Week 6: Instrumental Variables

PLSC 30600 - Causal Inference

Last two weeks

- Identification under **conditional ignorability**
  - Treatment assignment is independent of the potential outcomes given observed confounders $\mathbf{X}$
  - "Selection-on-observables"
- "Selection-on-observables" isn't a testable assumption
  - Relies on theory to decide which $\mathbf{X}$ to include.
  - DAGs can help here.
- Lots of estimation strategies
  - Stratify with low-dimensional $\mathbf{X}$
  - IPTW to eliminate treatment-covariate relationship, regression to model the outcome-covariate relationship.
  - Matching to reduce model dependence.
    - Or consider more modern flexible modelling techniques for $E[Y_i(d)|X_i]$

This week

- Can we estimate a treatment effect when neither ignorability nor conditional ignorability hold for treatment?
  - Can we get rid of *unobserved* confounding?
- "Instrumental variables" designs are one way of dealing with this
- We can identify *some* average of treatment effects if...
  - There *does* exist an ignorable or conditionally ignorable **instrument** which...
  - ...has a monotonic effect on the treatment...
  - ...and has no effect on the outcome *except* through its effect on the treatment.
- What's the average? The "Local Average Treatment Effect"
  - Average effect among those who are *moved* to take treatment by the instrument

Instrumental Variables

Treatment non-compliance

- Often experiments suffer from treatment **non-compliance**
  - Participants randomized to receive a phone call don't pick up.
  - Participants randomized to wear surgical masks choose not to.
- New notation!
  - Let Z_i denote whether i is assigned to receive a treatment.
  - Let D_i denote the treatment actually *taken* by an individual.
- Can we just take the simple difference-in-means between $D_i = 1$ and $D_i = 0$?
  - No! Non-compliance affected by other factors which might also affect the outcome.
  - We're stuck with an observational design.

- Unless...

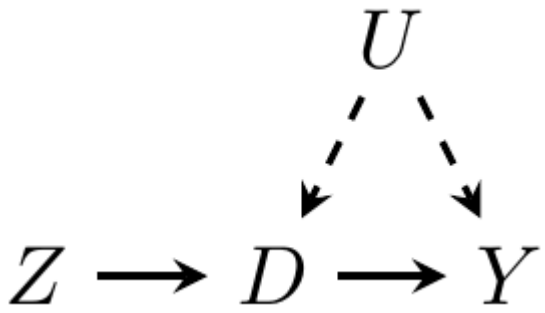
Intent-to-treat effect

- We can first just change the question - instead of the effect of **treatment**, we can make our estimand the effect of being **assigned to treatment**.
- Our estimator for the ITT is just the difference in means between the $Z_i = 1$ and $Z_i = 0$ arms

$$\hat{\tau}_{\text{ITT}} = \hat{E}[Y_i | Z_i = 1] - \hat{E}[Y_i | Z_i = 0]$$

- Identified under randomization of Z_i even if D_i is not randomized.
  - But combines two effects: the actual effect of D_i and the effect of Z_i on D_i.

Instrumental variables



- Suppose though that we're interested in the *actual* effect of receiving treatment (the effect of D_i). What can we do?

Instrumental variables

- Start by writing down potential outcomes for D_i along with joint potential outcomes of Y_i in terms of Z_i and D_i

$$D_i(z) = D_i \text{ if } Z_i = z$$

$$Y_i(d, z) = Y_i \text{ if } D_i = d, Z_i = z$$

- Observed treatment D_i is a function of treatment assignment (Z_i) - it's a post-treatment quantity (and so has potential outcomes).

Assumptions

1. Randomization of instrument
2. Exclusion restriction
3. Non-zero first-stage relationship
4. Monotonicity

Assumption 1: Randomization

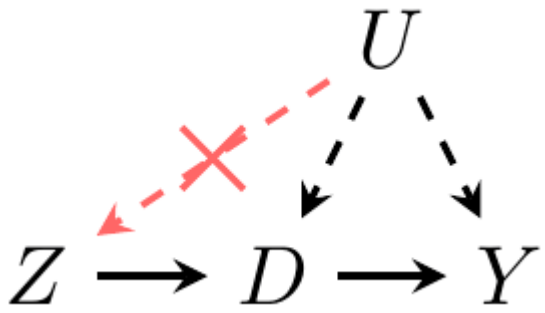
- Z_i is is independent of both sets of potential outcomes (potential outcomes for the treatment and potential outcomes for the outcome).

$$\{D_i(1), D_i(0)\} \perp\!\!\!\perp Z_i$$

$$\{Y_i(d, z) \forall d, z\} \perp\!\!\!\perp Z_i$$

- We can weaken this to conditional ignorability (where Z_i is randomized conditional on X_i), which is common in observational settings.
  - But if we don't believe conditional ignorability for the treatment, why would we believe it for the instrument?
- Sufficient to identify the **intent-to-treat (ITT)** effect

Assumption 1: Randomization



- The randomization assumption eliminates any arrows from U to Z.

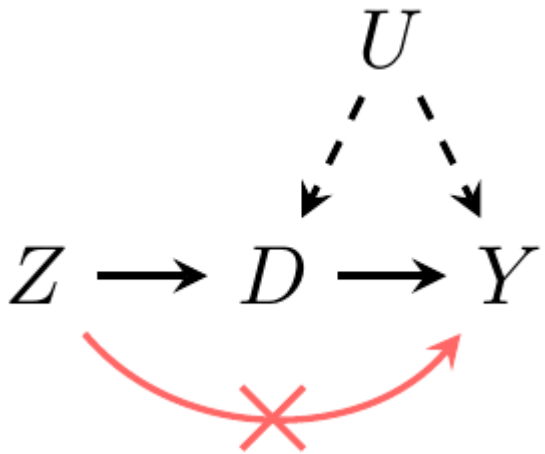
Assumption 2: Exclusion restriction

- Z_i **only affects** Y_i by way of its effect on D_i.
- In other words, if D_i were set at some level d, the potential outcome for $Y_i(d, z)$ does not depend on z.

$$Y_i(d, z) = Y_i(d, z') \text{ for any } z \neq z'$$

- **Not a testable assumption!** -- we have to justify this with substantive knowledge.
  - Easiest in the treatment non-compliance case
  - But consider what might happen in a non-blinded situation where respondents knew their treatment assignments.
- "Surprise" factor -- If I told you Z was associated with Y, would you think "that's odd"?

Assumption 2: Exclusion restriction



- The exclusion restriction eliminates any causal paths from Z to Y **except** for $Z \to D \to Y$.

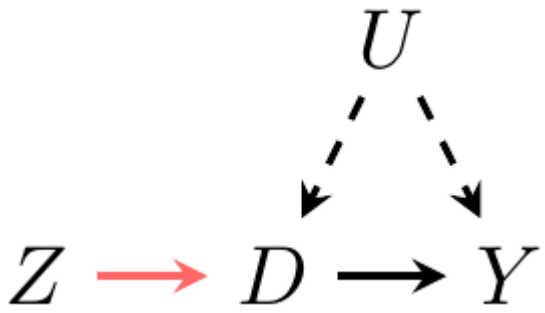
Assumption 3: Non-zero first stage

- Z_i has an effect on D_i

$$E[D_i(1) - D_i(0)] \neq 0$$

- Seems trivial, but we need this to make the estimator work.
- Magnitude matters for estimator performance - a "weak" first-stage $\rightsquigarrow$ heavily biased IV estimator
  - IV estimators are *consistent* but not *unbiased*.

Assumption 3: Non-zero first stage



- The non-zero first stage assumption requires a path from Z to D.

Assumption 4: Monotonicity

- Z_i's effect on D_i only goes in one direction **at the individual level**

$$D_i(1) - D_i(0) \geq 0$$

- If it goes the other way, we can always flip the direction of the treatment to make this hold
  - The key is that the instrument does not have a positive effect on D_i for some units and a negative effect for others.
- **Not a testable assumption**

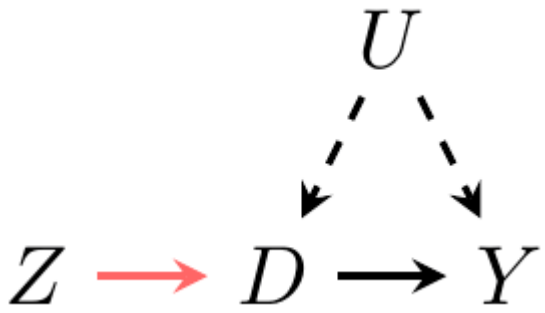
Assumption 4: Monotonicity

- In binary instrument/binary treatment world, this is sometimes called a "no defiers" assumption.

Stratum	$D_i(1)$	$D_i(0)$
"Always-takers"	1	1
"Never-takers"	0	0
"Compliers"	1	0
"Defiers"	0	1

- Under no defiers, every unit with $D_i = 1$ and $Z_i = 0$ is an always-taker, every unit with $D_i = 0$ and $Z_i = 1$ is a never-taker.

Assumption 4: Monotonicity



- Can't represent the monotonicity assumption in a DAG - it's an assumption about the form of the relationship between Z and D.

Interpreting the IV estimand

- The classic IV estimand with one instrument is a ratio of sample covariances.

$$\tau_{\text{IV}} = \frac{Cov(Y, Z)}{Cov(D, Z)}$$

- With a binary instrument, this is sometimes called the "Wald" estimand - a ratio of differences in means

$$\tau_{\text{IV}} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

Interpreting the IV estimand

- What does the Wald estimand correspond to in terms of causal effects?

$$\tau_{\text{IV}} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

- Under our identification assumptions:
  - The numerator is the ITT
  - The denominator is the first-stage effect

Interpreting the IV estimand

- Let's decompose the denominator first - under randomization:

$$\begin{aligned} E[D_i|Z_i = 1] - E[D_i|Z_i = 0] &= E[D_i(1)|Z_i = 1] - E[D_i(0)|Z_i = 0] \\ &= E[D_i(1)] - E[D_i(0)] \\ &= E[D_i(1) - D_i(0)] \end{aligned}$$

- With binary treatment/binary instrument, we can use law of total expectation to decompose by principal stratum

$$\begin{aligned} E[D_i(1) - D_i(0)] = E[D_i(1) - D_i(0)|D_i(1) = D_i(0)] \times P(D_i(1) = D_i(0)) + \\ E[D_i(1) - D_i(0)|D_i(1) > D_i(0)] \times P(D_i(1) > D_i(0)) + \\ E[D_i(1) - D_i(0)|D_i(1) < D_i(0)] \times P(D_i(1) < D_i(0)) \end{aligned}$$

- The first term is 0
- And by no defiers, the last term is 0 since $P(D_i(1) < D_i(0)) = 0$

$$E[D_i(1) - D_i(0)] = Pr(D_i(1) > D_i(0))$$

Interpreting the IV estimand

- Next, the numerator (the ITT). Under the exclusion restriction and randomization:

$$E[Y_i|Z_i = 1] = E\Big[Y_i(0) + \Big(Y_i(1) - Y_i(0)\Big)D_i(1)\Big]$$

$$E[Y_i|Z_i = 0] = E\Big[Y_i(0) + \Big(Y_i(1) - Y_i(0)\Big)D_i(0)\Big]$$

- The difference (with some algebra) is

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E\Big[\Big(Y_i(1) - Y_i(0)\Big) \times \Big(D_i(1) - D_i(0)\Big)\Big]$$

Interpreting the IV estimand

- Conditioning on the principal strata:

$$= E\Big[(Y_i(1) - Y_i(0)) \times (0) | (D_i(1) = D_i(0))\Big] \times P(D_i(1) = D_i(0)) +$$
$$E\Big[(Y_i(1) - Y_i(0)) \times (1) | (D_i(1) > D_i(0))\Big] \times P(D_i(1) > D_i(0)) +$$
$$E\Big[(Y_i(1) - Y_i(0)) \times (-1) | (D_i(1) < D_i(0))\Big] \times P(D_i(1) < D_i(0))$$

- Again, first term is zero because $D_i(1) - D_i(0) = 0$, third is zero by "no defiers" and we have

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E\Big[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)\Big] \times P(D_i(1) > D_i(0))$$

- The ITT is the product of a conditional average treatment effect and the proportion of compliers.

The LATE Theorem

- The IV estimand, under our identification assumptions, is a Local Average Treatment Effect (LATE):

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]$$

- The LATE is a conditional average treatment effect within the *subpopulation* of **compliers**
- If treatment effects are constant, we can generalize this to the whole sample.
  - But if effects are heterogeneous, we are not necessarily getting a "representative" treatment effect.

Better LATE than never?

- How should we interpret the LATE?
  - It's not necessarily the quantity we care about - we care about the effect of the treatment in the entire sample.
- Compliers are those compelled to take treatment by our encouragement. Would estimates generalize to those who are less encourageable?
  - The LATE is design-specific. If we came up with a different instrument, that changes the population on which we're estimating an effect!
  - What can we do?
  - We could describe the distribution of covariates among compliers vs. the population as a whole (Abadie's kappa-weighting).

Example: The effect of media on voting

- Gerber, Karlan and Bergan (2009, AEJ:AE) estimate the effect of reading the Washington Post (or Washington Times) on political attitudes and voting behavior.
  - Z_i: Random assignment to receive a free subscription to the Washington Post
  - D_i: Actually subscribing to the Washington Post (as measured by a post-encouragement survey)
  - Y_i: 2005 Turnout (measured in the survey)
- Assumptions:
  - Assignment to get the free subscription offer is ignorable/exogenous
  - Getting the free subscription offer affects actual subscriptions (non-zero first stage)
  - No one would subscribe to the Post if they *didn't* receive the offer but not subscribe if they *did*. (monotonicity/no defiers)
  - Assignment to get the free subscription offer doesn't affect voting *except through* actually subscribing to the Post (exclusion restriction)

Example: The effect of media on voting

- First, subset the data to WaPo or control observations that completed the follow-up survey

```
green <- read_dta("assets/publicdata.dta")
wapost <- green %>% filter(treatment != "TIMES"&!is.na(getpost)&!is.na(voted))
```

Example: The effect of media on voting

- Is there a first-stage effect?

```
lm_robust(getpost ~ post, data=wapost)
```

```
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    0.203     0.0189   10.73 4.06e-25    0.166    0.240 760
## post           0.341     0.0341    9.99 3.82e-22    0.274    0.408 760
```

- About 34 percent of the sample is a "complier" - quite substantial!

Example: The effect of media on voting

- Is there an ITT?

```
lm_robust(voted ~ post, data=wapost)
```

```
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  0.72627      0.021 34.6303 1.91e-158   0.6851   0.7674 760
## post        -0.00135      0.033 -0.0409  9.67e-01  -0.0661   0.0634 760
```

- ITT is essentially zero.

Example: The effect of media on voting

- Compare with the naive OLS estimate

```
lm_robust(voted ~ getpost, data=wapost)
```

```
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)    0.703     0.0204   34.45 2.11e-157 0.663119    0.743 760
## getpost        0.066     0.0332    1.99  4.70e-02 0.000877    0.131 760
```

- Post subscribers are 6pp more likely to vote in the 2005 VA gubernatorial election.
  - But is this causal? No!

Example: The effect of media on voting

- Let's estimate the LATE using the Wald estimator

```
(mean(wapost$voted[wapost$post == 1]) - mean(wapost$voted[wapost$post == 0]))/(mean(wapost$getp
```

```
## [1] -0.00396
```

- Equivalent to a ratio of regression coefficients

```
coef(lm_robust(voted ~ post, data=wapost))[2]/coef(lm_robust(getpost ~ post, data=wapost))[2]
```

```
##     post
## -0.00396
```

Example: The effect of media on voting

- We'll talk about inference later, but take note: the SE for the LATE can be much larger than the SE for the ITT

```
iv_robust(voted ~ getpost | post, data=wapost)
```

```
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  0.72707     0.0368 19.7765 1.51e-70    0.655    0.799 760
## getpost     -0.00396     0.0968 -0.0409 9.67e-01   -0.194    0.186 760
```

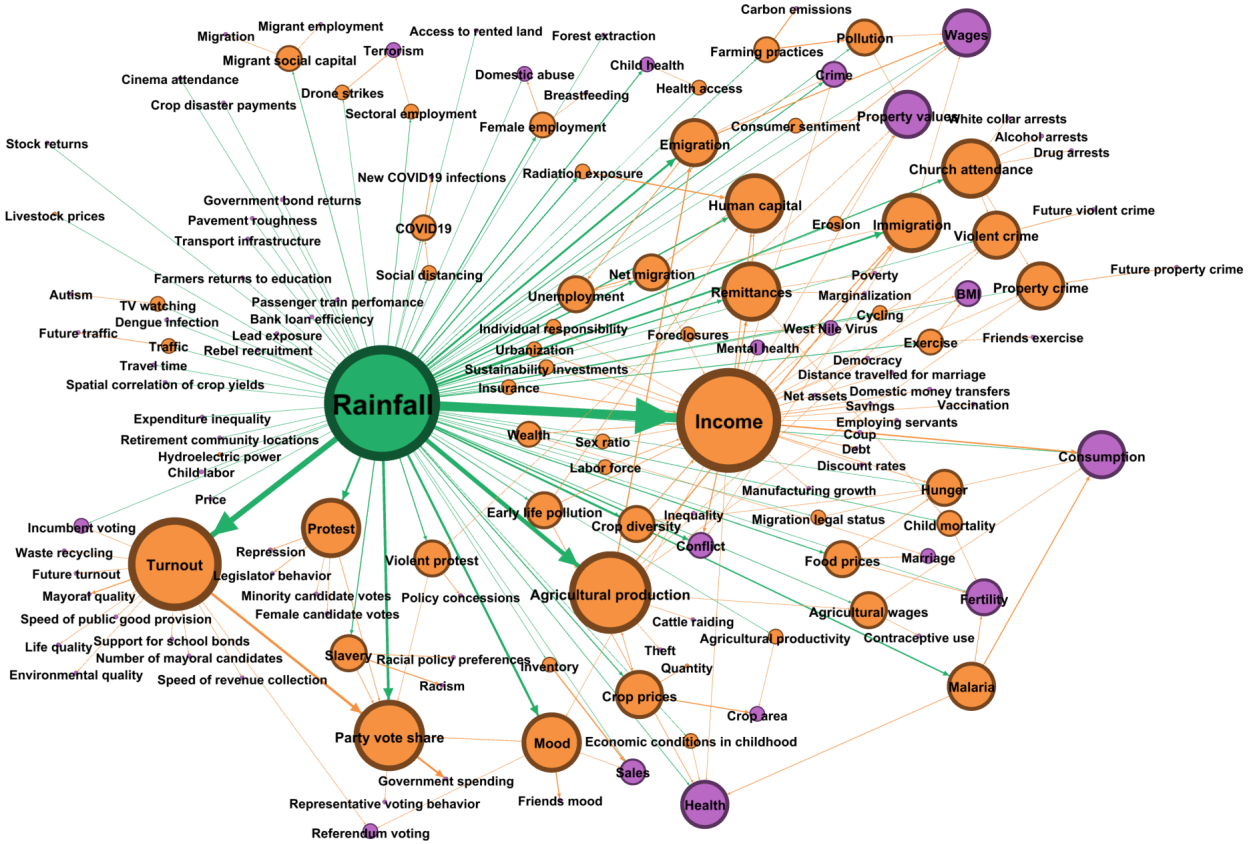
IV in observational studies

- Most applications of IV are not treatment non-compliance.
- But all follow the same underlying logic.
  - Treatment of interest is not randomized...but there exists a real or "natural" experiment that is.
  - And this natural experiment affects the outcome only through its effect on the treatment of interest.
- Examples:
  - Angrist (1990) - Vietnam draft lottery number as an instrument for the effect of military service on income.
  - Angrist and Krueger (1991) - Birth quarter as an instrument for education's effect on income.
  - Acemoglu et. al. (2001) - European settler mortality as an instrument for the effect of institutional quality on GDP per capita.
  - Kern & Hainmueller (2009) - West German TV signal strength as an instrument for the effect of watching West German TV on support for the East German regime
- Challenges
  - Exogeneity/ignorability isn't guaranteed
  - If an instrument has a large effect on your treatment of interest, it probably has an effect on other stuff that could affect the outcome as well (violating the exclusion restriction)

Discussion: The rainfall instrument

- **Miguel, Satyanth, and Sergenti (2004, JPE)** look at the effect of economic growth on civil conflict in 41 African countries.
  - Growth and conflict are confounded (e.g. by political institutions).
  - Instrument for GDP growth using the annual change in rainfall -- for heavily agrarian countries, rainfall fluctuations determine crop yields which are a large component of GDP.
  - Observe that changes in rainfall are associated with changes in GDP and negative GDP shocks (instrumented by rainfall) increase civil conflict.
- Does this satisfy the IV identification assumptions?
  - Exogeneity? Is rainfall as-good-as randomly assigned?
  - Monotonicity? Do positive rainfall shocks *strictly* boost GDP per capita?
  - Exclusion restriction? Is rainfall's effect transmitted only through the mechanism the authors define?

Discussion: The rainfall instrument



Mellon (2023) "Rain, Rain, Go Away: 195 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable"

Estimation and inference for IV

IV with constant effects

- Let's consider a linear model for the potential outcomes

$$Y_i(d) = \alpha + \tau d + \gamma U_i + \eta_i$$

- If we could control for U_i, we could estimate the regression to get an estimate of τ

$$Y_i = \alpha + \tau D_i + \gamma U_i + \eta_i$$

- But we can't - and regressing Y_i on D_i alone will not give a consistent estimator of τ since $Cov(\gamma U_i + \eta_i, D_i) \neq 0$

IV with constant effects

- Suppose Z_i is an instrument that is exogeneous and satisfies the exclusion restriction

$$Cov(\gamma U_i + \eta_i, Z_i) = 0$$

- Then, we can identify τ

$$\begin{aligned} Cov(Y_i, Z_i) &= Cov(\alpha + \tau D_i + \gamma U_i + \eta_i, Z_i) \\ &= Cov(\alpha, Z_i) + Cov(\tau D_i, Z_i) + Cov(\gamma U_i + \eta_i, Z_i) \\ &= \tau Cov(D_i, Z_i) \end{aligned}$$

- Which gives us our IV estimand

$$\tau = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)}$$

IV estimator

- We can estimate τ by plugging in the sample quantities.

$$\hat{\tau_{\text{IV}}} = \frac{\widehat{Cov(Y_i, Z_i)}}{\widehat{Cov(D_i, Z_i)}}$$

- This also can be written as a ratio of two regression coefficients

$$\hat{\tau_{\text{IV}}} = \frac{\widehat{Cov(Y_i, Z_i)}/\widehat{Var(Z_i)}}{\widehat{Cov(D_i, Z_i)}/\widehat{Var(Z_i)}}$$

- Denominator: "First stage": Regression of D_i on Z_i
- Numerator: "Reduced form": Regression of Y_i on Z_i

2SLS - Including covariates

- What if ignorability of Z_i only holds conditional on X_i -- or we want to include X_i as predictors to improve precision.
- We'll assume a particular structure for the outcome and treatment models

$$Y_i = X_i'\beta + \tau D_i + \epsilon_i$$

$$D_i = X_i'\alpha + \gamma Z_i + \nu_i$$

- Assume the X_i are exogenous but not excluded (appear in both equations). D_i is still endogenous so we can't get the treatment effect by just regressing outcome on treatment and covariates.

- Can we get an expression for Y_i in the form of Z_i alone?

2SLS - Including covariates

- Substitute in for D_i

$$\begin{aligned} Y_i &= X_i'\beta + \tau\left[X_i'\alpha + \gamma Z_i + \nu_i\right] + \epsilon_i \\ &= X_i'\beta + \tau\left[X_i'\alpha + \gamma Z_i\right] + \left[\tau\nu_i + \epsilon_i\right] \\ &= X_i'\beta + \tau E[D_i|X_i, Z_i] + \epsilon_i^* \end{aligned}$$

- We can identify τ by regressing Y_i on X_i and the **fitted values** from a regression of D_i on X_i and the instrument Z_i.
- **Intuition** -- we want to only use the variation in D_i that is driven by the **exogenous** factor Z_i.

2SLS - Including covariates

- You can also still get the ratio form of the IV estimator

$$\begin{aligned} Y_i &= X_i'\beta + \tau\left[X_i'\alpha + \gamma Z_i + \nu_i\right] + \epsilon_i \\ &= X_i'(\beta + \tau\alpha) + \tau\gamma Z_i + \left[\tau\nu_i + \epsilon_i\right] \end{aligned}$$

- The coefficient on Z_i in the reduced form regression is $\tau\gamma$
- The coefficient on Z_i in the first-stage is γ.
- So the ratio of the reduced form to the first-stage regression is τ.

Two-stage least squares

- **First stage** - Regress D_i on X_i and Z_i. Get the fitted values $\hat{D}_i$

$$\hat{D}_i = X_i'\hat{\alpha} + \hat{\gamma} Z_i$$

- **Second stage** - Regress Y_i on X_i and fitted values $\hat{D}_i$

$$\hat{Y}_i = X_i'\hat{\beta} + \hat{\tau}\hat{D}_i$$

- The coefficient on the fitted values is the IV estimate
  - But, the standard errors will be wrong - Why?

Illustrating 2SLS

- Recall our Gerber, Karlan and Bergan (2009, AEJ:AE) experiment
    - Z_i: Random assignment to receive a free subscription to the Washington Post
    - D_i: Actually subscribing to the Washington Post (as measured by a post-encouragement survey)
    - Y_i: 2005 Turnout (measured in the survey)
    - X_i: Gender, Age
- Let's load and subset

```
green <- read_dta("assets/publicdata.dta")
wapost <- green %>% filter(treatment != "TIMES"&!is.na(getpost)&!is.na(voted)&!is.na(Bfemale)&!
```

Illustrating 2SLS

- Our first stage regresses subscription on assignment + covariates

```
first_stage <- lm_robust(getpost ~ post + Bfemale + reportedage , data= wapost)
summary(first_stage)
```

```
##
## Call:
## lm_robust(formula = getpost ~ post + Bfemale + reportedage, data = wapost)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  CI Lower CI Upper  DF
## (Intercept)  0.12097    0.06406   1.889 5.94e-02 -0.004786  0.24673 729
## post         0.35233    0.03482  10.118 1.31e-22  0.283965  0.42069 729
## Bfemale     -0.00435    0.03505  -0.124 9.01e-01 -0.073156  0.06445 729
## reportedage  0.00170    0.00125   1.360 1.74e-01 -0.000756  0.00416 729
##
## Multiple R-squared:  0.134 ,    Adjusted R-squared:  0.13
## F-statistic: 35.4 on 3 and 729 DF,  p-value: <2e-16
```

Illustrating 2SLS

- Let's actually run 2SLS - I like two routines: **`iv_robust`** in **`estimatr`** (does 2SLS with robust SEs) and **`ivmodel`** in **`ivmodel`** (does robust 2SLS *and* weak-instrument robust tests + other diagnostics)

```
wapo_2sls <- iv_robust(voted ~ getpost  + Bfemale + reportedage | post + Bfemale + reportedage,
summary(wapo_2sls)
```

```
##
## Call:
## iv_robust(formula = voted ~ getpost + Bfemale + reportedage |
##     post + Bfemale + reportedage, data = wapost)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  0.22562    0.07334  3.0766 2.17e-03  0.08165   0.3696 729
## getpost      0.00428    0.09086  0.0471 9.62e-01 -0.17410   0.1827 729
## Bfemale     -0.03495    0.03354 -1.0423 2.98e-01 -0.10079   0.0309 729
## reportedage  0.01040    0.00127  8.1879 1.19e-15  0.00791   0.0129 729
##
## Multiple R-squared:  0.093 ,    Adjusted R-squared:  0.0893
## F-statistic: 23.3 on 3 and 729 DF,  p-value: 2.15e-14
```

Illustrating 2SLS

```
wapo_2sls2 <- ivmodelFormula(voted ~ getpost  + Bfemale + reportedage | post + Bfemale + report
summary(wapo_2sls2)
```

```
##
## Call:
## ivmodel(Y = Y, D = D, Z = Z, X = X, intercept = intercept, beta0 = beta0,
##     alpha = alpha, k = k, manyweakSE = manyweakSE, heteroSE = heteroSE,
##     clusterID = clusterID, deltarange = deltarange, na.action = na.action)
## sample size: 733
## _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
##
## First Stage Regression Result:
##
## F=111, df1=1, df2=729, p-value is <2e-16
## R-squared=0.132,   Adjusted R-squared=0.131
## Residual standard error: 0.444 on 730 degrees of freedom
## _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
##
## Coefficients of k-Class Estimators:
##
##              k Estimate Std. Error t value Pr(>|t|)
## OLS    0.00000  0.04708    0.03247    1.45     0.15
## Fuller 0.99863  0.00472    0.08979    0.05     0.96
## TSLS   1.00000  0.00428    0.09060    0.05     0.96
## LIML   1.00000  0.00428    0.09060    0.05     0.96
##
```

The weak instrument problem

- Our ratio estimator is consistent

$$\hat{\tau}_{\text{IV}} = \frac{\widehat{Cov(Y_i, Z_i)}}{\widehat{Cov(D_i, Z_i)}} \xrightarrow{p} \tau + \frac{Cov(U_i, Z_i)}{Cov(D_i, Z_i)}$$

- Under exogeneity $Cov(Z_i, U_i)$ is zero.
- However, when there are small violations of exogeneity, a weak instrument will amplify them.
- More generally, with a weak instrument, our t-ratio hypothesis tests assuming asymptotic normality will have **incorrect** type-1 error rates.
  - Why? Distributions of ratios are poorly behaved.

The weak instrument problem

- Let's use a simulation to see how bad the bias can be in IV versus just a simple OLS regression of outcome on treatment under unobserved confounding.
- Let $U_i \sim \mathcal{N}(0, 1)$ be an unobserved confounder. $Z_i \sim \text{Bern}(.5)$ is an **exogenous** instrument.

- The probability of treatment is modeled via a logit

$$\log\left(\frac{P(D_i = 1|Z_i, U_i)}{1 - P(D_i = 1|Z_i, U_i)}\right) = \gamma Z_i + U_i$$

- γ here captures the relationship between the exogenous instrument Z_i and the treatment

- The outcome is a function of U and a mean zero error term ϵ_i only, so the true treatment effect is 0

$$Y_i = U_i + \epsilon_i$$

The weak instrument problem

- Let's see how the Wald estimator performs when we have a pretty large effect of Z_i on D_i: $\gamma = 2$ and $N = 1000$

```
## First stage effect
mean(firststage)
```

```
## [1] 0.345
```

```
## F-statistic from the first stage
mean(firststageF)
```

```
## [1] 157
```

```
## Bias of the naive OLS Y ~ X
mean(naive)
```

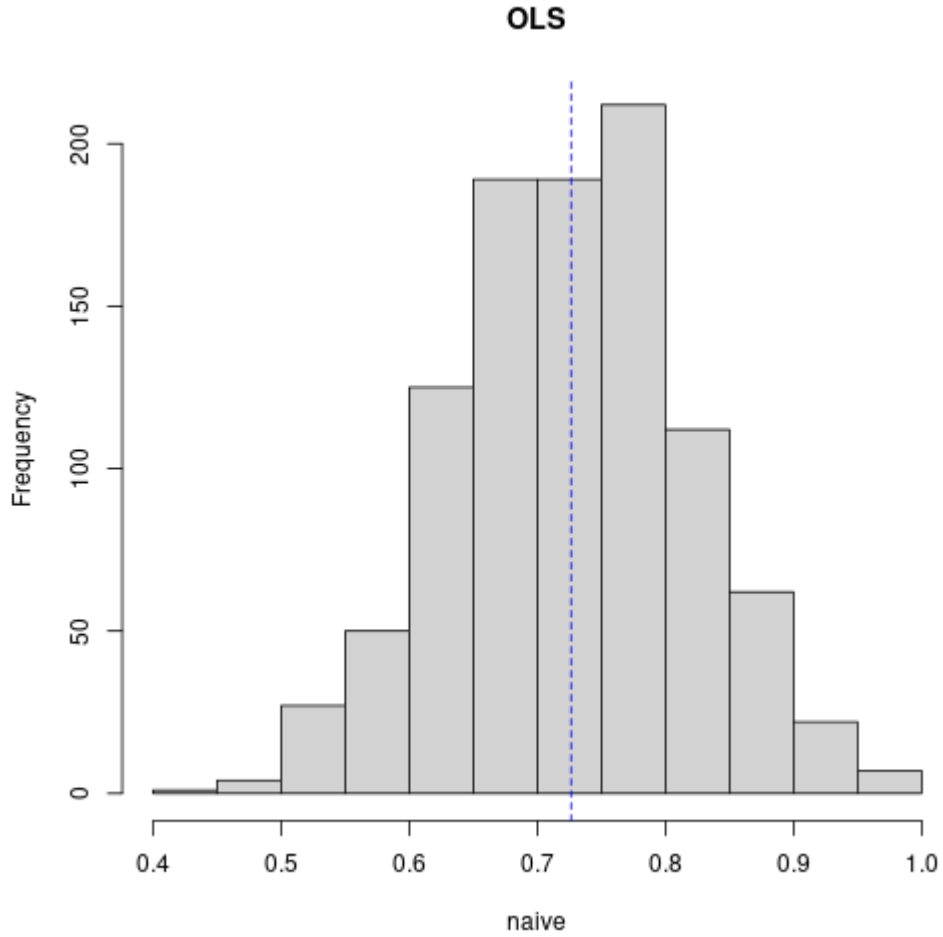
```
## [1] 0.726
```

```
## Bias of IV
mean(IV)
```

```
## [1] -0.0196
```

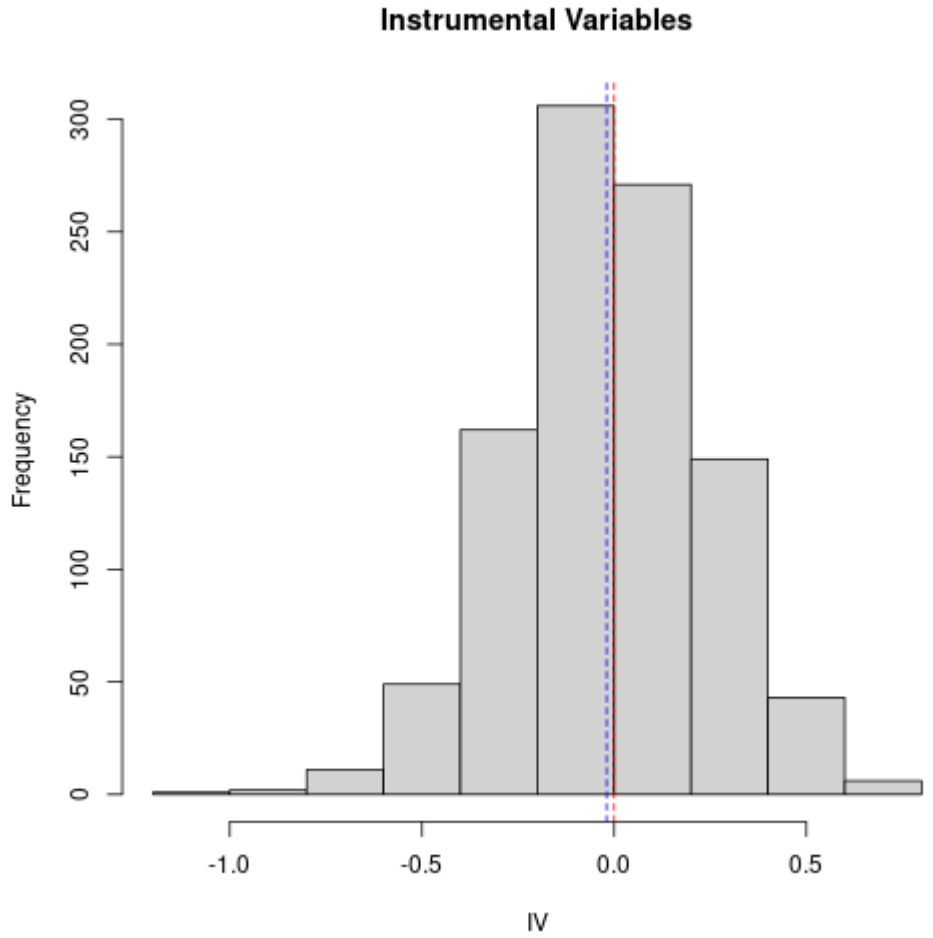
The weak instrument problem

- Sampling distribution of the naive OLS estimator



The weak instrument problem

- Sampling distribution of the IV estimator



The weak instrument problem

- Now, what happens when our instrument is weak: $\gamma = .2$ and $N = 1000$

```
## First stage effect
mean(firststage)
```

```
## [1] 0.041
```

```
## F-statistic from the first stage
mean(firststageF)
```

```
## [1] 2.6
```

```
## Bias of the naive OLS Y ~ X
mean(naive)
```

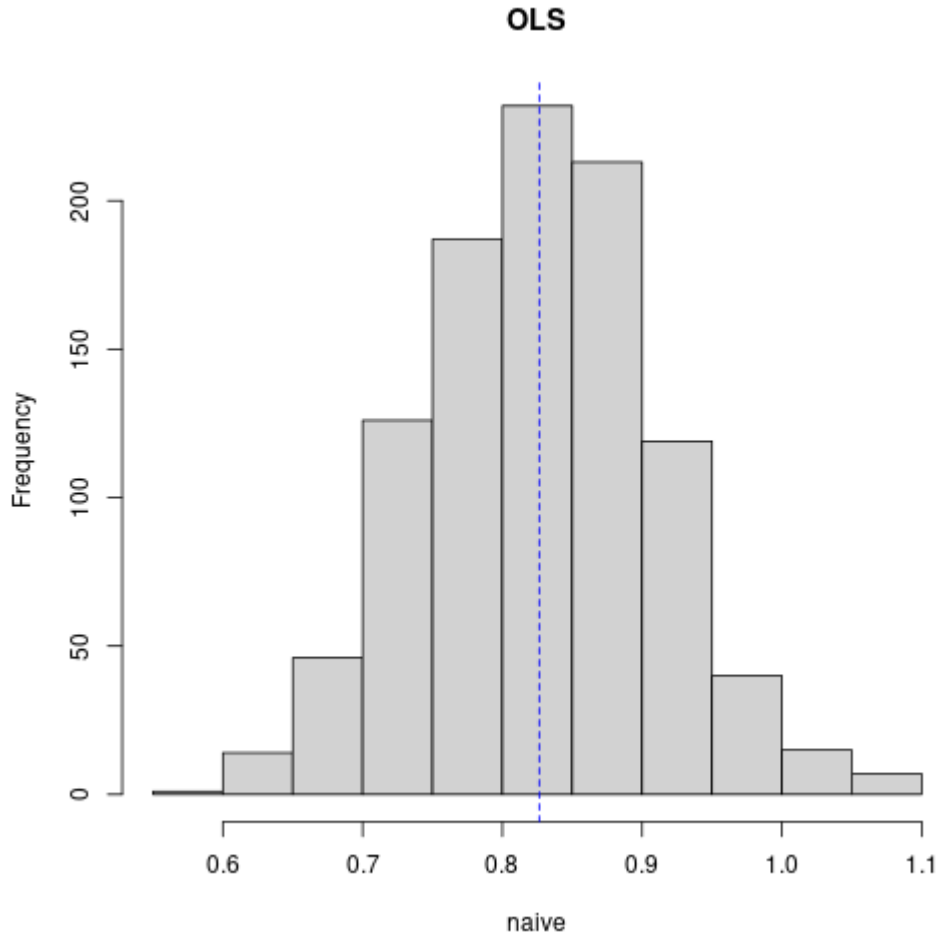
```
## [1] 0.826
```

```
## Bias of IV
mean(IV)
```

```
## [1] -1.04
```

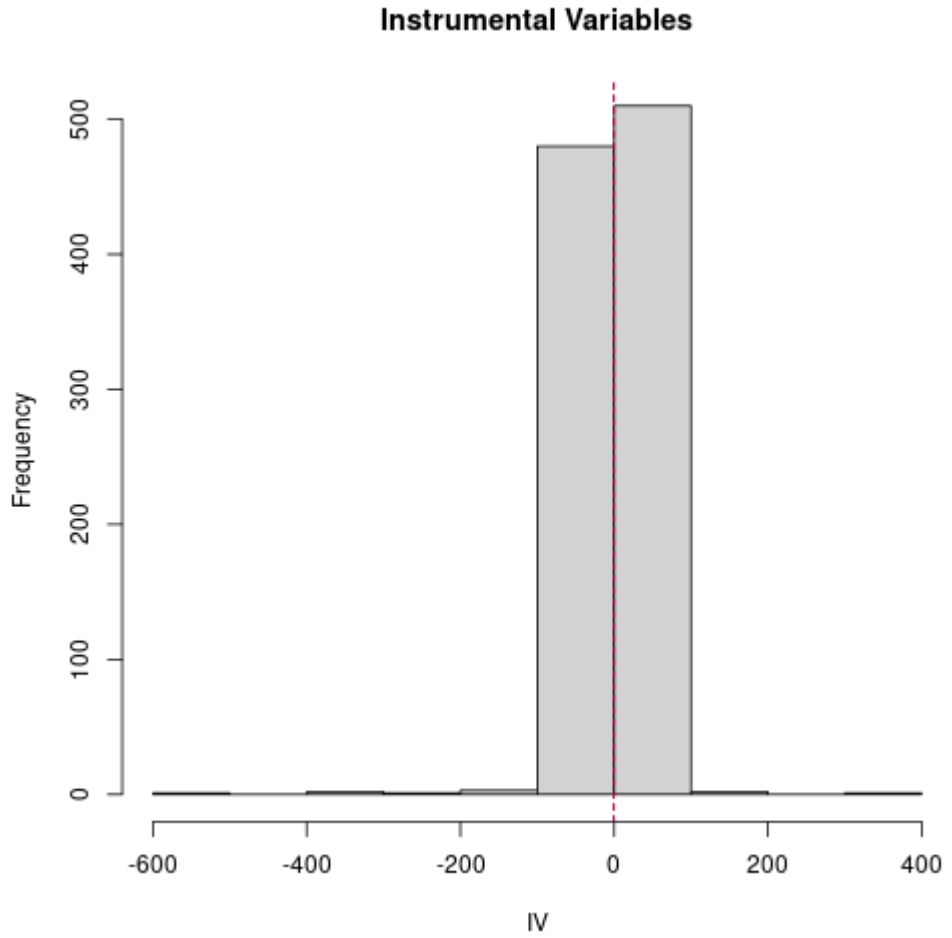
The weak instrument problem

- Sampling distribution of the naive OLS estimator



The weak instrument problem

- Sampling distribution of the IV estimator



The weak instrument problem

- When is an instrument too weak?
- **Classic result:** Staiger and Stock (1997), Stock and Yogo (2005) use first stage F-statistic thresholds
  - $\rightsquigarrow$ heuristic of first-stage F-statistic below 10.
- **Recently:** Lee, Moreira, McCrary, Porter (2020) -- If we want to use the F-statistic as a screen then we actually need $F > 104.7$
- Suggestions:
  - Permutation tests using a test statistic that does not depend on the first stage.
  - **Anderson-Rubin (1949)** approach
  - Angrist and Pischke (2009) - with "just-identified" IV (number of instruments = number of endogenous variables) bias is usually overwhelmed by the large standard errors.

Permutation test

- When the assignment process of Z_i is known, we can construct hypothesis tests using permutation inference assuming a constant treatment effect τ **(Imbens and Rosenbaum, 2005)**.
  - With a single, de-meaned instrument $\tilde{Z}_i = Z_i - \bar{Z}$, we can construct a test statistic based on the sample covariance between Z_i and Y_i with the effect removed:

$$T(\tau) = \frac{1}{N} \sum_{i=1}^{N} \tilde{Z}_i \times (Y_i - \tau D_i)$$

- If the instrument is valid, under the null hypothesis that $\tau = \tau_0$, we can get the **randomization distribution** of the test statistic by simply re-randomizing treatment according to the known assignment process.
  - Construct confidence intervals by "inverting the test" - what values of τ_0 does the test fail to reject?
- Alternative test statistics based on ranks of $Y_i - \tau D_i$) (possibly within strata) can also be used.

Anderson-Rubin Test

- Even when the assignment process is not known, the IV assumptions allow us to construct a test statistic that does not depend on the first stage.
  - This is the **Anderson-Rubin (1949)** approach - **Andrews, Stock and Sun (2019)** provide a good explanation especially for the "just-identified" case
- Let $\hat{\delta}_{\text{ITT}}$ be the **reduced form** or intent-to-treat estimate.
- The instrumental variables assumptions that the reduced form is related to the first stage π and the treatment effect τ

$$\delta_{\text{ITT}} = \gamma \times \tau$$

- Assuming a particular null $H_0 : \tau = \tau_0$ implies that

$$\delta_{\text{ITT}} - \gamma \times \tau_0 = 0$$

Anderson-Rubin Test

- And so we can construct a test statistic based on the difference between the estimated ITT and the estimated first stage adjusted by the null which we know is normal in large samples.

$$g(\tau_0) = \hat{\delta}_{\text{ITT}} - \hat{\gamma}\tau_0 \sim \mathcal{N}(0, \Omega(\tau_0))$$

- The Anderson-Rubin (1949) test statistic is:

$$AR(\tau) = g(\tau)'\Omega(\tau_0)^{-1}g(\tau)$$

Under the null $H_0 : \tau = \tau_0$, this has a chi-squared distribution which does not depend on the value of the first stage.

- **Intuitively**: Statistical properties of **differences** in two normal random variables are well-known and easy. Statistical properties of **ratios** are much more complicated!
  - Again, invert the test to get a confidence interval
  - Can get **infinite** confidence bounds with a weak instrument - the test **never rejects** for any value of τ_0

Example: Strong instrument

```
wapo_iv <- ivmodelFormula(voted ~ getpost  | post , data= wapost, heteroSE=T)
print(AR.test(wapo_iv))
```

```
## $Fstat
## [1] 0.0171
##
## $df
## [1]   1 731
##
## $p.value
## [1] 0.896
##
## $ci.info
## [1] "[-0.2058034696935, 0.175035583124674]"
##
## $ci
##       lower upper
## [1,] -0.206 0.175
```

Example: Weak instrument

```
weak_iv_data <- data.frame(Y = Y, D= D, Z=Z)
weak_iv <- ivmodelFormula(Y ~ D | Z , data= weak_iv_data, heteroSE=T)
print(AR.test(weak_iv))
```

```
## $Fstat
## [1] 2.26
##
## $df
## [1]   1 998
##
## $p.value
## [1] 0.133
##
## $ci.info
## [1] "Whole Real Line"
##
## $ci
##      lower upper
## [1,]  -Inf   Inf
```

Conclusion

- Instrumental variables lets us leverage *alternative* sources of randomness to learn about an otherwise confounded causal relationship.
- An instrument:
  - Affects treatment
  - Doesn't affect the outcome except through treatment
  - Is ignorable w.r.t the outcome.

- LATE theorem: The IV estimand is the ATE among those who would take treatment due to the instrument.

  - With continuous treatment/instrument - a weighted average of LATEs (Angrist and Imbens, 1995)
  - With covariates - a weighted average of covariate-specific LATEs
  - But be careful with this interpretation when the model is not fully saturated (Słoczyński, 2022)

- Statistical inference is tricky

  - Beware weak instruments - typical large-sample asymptotics do poorly when instruments are irrelevant.
  - Consider weak-instrument robust tests (Anderson-Rubin)
  - If it's not in the reduced form, it's not real.