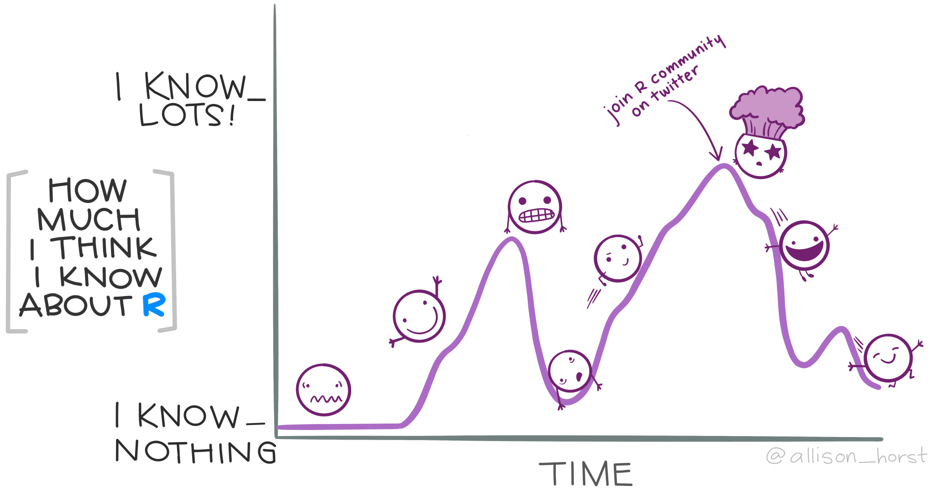
Remember to do the readings

- It's clear when someone doesn't do the readings. **Those submissions will not be counted**.

About JITTs: How comfortable are you with R?

About JITTs: How comfortable are you with R?





Roadmap

Last class:

- Introduction to instrumental variables (IV)
- How we can use **noncompliance in an RCT** as a good setting for IV

Today's class:

- Recap of IV
- Fuzzy RD
- Model selection (new chapter!)

Recap of instrumental variables

- An **instrumental variable** (or instrument) Z is a variable that allows us to **separate the endogenous part from the exogenous part** of our treatment variable D (or the variable for which we want to estimate an effect for).

Conditions for an IV:

Relevance: Cor(D,Z) != 0

Exclusion: Cor(Z,Y|D) = 0

Exogeneity: Cor(Z,U) = 0

Poll time!

If our treatment is "going to college", and we want to estimate an effect on future earnings:

Is "distance to college" a good instrument?

How do we use IVs to estimate LATEs?

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?

Let's go to R

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

```r
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

```r
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
```

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

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

GOTV: Intention to Treat

```r
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
```

```r
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

Poll time!

What is the treatment assignment variable and the treatment variable?

Do you think the ITT>LATE or LATE>ITT?

Use above/below cutoff as instrument: A parametric approach

```
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*.