ECONOMETRICS III

# Causal Inference

From policy questions to credible designs

**Štěpán Jurajda**

CERGE-EI • Doctoral Econometrics

Student lecture deck

# The course is organized around identification

PART I

## Causal inference

- Potential outcomes and policy estimands
- Selection on observables and DML
- Experiments, IV, DiD, and RD

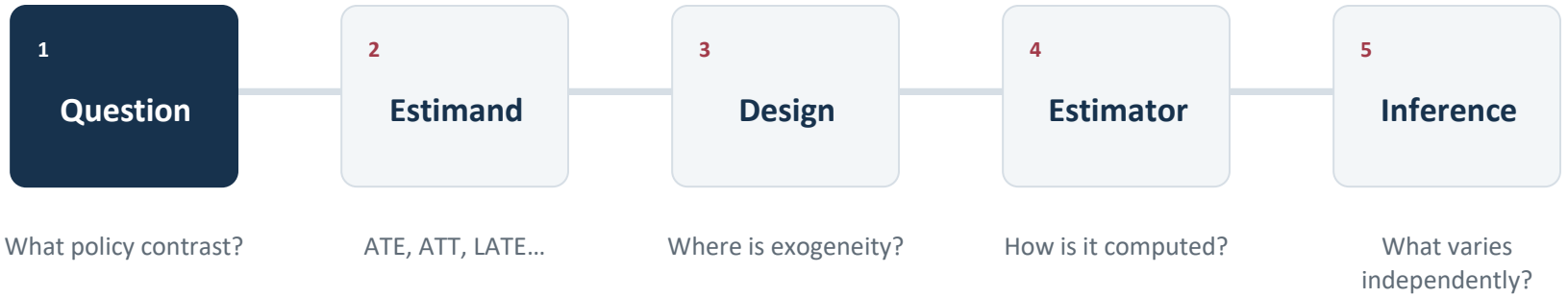
PART II

## Supporting models

- Panel data, clustering, and GMM
- Limited outcomes and selection
- Duration, count, and choice models

**The organizing question is always: which variation identifies which causal parameter?**

# A disciplined causal workflow




**If one link is vague, the causal claim is vague.**

PART I

# 01 Causal questions and potential outcomes

- Separate prediction from intervention
- Define the missing counterfactual
- Name the population whose effect matters

# Prediction is not causality

A regression coefficient acquires causal meaning only through a model and an identification argument.

## Prediction

- **Object:** $E[Y|X]$ or its best linear approximation
- **Variation:** all observed variation in X
- **Question:** what outcome is expected at $X=x$?

## Causal effect

- **Object:** a contrast of potential outcomes
- **Variation:** a defended exogenous component of treatment
- **Question:** what changes under intervention?

# Potential outcomes make the missing counterfactual explicit

Each unit has two conceptually well-defined outcomes; only one is observed.

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$$

- **Observed:** $Y_{1i}$ for treated units and $Y_{0i}$ for untreated units.
- **Missing:** the untreated outcome for the treated—and vice versa.
- **Identification:** construct a credible substitute for that missing outcome.

# ATE, ATT, and ATU answer different policy questions

| Estimand | Definition | Policy question |
|---|---|---|
| **ATE** | $E[Y_1 - Y_0]$ | What if everyone in the population were treated? |
| **ATT** | $E[Y_1 - Y_0 \mid T=1]$ | What did treatment do for participants? |
| **ATU** | $E[Y_1 - Y_0 \mid T=0]$ | What would treatment do for nonparticipants? |

**The population index is part of the estimand—not a reporting detail.**

# The raw treated–control gap mixes effect and selection

$$E[Y|T=1] - E[Y|T=0] = ATT + \{E[Y_0|T=1] - E[Y_0|T=0]\}$$

- **ATT:** the causal effect for those who selected into treatment.
- **Selection bias:** how treated and controls would differ even without treatment.
- **Design goal:** make the bracketed counterfactual difference zero or estimable.

# Heterogeneous effects change what every estimator means

When $\beta_i = Y_{1i} - Y_{0i}$ varies, selection may occur on both levels and gains.

**Never treated**

Effects may be irrelevant to observed participation.

**Marginal**

Treatment choice is changed by the instrument or policy.

**Always treated**

Effects may dominate participant averages.

**OLS, IV, ATT, and ATE generally weight the distribution of $\beta_i$ differently.**

# CIA, SUTVA, and overlap carry the identification argument

| Assumption | Operational meaning | Failure mode |
|---|---|---|
| **CIA** | $(Y_1, Y_0) \perp T \mid X$ | Important unobservables still drive treatment |
| **SUTVA** | One treatment version; no interference | Spillovers or hidden versions of treatment |
| **Overlap** | $0 < P(T=1 \mid X) < 1$ | No comparable treated/control units in some cells |

**Assumptions should be translated into features of the institutional setting.**

# 02 Selection on observables

- Design comparisons inside common support
- Use the propensity score to balance covariates
- Diagnose failure before estimating effects

# Unconfoundedness identifies effects within covariate cells

$$E[Y_1 - Y_0 \mid X=x] = E[Y \mid T=1, X=x] - E[Y \mid T=0, X=x]$$

- **Claim:** within X-cells, treatment assignment is as good as random.
- **Aggregation:** average conditional effects over the target population.
- **Risk:** functional-form extrapolation can hide lack of common support.

# Matching is a design exercise before it is an estimator

1. **Choose pre-treatment covariates.** Include confounders; do not condition on post-treatment variables.
2. **Enforce common support.** Discard comparisons that require extrapolation.
3. **Construct comparable sets.** Match, stratify, or weight using a stated distance rule.
4. **Inspect balance.** Evaluate standardized differences and the full distributions.
5. **Estimate and stress-test.** Vary specifications and report sensitivity.

# The propensity score compresses X while preserving balance

## What it buys

- A scalar balancing score $p(X)=P(T=1|X)$
- A practical route to matching, stratification, and weighting
- A visible diagnostic for common support

## What it does not buy

- It does not make unconfoundedness true
- It does not repair omitted pre-treatment confounders
- It does not justify extreme extrapolation

# IPW reconstructs a target population

$$\text{ATE: } E\left[ TY/p(X) - (1-T)Y/\{1-p(X)\} \right]$$

- **Treated units:** receive weight 1/p(X).
- **Control units:** receive weight 1/{1−p(X)}.
- **Practical warning:** extreme propensities create unstable weights and high variance.

# AIPW and DML protect the target moment from nuisance error

Orthogonal scores reduce first-order sensitivity to imperfect nuisance estimates; cross-fitting limits own-observation bias.

1. **Split the sample.** Reserve observations for honest target estimation.
2. **Learn nuisance functions.** Estimate outcomes and/or treatment propensities flexibly.
3. **Residualize or augment.** Build a Neyman-orthogonal score.
4. **Rotate folds.** Cross-fit so every observation contributes to both stages.
5. **Aggregate with valid inference.** Report the target parameter, not predictive fit.

# Diagnostics decide whether the design is credible

**Estimation comes after the design passes observable stress tests.**

- **Overlap:** plot propensity distributions; identify regions without comparisons.
- **Balance:** compare means and distributions after matching or weighting.
- **Placebos:** test for effects on lagged outcomes or predetermined covariates.
- **Sensitivity:** show how results move across design choices.

**A precise standard error cannot rescue an implausible comparison.**

# 03 Experiments, noncompliance, and IV

- Separate assignment from treatment received
- Interpret the Wald ratio through compliance types
- Match the local effect to the policy question

# Random assignment identifies ITT even when treatment is not taken

Keep the two causal questions distinct: the effect of an offer and the effect of treatment received.

### Assignment Z

- Randomized by design
- Often known for every unit
- Direct comparison identifies ITT

### Treatment D

- May be chosen or imperfectly complied with
- Potentially endogenous
- Needs IV assumptions for a treatment effect

# The Wald ratio rescales assignment effects by compliance

$$\text{LATE} = \{E[Y|Z=1] - E[Y|Z=0]\} / \{E[D|Z=1] - E[D|Z=0]\} = \text{ITT} / \text{first stage}$$

- **Numerator:** reduced-form effect of assignment on the outcome.
- **Denominator:** change in treatment take-up caused by assignment.
- **Interpretation:** average effect for units whose treatment is changed by Z.

# Monotonicity partitions the population into four response types

| Type | D(0) | D(1) | Response to encouragement |
|---|---|---|---|
| **Never-taker** | 0 | 0 | Never takes treatment |
| **Complier** | 0 | 1 | Follows the encouragement |
| **Defier** | 1 | 0 | Moves opposite to encouragement |
| **Always-taker** | 1 | 1 | Always takes treatment |

**Monotonicity rules out defiers; only compliers generate the first stage.**

# A valid instrument needs four linked assumptions

| Requirement | Formal idea | Institutional question |
|---|---|---|
| **Relevance** | $\text{Cov}(Z,D) \neq 0$ | Does assignment materially shift treatment? |
| **Independence** | $Z \perp$ potential outcomes | Why is the instrument as-if random? |
| **Exclusion** | Z affects Y only through D | Could the instrument change outcomes directly? |
| **Monotonicity** | $D(1) \geq D(0)$ | Could anyone systematically move the other way? |

# LATE is local to the compliers moved by this instrument

**Different instruments can identify different averages—even for the same treatment and outcome.**

- **Describe compliers:** use first-stage heterogeneity across observed characteristics.
- **Assess transport:** explain why complier effects should—or should not—generalize.
- **Align policy:** ask whether the planned intervention moves the same margin.

**External validity begins with a clear description of the identifying margin.**

# MTE connects local margins to policy-relevant averages

$MTE(x,u_D)=E[\beta \mid X=x,U_D=u_D]$ is the effect for people at a treatment-choice margin.

**High resistance**

Requires a strong inducement to enter treatment.

**Marginal unit**

Just indifferent at the current policy or instrument value.

**Low resistance**

Would enter treatment under weak inducement.

**ATE, ATT, LATE, and policy-relevant effects integrate the MTE with different weights.**

# Weak instruments demand identification-robust inference

## Warning signs

- Small or unstable first-stage effects
- Large standard errors and sensitivity to controls
- 2SLS pulled toward OLS in finite samples

## Responses

- Report first-stage diagnostics and reduced form
- Use Anderson–Rubin or related robust procedures
- Consider LIML/JIVE and simplify the instrument set

# Shift-share IVs require a precise source-of-randomness story

$$Z_i = \Sigma_k \, share_{ik,0} \times shift_k$$

- **Shock-based design:** many independent, exogenous shifts drive identification.
- **Share-based design:** initial exposure shares are the defended exogenous objects.
- **Inference:** account for common exposure and concentration of identifying shocks.

# 04 Difference-in-differences

- Construct the missing trend, not just a before–after contrast
- Treat event-study comparisons as designed 2×2 experiments
- Make staggered-adoption controls explicit

# The 2×2 DiD subtracts the counterfactual trend

$$\mathrm{ATT} = (\bar{Y}_{T,\mathrm{post}} - \bar{Y}_{T,\mathrm{pre}}) - (\bar{Y}_{C,\mathrm{post}} - \bar{Y}_{C,\mathrm{pre}})$$

- **First difference:** removes group-specific time-invariant levels.
- **Second difference:** removes the common untreated time shock.
- **Target:** the treated group in the post-treatment period.

# Parallel trends and no anticipation are the design assumptions

## Parallel trends

- Untreated outcome changes would match across groups
- May hold only conditional on covariates
- Is about potential outcomes, not observed pre-trends alone

## No anticipation

- Units do not respond before treatment starts
- Policy announcements can shift behavior early
- Leads can reveal—but not fully rule out—violations

# Event studies reveal dynamics but do not prove identification

**Event time traces effects relative to a stated reference period.**

- **Pre-treatment coefficients:** diagnose gross differences in trends and anticipation.
- **Post-treatment coefficients:** show onset, persistence, and possible adaptation.
- **Inference:** account for multiple testing when reading an entire path.
- **Design:** define valid controls separately for each cohort and event time.

# Staggered TWFE can compare treated units with already-treated units

## Valid comparisons

- Newly treated versus never treated
- Newly treated versus not-yet treated
- Cohort-specific event-time contrasts

## Contaminated comparisons

- Newly treated versus already treated
- Effects with opposite or opaque weights
- Compositional changes hidden by one coefficient

# Modern DiD estimators make the control group explicit

| Approach | Control group | Key trade-off |
|---|---|---|
| **Sun–Abraham** | Never treated | Stable controls; may be less comparable |
| **Callaway–Sant'Anna** | Never or not-yet treated | Flexible group-time ATT aggregation |
| **Borusyak–Jaravel–Spiess** | Untreated observations under stronger trend structure | Efficient imputation; stronger assumptions |

**Start from the group-time ATT, then aggregate transparently.**

# Covariate adjustment belongs inside the DiD design

1. **Define cohort and event time.** State the treatment timing and target ATT.
2. **Choose valid controls.** Never treated and/or not-yet treated.
3. **Enforce overlap.** Compare units with support in pre-treatment X.
4. **Adjust flexibly.** Use regression adjustment, IPW, AIPW, or DML.
5. **Aggregate transparently.** Show the weights across cohorts and event times.

# Continuous treatment and spillovers change the estimand

### Treatment intensity

- Parallel trends need not rule out selection on dose
- Define a causal response curve, not an opaque slope
- Generalized propensity scores can rebalance intensity

### Interference

- Untreated units may receive indirect exposure
- Control outcomes depend on neighbors' treatment
- Specify an exposure mapping and a new estimand

# 05 Regression discontinuity

- Exploit a known treatment assignment threshold
- Estimate a local effect with transparent smoothing choices
- Use manipulation and balance checks as design evidence

# RD compares units at a treatment threshold

**If potential outcomes vary smoothly in the running variable, units just above and below the cutoff are comparable.**

- **Running variable R:** determines treatment through a known cutoff c.
- **Locality:** the design identifies effects for units near c.
- **Continuity:** counterfactual outcomes should not jump at c absent treatment.

**The cutoff generates the comparison; the regression only estimates it.**

# Sharp and fuzzy RD identify different local effects

| Design | Treatment rule | Estimand |
|---|---|---|
| **Sharp RD** | $D = 1\{R \geq c\}$ | Jump in $E[Y \mid R]$ at $c$ |
| **Fuzzy RD** | $P(D=1 \mid R)$ jumps at $c$ | Outcome jump / treatment jump: local Wald effect |

**Fuzzy RD is an IV design for compliers at the threshold.**

# Bandwidth choice trades variance against approximation bias

Local polynomial estimation uses observations within h of the cutoff.

**Narrow h**

Low approximation bias; few observations and high variance.

**Balanced h**

Data-driven bandwidth targets an explicit loss criterion.

**Wide h**

More precision; greater reliance on functional form.

**Report robust bias-corrected inference and sensitivity to reasonable bandwidths.**

# Credible RD evidence includes density, balance, and sensitivity checks

- **Density at the cutoff:** look for sorting or manipulation of the running variable.
- **Predetermined covariates:** verify continuity at the threshold.
- **Alternative bandwidths/orders:** show the result is not a tuning artifact.
- **Placebo cutoffs:** check for jumps where no treatment change occurs.

**A smooth fitted line is not evidence of a valid RD design.**

PART II

# 06 Panel data and inference

- Use repeated observations to remove fixed heterogeneity
- Identify which units and periods generate the coefficient
- Cluster at the level of independent treatment variation

# Fixed effects remove time-invariant omitted variables

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- **Removed:** all unit-specific factors that are constant over time.
- **Retained:** within-unit changes in x and y.
- **Still required:** strict or appropriately sequential exogeneity of the within variation.

# First differences reveal that movers identify the coefficient

$$\Delta y_{it} = \Delta x_{it}\,\beta + \Delta \varepsilon_{it}$$

- **Binary treatment:** units with ΔD=0 do not identify β in the differenced equation.
- **Interpretation:** the effect is local to switchers or movers.
- **Threat:** switching may respond to transitory shocks in Δε.

# Random effects gain efficiency by adding an orthogonality restriction

## Fixed effects

- Allows $\text{Cov}(a_i, x_{it}) \neq 0$
- Uses within-unit variation
- Cannot estimate time-invariant regressors directly

## Random effects

- Requires $E[a_i \mid x_{i1}, \ldots, x_iT] = 0$
- Combines within and between variation
- More efficient when the restriction is credible

# Standard errors must follow the level of identifying variation

1. **Locate the treatment variation.** Individual, firm, region, cohort, or time?
2. **Locate dependence.** Which observations share shocks or assignment?
3. **Choose clusters.** Cluster at or above the independent assignment level.
4. **Check cluster count.** Use small-cluster corrections or randomization inference when needed.
5. **Explain the design.** A large row count does not create independent treatment variation.

# Two-way fixed effects are a panel model—and a DiD estimator

## Panel interpretation

- Unit FE absorb time-invariant levels
- Time FE absorb common shocks
- β comes from residual treatment changes

## DiD interpretation

- With one adoption date, β reproduces a DiD contrast
- With staggered timing, β aggregates many 2×2 comparisons
- Heterogeneous effects can make that aggregation misleading

PART II

# 07 IV mechanics, GMM, and dynamic panels

- Project treatment onto defended instruments
- Treat moment conditions as economic assumptions
- Use internal instruments carefully in dynamic panels

# 2SLS uses only the component of treatment projected on instruments

$$\text{Stage 1: } D = Z\pi + X\gamma + v \qquad \text{Stage 2: } Y = \hat{D}\beta + X\theta + u$$

- **Relevance:** Z must explain D conditional on X.
- **Exogeneity:** Z must be orthogonal to the structural error.
- **One coefficient:** may still summarize heterogeneous effects with instrument-specific weights.

# Overidentification tests are specification checks, not proof of validity

## What rejection says

- At least one moment condition is inconsistent with the others
- The model or instrument set needs investigation
- Heterogeneous effects may also create unequal IV estimates

## What non-rejection does not say

- It does not prove every instrument is exogenous
- Low power can conceal violations
- Common invalidity can pass undetected

# Measurement error creates endogeneity

$$x^* = x + \nu \quad \Rightarrow \quad \text{plim } \beta\text{OLS} = \beta \cdot \text{Var}(x) \,/\, \{\text{Var}(x)+\text{Var}(\nu)\}$$

- **Classical error:** attenuates the slope toward zero.
- **Panel data:** differencing can amplify noise relative to signal.
- **Remedies:** validation data, repeated measures, reliability corrections, or valid IV.

# GMM turns economic restrictions into sample moments

$$\text{Choose } \hat{\theta} \text{ to make } \bar{g}(\theta) = n^{-1}\Sigma_i\, g(W_i,\theta) \text{ close to zero}$$

- **Moments:** encode orthogonality or other model restrictions.
- **Weighting matrix:** determines efficiency and the geometry of the fit.
- **Diagnostics:** use Hansen/Sargan tests cautiously and inspect instrument proliferation.

# Arellano–Bond instruments differenced dynamics with lagged levels

1. **Start with dynamics.** $y_{it} = \rho y_{i,t-1} + x_{it}\beta + a_i + \varepsilon_{it}$.
2. **Difference out $a_i$.** The lagged dependent variable becomes endogenous.
3. **Use deeper lags.** Levels dated t−2 and earlier can instrument $\Delta y_{i,t-1}$ under serial-correlation restrictions.
4. **Test the error structure.** AR(1) is expected after differencing; AR(2) is a warning.
5. **Limit instruments.** Too many instruments overfit endogenous variables and weaken diagnostics.

PART II

# 08 Limited outcomes, selection, and duration

- Match the model to the support of the outcome
- Translate latent-index coefficients into observable effects
- Model selection when the observed sample is non-random

# LPM, logit, and probit encode different functional-form commitments

| Model | Conditional mean | Primary advantage |
|---|---|---|
| **LPM** | $P(Y=1\mid X)=X\beta$ | Transparent slopes; easy FE/IV; robust interpretation |
| **Logit** | $\Lambda(X\beta)$ | Probabilities in [0,1]; convenient odds structure |
| **Probit** | $\Phi(X\beta)$ | Latent-normal framework; useful in selection models |

**Model choice should follow the estimand and identification problem—not habit.**

# Report marginal effects on probabilities, not latent-index coefficients

$$\partial P(Y=1|X)/\partial x_k = f(X\beta) \cdot \beta_k$$

- **Nonlinearity:** the effect varies with X even when $\beta_k$ is constant.
- **Summary:** state whether you report average marginal effects or effects at chosen X.
- **Comparisons:** coefficient scales can differ across groups when latent variances differ.

# Censoring and truncation discard different information

## Censoring

- The unit remains in the sample
- The outcome is only known to lie beyond a threshold
- Likelihood combines mass at the limit with a density above it

## Truncation

- Units outside the selection rule are absent
- Their covariates and outcomes may be unobserved
- Likelihood conditions on sample inclusion

# Heckman correction models selection through an inverse Mills ratio

1. **Selection equation.** Estimate $P(S=1|Z)$ with a probit model.
2. **Compute λ.** The inverse Mills ratio summarizes selection under joint normality.
3. **Outcome equation.** Regress Y on X and λ in the selected sample.
4. **Identify carefully.** A credible exclusion from the outcome equation is strongly preferred.
5. **Interpret structurally.** The correction is only as credible as the distribution and selection model.

# Roy and switching models put comparative advantage at the center

## Outcome states

- $Y_o = X\theta_o + U_o$
- $Y_1 = X\theta_1 + U_1$
- Individual gain $\beta_i = Y_1 - Y_o$

## Selection rule

- Choose treatment when expected gain exceeds cost
- Selection may depend on both baseline outcomes and gains
- Observed groups are positively selected into their comparative advantage

# Duration models separate survival from the instantaneous hazard

$$h(t|X) = \lim_{\Delta \to 0} P(t \le T < t+\Delta \mid T \ge t, X) / \Delta$$

- **Survival S(t):** probability the spell lasts beyond t.
- **Hazard h(t):** instantaneous exit rate conditional on survival to t.
- **Censoring:** contributes exposure information without a completed duration.

# Count and multiple-choice models must match the outcome support

| Outcome | Common model | Key diagnostic |
|---|---|---|
| **Counts** | Poisson QML / negative binomial | Mean specification, zeros, overdispersion |
| **Unordered choices** | Multinomial or nested logit | IIA and substitution patterns |
| **Ordered choices** | Ordered logit/probit | Threshold structure; marginal effects by category |

**Coefficient signs alone may not reveal probability effects in nonlinear multi-outcome models.**

# Estimands are weighted averages of heterogeneous effects

The question is not only whether effects differ—but how the design weights them.

**ATE**

Population weights across all units.

**ATT**

Weights the treated population.

**LATE**

Weights compliers moved by a specific instrument.

**Policy effect**

Weights margins changed by the proposed policy.

**Report the weighting population and the identifying margin alongside the estimate.**

# Choose the method from the source of identifying variation

| Variation | Natural design | Typical estimand |
|---|---|---|
| **As-if random assignment** | RCT / IV | ITT or LATE |
| **Selection on observed X** | Matching / IPW / AIPW | ATE or ATT on overlap |
| **Differential policy timing** | DiD / event study | Group-time ATT |
| **Threshold rule** | Sharp or fuzzy RD | Local ATE or local LATE |
| **Within-unit change** | FE / first differences | Effect for movers under exogeneity |

# A credible empirical design survives four audits

- **Estimand audit:** whose effect, under which intervention, over what horizon?
- **Identification audit:** which variation is exogenous, and why?
- **Estimation audit:** what weights, functional forms, tuning choices, and diagnostics?
- **Inference audit:** what is the independent unit of variation, and are clusters sufficient?

**A transparent limitation is stronger than an overbroad causal claim.**

# Close the loop: answer the policy question you actually identified

**A strong empirical conclusion has three sentences.**

- **Design:** “Variation in treatment comes from…”
- **Estimand:** “Under these assumptions, the estimate is the effect for…”
- **Policy meaning:** “This evidence supports—or does not support—extrapolation to…”

**Credibility comes from alignment: question → estimand → design → estimator → inference.**