9. Two Stage Least Squares and Modern IV Estimators

ISS5096 || ECI
Jaewon (“Jay-one”) Yoo
National Tsing Hua University

Outline

1. Two-Stage Least Squares

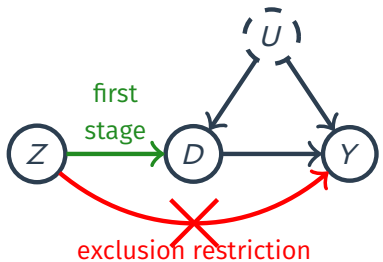
2. Modern IV via Machine Learning

3. Estimating IV: AJR Walkthrough + Modern IV

4. Notable IV Designs in Recent Empirical Work

Where are we? Where are we going?

- Last time:
  - **Instrumental variable** under noncompliance in randomized experiments.
    - The local ATE (or CACE): $\tau_{\text{LATE}} = \text{ITT}_{Y,\text{co}} = \frac{\text{ITT}_Y}{\text{ITT}_D}$
    - The Wald/IV estimator: $\widehat{\tau}_{\text{iv}} = \widehat{\text{ITT}}_Y / \widehat{\text{ITT}}_D$
    - Intent-to-treat analysis, compliance types, identification assumptions..



- Today:
  1. **Two-Stage Least Squares** (TSLS): estimands, weak IVs, multivalued D, general 2SLS.
  2. **Modern IV via Machine Learning**: PLIV (constant τ), DRIV (heterogeneous $\tau(\mathbf{X})$), DML.
  3. **Estimating IV in practice**: AJR walkthrough + modern IV in R.
  4. **Notable IV designs**: treatment assignments, regional characteristics, peers' environments, shift-share.

Source: *Chapter 4* of Mostly Harmless Econometrics (Textbook 1) by J. Angrist & J. Pischke

1/ Two-Stage Least Squares

Two Stage Least Squares

- **Two-stage least squares** (TSLS) is the classical approach to IV which assumes two linear models:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \eta_i$$

- Here, the treatment D_i is **endogenous** so $\mathbb{E}[\varepsilon_i | D_i] \neq 0$.
  - Canonical source: an omitted variable / unobserved confounder U opening a back-door path $D \leftarrow U \rightarrow Y$.
- But we have an **instrument** Z_i that is exogenous $\mathbb{E}[\varepsilon_i | Z_i] = 0$.
  - It's also exogenous for treatment uptake, so $\mathbb{E}[\eta_i | Z_i] = 0$.
- This implies the following CEF form for Y_i conditional on Z_i:

$$\mathbb{E}[Y_i | Z_i] = \alpha + \tau \mathbb{E}[D_i | Z_i] = \alpha + \tau \cdot (\gamma Z_i)$$

  - α is reused loosely (technically $\alpha + \tau\delta$ here); not the parameter of our interest.

TSLS Estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i|Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ A regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, then we have this simple relationship slopes:

$$\mathbb{E}[D_i|Z_i] = \delta + \gamma Z_i \qquad \rightsquigarrow \qquad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}(Z_i)}$$

- Applying this to above CEF, we have:

$$\tau = \frac{\text{cov}(Y_i, \gamma Z_i)}{\mathbb{V}(\gamma Z_i)} = \frac{\text{cov}(Y_i, Z_i)}{\gamma \mathbb{V}(Z_i)} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

- TSLS estimator:
  - Estimate $\widehat{\gamma}$ from regression of treatment D_i on instrument Z_i
  - Estimate $\widehat{\tau}_{\text{2SLS}}$ as the slope of a regression of Y_i on $\widehat{\gamma} Z_i$.
    - $\widehat{D}_i = \widehat{\delta} + \widehat{\gamma} Z_i$; the Z-driven $\widehat{\gamma} Z_i$ is exogenous, uncorrelated with ε_i (constant $\widehat{\delta}$ doesn't affect the slope).
  - Under this model, $\widehat{\tau}_{\text{2SLS}} \xrightarrow{p} \tau$ (but don't use SEs from second stage; see MHE section 4.6.1. *2SLS Mistakes*)

Binary Treatment and Instrument

- Under binary treatment/instrument, TSLS estimand is the LATE:

$$\tau = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]} = \frac{\text{ITT}_Y}{\text{ITT}_D} = \tau_{\text{LATE}}$$

  - $\text{cov}(Y, Z)/\text{cov}(D, Z) = \beta_Y/\beta_D$ ($\mathbb{V}(Z)$ cancels, regardless of Z). For binary Z, OLS slope $\beta_\cdot = \mathbb{E}[\cdot|Z{=}1] - \mathbb{E}[\cdot|Z{=}0]$ (W4), giving $\text{ITT}_Y/\text{ITT}_D$.
  - LATE defined over compliers $\{D_i(1) > D_i(0)\}$, single group only for binary D.

- And that the TSLS estimator is the Wald estimator:

$$\widehat{\tau}_{\text{2SLS}} = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} = \frac{\widehat{\text{ITT}}_Y}{\widehat{\text{ITT}}_D} = \widehat{\tau}_{\text{iv}}$$

  $\rightsquigarrow$ Constant effects model is not required for TSLS in this setting.

TSLS with Covariates

- Adding covariates linearly:

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i'\beta_y + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \mathbf{X}_i'\beta_d + \eta_i$$

  - This *uninteracted* form imposes constant τ across $\mathbf{X}$. Otherwise τ is an odd weighted function of causal effects and $\tau \neq \tau_{\text{LATE}}$.

- Why not Lin (W4) fully-interacted models ($\tau + \gamma'\widetilde{\mathbf{X}}_i$ varies) in IV?
  1. Each $D_i\widetilde{X}_{ij}$ is endogenous and requires its own instrument, e.g., $Z_i\widetilde{X}_{ij}$.
  2. Strong prediction of D by Z does not imply strong prediction of $D_i\widetilde{X}_{ij}$ by $Z_i\widetilde{X}_{ij}$.
  3. Many interaction terms can amplify weak-IV problems and complicate interpretation.

- $\rightsquigarrow$ **PLIV** (§2) inherits constant τ (ML-flexible $g(\mathbf{X})$, $m(\mathbf{X})$ only). **DRIV** (§2) relaxes to $\tau(\mathbf{X})$.

Weak Instruments

- IV is unstable if instrument weakly affects treatment; $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i, & \mathbb{E}[\varepsilon_i | D_i] &\neq 0 \\ D_i &= 0 \times Z_i + \eta_i, & \mathbb{E}[\epsilon_i | Z_i] &= \mathbb{E}[\eta_i | Z_i] = 0 \end{aligned}$$

- The bias of the Wald estimator:

$$\widehat{\tau}_{\text{iv}} - \tau = \frac{\widehat{\text{cov}}(\tau D_i + \varepsilon_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} - \tau = \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i Z_i}{\frac{1}{n}\sum_{i=1}^n \eta_i Z_i} \xrightarrow{d} \underbrace{\frac{\text{cov}(\varepsilon_i, \eta_i)}{\mathbb{V}[\varepsilon_i]}}_{\text{bias}} + \underbrace{W_i}_{\text{Cauchy r.v.}}$$

- Inconsistent and asymptotically heavy tails (b/c of Cauchy).
  - When $Z \to D$ effect is small but non-zero, similar behavior.
- **Relevant IV**: noise divided by signal $\to 0$; **Irrelevant IV**: noise divided by noise does not.

What to Do About Weak Instruments?

- **Detect**: first-stage F-test on excluded instruments ($H_0 : \gamma = 0$).
  - $F \geq 10 \Rightarrow$ bias is small (Stock & Yogo 2005); still the dominant convention in empirical IV.
  - $F \geq 104.7$ for correct CI coverage in the *worst-case* DGP (Lee et al. 2022).
    - Lee et al. show the standard 1.96 gives true 95% coverage *only* at $F \geq 104.7$; below that, actual coverage drops monotonically. At the conventional $F = 10$, worst-case actual coverage is around **85%**.
- **Weak-IV robust inference**: Anderson-Rubin (1949) test (simplified setting, binary Z, D).
  - **Trick**: rewrite $H_0 : \tau = \tau_0$ as $H_0 : \text{ITT}_Y - \tau_0 \, \text{ITT}_D = 0$: **linearises the null**, no weak-IV pathology.
  - Under the null, asymptotically:

$$g(\tau_0) = \widehat{\text{ITT}}_Y - \tau_0 \, \widehat{\text{ITT}}_D \sim N(0, \Omega(\tau_0))$$
$$\Omega(\tau_0) = \mathbb{V}[\widehat{\text{ITT}}_Y] + \tau_0^2 \, \mathbb{V}[\widehat{\text{ITT}}_D] - 2\tau_0 \, \text{cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_D)$$

  - AR statistic: $g(\tau_0)^2 / \Omega(\tau_0) \sim \chi^2_1$ regardless of first-stage strength; CI by analytical inversion.

```
> library(ivmodel)
> m <- ivmodel(Y = y, D = D, Z = Z)
> m$AR    # AR test (F-stat), p-value, df, and 95% CI by inversion
```

- Increasingly reported alongside the F-stat; handles weak-IV uncertainty correctly when first stage is borderline.

Multi-Valued Treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \ldots, K-1\}$ (*intensity / dosage*)
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(z,d),\ \forall z, d\},\ D_i(1),\ D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
  - Exclusion restriction: $Y_i(1,d) = Y_i(0,d)$ for all $d = 0, 1, \ldots, K-1$
  - Relevance: $\mathbb{P}(D_i(1) \geq j > D_i(0)) > 0$ for at least one j (some threshold has compliers)
- Example: $K = 3 \rightsquigarrow 9$ principal strata
  - Affected: $(D_i(0), D_i(1)) \in \{(0,1), (0,2), (1,2)\}$
  - Unaffected: $(D_i(0), D_i(1)) \in \{(0,0), (1,1), (2,2)\}$
  - Negatively affected: $(D_i(0), D_i(1)) \in \{(1,0), (2,0), (2,1)\}$
  - Last ruled out by monotonicity.

TSLS with Multivalued Treatments: ACR

- Under the assumptions on the previous slide, Angrist–Imbens 1995, JASA, Theorem 1:

$$\widehat{\tau}_{2SLS} \xrightarrow{p} \sum_{j=1}^{K-1} \omega_j \cdot \mathbb{E}\Big[Y_i(j) - Y_i(j-1) \,\Big|\, D_i(1) \geq j > D_i(0) \Big]$$

$$\omega_j = \frac{\mathbb{P}[D_i(1) \geq j > D_i(0)]}{\mathbb{E}[D_i(1) - D_i(0)]}, \qquad \sum_{j=1}^{K-1} \omega_j = 1.$$

- Intuition: a weighted average of effects per dose for each affected type.
  - Weights are proportional to size of the strata and how big the effect of the instrument is for that strata.
  - If instrument can only increase by 1 dose, then simplifies to weighted average of principal strata effects.
- This object is the Average Causal Response (ACR), the multi-valued analogue of LATE (named in MHE Th 4.5.3). Binary case ($K = 2$) collapses to standard LATE.

2/ Modern IV via Machine Learning

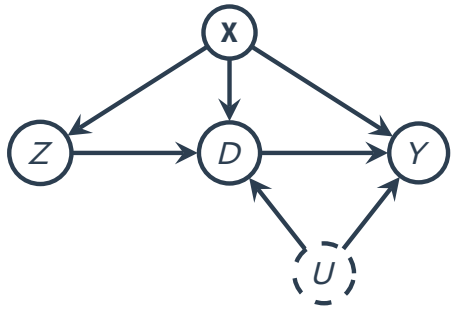
Cf. Chernozhukov et al. (2018), *Double/Debiased Machine Learning*, Econometrics Journal

From Classical TSLS-with-X_i to Modern IV

- **Setting**: empirical IV often needs to condition on many confounders, often unstructured / high-dim $\mathbf{X}_i$ (text, embeddings); theory does not pin down features or functional form.
- **Classical TSLS-with-X_i**: enters $\mathbf{X}_i$ linearly with constant τ. Two limits:
  - Linear/parametric nuisance is fragile when $\mathbf{X}_i$ is rich.
  - $D_i \cdot \mathbf{X}_i$ interactions for HTE $\rightsquigarrow$ Winston Lin fully-interacted IV trap (recall §1).
- **Modern IV ladder**, two steps out:
  - **PLIV** (Chernozhukov et al. 2018; partialling-out idea from Robinson 1988 PLR): keep constant τ; ML handles $\mathbf{X}_i$ flexibly.
  - **DRIV** (Syrgkanis et al. 2019, NeurIPS): drop constant τ; $\tau(\mathbf{X}_i)$ ML-learned $\rightsquigarrow$ HTE.
- Both ride on **Double/Debiased Machine Learning (DML)**, the engine introduced next slide.

What is DML? Three Nuisances + Two Protections

- DAG (covariates $\mathbf{X}_i$, unobserved confounder U_i):



- **Three nuisance functions** = parts of Y, D, Z explained by $\mathbf{X}_i$:

$$\ell_0(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i], \quad r_0(\mathbf{X}_i) = \mathbb{E}[D_i \mid \mathbf{X}_i], \quad m_0(\mathbf{X}_i) = \mathbb{E}[Z_i \mid \mathbf{X}_i].$$

- **Naive approach**: estimate the nuisance functions with ML, residualize using the same data, and treat the resulting residuals as if they were based on the true CEFs.
- **DML-IV keeps the residual-IV idea, but makes it valid for flexible ML**:
  1. **Cross-fitting**: residuals are *out-of-fold*, not in-sample fitted residuals.
  2. **Orthogonalized moment**: Y, D, Z are all residualized, so first-step errors (which are functions of $\mathbf{X}_i$) have limited first-order impact on $\widehat{\tau}$.

Both baked into the algorithm next slide. See Chernozhukov et al. 2018.

PLIV and DRIV via DML

- **PLIV model**: $Y_i = \tau D_i + g(\mathbf{X}_i) + \varepsilon_i$, $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i, Z_i] = 0$. Constant τ; $g(\mathbf{X}_i)$ ML-flexible.
- **Algorithm** (K-fold cross-fitting; non-linear Frisch-Waugh-Lovell):
  1. Partition into folds $\{I_k\}_{k=1}^K$. On data *outside* I_k, fit ML estimates $\widehat{\ell}_{[k]}, \widehat{r}_{[k]}, \widehat{m}_{[k]}$.
  2. For $i \in I_k$: $\widetilde{Y}_i = Y_i - \widehat{\ell}_{[k]}(\mathbf{X}_i)$, $\widetilde{D}_i = D_i - \widehat{r}_{[k]}(\mathbf{X}_i)$, $\widetilde{Z}_i = Z_i - \widehat{m}_{[k]}(\mathbf{X}_i)$.
  3. Pooled Wald on residuals: $\widehat{\tau} = \frac{\sum_i \widetilde{Z}_i \widetilde{Y}_i}{\sum_i \widetilde{Z}_i \widetilde{D}_i}$.

  $\rightsquigarrow$ *PLIV = Wald estimator on ML-residualized variables.* Both DML protections (cross-fitting + orthogonalization) baked in.
- **DRIV** (Syrgkanis et al. 2019, NeurIPS): drop the constant-τ assumption $\rightsquigarrow$ $\tau(\mathbf{X}_i)$ ML-learned.

$$\text{PLIV: } Y_i = \tau D_i + g(\mathbf{X}_i) + \varepsilon_i \quad \rightsquigarrow \quad \text{DRIV: } Y_i = \theta(\mathbf{X}_i) D_i + g(\mathbf{X}_i) + \varepsilon_i.$$

Same Wald ratio on residuals, but kept *conditional on* $\mathbf{X}_i$ instead of pooled:

$$\theta(\mathbf{X}_i) = \frac{\mathbb{E}[\widetilde{Z}_i \widetilde{Y}_i \mid \mathbf{X}_i]}{\mathbb{E}[\widetilde{Z}_i \widetilde{D}_i \mid \mathbf{X}_i]}.$$

PLIV averages over $\mathbf{X}_i$ $\rightsquigarrow$ one number τ. DRIV keeps the conditioning $\rightsquigarrow$ function $\theta(\mathbf{X})$ learned by ML. HTE built into the spec.

Choosing the ML Learner: How Sensitive Is DML?

- **Natural question**: we said "use ML" for the nuisance functions, but *which* ML? Lasso, RF, XGBoost, neural net are all candidates, and different choices can yield very different DML point estimates on the same data.
- Open question until Ahrens, Chernozhukov, Hansen, Kozbur, Schaffer, Wiemann (2025) took it on. They re-examine Dube, Jacobs, Naidu, Suri (2020) (*AER: Insights*, online labor monopsony, $n \approx 258$k): **18 candidate learners $\Rightarrow$ point estimates from** -0.39 **to** $+0.15$**.**
- **Two remedies they propose**:
  - **CVC test** (Cross-Validation with Confidence): H_0 : "this learner has the lowest predictive risk among candidates." Keep learners with $p > 0.1{\sim}0.2$; report DML using that subset.
  - **Stacking / model averaging** (*super learner*): data-driven weighted average over the full set of candidate learners. Single point estimate, no manual learner pick.
- **Practical**: `ddml` R package implements both. Report cross-fitted R^2 for each nuisance regression as a sanity check before using its DML estimate.

Open Question: $\mathbf{X}_i$ Selection Itself

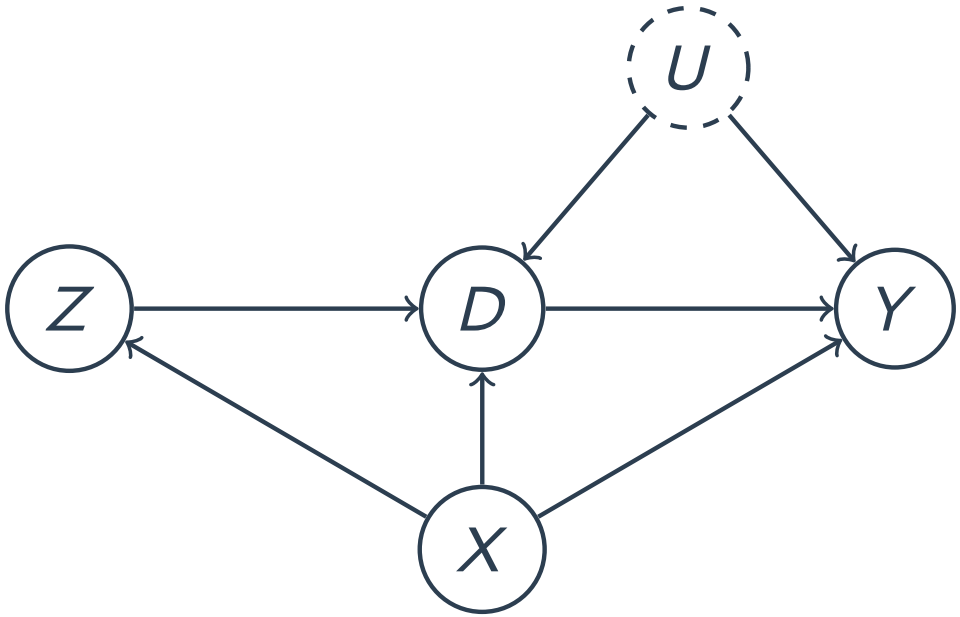
- DML / DRIV assumes the analyst hands the algorithm a "right" set of $\mathbf{X}_i$ to condition on. Which $\mathbf{X}_i$? is itself a causal-graph question.
- **Canonical pitfall**: conditioning on a *collider* opens a spurious D_i-Y_i path that does not exist marginally. Classical FWL/OLS practice catches this via DAG reasoning before any regression (recall ECI W6).
- **ML cannot recognize colliders**: a tree or Lasso fitting $\mathbf{X}_i \longrightarrow Y_i$ uses any predictive signal, including signal operating through a collider. Variable importance is *not* causal validity. See Hünermund, Louw & Caspi (2023, *JCI*).
- Active research direction: combining ML's prediction strength with causal-graph reasoning so that the analyst (or the algorithm) does not silently introduce collider bias when assembling $\mathbf{X}_i$. *Work in progress.*

3/ Estimating IV: AJR Walkthrough + Modern IV

From Theory to Practice: Classical TSLS Setup

- Modern IV (§2) shines when **X** is high-dim or text-scale. AJR sits in the **classical corner**: a single binary IV, modest **X**; classical TSLS suffices.
- Classical-corner ingredients:
  1. observational data with *unobserved* confounders U;
  2. a single instrumental variable Z with a qualitatively defensible exclusion restriction;
  3. low-dim observed covariates $\mathbf{X}_i$ (linearly conditioned).
- DAG (recap from §2): see next slide.

Classical TSLS Setup: DAG (recap from §2)



- $Z \to D$: first stage. $D \to Y$: causal effect of interest. U: unobserved confounder. X: observed covariates.
- Exclusion restriction: Z affects Y only through D (no direct $Z \to Y$ arrow).

Property Rights & Economic Development

Created using DALL·E 3 motivated by Ratna Sagar Shrestha

Recognize the person on the right?

- Q: Do property rights (i.e., institutions) promote economic development?
    - Famous paper on this: Acemoglu, Johnson, and Robinson (2001) AER
    - Relationship between strength of property rights in a country and GDP.

The AJR Data

Name	Description
shortnam	three-letter country code
africa	indicator for if the country is in Africa
asia	indicator for if country is in Asia
logem4	log mortality rates faced by European settlers (IV)
avexpr	strength of property rights (protection against expropriation)
logpgp95	log GDP per capita

```
> ajr <- read_csv("https://bit.ly/3RUJDWK"); ajr

# A tibble: 163 × 15
   shortnam africa lat_abst malfal94 avexpr logpgp95 logem4  asia yellow baseco leb95
   <chr>     <dbl>    <dbl>    <dbl>  <dbl>    <dbl>  <dbl> <dbl>  <dbl>  <dbl> <dbl>
 1 AFG           0   0.367   0.00372  NA       NA      4.54     1      0     NA  NA
 2 AGO           1   0.137   0.950     5.36     7.77   5.63     0      1      1  46.5
 3 ARE           0   0.267   0.0123    7.18     9.80  NA        1      0     NA  NA
 4 ARG           0   0.378   0         6.39     9.13   4.23     0      0      1  72.9
 5 ARM           0   0.444   0        NA        7.68  NA        1      0     NA  NA
 6 AUS           0   0.300   0         9.32     9.90   2.15     0      1      1  78.2
 7 AUT           0   0.524   0         9.73     9.97  NA        0      0     NA  NA
 8 AZE           0   0.448   0        NA        7.31  NA        1      0     NA  NA
 9 BDI           1   0.0367  0.950    NA        6.57   5.63     0      1     NA  NA
10 BEL           0   0.561   0         9.68     9.99  NA        0      0     NA  NA
# i 153 more rows
# i 4 more variables: imr95 <dbl>, meantemp <dbl>, lt100km <dbl>, latabs <dbl>
# i Use `print(n = ...)` to see more rows
```

In R: Example Code Using ajr Data

```
# Center (i.e., demeaning) the variables
> ajr <- ajr |>
    mutate(
      D_cnt = avexpr - mean(avexpr, na.rm = TRUE),
      Y_cnt = logpgp95 - mean(logpgp95, na.rm = TRUE),
      Z_cnt = logem4 - mean(logem4, na.rm = TRUE)
    ) |>
    na.omit()

# Compute the ITTs on D and on Y:
> ITT_D <- cov(ajr$Z_cnt, ajr$D_cnt)
> ITT_Y <- cov(ajr$Z_cnt, ajr$Y_cnt)
> wald_estimate <- ITT_Y / ITT_D; round(wald_estimate, digits = 4)
[1] 0.9242

# Same as reg Y~Z / reg D~Z:
> ITT_Y <- coef(lm(Y_cnt ~ Z_cnt, data = ajr))[[2]]
> ITT_D <- coef(lm(D_cnt ~ Z_cnt, data = ajr))[[2]]
> wald_estimate <- ITT_Y / ITT_D; round(wald_estimate, digits = 4)
[1] 0.9242
```

- The Wald estimate from manual calculation: 0.9242

In R: Example Code Using ajr Data

```
# Compare with ivreg
> ivreg_result <- AER::ivreg(Y_cnt ~ D_cnt | Z_cnt, data = ajr); summary(ivreg_result)

Call:
AER::ivreg(formula = Y_cnt ~ D_cnt | Z_cnt, data = ajr)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40175 -0.54950  0.01792  0.68944  1.67361

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.031e-16  1.231e-01   0.000        1
D_cnt       9.242e-01  1.547e-01   5.974 1.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9458 on 57 degrees of freedom
Multiple R-Squared: 0.1107,        Adjusted R-squared: 0.09506
Wald test: 35.68 on 1 and 57 DF,  p-value: 1.589e-07

> round(coef(ivreg_result), digits = 4)
(Intercept)       D_cnt
     0.0000      0.9242
```

- The 2sls estimate from AER::`ivreg()`: 0.9242
  - Caveat: compute robust SE separately! (don't use SE from 2nd stage)

In R: Example Code Using ajr Data

```
# Estimate robust SE with estimatr::iv_robust()
> iv_rob <- estimatr::iv_robust(Y_cnt ~ D_cnt | Z_cnt, data = ajr, se_type = "HC2"); iv_rob
               Estimate Std. Error      t value     Pr(>|t|)   CI Lower CI Upper DF
(Intercept) 4.148820e-16  0.1234908 3.359619e-15 1.000000e+00 -0.2472860 0.247286 57
D_cnt       9.242173e-01  0.1791511 5.158871e+00 3.262823e-06  0.5654735 1.282961 57
```

- Robust SE with HC2 using `estimatr::iv_robust()`: 0.1791

```
> iv_rob_cov <- estimatr::iv_robust(Y_cnt ~ D_cnt + lat_abst + asia + africa |
                                    Z_cnt  + lat_abst + asia + africa, data = ajr,
                                  se_type = "HC2")
> var_labels <- c(
   "D_cnt" = "Avg. Expropriation Risk (D_cnt)", "lat_abst" = "Abs. Value of Latitude",
   "asia" = "Asian country", "africa" = "African country", "Z_cnt" = "log Mortality Rate (D_cnt)"
 )
```

- When adjusting for covariates in TSLS, include them both in the 1st and 2nd stage.

IV Regression Table with `modelsummary`

```
> modelsummary::modelsummary(
    models = list(ivreg_result, iv_robust, iv_rob_cov),
    coef_map = var_labels, gof_map = c("nobs", "r.squared", "adj.r.squared"), stars = T,
    notes = "Note: See appendix for other model statistics.", output = "modelsummary_tab.tex")
```

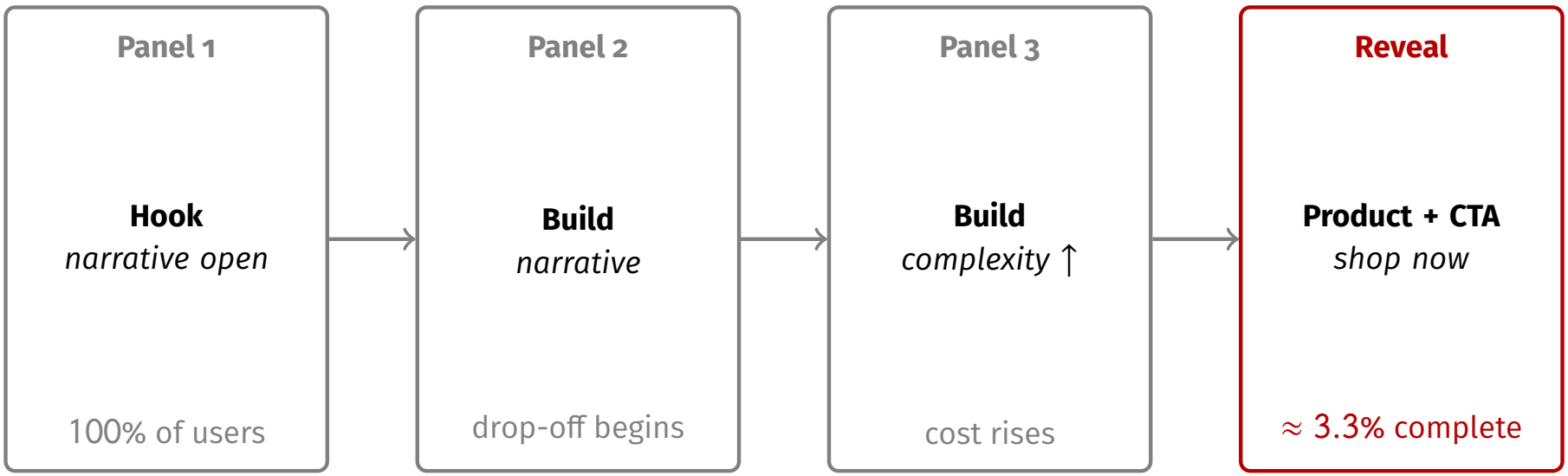
	(1)	(2)	(3)
Avg. Expropriation Risk (D_cnt)	0.924***	0.924***	1.015**
	(0.155)	(0.179)	(0.351)
Abs. Value of Latitude			-1.596
			(1.529)
Asian country			-1.048*
			(0.425)
African country			-0.390
			(0.342)
Num.Obs.	59	59	59
R2	0.111	0.111	0.045
R2 Adj.	0.095	0.095	-0.026

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: See appendix for other model statistics.

What Does a BLI Ad Look Like?

- Blind Lead-In (BLI) ads are multi-stage interactive sliders. The product is revealed *only* on the last panel; users must advance through each panel themselves.



- In passive media (TV, pre-roll), exposure $\Longrightarrow$ reveal automatically. In BLI, only $\approx 3.3\%$ of exposed users reach the reveal.
- $\Longrightarrow$ The behavioral object of interest is D (completion), not Z (exposure). Standard ITT-on-Z averages over 96.7% never-takers.

Modern IV Demo: A Digital Advertising Experiment

- Online book retailer A/B-tests two ad formats ($n \approx 1.4$M user-impressions, 76 ad creatives).
  - Z: random assignment to interactive (Blind Lead-Ins, BLI) vs. static banner.
  - D: BLI *completion* (treated-side completion rate $\approx 3.32\%$; **one-sided noncompliance**: $D = 0$ for all $Z = 0$).
  - Y: 30-day purchase conversion (binary).
- Why this is a useful DML-IV testbed:
  - High-dim creative features (face presence, slide count, character count, color, area ratios).
  - Rich user covariates (spending, visits, age, gender, B2B, genre fit).
  - Plausible $\tau(\mathbf{X})$ heterogeneity along creative-content $\times$ user-engagement axes.
- **Plan**: (i) classical pooled 2SLS LATE (`fixest`); (ii) DML PLIV pooled $\widehat{\tau}$; (iii) DRIV via reticulate (`econml.LinearIntentToTreatDRIV`) for $\widehat{\theta}(\mathbf{X}_i)$ and subgroup CATE; (iv) lift translation.

Pooled 2SLS LATE: Classical Baseline

- 2SLS LATE on 30-day conversion (date FE, user covariates, cluster-robust SE at the user level):

```
> library(fixest)
> late <- feols(lngt_conv ~ ctrl_vars | date | tot ~ treatment,
+               data = d, cluster = ~ memno)
> summary(late)
TSLS estimation, Dep. Var.: lngt_conv
Observations: 1,408,121 / Fixed-effects: date (80)
Standard-errors: Clustered (memno)
        Estimate Std. Error t value Pr(>|t|)
fit_tot 0.002326   0.001210   1.922   0.0548 .
... (control coefficients omitted)
First-stage F (Kleibergen-Paap): ~1.5e4   # very strong instrument
```

- **LATE**: BLI completion raises 30-day conversion by ≈ 0.23 **percentage points** among compliers (`fit_tot` = 0.00233, $p \approx 0.055$).
- **Weak-IV diagnostic**: first-stage $F = 10{,}007$ (Kleibergen-Paap, cluster-robust). Compliance is small (3.32%) but big-tech ad RCTs are typically large-N enough that the first stage is overwhelming; Stock-Yogo $F \geq 10$ **vastly** cleared.

DML PLIV on BLI: Pooled $\widehat{\tau}$

- PLIV via DoubleML R package: ML residualization of Y, D, Z on the rich $\mathbf{X}_i$ vector, cross-fitting + Neyman-orthogonal moment baked in (§2 algorithm). Single call:

```
> data <- DoubleMLData$new(df_bli, y_col = "Y", d_cols = "D",
+                          z_cols = "Z", x_cols = X_names)
> pliv <- DoubleMLPLIV$new(data, ml_l, ml_m, ml_r, n_folds = 5)
> pliv$fit(); pliv$summary()
```

- Pooled $\widehat{\tau}$ comparison on BLI ($n \approx 1.4$M):

Estimator	$\widehat{\tau}$	s.e.
2SLS pooled (above)	0.00233	0.00121
DML PLIV (Lasso nuisance)	0.00217	0.00116

$\Longrightarrow$ DML PLIV reproduces the classical 2SLS pooled estimate; flexible ML handling of $\mathbf{X}_i$ doesn't shift the constant-τ estimate. *HTE next.*

DRIV via Reticulate (Syrgkanis et al. 2019)

- DRIV: heterogeneous LATE $\theta(\mathbf{X}_i) = \alpha + \beta'\mathbf{X}_i$ via cross-fitted nuisance + Neyman-orthogonal score (§2). Implementation: `econml.LinearIntentToTreatDRIV` (Python only). Call from R via `reticulate`:

```
library(reticulate)
econml_iv <- import("econml.iv.dr"); sk <- import("sklearn.ensemble")
est <- econml_iv$LinearIntentToTreatDRIV(
  model_y_xw   = sk$RandomForestRegressor(n_estimators = 100L),
  model_t_xwz  = sk$RandomForestClassifier(n_estimators = 100L),
  flexible_model_effect = "auto",  cv = 5L)
est$fit(Y = Y, T = D, Z = Z, X = X_paper, W = W_paper)
est$ate__inference()$summary_frame()
```

- Pooled ATE: $\widehat{\tau} = 0.00247$ (s.e. 0.00107, 95% CI $[+0.0004, +0.0046]$); same band as 2SLS / DML PLIV (lift table next slide).
- **Subgroup CATE** (average $\widehat{\theta}(\mathbf{X}_i)$ within subgroup) exposing where the effect concentrates:

Subgroup	n	CATE
Engaged (high recent visit-days)	686,903	0.00222
Disengaged (low recent visit-days)	721,218	0.00114
`face_early` present	586,313	0.00235
`face_early` absent	821,808	0.00118

Engaged × `face_early` users absorb *nearly all* the BLI completion effect (≈ 2× pooled); disengaged + no early face: near-zero. The split lines up with *information foraging theory*: engaged users plus an early visual anchor commit more attention to scanning the funnel. Modern IV (DRIV here) is what enables HTE estimation: testing such patterns causally, not just describing them.

Lift Translation: from $\widehat{\tau}$ to Industry Magnitude

- Gordon, Zettelmeyer, Bhargava & Chapsky 2019, *Marketing Science* eq. (8): lift = incremental conversion among treated, expressed as a percentage of the treated group's *counterfactual* (untreated) conversion.

$$\tau_\ell = \frac{\Delta\,\text{Conversion}_{\text{treated}}}{\mathbb{E}[Y^{obs} \mid Z=1, W^{obs}=1] - \tau} = \frac{\tau}{\mathbb{E}[Y^{obs} \mid Z=1, W^{obs}=1] - \tau}$$

  - Under random assignment, the denominator is identified by the control-group conversion. **NNE** (Number Needed to Expose) $= 1/\text{ITT}_Y = 1/(\text{compliance rate} \times \widehat{\tau}_{\text{LATE}})$: ad impressions per incremental purchase.

- BLI lift table (control conversion 0.051% as estimator of treated-counterfactual baseline):

Method	$\widehat{\theta}$	Compliers conv.	**Lift (%)**	NNE
2SLS pooled (paper)	0.00229	0.280%	+449%	13,200
DML PLIV (Lasso)	0.00217	0.268%	+426%	13,900
DRIV (econml)	0.00247	0.298%	+484%	12,300

- Industry benchmark: typical online display-ad NNE is 50,000–200,000. BLI's 12,300–18,400 is order-of-magnitude better.
- **Back-of-envelope**: 100M annual impressions $\times$ 3.32% completion $\times \widehat{\tau}_{\text{LATE}}$ (≈ 0.0023) $\approx$ 7,640 **extra purchases/year** causally attributable to BLI completion ($\approx$ 114M KRW at the retailer's average book price).

4/ Notable IV Designs in Recent Empirical Work

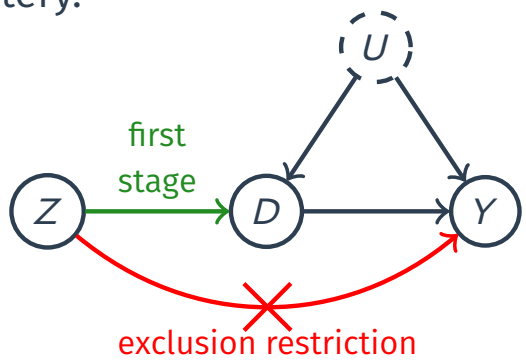
When you have a great idea for an IV but the first stage turns out to be not significant.

Source: Causal Inference for the Brave and True By Matheus Facure Alves

Treatment Assignments as IVs

Bloom et al. (*QJE* 2015)[1]

- Context: 9-month RCT in Ctrip's Shanghai call-centre; 249 volunteers entered the lottery.



Instrument Z: even-birthday lottery eligibility.

Treatment D: actually working from home ≥ 4 days/week.

Outcome Y: weekly productivity index (calls, minutes logged-in, calls/minute).

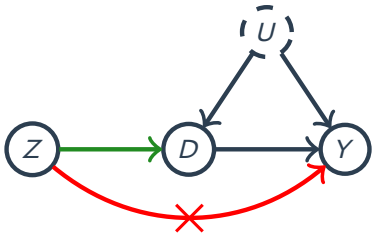
- Identification issue: Endogenous uptake; one-sided non-compliance (some winners revert, losers cannot WFH).
- IV logic:
  1. Randomization: Balance tests on 18 baseline characteristics show no joint differences; p=0.466.
  2. Relevance: 80-90% of winners comply (1st-stage F $\approx$ 23)
  3. Exclusion: tasks, pay, IT identical – only location changes.
  4. Monotonicity: winners may or may not take WFH, losers cannot $\rightsquigarrow$ $D_i(1) \geq D_i(0)$ trivially (one-sided non-compliance).

[1] Bloom, N., Liang, J., Roberts, J. & Ying, Z. "Does Working from Home Work?" *QJE* 2015. 130(1): 165–218.

Regional Characteristics as IVs

Narang and Shankar (*Marketing Sci.* 2019)[2]

- Context: Observational panel on 32 mil. customers of a U.S. video-game / electronics retailer, Jan 2013 – Dec 2015; firm's mobile-shopping app launched July 2014.



Instrument Z: number of cell towers in shopper's ZIP.

Treatment D: adoption of the retailer's mobile app.

Outcome Y: Monthly on/offline purchase & return amounts.

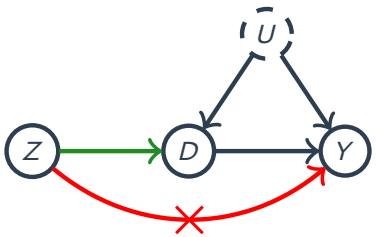
- Identification issue: Self-selection. Tech-savvy or high-value shoppers more likely to install the app.
- IV logic:
  1. Randomization: 90% of cell-towers built before 2010 and shows only weak correlations with 2010–15 population growth and store count, included in $\mathbf{X}_i$.
  2. Relevance: Each cell tower raises the log-odds of adoption by 0.0022 (z=3.7).
  3. Exclusion: controlled for ZIP income, store distance, and competitor presence to block direct demand channels.
  4. Monotonicity: Better signal only increases the prob. of downloading the app (no defiers).

[2] Narang & Shankar. "Mobile App Introduction and On- /Off-line Purchases and Product Returns." Marketing Sci. 2019. 38(5): 756 – 772.

Peers' Environments as IVs

Aral and Nicolaides (*Nature Comm.* 2017)[3]

- Context: Global fitness-tracking network, 1.1 Mil. runners, 3.4 Mil. ties, 350 Mil. km logged over 5 years.



Instrument Z: daily weather (rain/extreme temperature) in friend's city.

Treatment D: friend's running distance that day.

Outcome Y: user/ego's running distance (same or next day).

- Identification issue: Homophily & shared shocks. Similar friends exercise together or face the same local weather.
- IV logic:
  1. Randomization: Use only friend pairs in different cities whose weather paths are uncorrelated; ego's own weather + date FE included.
  2. Relevance: first-stage F = 216–430 well above Stock-Yogo cutoff.
  3. Exclusion: Weather in friend's city is uncorrelated with ego's weather by construction.
  4. Monotonicity: bad weather never makes the friend run more.

[3] Sinan A. & Nicolaides C. "Exercise contagion in a global social network." Nature Communications. 2017. 8: 14753.

Onto the presentations & discussions!

Contact Information:
jaewon.yoo@iss.nthu.edu.tw
https://j1yoo.github.io/

Appendix

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ and now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$

- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.
  - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.
  - Rank condition: $\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']$ and $\mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$ have full rank.

- Identification:
  - $k = \ell$: just-identified.
  - $k < \ell$: over-identified (can test the exclusion restriction, kinda)
  - $k > \ell$: unidentified (fails rank condition)

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\boldsymbol{\Pi} = \left(\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']\right)^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \quad \text{(projection matrix, i.e., pi/}\boldsymbol{\Pi}\text{)}$$

$$\tilde{\mathbf{X}}_i = \boldsymbol{\Pi}'\mathbf{Z}_i \quad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\boldsymbol{\Pi}'\mathbf{Z}_i$ and multiply both sides of the outcome equation by them:

$$\begin{aligned}
Y_i &= \mathbf{X}_i'\beta + \varepsilon_i \\
\boldsymbol{\Pi}'\mathbf{Z}_i Y_i &= \boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \boldsymbol{\Pi}'\mathbf{Z}_i\varepsilon_i \\
\mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\varepsilon_i] \\
\mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \boldsymbol{\Pi}'\mathbb{E}[\mathbf{Z}_i\varepsilon_i] \\
\mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta \\
\mathbb{E}[\tilde{\mathbf{X}}_i Y_i] &= \mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i']\beta \\
\beta &= \left(\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\tilde{\mathbf{X}}_i Y_i]
\end{aligned}$$

How to Estimate the Parameters

- Collect $\mathbf{X}_i$ into an $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into an $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \ldots, \mathbf{Z}_n')$

- In-sample projection matrix produces fitted values:

$$\widehat{\mathbb{X}} \;=\; \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$$

  - Fitted values of the regression of $\mathbb{X}$ on $\mathbb{Z}$.
  - Matrix party trick: $\mathbb{X}'\mathbb{Z}/n = (1/n)\sum_i^n \mathbf{X}_i\mathbf{Z}_i' \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$.

- Take the population formula for the parameters:

$$\beta \;=\; \left(\mathbb{E}[\tilde{\mathbf{x}}_i\mathbf{x}_i']\right)^{-1}\mathbb{E}[\tilde{\mathbf{x}}_i\, Y_i]$$

- And plug in the sample values (the n cancels out):

$$\widehat{\beta}_{\text{2SLS}} = (\widehat{\mathbb{X}}'\mathbb{X})^{-1}\widehat{\mathbb{X}}'\mathbf{y} \xrightarrow{p} \beta$$

- This is how R/`Stata` estimate the 2SLS parameters.

Asymptotic Variance for 2SLS

- We can write the centered, normalized TSLS estimator as:

$$\sqrt{n}\left(\widehat{\beta}_{2SLS} - \beta\right) = \underbrace{\left(n^{-1}\sum_i \widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'\right)^{-1}}_{\xrightarrow{p} \left(\mathbb{E}[\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i']\right)^{-1}} \underbrace{\left(n^{-1/2}\sum_i \widehat{\mathbf{X}}_i\varepsilon_i\right)}_{\xrightarrow{d} N\left(0, \mathbb{E}[\widehat{\mathbf{X}}_i'\varepsilon_i'\varepsilon_i\widehat{\mathbf{X}}_i]\right)}$$

- Thus, $\sqrt{n}\left(\widehat{\beta}_{2SLS} - \beta\right)$ has asymptotic variance:

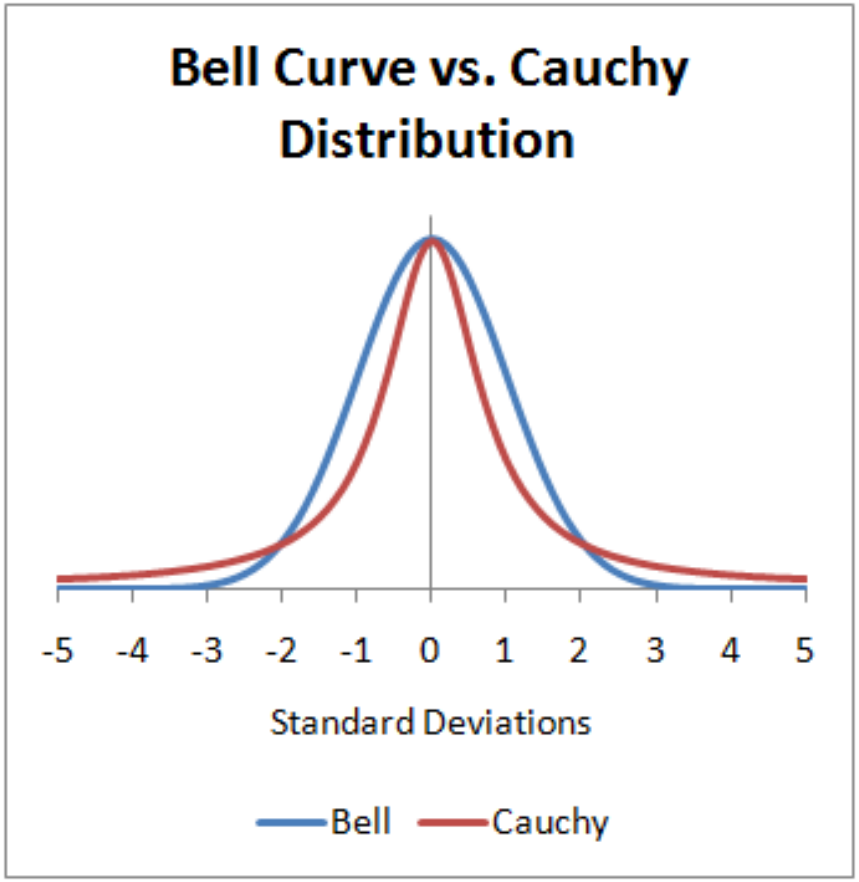
$$\left(\mathbb{E}[\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i']\right)^{-1}\mathbb{E}[\widehat{\mathbf{X}}_i'\varepsilon_i'\varepsilon_i\widehat{\mathbf{X}}_i]\left(\mathbb{E}[\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i']\right)^{-1}$$

- **Robust 2SLS variance estimator** with residuals $\widehat{u}_i = Y_i - \mathbf{X}_i'\widehat{\beta}$:

$$\widehat{\text{Var}}\left(\widehat{\beta}_{2SLS}\right) = (\widehat{\mathbb{X}}'\widehat{\mathbb{X}})^{-1}\left(\sum_i \widehat{u}_i^2\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'\right)(\widehat{\mathbb{X}}'\widehat{\mathbb{X}})^{-1}$$

- HC2, clustering, and autocorrelation versions exist.

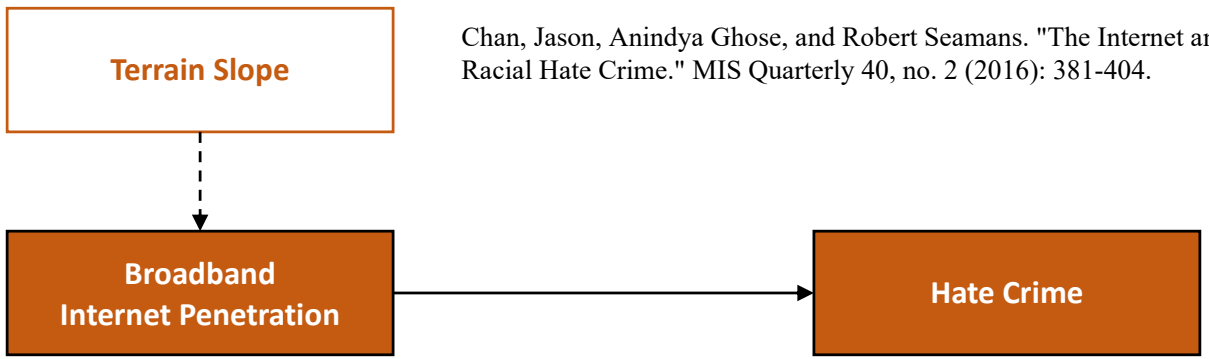
Cauchy vs. Normal Distribution



Source: https://stats.stackexchange.com/questions/36027/why-does-the-cauchy-distribution-have-no-mean

Regional Characteristics as IVs

- Chan, Ghose, Seamans, MISQ, 2016:

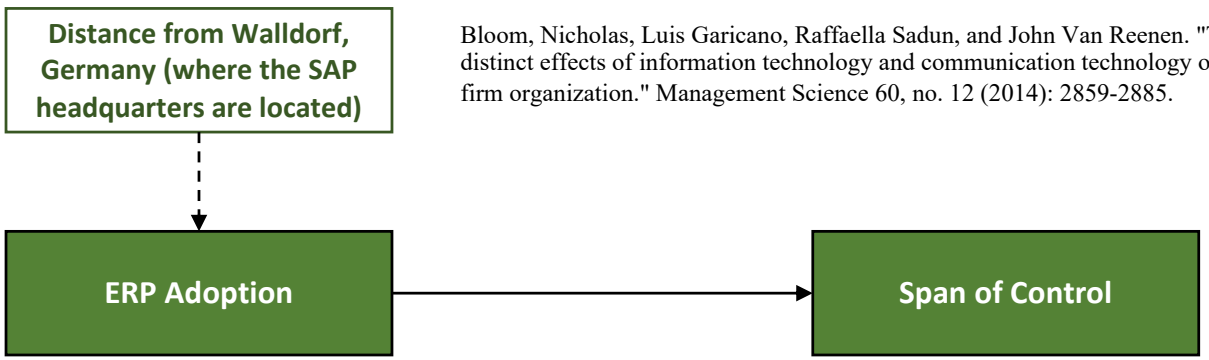


Chan, Jason, Anindya Ghose, and Robert Seamans. "The Internet and Racial Hate Crime." MIS Quarterly 40, no. 2 (2016): 381-404.

  - Observational study using *Hate Crime Statistics* from FBI.
    - RQ: Does the spread of Internet increase racial hate crime?
  - Issue? Confounding
  - IV? terrain slope/steepness (i.e., how many hills in a given region?)

Geographical Proximity-Based IVs

- RQ: Does the adoption of an ERP system impact plant manager autonomy (i.e., span of control)?
  - Span of control: the number of employees managed by supervisors or managers in an organization (high SoC = centralized).

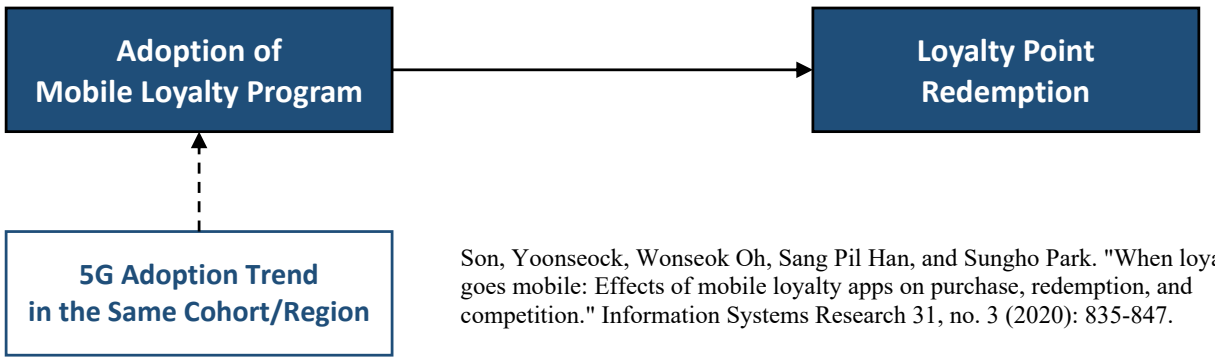


Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen. "The distinct effects of information technology and communication technology on firm organization." Management Science 60, no. 12 (2014): 2859-2885.

  - Observational study: the CEP management and organization survey and the Harte-Hanks ICT panel (Bloom, Garicano, Sadun, Van Reenen, MgmtSci, 2014).
  - IV: Distance from the ERP market leader (i.e., SAP) w/ 25%+ market share ⇝ likely more established connections with SAP (German firms vs. French, England firms).

Macro/Cohort Trends as IVs

- RQ: How does loyalty app adoption affect customer point redemption behavior?

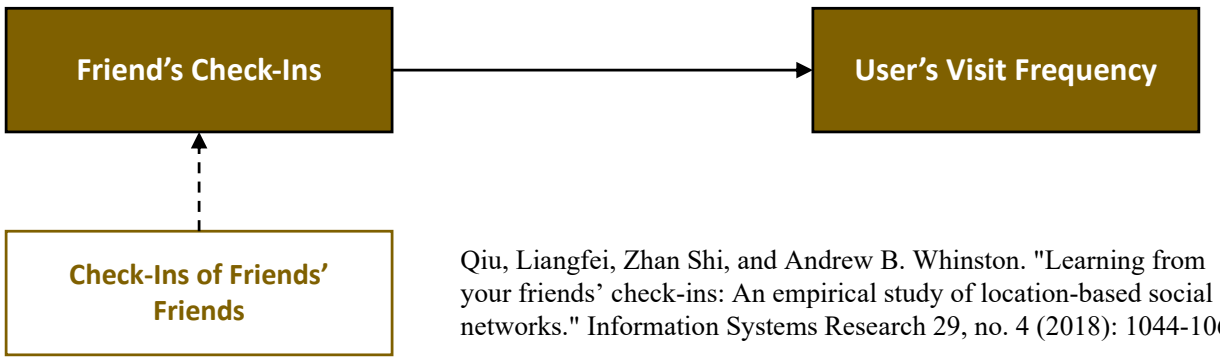


Son, Yoonseock, Wonseok Oh, Sang Pil Han, and Sungho Park. "When loyalty goes mobile: Effects of mobile loyalty apps on purchase, redemption, and competition." Information Systems Research 31, no. 3 (2020): 835-847.

  - Observational study: data on loyalty app adoption status, loyalty point redemption patterns, and purchase behaviors in multivendor loyalty program (MVLP) context (Son, Oh, Han, Park, ISR, 2020).
  - IV: 5G adoption rate in the same cohort (e.g., age group).

Peers' Environments as IVs

- Qiu, Shi, Whinston, ISR, 2018:



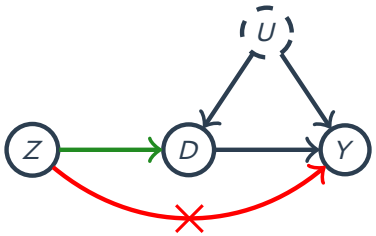
Qiu, Liangfei, Zhan Shi, and Andrew B. Whinston. "Learning from your friends' check-ins: An empirical study of location-based social networks." Information Systems Research 29, no. 4 (2018): 1044-1061.

  - Observational study using data on restaurant check-in information and the users' social network ties from a major SNS in China.
    - RQ: Is there observational learning/herding effect for restaurant discovery?
  - Issue? Homophily.
  - IV: Check-in activities of friend's friends.

Bartik (Shift-Share) IVs

Barron et al. (*Marketing Sci.* 2020)[4]

- Context: Panel of 43,000 ZIP codes (100 largest U.S. metro areas), monthly 2011-2016. Airbnb listings scraped; Zillow rent & price indices matched at ZIP-month.



Instrument Z: Google-Trends "Airbnb" (global shock) × 2010 ZIP "touristiness" (# food/lodging firms)

Treatment D: $\ln(1 + \text{Airbnb listings})$ in ZIP-month.

Outcome Y: Zillow rent index, house price index , price-to-rent.

- Identification issue: Hot ZIPs attract both Airbnb supply and rising housing costs ⇝ upward bias.
- IV logic:
  1. Conditional exogeneity: 2010 touristiness fixed before Airbnb's national growth; Google-Trends shock is global, not driven by ZIP-specific factors. **(Hardest assumption for Bartik IVs; defended in detail next slide.)**
  2. Relevance: first-stage F = 650–820 well above Stock-Yogo cutoff.
  3. Exclusion: global search shocks unrelated to local housing; touristiness fixed in 2010. City×month FE + placebo ZIPs (no listings) show no direct price effect.
  4. Monotonicity: more global Airbnb awareness cannot lower listings in touristy ZIPs.

[4] Barron, K., Kung, E., & Proserpio, D. "The Effect of Home-Sharing on House Prices and Rents" Marketing Sci. 2020. 40(2): 283–304.

Bartik (Shift–Share) IVs

Barron et al. (*Marketing Sci.* 2020)

- IV logic 1 (continued): defending **conditional exogeneity**, the hardest assumption for Bartik IVs:
  - Parallel pre-trends: prior to 2012, rents and prices evolve identically across touristiness quartiles (Fig. 5) $\rightsquigarrow$ no pre-existing divergence.
  - Touristiness $\times$ time test: adding a ZIP-specific trend term ($h_{i,2010} \times t$) yields an insignificant coefficient; IV effect unchanged.
  - Rich time-varying controls: results robust after adding ZIP income, population, hotel rooms, TripAdvisor reviews, and airport arrivals $\rightsquigarrow$ IV not picking up gentrification or tourism shocks.