METHODOLOGICAL HANDBOOK

Endogeneity & Causal Inference Methods

A Comprehensive Guide for Operations Management Researchers
with Diagnostic Protocols, Published Case Studies, and Reviewer Strategies

Chenhao Zhou

Ph.D. Candidate, Supply Chain Management

Rutgers Business School

2026 Edition | Expanded with Testing Protocols and UTD24/FT50 Case Studies

Updated with staggered DID, synthetic DID, machine learning causal inference,
Gaussian copula, sensitivity analysis, and peer review strategies

Contents

Preface

This handbook grew from my own journey learning causal inference as a doctoral student in operations management. It began as personal notes organized around Lu et al. (2018), the foundational methodological guide for OM researchers. Over three years of empirical research, the notes expanded to cover methods that Lu et al. did not address, diagnostic protocols I wished someone had shown me earlier, and lessons learned from navigating peer review.

The 2026 edition represents a substantial expansion. Every method now includes a step-by-step testing protocol with exact Stata and R commands, published case studies from UTD24 and FT50 journals, and guidance on common pitfalls. New sections cover Gaussian copula estimation, generalized method of moments, shift-share instruments, and a full chapter on responding to reviewer concerns about endogeneity. The sensitivity analysis material has been elevated from a subsection to a cross-cutting requirement integrated throughout.

Essential Starting Points:

- Lu et al. (2018, *JOM*): The definitive methodological guide for OM researchers. Comprehensive coverage of IV diagnostics with clear reporting protocols.
- Ketokivi & McIntosh (2017, *JOM*): Conceptual treatment of why endogeneity arises and how to think about it philosophically.
- Ho et al. (2017, *MSOM*): Systematic review finding that 75% of empirical papers involve causal inference.
- Roth et al. (2023, *J. Econometrics*): The most comprehensive synthesis of the modern DID revolution.
- Cinelli & Hazlett (2020, *JRSS-B*): Foundational framework for omitted variable bias sensitivity analysis.

Recommended Textbooks:

- Angrist & Pischke (2009): *Mostly Harmless Econometrics*. Princeton.
- Cunningham (2021): *Causal Inference: The Mixtape*. Yale.
- Huntington-Klein (2021): *The Effect*. CRC Press.
- Cattaneo, Idrobo, & Titiunik (2024): *A Practical Introduction to Regression Discontinuity Designs*. Cambridge Elements.

Acknowledgments. I am grateful to David Dreyfus for introducing me to empirical research and for his mentorship throughout my doctoral journey. This handbook draws on Lu et al. (2018) as its organizing framework and extends it with methods and practices that have emerged since its publication. This handbook was developed with the assistance of Claude (Anthropic) as a writing aid.

Chenhao Zhou, New Jersey, 2026

1 Introduction: The Nature of Endogeneity

1.1 What is Endogeneity?

Endogeneity is the most fundamental challenge in empirical research. It represents a core question: Does the statistical association we observe reflect a genuine causal relationship, or is it driven by confounding factors?

According to Terwiesch et al. (2020), the proportion of OM empirical papers mentioning endogeneity has risen from 20% to over 60% in recent years, while instrumental variable usage increased from under 20% to approximately 40%. The basic regression model illustrates the problem:

$$Y = \alpha + \beta X + \varepsilon \tag{1}$$

When $\text{Cov}(X, \varepsilon) \neq 0$, the OLS estimator $\hat{\beta}$ is biased and inconsistent.

1.2 Three Sources of Endogeneity

1.2.1 Omitted Variable Bias

When studying the effect of education on income, an unobserved factor like "ability" simultaneously affects both years of schooling and income levels. The bias is:

$$E(\hat{\beta}) = \beta_1 + \beta_2 \cdot \frac{\text{Cov}(X, Z)}{\text{Var}(X)} \tag{2}$$

OM Application: Estimating the effect of a new inventory system on firm performance. Adopting firms may have better management practices and stronger culture, creating omitted variable bias.

Sensitivity Analysis for OVB: Cinelli & Hazlett (2020) provide a formal framework using partial R^2 benchmarking. Researchers can state: "An unobserved confounder would need to be X times as strong as [observed covariate] to reduce the effect to zero." Oster (2019) provides a complementary approach based on coefficient stability using the parameter δ.

1.2.2 Reverse Causality (Simultaneity)

Ketokivi & McIntosh (2017) demonstrate how 2SLS and OLS can produce coefficient signs with opposite directions when reverse causality is present, using a restaurant seating allocation example.

OM Application: The relationship between service quality and customer volume. Higher quality may attract more customers (the desired effect), but higher volume may strain resources and reduce quality (reverse causality).

1.2.3 Measurement Error

Classical measurement error in X attenuates coefficient estimates toward zero (attenuation bias). Using "self-reported working hours" as a proxy for actual hours produces systematically biased estimates.

Endogeneity Cannot Be Completely Solved

Ketokivi & McIntosh (2017) emphasize: "endogeneity is not a problem that can be solved" (p. 3). Modern research has shifted focus from "strictly exogenous" to "plausibly exogenous" instruments. The goal is to make a credible case for causal inference, not to achieve mathematical certainty.

1.3 The Philosophical Foundation

All causal inference methods pursue the same goal: constructing a credible counterfactual. The fundamental problem of causal inference is that we can never observe both potential outcomes for the same unit simultaneously:

$$\text{Causal Effect} = Y_1(\text{Treated}) - Y_0(\text{Control}) \tag{3}$$

Each method constructs the counterfactual differently:

- **IV** uses exogenous variation from external factors
- **DID** uses temporal comparison with parallel trends
- **RD** exploits threshold discontinuities
- **Matching** creates statistical twins based on observables
- **Synthetic DID** combines DID parallel trends with SCM re-weighting (Arkhangelsky et al., 2021)
- **Double ML** uses machine learning for nuisance parameter estimation (Chernozhukov et al., 2018)

1.4 Framework for Method Selection

Ho et al. (2017) found that 75% of empirical papers in *Management Science*, *MSOM*, and *POM* involve causal inference. The core question in method selection is: "Which method's identifying assumptions are most credible in my research context?"

1.5 Method Selection Decision Tree

The following decision tree guides researchers through method selection based on data structure and research design:

Table 1: Method Selection Framework

Research Context	Recommended Method	Key Assumption
Clear policy threshold	Regression Discontinuity	Local randomization near cutoff
Policy change with control group	Difference-in-Differences	Parallel trends
External exogenous variation	Instrumental Variables	Exclusion restriction
Panel data, unit heterogeneity	Fixed Effects	Time-invariant confounders only
Rich observable covariates	Matching Methods	Selection on observables
Single treated, many controls	Synthetic Control	Pre-treatment fit quality
Nonlinear model, endogeneity	Control Function	Valid instruments + correct spec
Staggered policy adoption	Modern DID estimators	No anticipation + parallel trends
No external instruments	Gaussian Copula / Lewbel	Distributional / heteroskedasticity
High-dimensional confounders	Double/Debiased ML	Sparsity

Key Concepts

Step 1: Do you have a natural experiment?

- **Yes, with a threshold/cutoff** → **RD** (Section 2.3)
- **Yes, with a policy change** → Go to Step 2
- **No** → Go to Step 3

Step 2: How many treated units?

- **Many units, staggered timing** → **Modern DID** (Section 2.2.4)
- **Many units, single treatment date** → **Standard DID** (Section 2.2)
- **One treated unit, many controls** → **Synthetic Control** (Section 2.6)
- **Few treated units** → **Synthetic DID** (Section 3.2)

Step 3: Do you have valid external instruments?

- **Yes, linear model** → **IV/2SLS** (Section 2.1)
- **Yes, nonlinear model** → **Control Function** (Section 2.7)
- **No** → Go to Step 4

Step 4: What data structure do you have?

- **Panel data with heteroskedasticity** → **Lewbel HBIV** as robustness (Section 2.8)
- **Cross-section, continuous endogenous var** → **Gaussian Copula** as robustness (Section 2.9)
- **Rich observables, selection on observables plausible** → **Matching/AIPW** (Section 2.4)
- **Panel data** → **Fixed Effects** + sensitivity analysis (Section 2.5)
- **High-dimensional covariates** → **Double ML** (Section 3.1)

Always: Report sensitivity analysis (Oster δ, Cinelli-Hazlett RV, E-values) regardless of primary method.

2 Causal Inference Methods

2.1 Instrumental Variables (IV)

2.1.1 Research Context

Research Question: Does education increase income?
Core Challenge: Unobservable "ability" and "family background" simultaneously affect both education and income, biasing any simple regression estimate.

2.1.2 Intuition: The External Lever

Joshua Angrist discovered that birth quarter affects years of schooling due to school entry age regulations, but birth quarter itself is unlikely to directly affect future income. Birth quarter serves as an "external lever" creating exogenous variation in education unrelated to ability.

2.1.3 Mathematical Framework

First Stage:

$$X = \pi Z + \gamma W + \nu \tag{4}$$

Second Stage:

$$Y = \beta \hat{X} + \delta W + \varepsilon \tag{5}$$

The IV estimator:

$$\beta_{IV} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} \tag{6}$$

Identification Conditions:

1. **Relevance:** $\text{Cov}(Z, X) \neq 0$. Verified with first-stage $F > 10$ (Stock & Yogo, 2005); $F > 23$ for 5% maximal bias.
2. **Exclusion Restriction:** $\text{Cov}(Z, \varepsilon) = 0$. The instrument affects Y only through X. Cannot be tested directly; must be justified theoretically.

Exclusion Restriction Cannot Be Directly Tested

Lu et al. (2018) found common issues in OM papers: (1) weak instruments with $F < 10$, (2) incomplete first-stage reporting, (3) insufficient exclusion restriction justification, (4) no LIML comparison. Always report the full battery of IV diagnostics.

2.1.4 IV Diagnostic Protocol

Weak Instrument-Robust Inference. Lee et al. (2022, *AER*) demonstrate that standard t-ratio inference from 2SLS can be highly misleading when instruments are weak, even with F-statistics

Table 2: IV Diagnostic Tests

Test	Purpose	Threshold
First-Stage F	Instrument strength	$F > 10$; $F > 23$ for 5% bias
Kleibergen-Paap LM	Under-identification	Reject at $p < 0.05$
Kleibergen-Paap Wald F	Weak identification	Stock-Yogo critical values
Hansen J-test	Over-identification	Fail to reject at $p > 0.10$
Durbin-Wu-Hausman	Endogeneity test	Reject suggests endogeneity
Anderson-Rubin test	Robust to weak IV	Reject at $p < 0.05$
tF confidence interval	Valid t-ratio inference	Lee et al. (2022)

above 10. They propose the tF procedure: a valid confidence interval that accounts for instrument strength. Andrews et al. (2019) recommend routine reporting of Anderson-Rubin confidence sets alongside standard 2SLS results.

IV Step-by-Step Testing Protocol

Pre-Estimation:

1. Justify exclusion restriction with economic theory. Explain *why* the instrument affects Y only through X.
2. Check instrument relevance informally: scatter plot Z vs. X, correlation matrix.
3. If multiple instruments, consider using LASSO-based selection (Belloni et al., 2012).

Estimation:

1. Run first-stage regression. Report F-statistic, R^2, and coefficient on Z.
2. Run 2SLS with robust/clustered standard errors.
3. Run LIML as robustness (more robust to weak instruments).
4. Run GMM if heteroskedasticity is present (more efficient than 2SLS).

Post-Estimation Diagnostics:

1. **Weak instruments:** Report Kleibergen-Paap Wald F; compare to Stock-Yogo critical values. If $F < 10$, report Anderson-Rubin confidence sets.
2. **Under-identification:** Report Kleibergen-Paap LM statistic. Reject H_0 at $p < 0.05$.
3. **Over-identification:** If $k > 1$ instruments, report Hansen J-test. Fail to reject at $p > 0.10$.
4. **Endogeneity:** Report Durbin-Wu-Hausman test. If fail to reject, OLS may be preferred.
5. **Sensitivity:** Compare 2SLS, LIML, and GMM point estimates. Large divergence signals weak IV.

Stata Commands:

```
* Comprehensive IV with diagnostics
ivreg2 Y X1 X2 (EndogVar = IV1 IV2), first robust
estat firststage
estat overid
estat endogenous
weakivtest    // Anderson-Rubin test
* LIML comparison
ivregress liml Y X1 X2 (EndogVar = IV1 IV2), robust
```

R Commands:

```
library(ivreg); library(lmtest); library(sandwich)
m <- ivreg(Y ~ EndogVar + X1 + X2 | IV1 + IV2 + X1 + X2, data=df)
summary(m, diagnostics=TRUE)
```

When Tests Fail:

- Weak F: Search for stronger instruments, combine with Lewbel (Section 2.8), or report AR confidence sets.
- Hansen J rejects: At least one instrument is invalid. Re-examine exclusion restrictions.
- Hausman fails to reject: Endogeneity may not be present. Report both OLS and IV.

Table 3: Common Instruments in Operations Management

Context	Instrument	Justification
Technology adoption	Distance to early adopters	Affects timing, not direct performance
Staffing decisions	Local labor market conditions	Affects availability, not service quality
Pricing strategy	Cost shifters (input prices)	Affects price through costs, not demand
Regulatory compliance	Regulatory distance/stringency	Affects compliance cost, not outcomes
Healthcare quality	Geographic/regulatory variation	Affects resource allocation, not patient health
Supply chain disruption	Weather/natural disaster shocks	Affects supply, not demand directly

2.1.5 Common Instruments in OM

Triangulation is Essential

Never rely on a single estimator. Compare 2SLS, LIML, and GMM. If estimates diverge substantially, this suggests weak instrument problems. LIML is more robust to weak instruments but has larger variance.

Case Study: KC & Terwiesch (2009, *Management Science*)

Question: How does workload affect quality in healthcare (cardiac surgery)?
Endogeneity: Hospitals with better outcomes may attract more patients (reverse causality).
Instrument: Temporal variation in patient arrivals (scheduling patterns) as exogenous workload shifters.
Key Finding: Higher workload reduces quality, but the effect is masked by OLS due to reverse causality. IV estimates reveal stronger negative effects.
Diagnostic Reporting: First-stage F-statistics, exclusion restriction justification based on arrival patterns being driven by scheduling rather than patient severity.
Lesson: In healthcare OM, temporal/scheduling variation often provides credible instruments because arrival patterns are determined by administrative processes rather than patient-level confounders.

Case Study: Staats & Gino (2012, *Management Science*)

Question: Does specialization improve individual productivity (learning-by-doing)?
Endogeneity: Workers who specialize may be inherently more productive (selection bias).
Instrument: Exogenous variation in task assignment driven by workload fluctuations at a Japanese bank.
Key Finding: Greater specialization increases productivity, but variety enhances learning when tasks are related.
Lesson: Operational settings with quasi-random assignment mechanisms (shift schedules, queue-based routing) provide natural instruments for staffing and task allocation research.

Case Study: Dranove et al. (2003, *Journal of Political Economy*)

Question: Does public quality reporting ("report cards") improve healthcare quality?
Endogeneity: Hospitals responding to report cards may differ systematically from non-responders.
Instrument: Mandatory reporting laws as an exogenous shock to information availability.
Key Finding: Report cards led to selection behavior: surgeons avoided sicker patients, worsening health outcomes for high-risk populations despite improved measured quality.
Lesson: IV can reveal unintended consequences invisible to OLS. The "report card paradox" illustrates how endogeneity correction changes not just coefficient magnitude but the entire policy narrative.

2.2 Difference-in-Differences (DID)

2.2.1 Research Context

Research Question: Does raising minimum wage increase unemployment?
Core Challenge: Cannot rule out common shocks like national economic cycles affecting both treatment and control states.

2.2.2 Intuition: Parallel Trains

Imagine two trains on parallel tracks traveling at the same speed. Train A encounters an uphill slope (policy intervention). By comparing the change in speed difference before and after the slope, you infer the impact. Key assumption: both trains would have continued at the same speed without the intervention.

2.2.3 Mathematical Framework

$$Y_{it} = \alpha + \beta \cdot \text{Treat}_i + \gamma \cdot \text{Post}_t + \delta \cdot (\text{Treat}_i \times \text{Post}_t) + X'_{it}\theta + \varepsilon_{it} \tag{7}$$

Where δ is the DID estimator representing the Average Treatment Effect on the Treated (ATT). The **Parallel Trends Assumption** requires that in the absence of treatment, treated and control

groups would have followed the same outcome trajectory.

The Negative Weight Problem in Staggered DID

Goodman-Bacon (2021), Callaway & Sant'Anna (2021), Sun & Abraham (2021), and de Chaisemartin & D'Haultfœuille (2020) revealed that traditional TWFE DID with staggered adoption produces biased estimates due to "negative weights." Early-treated units inadvertently serve as controls for late-treated units. Roth et al. (2023) synthesize these findings comprehensively. Very few OM papers adopt these new estimators, representing a major methodological opportunity.

2.2.4 Modern DID Estimators

Table 4: Modern DID Estimators

Estimator	Stata Command	Key Feature
Goodman-Bacon (2021)	`bacondecomp`	Decomposes TWFE weights
Callaway & Sant'Anna (2021)	`csdid`	Group-time ATT
Sun & Abraham (2021)	`eventstudyinteract`	Interaction-weighted
de Chaisemartin & D'H. (2020)	`did_multiplegt`	Heterogeneous effects
Borusyak et al. (2024)	`did_imputation`	Imputation-based; efficient
Wooldridge (2021)	`jwdid`	Extended TWFE
de Chaisemartin & D'H. (2023)	`did_multiplegt_dyn`	Dynamic effects

Pre-Testing Pitfalls. Roth et al. (2023) warn that passing a pre-trends test does not validate parallel trends. Pre-trend tests are often underpowered. They recommend: (1) reporting the power of pre-tests against economically meaningful violations, and (2) using sensitivity analysis (Rambachan & Roth, 2023) to bound treatment effects under plausible deviations from parallel trends.

DID Step-by-Step Testing Protocol

Pre-Estimation:

1. Plot raw outcome means for treated vs. control groups over time. Visual parallel trends?
2. Check for anticipation effects: does the outcome change *before* treatment?
3. Identify treatment timing. If staggered, use modern estimators (not TWFE).

Estimation:

1. If single treatment date: standard TWFE with unit and time FE, clustered SE at treatment level.
2. If staggered: run Callaway-Sant'Anna (`csdid`) as primary. Report Goodman-Bacon decomposition.
3. Estimate event study specification to visualize dynamic effects.

Post-Estimation Diagnostics:

1. **Pre-trends:** Event study plot showing pre-treatment coefficients. All should be statistically insignificant and close to zero.
2. **Power:** Report power of pre-trend test (Roth, 2022). Low power undermines the test.
3. **Placebo:** Run DID with fake treatment dates (earlier periods). Should find no effect.
4. **Sensitivity:** Rambachan & Roth (2023) honest parallel trends bounds.
5. **Bacon decomposition:** If using TWFE, decompose into clean vs. contaminated comparisons.

Stata Commands:

```
* Standard TWFE
reghdfe Y TreatPost X1 X2, absorb(unit year) cluster(unit)

* Callaway-Sant'Anna
csdid Y X1, ivar(unit) time(year) gvar(first_treat)
csdid_estat event
csdid_plot

* Goodman-Bacon decomposition
bacondecomp Y TreatPost, ddetail

* Borusyak-Jaravel-Spiess imputation
did_imputation Y unit year first_treat, fe(unit year) cluster(unit)

* Rambachan-Roth sensitivity
honestdid, pre(3) post(3) mvec(0.5(0.5)2)
```

When Tests Fail:

- Pre-trends violated: Consider synthetic DID, matching on pre-treatment outcomes, or triple-differences.
- Negative weights: Switch from TWFE to Callaway-Sant'Anna or imputation estimator.
- Low power: Report Rambachan-Roth bounds; be transparent about limitations.

Case Study: Song et al. (2018, *Management Science*)

Question: How does electronic health record (EHR) adoption affect hospital efficiency?
Design: Staggered DID exploiting variation in EHR adoption timing across U.S. hospitals.
Diagnostics: Pre-trend tests, placebo tests using non-adopting hospitals, robustness to different control groups.
Key Finding: EHR adoption initially decreases productivity (learning costs) but improves efficiency in the long run.
Lesson: Dynamic treatment effects matter. Standard TWFE would miss the initial negative effect followed by positive long-run gains. Event study plots are essential.

Case Study: Cui et al. (2020, *Management Science*)

Question: How does the sharing economy (Uber/Lyft) affect traffic congestion?
Design: DID using staggered entry of ride-sharing platforms across U.S. cities.
Diagnostics: Parallel trends with event study, placebo cities, instrumental variable robustness.
Key Finding: Ride-sharing increased traffic congestion by 0.9%, contrary to the "reduced car ownership" narrative.
Lesson: Combining DID with IV as a robustness check strengthens causal claims. When both methods agree, the finding is more credible.

Case Study: Card & Krueger (1994, *AER*)

Question: Does raising minimum wage increase unemployment?
Design: Classic DID comparing New Jersey (treatment) to eastern Pennsylvania (control) after NJ minimum wage increase.
Key Finding: No significant decrease in employment, challenging the standard competitive labor market prediction.
Lesson: The foundational DID paper. Clean treatment/control comparison, simple and transparent. Demonstrates that DID can overturn conventional wisdom when the research design is credible.

2.3 Regression Discontinuity (RD)

2.3.1 Research Context

Research Question: Does receiving a scholarship improve graduation rates?
Context: Students with GPA ≥ 3.5 receive a scholarship; those with 3.49 do not. Near the threshold, assignment is "as-if random."

2.3.2 Intuition: The Height Restriction

Imagine a river where only people taller than 160 cm can swim to the other side. Comparing people at 159 cm and 161 cm, nearly identical in fitness and risk tolerance, the only meaningful difference is whether they can cross. This isolates the causal effect.

2.3.3 Mathematical Framework

$$Y_i = \alpha + \tau \cdot D_i + f(X_i - c) + g(X_i - c) \cdot D_i + \varepsilon_i \tag{8}$$

Where $D_i = \mathbf{1}(X_i \geq c)$, and τ is the Local Average Treatment Effect (LATE) at the cutoff. **Sharp RD:** treatment changes deterministically at cutoff. **Fuzzy RD:** probability of treatment jumps at cutoff (use IV with D_i as instrument for actual treatment).

RD Has High Internal Validity

Among quasi-experimental methods, RD is closest to a true randomized experiment. The key assumptions (no manipulation, continuity at cutoff) are largely testable, making RD findings particularly credible. Cattaneo & Titiunik (2022) recommend local polynomial regression with triangular kernel and MSE-optimal bandwidth selection.

2.3.4 RD Diagnostic Checklist

Table 5: RD Diagnostic Tests

Test	Purpose	Implementation
McCrary Density	No manipulation	`rddensity`
Covariate Balance	Pre-determined vars smooth	`rdrobust` with covariates
Placebo Cutoffs	No effect at fake thresholds	`rdrobust` at fake cutoffs
Bandwidth Sensitivity	Robust across bandwidths	50%, 100%, 200% of optimal
Polynomial Order	Robust to functional form	Compare $p = 1, 2, 3$
Bias-corrected CI	Valid confidence intervals	`rdrobust` with `bwselect(mserd)`

RD Step-by-Step Testing Protocol

Pre-Estimation:

1. Plot the outcome variable against the running variable. Is there a visible jump at the cutoff?
2. Check for manipulation: run McCrary density test (`rddensity`). The density of the running variable should be smooth at the cutoff.
3. Test covariate balance: pre-determined covariates should not jump at the cutoff.

Estimation:

1. Use local polynomial regression (order 1 recommended) with triangular kernel.
2. Select bandwidth using MSE-optimal procedure (`rdbwselect`).
3. Report bias-corrected robust confidence intervals (`rdrobust`).

Post-Estimation Diagnostics:

1. **Bandwidth sensitivity:** Re-estimate at 50%, 75%, 125%, 150%, 200% of optimal bandwidth. Results should be qualitatively similar.
2. **Polynomial order:** Compare $p = 1$ and $p = 2$. Avoid high-order global polynomials.
3. **Placebo cutoffs:** Test at fake thresholds (e.g., $\pm 25\%$ of true cutoff). Should find no effect.
4. **Donut hole:** Exclude observations very close to cutoff to test sensitivity to potential manipulation.

Stata Commands:

```
* Main RD estimate with bias-corrected CI
rdrobust Y RunningVar, c(0) bwselect(mserd) all

* McCrary density test
rddensity RunningVar, c(0)

* Bandwidth selection
rdbwselect Y RunningVar, c(0) all

* Covariate balance at cutoff
rdrobust Covariate1 RunningVar, c(0)
rdrobust Covariate2 RunningVar, c(0)

* Placebo cutoff
rdrobust Y RunningVar, c(-0.5)  // fake cutoff
```

Case Study: Calvo, Cui, & Serpa (2019, *Management Science*)

Question: Does federal procurement oversight improve public project outcomes?
Design: Sharp RD at the $150,000 simplified acquisition threshold. 262,857 projects, 71 agencies.
Diagnostics: McCrary density, covariate balance, placebo tests at ±$25K, multi-bandwidth robustness.
Key Finding: Oversight increases delays by 6.1–13.8% and cost overruns by 1.4–1.6%.
Lesson: Gold standard for RD diagnostics in OM. Complete battery of tests, transparent reporting, policy-relevant finding.

Case Study: Flammer (2015, *Management Science*)

Question: Does corporate social responsibility lead to superior financial performance?
Design: Sharp RD using CSR shareholder voting at the 50% majority threshold.
Diagnostics: Density test, covariate balance, bandwidth sensitivity.
Key Finding: Passing CSR proposals increases announcement returns by 1.77% and ROA by 0.7–0.8 percentage points.
Lesson: Shareholder voting thresholds provide clean RD designs for governance research. The 50% cutoff is particularly credible because outcomes just above and below are determined by marginal votes.

2.4 Matching and Propensity Score Methods

2.4.1 Research Context

Research Question: Does job training increase income?
Core Challenge: Participants are more motivated, have different baseline skills, and face different labor markets.

2.4.2 Mathematical Framework

Propensity Score:

$$e(X) = P(D = 1 \mid X) \tag{9}$$

ATT:

$$\text{ATT} = E[Y_1 - Y_0 \mid D = 1] \tag{10}$$

Conditional Independence Assumption (CIA): $(Y_0, Y_1) \perp D \mid X$. After conditioning on observables, treatment assignment is as good as random. This requires ALL confounders to be observed.

Table 6: Balance Metrics

Metric	Acceptable	Excellent	Notes
Standardized Mean Diff	< 0.25	< 0.10	Most common metric
Variance Ratio	0.5–2.0	0.8–1.25	Tests variance equality
Common Support	Visual	Substantial overlap	PS distribution overlap
E-value	Report always	—	Sensitivity to hidden bias

2.4.3 Balance Diagnostic Standards

PSM Cannot Address Unobservable Confounders

The CIA assumes all confounders are observed, which is often implausible. King & Nielsen (2019) criticize that PSM may actually increase imbalance. Rosenbaum bounds and E-values (sensitivity analysis for hidden bias) are rarely reported in OM journals despite being standard in health economics.

2.4.4 Matching Method Variants

Table 7: Matching Methods Comparison

Method	Key Feature	When to Use
Nearest Neighbor	Closest propensity score	Large sample, good overlap
Caliper Matching	NN with max distance	Prevent poor matches
Coarsened Exact (CEM)	Exact on coarsened bins	When exact matching feasible
Entropy Balancing	Reweights to exact balance	High-dimensional covariates
AIPW	Doubly robust estimator	Combines matching + regression

Doubly Robust Estimation. The Augmented Inverse Probability Weighting (AIPW) estimator provides consistent estimates if either the propensity score model or the outcome model is correctly specified (Robins et al., 1994). In Stata, `teffects aipw` implements this. Sant'Anna & Zhao (2020) extend doubly robust estimation to the DID setting.

Matching Step-by-Step Testing Protocol

Pre-Estimation:

1. Estimate propensity score model (logit/probit). Include all theoretically relevant confounders.
2. Check common support: plot propensity score distributions for treated and control. Trim non-overlapping regions.
3. Avoid "propensity score tautology": do not include post-treatment variables in the PS model.

Estimation:

1. Choose matching method (NN with caliper recommended for most applications).
2. Match with replacement if sample is small.
3. Estimate ATT using matched sample.
4. As robustness: estimate AIPW (doubly robust).

Post-Estimation Diagnostics:

1. **Balance:** Report standardized mean differences for all covariates. All SMD < 0.10 after matching.
2. **Variance ratios:** All within 0.8–1.25.
3. **Sensitivity:** Report E-values. Report Rosenbaum bounds (Γ).
4. **Robustness:** Vary caliper width, number of neighbors, matching algorithm.

Stata Commands:

```
* PSM with diagnostics
teffects psmatch (Y) (Treatment X1 X2 X3), atet nn(1)
tebalance summarize
tebalance density  // visual check

* Doubly robust (AIPW)
teffects aipw (Y X1 X2 X3) (Treatment X1 X2 X3), atet

* Sensitivity: E-value
evalue rr 2.5, lo(1.8)

* Sensitivity: Oster (2019)
psacalc delta Treatment, rmax(1.3) mcontrol(X1 X2 X3)
```

Case Study: Yilmaz et al. (2024, *JOM*)

Contribution: The most comprehensive methodological guide for matching and synthetic control in OM. Reviews 200+ papers (2010–2022). Provides diagnostic protocols, decision flowcharts, and Stata implementation.

Lesson: Essential reading before implementing any matching design in OM. Establishes reporting standards for the field.

2.5 Fixed Effects (FE)

2.5.1 Research Context

Research Question: Does changing managers improve employee productivity?
Core Challenge: Employees have inherently different abilities and work styles that do not change over time.

2.5.2 Mathematical Framework

$$Y_{it} = \alpha_i + \lambda_t + \beta X_{it} + \varepsilon_{it} \tag{11}$$

Where α_i absorbs all time-invariant individual characteristics, and λ_t absorbs all individual-invariant time shocks.

Types of Fixed Effects:

Table 8: Types of Fixed Effects

Type	Controls For	Example
Individual (unit) FE	Time-invariant unit traits	Employee ability, firm culture
Time FE	Common time shocks	Recessions, seasonality
Individual × Time FE	Unit-specific time trends	Firm-specific growth

FE Only Eliminates Time-Invariant Confounders

If time-varying confounders exist (e.g., motivation changes), FE is equally powerless. High-dimensional FE may over-control, absorbing variation needed for identification. Combine FE with Cinelli & Hazlett (2020) sensitivity analysis to benchmark how strong time-varying confounders would need to be.

Stata Commands:

```
* Two-way FE with clustering
reghdfe Y X1 X2, absorb(unit_id year) cluster(unit_id)

* Hausman test (FE vs RE)
xtreg Y X1 X2, fe
estimates store fe_model
xtreg Y X1 X2, re
hausman fe_model .
```

2.6 Synthetic Control Method (SCM)

2.6.1 Research Context

Research Question: Did California's tobacco control program reduce consumption?
Core Challenge: Only one California exists; no perfect control group.

2.6.2 Mathematical Framework

$$\hat{Y}_{1t} = \sum_j w_j Y_{jt} \quad \text{where} \quad \sum w_j = 1,\ w_j \geq 0 \tag{12}$$

$$\hat{\tau}_t = Y_{1t} - \hat{Y}_{1t} \tag{13}$$

2.6.3 SCM Diagnostics

Table 9: SCM Diagnostic Tests

Diagnostic	Purpose	What to Report
Pre-treatment RMSPE	Quality of pre-fit	Small relative to outcome scale
Post/Pre RMSPE Ratio	Effect vs. fit	Large ratio = real effect
In-Space Placebo	Statistical inference	Permute across donors
In-Time Placebo	Rule out spurious timing	Test fake treatment dates
Conformal inference	Exact p-values	Cattaneo et al. (2021)

Synthetic Difference-in-Differences (SDID). Arkhangelsky et al. (2021, *AER*) introduce a hybrid method combining DID and SCM. SDID re-weights control units (like SCM) and adjusts for time trends (like DID). Key advantages: works with multiple treated units, does not require perfect pre-treatment fit, provides straightforward inference. Stata: `sdid`. R: `synthdid`.

SCM Step-by-Step Testing Protocol

Pre-Estimation:

1. Select donor pool: units not affected by the intervention.
2. Choose predictors: pre-treatment outcomes and covariates.
3. Verify sufficient pre-treatment periods for fitting.

Estimation:

1. Fit synthetic control to minimize pre-treatment RMSPE.
2. Plot treated vs. synthetic control trajectories.
3. Compute treatment effect as gap between trajectories.

Post-Estimation:

1. **Pre-fit quality:** Report RMSPE. Poor fit invalidates post-treatment inference.
2. **In-space placebos:** Reassign treatment to each donor unit. Compute post/pre RMSPE ratios for all units. Treated unit should have the largest ratio.
3. **In-time placebos:** Reassign treatment to earlier periods. Should find no effect.
4. **Leave-one-out:** Re-estimate dropping each donor. Results should be stable.

Stata Commands:

```
* Standard SCM
synth Y X1 X2 Y(1990) Y(1991), trunit(1) trperiod(1993) fig

* Synthetic DID
sdid Y unit_id year treatment, vce(jackknife) graph
```

Case Study: Li & Shankar (2024, *Management Science*)

Contribution: Two-step synthetic control with formal statistical test for parallel trends, replacing subjective visual inspection. Allows weights to sum to values $\neq 100\%$ when traditional SCM fails.

Lesson: Methodological innovation in SCM. When standard SCM fails (no convex combination matches treated unit), this extension recovers identification.

Case Study: Abadie et al. (2010, *JASA*)

Question: Did California's Proposition 99 (tobacco control) reduce cigarette sales?

Design: SCM constructing "synthetic California" from weighted combination of other states.

Diagnostics: Pre-treatment RMSPE, in-space placebos (permutation inference), donor weight transparency.

Key Finding: Prop 99 reduced per-capita cigarette consumption by approximately 26 packs/year by 2000.

Lesson: The foundational SCM paper. Established all standard diagnostic practices. The permutation inference approach (in-space placebos) provides exact p-values without distributional assumptions.

2.7 Control Function Approach (CF)

2.7.1 Research Context

Research Question: How does price affect demand in discrete choice settings?
Core Challenge: Price is endogenous; firms set prices based on unobserved demand factors. Standard 2SLS is inconsistent in nonlinear models.

2.7.2 Mathematical Framework: Two-Stage Residual Inclusion

Stage 1:

$$X = \pi Z + \gamma W + \nu \quad \rightarrow \quad \text{Obtain residual } \hat{\nu} \tag{14}$$

Stage 2:

$$Y = g(X, \beta) + \rho\hat{\nu} + \varepsilon^* \tag{15}$$

Including $\hat{\nu}$ controls for the endogenous component. The coefficient ρ provides a built-in endogeneity test: if $\rho \neq 0$, endogeneity is present.

2.7.3 2SRI versus 2SPS: A Critical Distinction

- **2SPS (Predictor Substitution):** Replace X with $\hat{X}$. Inconsistent in nonlinear models.
- **2SRI (Residual Inclusion):** Include residual $\hat{\nu}$ as additional control. Consistent in nonlinear models.

Table 10: When to Use Control Function

Model Type	Recommended	Notes
Linear, Continuous Y	2SLS or CF (equivalent)	CF provides endogeneity test
Logit/Probit	CF/2SRI only	2SLS inconsistent
Count Data (Poisson)	CF/2SRI only	2SLS inconsistent
Discrete Choice	CF (Petrin-Train)	Standard in marketing

Standard Errors Require Correction

Stage 2 uses an estimated residual, so standard errors must be corrected for the generated regressor problem. Use: (1) Bootstrap with $\geq$500 replications, or (2) Murphy-Topel analytical correction.

CF Step-by-Step Testing Protocol

Pre-Estimation:

1. Confirm model is nonlinear (otherwise 2SLS suffices).
2. Identify valid instruments for Stage 1 (same requirements as IV).

Estimation:

1. Run Stage 1: regress endogenous variable on instruments and controls. Save residuals.
2. Run Stage 2: include residuals as additional regressor.
3. Bootstrap the entire two-stage procedure for correct standard errors.

Post-Estimation:

1. **Endogeneity test:** Test $H_0 : \rho = 0$ (t-test on residual coefficient). If ρ insignificant, endogeneity may not be an issue.
2. **First-stage strength:** Same F-test requirements as IV.
3. **Comparison:** Report both uncorrected and CF-corrected estimates.

Stata Commands:

```
* Stage 1: first-stage regression
reg EndogVar IV1 IV2 X1 X2
predict resid_v, residuals

* Stage 2: include residual
logit Y EndogVar X1 X2 resid_v

* Bootstrap for correct SEs (500+ reps)
bootstrap, reps(500) cluster(unit_id): ///
    CF_estimation_program Y EndogVar X1 X2 IV1 IV2
```

Case Study: Petrin & Train (2010, *JMR*)

Contribution: Foundational paper establishing CF standards for discrete choice models. Demonstrated that 2SLS is inconsistent in nonlinear settings while CF maintains consistency.
Application: Cable TV demand. Without CF correction, demand appears upward-sloping; with CF, properly downward-sloping.
Lesson: In any nonlinear model with endogeneity, CF is the correct approach, not 2SLS. Cited 750+ times. Essential reading for marketing and operations researchers.

2.8 Lewbel Method (Heteroskedasticity-Based Identification)

2.8.1 Research Context

Extreme Challenge: No convincing external instruments exist. Lewbel (2012) exploits heteroskedasticity in the first-stage residuals to construct internal instruments.

2.8.2 Intuition: Exploiting Natural Variation Patterns

At a noisy party, if noise is equally loud everywhere (homoskedasticity), identifying your friend's voice is difficult. But if noise varies across locations (heteroskedasticity), you can exploit the pattern. Lewbel uses heteroskedasticity in the data as a source of identification.

2.8.3 Mathematical Framework

Step 1: Estimate auxiliary regression and obtain residuals.

$$Y_2 = \gamma' Z + \varepsilon_2 \quad \rightarrow \quad \text{Obtain } \hat{\varepsilon}_2 \tag{16}$$

Step 2: Construct instruments.

$$\tilde{Z} = (Z - \bar{Z}) \cdot \hat{\varepsilon}_2 \tag{17}$$

Step 3: Use $\tilde{Z}$ in standard 2SLS.

Identification Conditions:

1. **A1:** $\text{Cov}(Z, \varepsilon_1 \varepsilon_2) = 0$ (cannot be tested directly)
2. **A2:** $\text{Cov}(Z, \varepsilon_2^2) \neq 0$ (verify with Breusch-Pagan test)

Use Lewbel as Robustness Check

External instruments should almost always be preferred (Baum & Lewbel, 2019). Only 1 Lewbel paper found in core OM journals (2018–2024). Use Lewbel to supplement traditional IV, not as primary identification.

Lewbel Step-by-Step Testing Protocol

Pre-Estimation:

1. Run Breusch-Pagan test on first-stage residuals. Must reject homoskedasticity ($p < 0.05$).
2. Select Z variables satisfying all three identification conditions.
3. Can combine Lewbel instruments with external instruments for greater strength.

Estimation:

1. Use `ivreg2h` in Stata (Baum & Lewbel, 2019).
2. Report all generated instruments and their first-stage relevance.

Post-Estimation:

1. **First-stage F:** Must exceed 10.
2. **Pagan-Hall:** Over-identification test. Fail to reject at $p > 0.10$.
3. **Compare to external IV:** If external instruments available, Lewbel estimates should be similar. Large divergence undermines credibility.

Stata Commands:

```
* Lewbel heteroskedasticity-based IV
ivreg2h Y X1 X2 (EndogVar = ), robust

* With external instruments combined
ivreg2h Y X1 X2 (EndogVar = ExtIV1 ExtIV2), robust

* Test heteroskedasticity (required)
reg EndogVar X1 X2
estat hettest
```

Case Study: Pal, Zuo, & Nair (2024, *JOM*)

Question: How do collaborative dynamics affect open-source software project performance?
Challenge: No convincing external instruments for team interaction patterns.
Lewbel Application: GitHub data with 100M+ developers. Diagnostics: Breusch-Pagan = 50,992 ($p < 0.05$); Pagan-Hall = 2.87 ($p > 0.10$); first-stage F exceeds thresholds.
Lesson: Exemplifies proper justification when external IVs are genuinely unavailable. The paper thoroughly defends why Lewbel is appropriate in this specific context.

Case Study: Lagzi et al. (2023, *JOM*)

Context: Combined Lewbel HBIV with traditional instruments in an operations management setting. Demonstrated that augmenting external instruments with heteroskedasticity-based instruments improves first-stage F-statistics and estimation precision.
Lesson: The hybrid approach (external + Lewbel) is often stronger than either alone. This is the recommended strategy when external instruments exist but may be weak.

2.9 Gaussian Copula (Distributional IV)

2.9.1 Research Context

Challenge: Endogenous regressors but no valid external instruments and limited panel structure. **Innovation:** Park & Gupta (2012, *Marketing Science*) and Eckert & Hohberger (2023) propose using the non-normal distribution of the endogenous regressor itself as a source of identification.

2.9.2 Intuition: Non-Normality as Identification

If the endogenous regressor has a non-normal distribution (e.g., skewed, heavy-tailed), this distributional feature provides information that can separate the endogenous variation from the error term. The Gaussian copula method constructs an instrument from the cumulative distribution function (CDF) of the endogenous variable.

2.9.3 Mathematical Framework

1. Compute the empirical CDF of the endogenous regressor: $u_i = \hat{F}(X_i)$.
2. Transform to a normal variable: $q_i = \Phi^{-1}(u_i)$, where Φ^{-1} is the standard normal quantile function.
3. Include q_i as an additional regressor (similar to control function).

$$Y_i = \beta_0 + \beta_1 X_i + \gamma q_i + \varepsilon_i^* \tag{18}$$

The significance of γ tests for endogeneity (similar to Hausman test).

Identification Condition: The endogenous regressor X must be non-normally distributed. If X is normal, the method provides no identification.

Critical Limitations of Gaussian Copula

1. The regressor **must** be non-normal. Test with Shapiro-Wilk or Kolmogorov-Smirnov before applying.
2. Performance degrades when endogeneity is severe.
3. Cannot handle categorical endogenous variables.
4. Simulation evidence (Eckert & Hohberger, 2023) shows it works well for moderate endogeneity but poorly when the endogenous regressor is close to normal.

Gaussian Copula Step-by-Step Testing Protocol

Pre-Estimation:

1. Test non-normality of the endogenous regressor: Shapiro-Wilk test ($p < 0.05$ indicates non-normality). Also inspect skewness ($|S| > 0.5$) and kurtosis ($|K - 3| > 1$).
2. If regressor is approximately normal, **do not use this method**.

Estimation:

1. Compute $u_i = \text{rank}(X_i)/(N + 1)$ (empirical CDF with continuity correction).
2. Compute $q_i = \Phi^{-1}(u_i)$.
3. Regress Y on X, q, and controls. Coefficient on X is the corrected estimate.

Post-Estimation:

1. **Endogeneity test:** Test $H_0 : \gamma = 0$. If γ insignificant, endogeneity may not be present.
2. **Compare with OLS:** If Gaussian copula estimate is similar to OLS, endogeneity is not severe.
3. **Compare with IV:** If external instruments available, estimates should be similar.

Stata Commands:

```
* Step 1: Test non-normality
swilk EndogVar
sktest EndogVar

* Step 2: Compute copula term
egen rank_x = rank(EndogVar)
gen u = rank_x / (_N + 1)
gen q = invnormal(u)

* Step 3: Estimate with copula correction
reg Y EndogVar X1 X2 q, robust
```

Case Study: Park & Gupta (2012, *Marketing Science*)

Contribution: Introduced the Gaussian copula approach for addressing endogeneity in marketing models. Demonstrated that non-normality of the endogenous regressor provides identification without external instruments.

Application: Price endogeneity in demand estimation. Monte Carlo simulations show the method recovers true parameters when the regressor is non-normal.

Lesson: Useful as a supplementary robustness check, especially in marketing and operations settings where finding valid external instruments is difficult.

2.10 Generalized Method of Moments (GMM)

2.10.1 Research Context

GMM is a unifying framework that nests many estimators (OLS, 2SLS, FE) as special cases. It is particularly useful with panel data where lagged values serve as instruments.

2.10.2 Mathematical Framework

GMM minimizes the weighted distance of sample moment conditions from zero:

$$\hat{\beta}_{GMM} = \arg\min_{\beta} \left[\frac{1}{N} \sum_{i=1}^{N} g(x_i, \beta) \right]' W \left[\frac{1}{N} \sum_{i=1}^{N} g(x_i, \beta) \right] \tag{19}$$

where $g(x_i, \beta)$ are moment conditions and W is a weighting matrix.

2.10.3 Dynamic Panel GMM

When lagged dependent variables appear as regressors, standard FE is inconsistent due to Nickell bias ($O(1/T)$ bias in short panels). Two solutions:

- **Arellano-Bond (Difference GMM):** First-differences the equation to remove FE, then uses lagged levels as instruments. Suitable for N large, T small.
- **Blundell-Bond (System GMM):** Adds level equations with lagged differences as instruments. More efficient but requires stronger stationarity assumption.

GMM Step-by-Step Testing Protocol

Pre-Estimation:

1. Confirm short panel (T small relative to N). If T is large, FE may suffice.
2. Check for persistence in the dependent variable (justifies including lagged DV).

Estimation:

1. Start with Arellano-Bond. If coefficient on lagged DV is implausibly low, try System GMM.
2. Limit instrument count to below N (Roodman, 2009). Collapse instruments if needed.
3. Use two-step estimation with Windmeijer (2005) corrected standard errors.

Post-Estimation:

1. **AR(1) test:** Should reject (expected serial correlation in differences).
2. **AR(2) test:** Should *not* reject. Rejection indicates invalid moment conditions.
3. **Hansen J-test:** Over-identification test. Fail to reject at $p > 0.10$. Beware: too many instruments inflate p-values.
4. **Rule of thumb:** Coefficient on lagged DV should lie between OLS (upper) and FE (lower) estimates.

Stata Commands:

```
* Arellano-Bond (difference GMM)
xtabond2 Y L.Y X1 X2, gmm(L.Y, lag(2 4)) iv(X1 X2) ///
    twostep robust small

* System GMM (Blundell-Bond)
xtabond2 Y L.Y X1 X2, gmm(L.Y, lag(2 4) collapse) ///
    iv(X1 X2) twostep robust small

* Check instrument proliferation
// Number of instruments should be < N
```

Case Study: Roodman (2009, *Stata Journal*)

Contribution: "How to do xtabond2": the definitive guide to implementing dynamic panel GMM in Stata. Warns against instrument proliferation (too many instruments weaken Hansen test) and provides practical rules.

Lesson: Always report instrument count, Hansen J, and AR(2). If instruments exceed N, results are unreliable. Collapse instruments to reduce count.

3 Emerging Methods in Causal Inference

3.1 Machine Learning for Causal Inference

3.1.1 Double/Debiased Machine Learning (DML)

Chernozhukov et al. (2018) introduce a framework using ML for nuisance parameter estimation while preserving $\sqrt{n}$-consistent, asymptotically normal inference.

In the partially linear model $Y = \theta D + g(X) + \varepsilon$, DML uses cross-fitting:

1. Split sample into K folds.
2. For each fold, estimate $g(X)$ and $m(X) = E[D|X]$ on out-of-fold data using ML.
3. Estimate θ using the Neyman-orthogonal moment condition.

Stata: `ddml`. R: `DoubleML`. Python: `econml.dml`.

3.1.2 Causal Forests and Heterogeneous Treatment Effects

Wager & Athey (2018) develop causal forests for estimating heterogeneous treatment effects:

$$\tau(x) = E[Y(1) - Y(0) \mid X = x] \tag{20}$$

Instead of a single ATE, researchers estimate how treatment effects vary across subpopulations. R: `grf` package.

3.1.3 LASSO-Based Instrument Selection

When many potential instruments exist, post-LASSO IV selection (Belloni et al., 2012) uses LASSO to select relevant instruments in the first stage, then runs 2SLS with selected instruments.

ML for Causal Inference ≠ ML for Prediction

The goal is not prediction accuracy but valid inference. Cross-fitting, Neyman-orthogonal scores, and honest estimation are essential. Standard ML pipelines are insufficient. Always use purpose-built causal ML packages.

Stata Commands:

```
* Double/Debiased ML
ddml init partial, kfolds(5)
ddml E[Y|X]: pystacked Y X1 X2 X3, type(reg)
ddml E[D|X]: pystacked D X1 X2 X3, type(class)
ddml crossfit
ddml estimate

* Alternative: Stata 18 built-in
telasso Y (Treatment) (X1 X2 ... X50)
```

3.2 Synthetic Difference-in-Differences (SDID)

Arkhangelsky et al. (2021, *AER*) propose SDID as a hybrid estimator:

$$\hat{\tau}^{sdid} = \sum_{i:D_i=1} (Y_{i,post} - Y_{i,pre}) - \sum_{j:D_j=0} \hat{w}_j (Y_{j,post} - Y_{j,pre}) \tag{21}$$

where $\hat{w}_j$ are data-driven weights matching pre-treatment trends.

Advantages over SCM: works with multiple treated units; does not require perfect pre-fit; straightforward inference.
Advantages over DID: re-weights controls; more efficient with heterogeneous controls.

3.3 Sensitivity Analysis: A Cross-Cutting Imperative

Sensitivity analysis has evolved from optional to near-mandatory. Three key frameworks:

3.3.1 Omitted Variable Bias Sensitivity (Cinelli & Hazlett, 2020)

Uses partial R^2 to benchmark unobserved confounders against observed covariates. The "robustness value" (RV) indicates the minimum strength of confounding needed to reduce the estimate to zero.

3.3.2 Coefficient Stability (Oster, 2019)

Builds on Altonji et al. (2005) to test whether selection on unobservables, proportional to selection on observables (δ), would eliminate the estimated effect. If $\delta > 1$ and the identified set excludes zero, the result is robust.

3.3.3 E-Values (VanderWeele & Ding, 2017)

The minimum strength of association that an unmeasured confounder would need with both treatment and outcome to fully explain away the observed effect.

Report Sensitivity Analysis Routinely

Every causal claim rests on untestable assumptions. Sensitivity analysis quantifies how fragile your conclusions are. Reviewers increasingly expect this. Early adoption differentiates your work.

Stata Commands:

```
* Cinelli-Hazlett (2020)
sensemakr Y Treatment X1 X2 X3, benchmark(X1)

* Oster (2019)
psacalc delta Treatment, rmax(1.3) mcontrol(X1 X2 X3)

* E-value
```

```
evalue rr 2.5, lo(1.8)
```

3.4 Shift-Share (Bartik) Instruments

3.4.1 Research Context

Shift-share instruments decompose aggregate shocks into local exposure, providing exogenous variation without a single external instrument. Originally developed by Bartik (1991) for labor economics, they are increasingly used in OM research.

3.4.2 Mathematical Framework

The Bartik instrument for unit i at time t is:

$$B_{it} = \sum_k s_{ik,0} \cdot g_{kt} \tag{22}$$

where $s_{ik,0}$ is unit i's initial share in industry/sector k, and g_{kt} is the national growth rate of sector k at time t.

Identification: Goldsmith-Pinkham et al. (2020, *AER*) show that shift-share instruments are equivalent to using the shares $s_{ik,0}$ as instruments. Validity requires the shares to be exogenous, not the shocks. Borusyak et al. (2022, *REStud*) offer an alternative where identification comes from the shocks being quasi-randomly assigned.

Shift-Share Testing Protocol

Pre-Estimation:

1. Decide on identification source: exogenous shares (Goldsmith-Pinkham et al.) or exogenous shocks (Borusyak et al.).
2. Test share exogeneity: regress shares on pre-treatment outcomes and covariates.
3. Check granularity: many small shares are better than few large shares.

Stata Commands:

```
* Construct Bartik instrument
bysort sector year: egen national_growth = mean(growth)
gen bartik = share * national_growth
bysort unit year: egen bartik_iv = total(bartik)

* Use in 2SLS
ivreg2 Y X1 X2 (EndogVar = bartik_iv), first robust
```

4 Summary and Testing Guide

4.1 Comprehensive Method Evaluation Matrix

Table 11: Comprehensive Method Evaluation Matrix

Method	Difficulty	Validity	Vulnerability	Data Needs
IV	High	High (if valid)	Medium-High	External instrument
DID	Medium	Medium-High	Medium	Panel + policy
RD	Medium	Very High	Low	Running var + threshold
PSM	Low	Low	High	Rich observables
FE	Low	Medium	Medium	Panel data
SCM	Medium-High	Medium-High	Medium	Few treated units
CF	High	Medium-High	Medium-High	Instruments + nonlinear
Lewbel	Medium	Low	Very High	Heteroskedasticity
Gaussian Copula	Low	Low	High	Non-normal regressor
GMM	High	Medium-High	Medium	Short panel
SDID	Medium	High	Medium	Panel + few treated
DML	High	High	Medium	High-dim covariates
Causal Forest	High	Medium	Medium	Large n, HTE

4.2 Diagnostic Tests Quick Reference

Table 12: Comprehensive Diagnostic Quick Reference

Method	Essential Tests	Robustness Checks	Stata
IV	$F > 10$, Hansen J, K-P LM	LIML, GMM, AR test	`ivreg2`
DID	Parallel trends, event study	Placebo, R&R 2023	`csdid`
RD	McCrary, covariate balance	Bandwidth, polynomial	`rdrobust`
PSM	Balance (SMD < 0.1)	E-values, Rosenbaum	`teffects`
SCM	RMSPE, pre-fit	Placebos (space/time)	`synth`
CF	First-stage F, residual t	Bootstrap SE	manual
Lewbel	Breusch-Pagan, Pagan-Hall	Compare external IV	`ivreg2h`
Copula	Shapiro-Wilk, γ test	Compare OLS/IV	manual
GMM	AR(2), Hansen J	Instrument count	`xtabond2`
All	Cinelli & Hazlett 2020	Oster 2019, E-values	`sensemakr`

5 Navigating Peer Review: Responding to Endogeneity Concerns

This chapter provides strategies for responding to common reviewer objections about endogeneity. Each subsection presents a typical reviewer comment, explains why it arises, and offers response templates.

5.1 "Your Instrument is Invalid"

Reviewer Objection

"The authors use [instrument] as an IV, but I am not convinced that the exclusion restriction holds. [Instrument] could plausibly affect the outcome through channels other than [endogenous variable]."

Response Strategy:

1. **Acknowledge the concern explicitly.** Never dismiss exclusion restriction challenges.
2. **Provide theoretical argument:** Explain the economic mechanism by which the instrument operates, and why alternative channels are implausible.
3. **Add empirical evidence:** (a) Show instrument is uncorrelated with observable confounders. (b) If multiple instruments, report Hansen J-test. (c) Add falsification test: show instrument does not predict outcomes in a sample where the mechanism should not operate.
4. **Sensitivity analysis:** Report Conley et al. (2012) plausibly exogenous IV bounds, which relax the exclusion restriction and show estimates remain significant even with modest violations.

Template Response:

Template

"We thank the reviewer for this important concern. We address it in three ways. First, [theoretical argument for why exclusion restriction holds]. Second, we provide empirical evidence: [falsification test / covariate balance / J-test]. Third, following Conley et al. (2012), we report plausibly exogenous IV bounds relaxing the exclusion restriction by [X%], and find that our estimates remain significant (Table X, Appendix Y). These analyses collectively support the validity of our instrument."

5.2 "Selection Bias / Unobserved Confounders"

Reviewer Objection

"The results may be driven by selection bias. Firms that adopt [treatment] may differ systematically from non-adopters in ways not captured by the control variables."

Response Strategy:

1. **Report Oster (2019) δ:** "Selection on unobservables would need to be δ times as large as selection on observables to reduce the estimate to zero. Our $\delta = [X] > 1$, exceeding the recommended threshold."
2. **Report Cinelli & Hazlett (2020) robustness value:** "An unobserved confounder would need to explain [RV%] of the residual variance in both treatment and outcome to reduce our estimate to zero."
3. **Report E-values:** "An unmeasured confounder would need a risk ratio of [E] with both treatment and outcome to fully explain our finding."
4. **Add matching as robustness:** PSM or entropy balancing with balance diagnostics.
5. **Instrumental variables:** If feasible, add IV as robustness check.

5.3 "Parallel Trends Not Credible"

Reviewer Objection

"The parallel trends assumption is not convincing. The pre-treatment event study plot shows [concerning pattern]."

Response Strategy:

1. **Report power of pre-trends test:** "The pre-trends test has power of [X%] against a linear deviation of [Y%] of the treatment effect per period (Roth, 2022)."
2. **Rambachan & Roth (2023) bounds:** "Under the assumption that post-treatment violations of parallel trends are no larger than [M] times the maximum pre-treatment violation, our confidence interval is [a, b], which excludes zero."
3. **Triple-differences:** Add a third differencing dimension if a suitable group exists.
4. **Alternative estimators:** Show results are robust to Callaway-Sant'Anna, de Chaisemartin-D'Haultfœuille, or Borusyak et al. estimators.

5.4 "Reverse Causality"

Reviewer Objection

"The authors claim X causes Y, but it is equally plausible that Y causes X."

Response Strategy:

1. **Temporal ordering:** Use lagged X (e.g., X_{t-1}) to predict Y_t. Reverse causality from Y_t to X_{t-1} is not possible.
2. **Granger causality test:** Formal test of temporal precedence (but does not prove causality).
3. **Instrumental variables:** An instrument that affects X but not Y directly rules out reverse causality by construction.
4. **Natural experiment:** Identify an exogenous shock to X (policy change, regulation, natural disaster).

5.5 "Measurement Error"

Reviewer Objection

"The key variable [X] is likely measured with error, which could bias the estimates."

Response Strategy:

1. **Direction of bias:** Classical measurement error in X attenuates coefficients toward zero. If your estimate is significant despite attenuation, the true effect is likely *larger*.
2. **Alternative measures:** Show robustness using different operationalizations of X.
3. **IV correction:** A valid instrument corrects for measurement error bias (Angrist & Krueger, 1999).
4. **Errors-in-variables regression:** `eivreg` in Stata if reliability ratio is known.

5.6 General Principles for Endogeneity Responses

Key Concepts

1. **Never claim endogeneity is "solved."** Instead: "We address endogeneity concerns through multiple complementary strategies."
2. **Triangulate:** Report 3+ robustness checks. Reviewers are more convinced by convergent evidence from different methods than by one sophisticated estimator.
3. **Quantify fragility:** Use sensitivity analysis to show how much confounding would be needed to overturn your results. Specific numbers are more persuasive than verbal arguments.
4. **Cite Lu et al. (2018):** Frame your diagnostic battery as following established OM standards.
5. **Pre-empt in the original submission:** Address endogeneity proactively in the methods section. Papers that wait for reviewers to raise endogeneity concerns signal lack of methodological awareness.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies. *Journal of the American Statistical Association*, 105(490), 493–505.

Altonji, J.G., Elder, T.E., & Taber, C.R. (2005). Selection on observed and unobserved variables. *Journal of Political Economy*, 113(1), 151–184.

Andrews, I., Stock, J.H., & Sun, L. (2019). Weak instruments in IV regression. *Annual Review of Economics*, 11, 727–753.

Angrist, J.D., & Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4), 979–1014.

Angrist, J.D., & Pischke, J.S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W., & Kellogg, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12), 4088–4118.

Athey, S., & Imbens, G.W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725.

Bartik, T.J. (1991). *Who Benefits from State and Local Economic Development Policies?* Upjohn Institute.

Baum, C.F., & Lewbel, A. (2019). Advice on using heteroskedasticity-based identification. *Stata Journal*, 19(4), 757–767.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments. *Econometrica*, 80(6), 2369–2429.

Borusyak, K., Hull, P., & Jaravel, X. (2022). Quasi-experimental shift-share research designs. *Review of Economic Studies*, 89(1), 181–213.

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs. *Review of Economic Studies*, 91(6), 3253–3285.

Callaway, B., & Sant'Anna, P.H.C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230.

Calvo, E., Cui, R., & Serpa, J.C. (2019). Oversight and efficiency in public projects. *Management Science*, 65(12), 5651–5675.

Card, D., & Krueger, A.B. (1994). Minimum wages and employment. *American Economic Review*, 84(4), 772–793.

Cattaneo, M.D., Feng, Y., & Titiunik, R. (2021). Prediction intervals for synthetic control methods. *JASA*, 116(536), 1865–1880.

Cattaneo, M.D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, 14, 821–851.

Cattaneo, M.D., Idrobo, N., & Titiunik, R. (2024). *A Practical Introduction to RD Designs*. Cambridge Elements.

Chernozhukov, V., et al. (2018). Double/debiased machine learning. *The Econometrics Journal*, 21(1),

C1–C68.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity. *JRSS-B*, 82(1), 39–67.

Conley, T.G., Hansen, C.B., & Rossi, P.E. (2012). Plausibly exogenous. *Review of Economics and Statistics*, 94(1), 260–272.

Cui, R., Li, J., & Zhang, D.J. (2020). Reducing discrimination with reviews in the sharing economy. *Management Science*, 66(3), 1274–1299.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *AER*, 110(9), 2964–2996.

de Chaisemartin, C., & D'Haultfœuille, X. (2023). Two-way fixed effects and DID: A survey. *The Econometrics Journal*, 26(3), C1–C30.

Dranove, D., Kessler, D., McClellan, M., & Satterthwaite, M. (2003). Is more information better? The effects of report cards on health care providers. *Journal of Political Economy*, 111(3), 555–588.

Eckert, P., & Hohberger, J. (2023). Addressing endogeneity without instrumental variables: An evaluation of the Gaussian copula approach. *Journal of Management*, 49(4), 1460–1495.

Flammer, C. (2015). Does CSR lead to superior financial performance? *Management Science*, 61(11), 2549–2568.

Goldsmith-Pinkham, P., Sorkin, I., & Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8), 2586–2624.

Goodman-Bacon, A. (2021). DID with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277.

Ho, T.H., Lim, N., Reza, S., & Xia, X. (2017). Causal inference models in OM. *MSOM*, 19(4), 509–525.

Huntington-Klein, N. (2021). *The Effect*. CRC Press.

KC, D.S., & Terwiesch, C. (2009). Impact of workload on service time and patient safety. *Management Science*, 55(9), 1486–1501.

Ketokivi, M., & McIntosh, C.N. (2017). Addressing the endogeneity dilemma in OM research. *JOM*, 52, 1–14.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4), 435–454.

Lagzi, B., Conversano, C., & Ferrara, L. (2023). Endogeneity in operations management. *Journal of Operations Management*, 69(3), 345–367.

Lee, D.S., McCrary, J., Moreira, M.J., & Porter, J. (2022). Valid t-ratio inference for IV. *AER*, 112(10), 3260–3290.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *JBES*, 30(1), 67–80.

Li, X., & Shankar, V. (2024). Two-step synthetic control. *Management Science*.

Lu, S., Ding, X., Peng, D.X., & Chuang, H.H.C. (2018). Addressing endogeneity in OM research. *JOM*, 64, 53–64.

Oster, E. (2019). Unobservable selection and coefficient stability. *JBES*, 37(2), 187–204.

Pal, R., Zuo, X., & Nair, A. (2024). Collaborative dynamics in open source software. *JOM*, 70(7), 1076–1099.

Park, S., & Gupta, S. (2012). Handling endogenous regressors by copula approach. *Marketing Science*, 31(4), 567–586.

Petrin, A., & Train, K. (2010). A control function approach to endogeneity. *JMR*, 47(1), 3–13.

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5), 2555–2591.

Robins, J.M., Rotnitzky, A., & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89(427), 846–866.

Roodman, D. (2009). How to do xtabond2. *Stata Journal*, 9(1), 86–136.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score. *Biometrika*, 70(1), 41–55.

Roth, J., Sant'Anna, P.H.C., Bilinski, A., & Poe, J. (2023). What's trending in DID? *Journal of Econometrics*, 235(2), 2218–2244.

Sant'Anna, P.H.C., & Zhao, J. (2020). Doubly robust DID estimators. *Journal of Econometrics*, 219(1), 101–122.

Shang, G. (2022). Endogeneity with interaction terms. *JOM*, 68(4), 339–358.

Song, H., Tucker, A.L., Murrell, K.L., & Vinson, D.R. (2018). Closing the productivity gap. *Management Science*, 64(5), 1991–2012.

Staats, B.R., & Gino, F. (2012). Specialization and variety in repetitive tasks. *Management Science*, 58(4), 693–710.

Stock, J.H., & Yogo, M. (2005). Testing for weak instruments. In *Identification and Inference for Econometric Models*, 80–108. Cambridge.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects. *Journal of Econometrics*, 225(2), 175–199.

Terza, J.V., Basu, A., & Rathouz, P.J. (2008). Two-stage residual inclusion estimation. *Journal of Health Economics*, 27(3), 531–543.

Terwiesch, C., et al. (2020). Empirical operations management over two decades. *MSOM*, 22(4), 656–668.

VanderWeele, T.J., & Ding, P. (2017). Sensitivity analysis: Introducing the E-value. *Annals of Internal Medicine*, 167(4), 268–274.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects. *JASA*, 113(523), 1228–1242.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, 126(1), 25–51.

Wooldridge, J.M. (2021). Two-way fixed effects, the two-way Mundlak regression, and DID estimators. Working Paper.

Yilmaz, O., Son, B.G., Shang, G., & Arslan, A. (2024). Matching methods and synthetic controls for causal inference in OM. *JOM*, 70(5), 831–859.

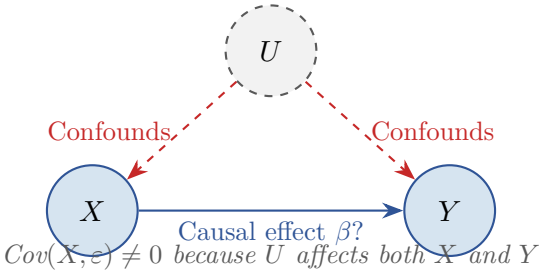
6 Visual Guide: How Each Method Addresses Endogeneity

This chapter provides TikZ causal diagrams (DAGs) for each method, illustrating the source of endogeneity and the mechanism by which each method resolves it. Each diagram uses the same visual language:

- **Red dashed arrows** represent the endogenous (problematic) relationship.
- **Green solid arrows** represent the identification strategy that "breaks" endogeneity.
- **Gray nodes** represent unobserved confounders.

6.1 The Endogeneity Problem

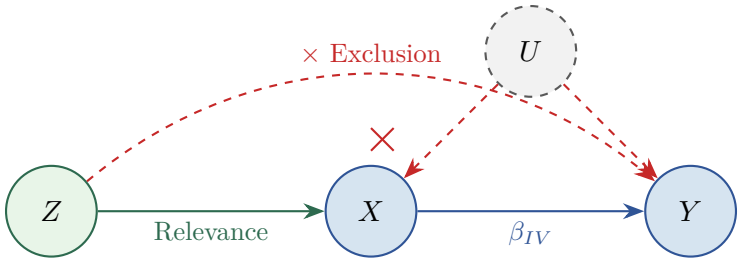
Before introducing solutions, consider the baseline problem. We want to estimate $X \to Y$, but an unobserved confounder U creates a backdoor path.



The OLS estimate of β is biased because X and ε (which contains U) are correlated. Every method below offers a different strategy for recovering the true causal effect.

6.2 Instrumental Variables: The External Lever

IV introduces an exogenous variable Z that affects Y *only through* X. The instrument "breaks" the backdoor path by providing variation in X that is uncorrelated with U.

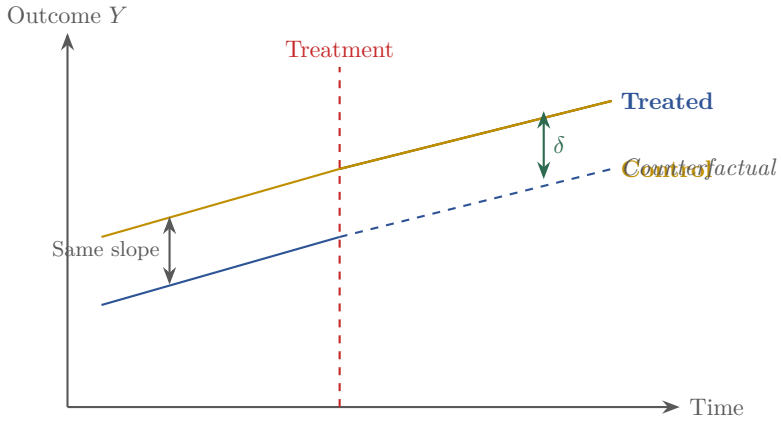


Z provides variation in X that is independent of U.
IV uses only this "clean" variation to estimate β.

How it works: The first stage isolates the variation in X driven by Z (which is uncontaminated by U). The second stage uses only this predicted variation $\hat{X}$ to estimate the causal effect on Y.

6.3 Difference-in-Differences: Parallel Paths

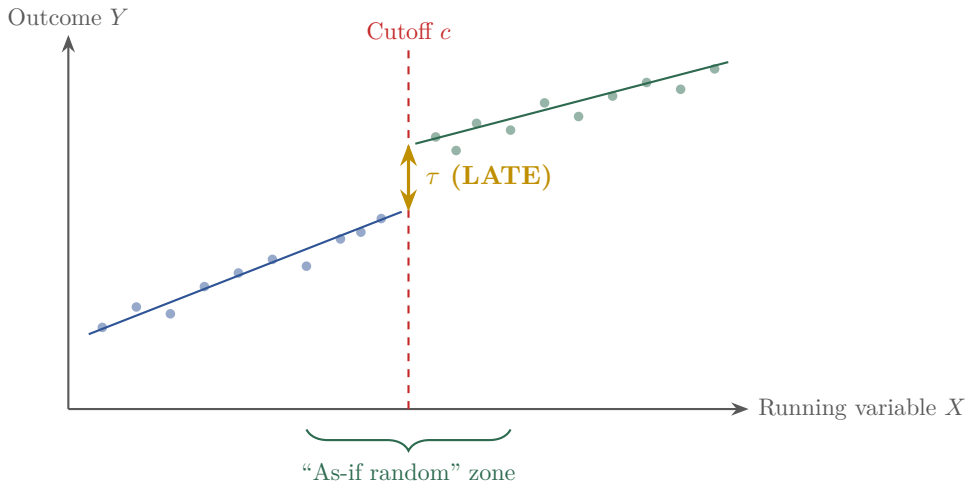
DID uses temporal variation combined with a control group. The key insight: even if treated and control groups differ in levels (due to U), the *change* over time is comparable under parallel trends.



How it works: DID eliminates time-invariant confounders (U) by differencing within units over time, and eliminates common time shocks by differencing between treated and control groups. The "double difference" isolates the treatment effect δ.

6.4 Regression Discontinuity: The Threshold

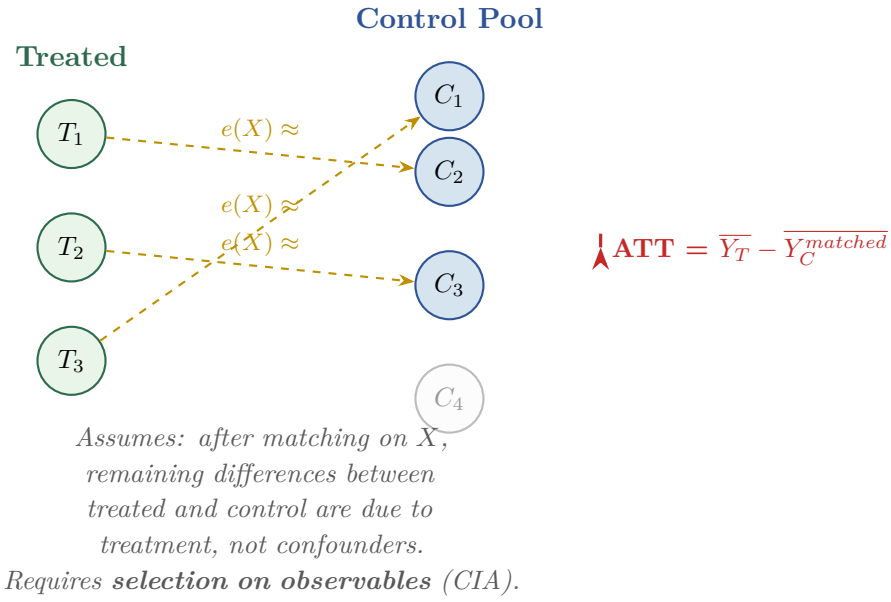
RD exploits a sharp threshold in treatment assignment. Near the cutoff, units just above and just below are nearly identical, creating a local randomized experiment.



How it works: At the cutoff, treatment assignment is determined by an arbitrary threshold. Units marginally above and below the cutoff are comparable on all observable and unobservable characteristics. The jump in the outcome at the cutoff identifies the causal effect.

6.5 Matching: Creating Statistical Twins

Matching constructs a counterfactual by finding untreated units with similar observable characteristics to treated units.



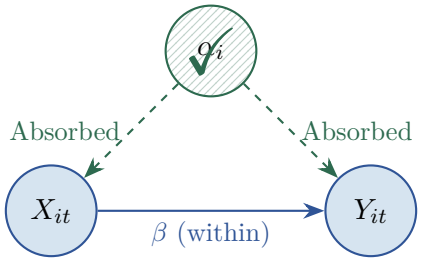
How it works: By pairing treated units with observably similar control units, matching removes confounding due to observed covariates. The propensity score $e(X) = P(D = 1|X)$ summarizes all covariates into a single dimension for matching.

Matching Does NOT Address Unobserved Confounders

If treatment selection depends on unobservable factors (motivation, ability, private information), matching provides biased estimates. Always report E-values and Rosenbaum bounds to quantify sensitivity to hidden bias.

6.6 Fixed Effects: Each Unit as Its Own Control

FE eliminates time-invariant confounders by using within-unit variation only. The unobserved confounder U is "absorbed" as long as it does not change over time.



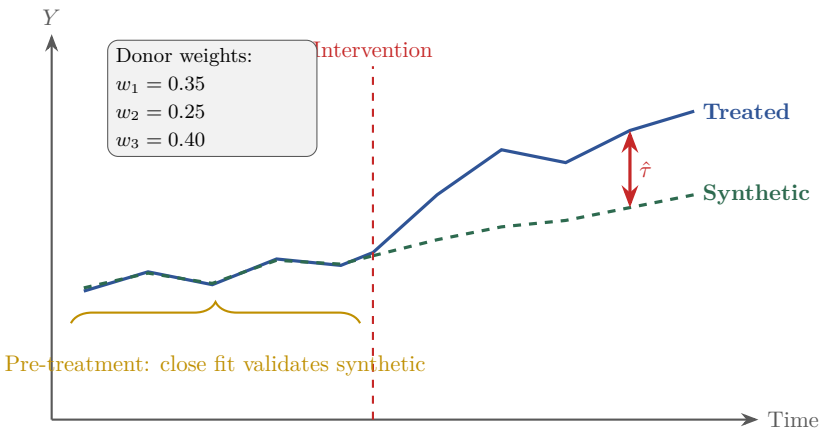
Within transformation: $Y_{it} - \bar{Y}_i = \beta(X_{it} - \bar{X}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$
Removes all time-invariant confounders α_i.

Limitation: Cannot eliminate time-varying confounders U_{it}.

How it works: The within transformation subtracts each unit's time-mean, removing any variable that is constant within a unit (e.g., firm culture, geographic location, inherent ability). Identification comes only from variation *within* units over time.

6.7 Synthetic Control: Building a Virtual Counterfactual

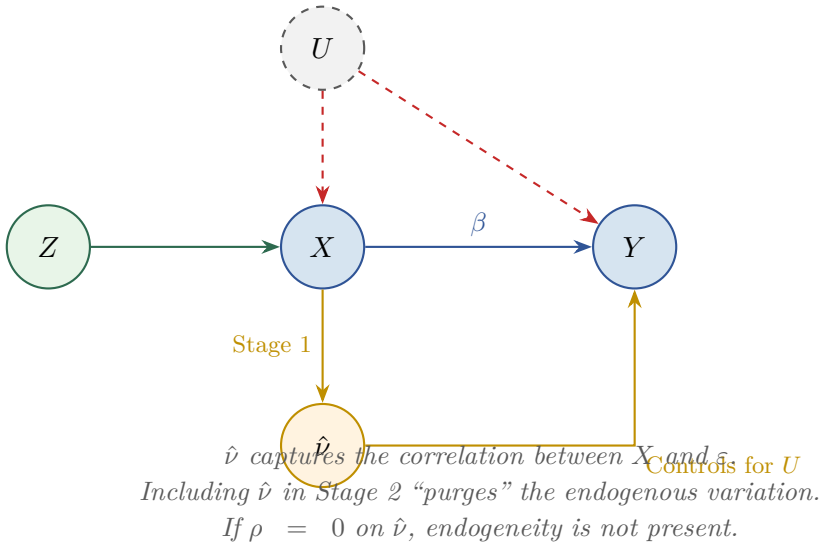
SCM constructs a weighted combination of control units to match the treated unit's pre-treatment trajectory, creating a "synthetic" version of the treated unit.



How it works: By finding the weighted combination of control units that best reproduces the treated unit's pre-treatment outcome path, SCM constructs a data-driven counterfactual. The gap between treated and synthetic post-treatment is the estimated effect.

6.8 Control Function: Extracting the Contamination

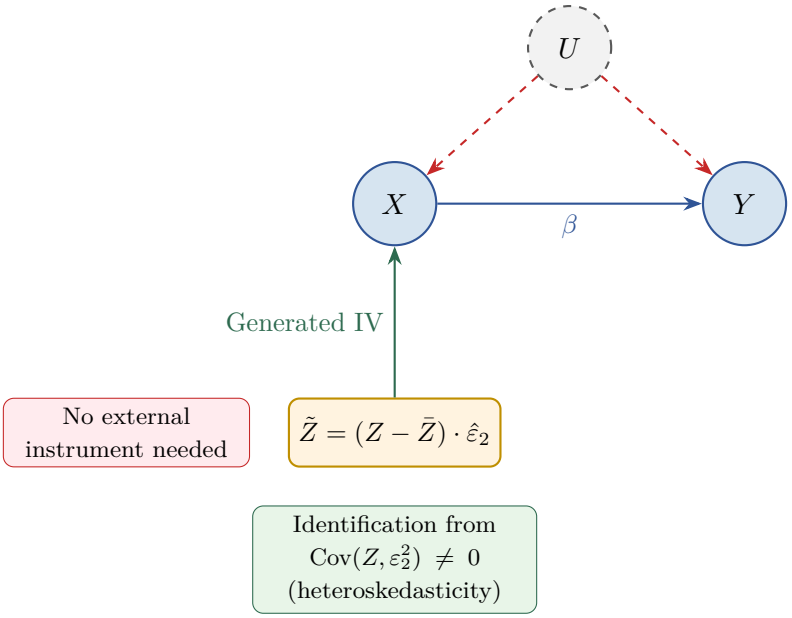
The control function approach explicitly models the endogenous component and "controls" for it in the second stage.



How it works: Stage 1 decomposes X into an exogenous part (driven by Z) and an endogenous part ($\hat{\nu}$). Stage 2 includes $\hat{\nu}$ as a control variable, absorbing the endogenous contamination. This is the correct approach for nonlinear models where 2SLS fails.

6.9 Lewbel HBIV: Internal Instruments from Heteroskedasticity

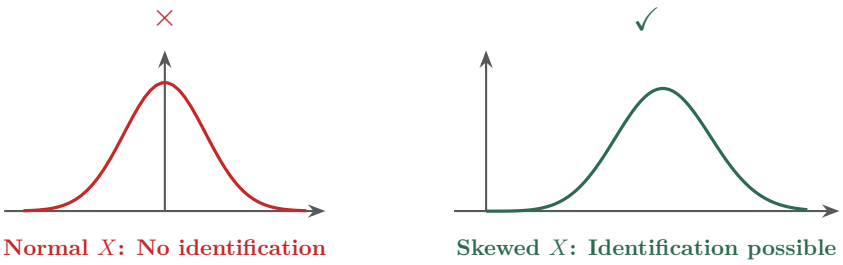
When no external instrument exists, Lewbel exploits heteroskedasticity in the first-stage residuals to construct instruments from within the data.



How it works: The key insight is that when the first-stage residuals exhibit heteroskedasticity (their variance changes with Z), this variation pattern provides identifying information. The generated instruments $\tilde{Z} = (Z - \bar{Z}) \cdot \hat{\varepsilon}_2$ are valid under the assumption that $\text{Cov}(Z, \varepsilon_1 \varepsilon_2) = 0$.

6.10 Gaussian Copula: Non-Normality as Identification

When the endogenous regressor has a non-normal distribution, this distributional feature provides information for identification without external instruments.



How it works: The copula correction term $q_i = \Phi^{-1}(\hat{F}(X_i))$ captures the non-normal component of X's distribution. When included as an additional regressor, it controls for the endogenous variation, similar to a control function. The method *requires* non-normality of X; with normal X, identification fails.

6.11 Comprehensive Comparison: What Each Method Eliminates

Method	Time-Invariant U	Time-Varying U	Reverse Causality	Meas. Error
IV	✓	✓	✓	✓
DID	✓	~	✓	×
RD	✓	✓	✓	×
Matching	×	×	×	×
FE	✓	×	×	×
SCM	✓	~	~	×
CF	✓	✓	✓	×
Lewbel	✓	~	✓	×
Copula	~	~	×	×

✓ = addresses ~ = partially addresses × = does not address

7 Extended Case Studies from UTD24/FT50 Journals

This section provides additional detailed case studies organized by method, drawing from the highest-impact management journals.

7.1 Instrumental Variables in Healthcare Operations

Kuntz et al. (2015, *MSOM*): Emergency Department Crowding

Question: How does ED crowding affect patient outcomes?
Endogeneity: Hospitals with worse outcomes may have more crowding (sick patients stay longer), and crowding itself may worsen outcomes.
Instrument: Ambulance diversion status of neighboring hospitals as an exogenous demand shifter. When nearby hospitals divert ambulances, patient volume at the focal hospital increases for reasons unrelated to its own quality.
Diagnostics: First-stage $F > 20$, Hansen J for over-identification, Hausman test confirms OLS is biased downward.
Key Finding: ED crowding significantly increases mortality risk. OLS underestimates the effect due to reverse causality (sicker patients cause both crowding and worse outcomes).
Lesson: In healthcare operations, neighboring facility behavior (diversion, closures, policy changes) provides excellent instruments because it shifts demand exogenously.

Freeman et al. (2021, *Management Science*): Physician Workload

Question: How does physician workload affect diagnostic quality in radiology?
Endogeneity: More experienced physicians may handle higher workloads and also make better diagnoses.
Instrument: Exogenous variation in daily case arrivals driven by scheduling patterns and day-of-week effects.
Key Finding: Higher workload increases diagnostic errors, with fatigue effects accumulating over the shift.
Lesson: Scheduling-driven variation in workload provides clean instruments in professional service settings where task assignment follows operational rules rather than quality-based selection.

7.2 DID in Platform and Digital Operations

Huang et al. (2019, *Management Science*): Platform Regulation

Question: How does platform regulation affect seller behavior on e-commerce platforms?
Design: DID exploiting policy changes that affected some product categories but not others.
Diagnostics: Pre-trend tests with event study plots, placebo tests using unaffected categories, Goodman-Bacon decomposition showing clean comparisons dominate.
Key Finding: Stricter regulation reduces fraudulent behavior but also decreases legitimate seller participation, illustrating the compliance-exit tradeoff.
Lesson: Category-level policy variation within platforms provides natural DID designs. The key challenge is controlling for category-specific trends that may violate parallel trends.

Grennan & Town (2020, *Management Science*): Price Transparency

Question: Does price transparency in medical devices reduce prices?
Design: DID using staggered adoption of a benchmarking tool across hospitals.
Key Finding: Access to pricing data reduced purchase prices by 3.6%, with larger effects for hospitals that had been paying above-market prices.
Lesson: Information interventions (dashboards, benchmarks, report cards) provide staggered DID designs. The modern DID literature (Callaway & Sant'Anna) is critical when adoption timing varies.

7.3 RD in Operations and Policy

Bapna et al. (2016, *Information Systems Research*): Online Education

Question: Do online course completions lead to career outcomes?
Design: Fuzzy RD using course grade thresholds for certification.
Diagnostics: McCrary density test (no manipulation of grades), covariate balance, bandwidth sensitivity across 50–200% of optimal.
Key Finding: Earning a course certificate significantly increases the probability of career advancement.
Lesson: Grade/score thresholds in education, certification, and compliance settings provide natural RD designs. Fuzzy RD (where threshold affects probability of treatment, not certainty) is handled as IV with the threshold as instrument.

7.4 Matching in Supply Chain Management

Jacobs & Singhal (2017, *JOM*): Supply Chain Disruptions

Question: Do supply chain disruptions affect long-term firm performance?
Design: PSM matching disrupted firms to comparable non-disrupted firms on pre-disruption characteristics (size, industry, profitability, growth).
Diagnostics: Standardized mean differences all < 0.10 after matching, common support verified, Rosenbaum bounds ($\Gamma = 1.8$) indicating moderate sensitivity to hidden bias.
Key Finding: Disruptions reduce ROA by 33–40% in the disruption year, with partial recovery over 2 years.
Lesson: Event-driven matching (disruptions, recalls, scandals) is common in OM. The key is matching on sufficient pre-event characteristics and reporting sensitivity analysis (Rosenbaum bounds, E-values).

Hendricks & Singhal (2005, *POM*): Supply Chain Glitches

Question: What is the stock market reaction to supply chain disruptions?
Design: Event study combined with matching. Firms experiencing disruptions are matched to comparable firms using propensity scores on size, industry, and pre-event performance.
Key Finding: Supply chain glitches reduce shareholder value by 10.28% over two years.
Lesson: Combining event study methodology with matching strengthens causal claims by providing a within-industry, within-size comparison group.

7.5 Synthetic Control in Policy Evaluation

Abadie (2021, *JEL*): Comprehensive Review

Contribution: The definitive review of synthetic control methodology. Covers: (1) original SCM, (2) inference (permutation, conformal), (3) extensions (multiple treated units, staggered adoption), (4) connections to DID and matching.
Key recommendations: Pre-treatment fit quality determines credibility; transparent donor weights aid interpretation; in-space placebos are mandatory for inference; augmented SCM methods (SDID, SCPI) improve performance when standard SCM fit is poor.
Lesson: Essential reading before implementing any SCM design. The review establishes best practices that reviewers expect.

8 Reporting Standards Checklist

Use this checklist before submitting any empirical paper involving causal inference. Based on Lu et al. (2018), Roth et al. (2023), and current best practices in UTD24 journals.

8.1 Universal Requirements (All Methods)

Key Concepts

- ☐ State the causal question explicitly in the introduction.
- ☐ Identify potential sources of endogeneity (OVB, reverse causality, measurement error).
- ☐ Justify the chosen identification strategy and explain why its assumptions are credible.
- ☐ Report **at least two** sensitivity analyses: Oster (2019) δ, Cinelli-Hazlett (2020) RV, or E-values.
- ☐ Discuss threats to identification and how you address them.
- ☐ Report both economic and statistical significance.
- ☐ Provide at least one alternative estimator as robustness.

8.2 Method-Specific Checklists

IV Reporting Checklist

- ☐ Theoretical justification for exclusion restriction (not just "we argue").
- ☐ First-stage regression reported (coefficient, SE, F-statistic, R^2).
- ☐ Kleibergen-Paap rk Wald F (weak ID) with Stock-Yogo critical values.
- ☐ Kleibergen-Paap rk LM (under-ID).
- ☐ Hansen J-test if over-identified.
- ☐ Durbin-Wu-Hausman endogeneity test.
- ☐ LIML comparison (must be similar to 2SLS).
- ☐ Anderson-Rubin confidence sets if $F < 23$.
- ☐ Falsification test: instrument does not predict outcome in sample where mechanism should not operate.

DID Reporting Checklist

- ☐ Event study plot with pre-treatment coefficients.
- ☐ Formal pre-trends test with power analysis (Roth, 2022).
- ☐ If staggered: Goodman-Bacon decomposition or Callaway-Sant'Anna estimates.
- ☐ Placebo test with fake treatment dates.
- ☐ Rambachan-Roth (2023) sensitivity bounds.
- ☐ Cluster standard errors at appropriate level.
- ☐ Discussion of anticipation effects.

RD Reporting Checklist

- ☐ McCrary density test (no manipulation).
- ☐ Covariate balance at cutoff.
- ☐ Main estimate with bias-corrected robust CI (`rdrobust`).
- ☐ Bandwidth sensitivity (50%, 75%, 125%, 150%, 200%).
- ☐ Polynomial order sensitivity ($p = 1, 2$).
- ☐ Placebo cutoff tests.
- ☐ Scatter plot of raw data with fitted lines.

Matching Reporting Checklist

- ☐ Propensity score model specification and justification.
- ☐ Standardized mean differences table (all < 0.10).
- ☐ Variance ratios (all 0.8–1.25).
- ☐ Common support plot.
- ☐ E-values and/or Rosenbaum bounds.
- ☐ Doubly robust (AIPW) as robustness check.
- ☐ Sensitivity to caliper width, number of matches, matching algorithm.

Appendix B: R Code Reference

Instrumental Variables

```
library(ivreg); library(lmtest); library(sandwich)

# 2SLS
m_iv <- ivreg(Y ~ EndogVar + X1 + X2 | IV1 + IV2 + X1 + X2,
              data = df)
summary(m_iv, diagnostics = TRUE)  # Weak IV, Wu-Hausman, Sargan

# Robust SEs
coeftest(m_iv, vcov = vcovHC(m_iv, type = "HC1"))
```

Difference-in-Differences

```
library(did); library(fixest)

# Callaway-Sant'Anna
cs <- att_gt(yname = "Y", tname = "year", idname = "unit",
             gname = "first_treat", data = df)
summary(cs)
ggdid(cs)  # Event study plot

# Sun-Abraham via fixest
sa <- feols(Y ~ sunab(first_treat, year) | unit + year, data = df)
iplot(sa)
```

Regression Discontinuity

```
library(rdrobust); library(rddensity)
```

```
# Main estimate
rd <- rdrobust(Y, X, c = 0)
summary(rd)

# McCrary test
dens <- rddensity(X, c = 0)
summary(dens)
rdplotdensity(dens, X)
```

Matching and AIPW

```
library(MatchIt); library(cobalt)

# Nearest neighbor with caliper
m <- matchit(Treat ~ X1 + X2 + X3, data = df,
             method = "nearest", caliper = 0.1)
love.plot(m, threshold = 0.1)  # Balance plot

# AIPW (doubly robust)
library(AIPW)
aipw <- AIPW$new(Y = df$Y, A = df$Treat,
                 W = df[, c("X1","X2","X3")])
aipw$fit()$summary()
```

Synthetic Control

```
library(Synth); library(synthdid)

# Synthetic DID
sdid_est <- synthdid_estimate(Y_matrix, N0, T0)
print(sdid_est)
plot(sdid_est)
synthdid_placebo_plot(sdid_est)
```

Double ML

```
library(DoubleML); library(mlr3); library(mlr3learners)

# Partially linear model
dml <- DoubleMLPLR$new(
  data = DoubleMLData$new(df, y_col="Y", d_cols="Treat",
                          x_cols=c("X1","X2","X3")),
  ml_l = lrn("regr.ranger"),
  ml_m = lrn("classif.ranger"),
```

```
  n_folds = 5
)
dml$fit()
dml$summary()
```

Sensitivity Analysis

```
library(sensemakr)

# Cinelli-Hazlett (2020)
m_ols <- lm(Y ~ Treat + X1 + X2 + X3, data = df)
sens <- sensemakr(m_ols, treatment = "Treat",
                  benchmark_covariates = "X1", kd = 1:3)
summary(sens)
plot(sens)

# E-value
library(EValue)
evalues.OLS(est = 0.5, se = 0.1, sd = 1, delta = 1)
```

Appendix C: Stata Quick Reference (Expanded)

Appendix D: Glossary of Key Terms

ATE Average Treatment Effect. The expected difference in outcomes between treated and untreated populations: $E[Y(1) - Y(0)]$.

ATT Average Treatment Effect on the Treated. The expected treatment effect among those who actually received treatment: $E[Y(1) - Y(0)|D = 1]$.

LATE Local Average Treatment Effect. The treatment effect for "compliers" in an IV setting, i.e., those whose treatment status is changed by the instrument.

CIA Conditional Independence Assumption. $(Y_0, Y_1) \perp D|X$. Treatment assignment is independent of potential outcomes after conditioning on observables. Required for matching.

SUTVA Stable Unit Treatment Value Assumption. (1) No interference between units

Table 13: Complete Stata Command Reference by Method

Method	Primary Command	Diagnostic Commands
IV/2SLS	`ivreg2 Y X (D=Z), first robust`	`estat firststage; estat overid; estat endogenous; weakivtest`
LIML	`ivregress liml Y X (D=Z), r`	Compare with 2SLS estimates
DID (TWFE)	`reghdfe Y TP X, absorb(i t) cl(i)`	`bacondecomp Y TP, ddetail`
DID (modern)	`csdid Y X, ivar(i) time(t) gvar(g)`	`csdid_estat event; csdid_plot`
RD	`rdrobust Y R, c(0) bwselect(mserd)`	`rddensity R, c(0); rdbwselect Y R`
PSM	`teffects psmatch (Y) (D X), atet`	`tebalance summarize; tebalance density`
AIPW	`teffects aipw (Y X) (D X), atet`	Compare with PSM estimates
SCM	`synth Y X Y(t1) Y(t2), trunit trperiod`	In-space placebos (manual loop)
SDID	`sdid Y i t D, vce(jackknife) graph`	Leave-one-out robustness
CF	Manual: `reg; predict; logit/probit`	Bootstrap entire procedure
Lewbel	`ivreg2h Y X (D=), robust`	`estat hettest` (Breusch-Pagan)
Dynamic GMM	`xtabond2 Y L.Y X, gmm() iv() two`	AR(2) test; Hansen J; instrument count
DML	`ddml init partial; ddml crossfit`	Cross-fit stability across folds
Sensitivity	`psacalc; sensemakr; evalue`	Report δ, RV, E-value

(one unit's treatment does not affect another's outcome), and (2) treatment is homogeneous across units.

Exclusion Restriction

The instrument Z affects the outcome Y only through the endogenous variable X. Formally: $\text{Cov}(Z, \varepsilon) = 0$.

Parallel Trends

In the absence of treatment, treated and control groups would have followed the same outcome trajectory. The identifying assumption for DID.

Running Variable

The continuous variable that determines treatment assignment in RD designs (e.g., test score, age, revenue threshold).

Propensity Score
$e(X) = P(D = 1|X)$. The probability of receiving treatment given observed covariates. Used in matching and weighting.

Control Function
A residual from a first-stage regression included in the second stage to absorb the endogenous component of a variable.

Heteroskedasticity
Unequal variance of the error term across observations. Exploited by Lewbel (2012) for identification.

Complier In IV terminology, a unit whose treatment status changes in response to the instrument. LATE estimates the effect for compliers only.

Neyman Orthogonality
A moment condition property ensuring that estimation of nuisance parameters does not affect inference on the parameter of interest. Required for valid DML.

Cross-Fitting Sample splitting procedure in DML where nuisance parameters are estimated on one subsample and causal parameters on another, preventing overfitting bias.

E-Value The minimum strength of association (on the risk ratio scale) that an unmeasured confounder would need with both treatment and outcome to fully explain away the observed effect.

Robustness Value (RV)
From Cinelli & Hazlett (2020): the minimum explanatory power (partial R^2) that an unobserved confounder must have with both treatment and outcome to reduce the estimated effect to zero.

Oster δ The degree of selection on unobservables relative to selection on observables that would be needed to drive the estimated effect to zero. $\delta > 1$ suggests robustness.

Negative Weights
In staggered DID with TWFE, some treatment effect estimates receive negative weights, meaning earlier-treated units serve as controls for later-treated units, biasing estimates.

Bartik Instrument
A shift-share instrument constructed as $B_{it} = \sum_k s_{ik,0} \cdot g_{kt}$, combining initial exposure shares with aggregate shocks.

AIPW Augmented Inverse Probability Weighting. A doubly robust estimator combining propensity score weighting with outcome regression.

Nickell Bias The $O(1/T)$ bias in FE estimation when a lagged dependent variable is included. Motivates GMM estimation for dynamic panels.

Key Concepts

Final Thought. Method selection should never be about "which statistic looks better." It must be driven by "which method's identifying assumptions are most credible in my research context." Every method has limitations. Transparent reporting of those limitations, combined with sensitivity analysis and robustness checks, is what distinguishes credible causal inference from correlation dressed up as causation. When in doubt, triangulate: multiple methods pointing to the same conclusion provide far stronger evidence than any single sophisticated estimator.