ENDOGENEITY AND CAUSAL INFERENCE METHODS

A Beginner's Handbook for Operations Management Researchers

Chenhao Zhou
Ph.D. Candidate, Supply Chain Management
Rutgers Business School

2025 Edition

Preface

Before diving into specific methods, I recommend beginning with these foundational papers. They shaped my understanding of causal inference in operations management and provide the essential framework for everything that follows.

Essential Starting Points

• Lu, Ding, Peng & Chuang (2018). Addressing Endogeneity in Operations Management Research. Journal of Operations Management, 64, 53–64. □□□□□

The definitive methodological guide for OM researchers. Provides comprehensive coverage of IV diagnostics with clear protocols for reporting standards.

• Ketokivi & McIntosh (2017). Addressing the Endogeneity Dilemma in Operations Management Research. Journal of Operations Management, 52, 1–14. □□□□□

Excellent conceptual treatment of why endogeneity arises and how to think about it philosophically.

• Ho, Lim, Reza & Xia (2017). Causal Inference Models in Operations Management. Manufacturing & Service Operations Management, 19(4), 509–525. □□□□□

Systematic review finding that 75% of empirical papers involve causal inference.

Method-Specific Guides

• Yilmaz, Son, Shang & Arslan (2024). Matching Methods and Synthetic Controls. JOM, 70(5), 831–859. □□□□□

• Petrin & Train (2010). Control Function Approach in Consumer Choice. Journal of Marketing Research, 47(1), 3–13. □□□□□

• Shang (2022). Endogeneity with Interaction Terms. JOM, 68(4), 339–358. □□□□□

Recommended Textbooks

• Angrist & Pischke (2009). Mostly Harmless Econometrics. Princeton. □□□□□

• Cunningham (2021). Causal Inference: The Mixtape. Yale. □□□□□

• Huntington-Klein (2021). The Effect. CRC Press. □□□□□

Acknowledgments

I am grateful to David Dreyfus for introducing me to empirical research and for his patience and guidance throughout my learning journey. His mentorship shaped my approach to thinking carefully about causal claims.

This handbook draws heavily on Lu, Ding, Peng & Chuang (2018)'s comprehensive review paper in the Journal of Operations Management, which provided an essential methodological framework for organizing this material.

This handbook was developed with the assistance of Claude (Anthropic) as a writing aid.

Chenhao Zhou

New Jersey, 2025

Contents

CHAPTER 1

Introduction: The Nature of Endogeneity

1.1 What is Endogeneity?

Endogeneity is the most fundamental challenge in empirical research. At its core, it represents a philosophical question: Does the statistical association we observe represent a genuine causal relationship, or is it merely a correlation driven by confounding factors?

According to Terwiesch et al. (2020)'s review of Manufacturing & Service Operations Management, 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%.

Consider the basic regression model:

$$Y = \alpha + \beta X + \varepsilon$$

When Cov(X, ε) ≠ 0, the OLS estimator is biased and inconsistent. The direction and magnitude of this bias depend on the correlation structure between the explanatory variable and the error term.

1.2 Three Sources of Endogeneity

Omitted Variable Bias

Classic Example: When studying the effect of education on income, an unobserved factor like "ability" simultaneously affects both years of schooling and income levels. More able individuals tend to obtain more education AND earn higher wages, regardless of their education.

$$E(\beta) = \beta_1 + \beta_2 \cdot Cov(X, Z) / Var(X)$$

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

Reverse Causality (Simultaneity)

Classic Example: Ketokivi & McIntosh (2017)'s restaurant seating allocation example demonstrates how 2SLS and OLS can produce coefficient signs with opposite directions

when reverse causality is present. Does advertising increase sales, or do higher sales lead to larger advertising budgets?

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

Measurement Error

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

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 at the same time.

$$Causal\ Effect = Y_1(Treated) - Y_0(Control)$$

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.

1.4 Framework for Method Selection

Ho, Lim, Reza & Xia (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?"

Research Context	Recommended Method	Key Assumption
Clear policy threshold exists	Regression Discontinuity	Local randomization near cutoff

Research Context	Recommended Method	Key Assumption
Policy change with control group	Difference-in-Differences	Parallel trends
External exogenous variation	Instrumental Variables	Exclusion restriction
Panel data with 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 with endogeneity	Control Function	Valid instruments + correct specification

CHAPTER 2

Causal Inference Methods

2.1 Instrumental Variables (IV)

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.

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.*

Mathematical Framework

$$\textit{First Stage: } X = \pi Z + \gamma W + \nu$$

$$\textit{Second Stage: } Y = \beta \hat{X} + \delta W + \varepsilon$$

$$\textit{IV Estimator: } \beta_{IV} = Cov(Y, Z) / Cov(X, Z)$$

Identification Conditions:
1. Relevance: Cov(Z, X) ≠ 0. Verified with first-stage F > 10 (Stock-Yogo); F > 23 for 5% maximal bias.
2. Exclusion Restriction: Cov(Z, ε) = 0. The instrument affects Y only through X. This cannot be tested directly and must be justified theoretically.

⚠ Exclusion Restriction Cannot Be Directly Tested

Lu et al. (2018)'s systematic review found common issues: (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.

IV Diagnostic Protocol

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

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.

Common Instruments in OM

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

2.2 Difference-in-Differences (DID)

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.

Intuition: Parallel Trains

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

Mathematical Framework

$$Y_{it} = \alpha + \beta \cdot Treat_i + \gamma \cdot Post_t + \delta \cdot (Treat_i \times Post_t) + X_{it}'\theta + \varepsilon_{it}$$

Where δ is the DID estimator representing the Average Treatment Effect on the Treated (ATT). The Parallel Trends Assumption requires that in absence of treatment, treated and control groups would have followed the same outcome trajectory.

⚠ Negative Weight Problem in Staggered DID

Goodman-Bacon (2021), Callaway & Sant'Anna (2021), Sun & Abraham (2021), and de Chaisemartin & D'Haultfoeuille (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. Very few OM papers adopt these new estimators—a major methodological gap.

Modern DID Estimators

Estimator	Key Feature	Stata Command
Goodman-Bacon (2021)	Decomposes TWFE weights	bacondecomp
Callaway-Sant'Anna (2021)	Group-time ATT aggregation	csdid
Sun-Abraham (2021)	Interaction-weighted estimator	eventstudyinteract
de Chaisemartin-D'H (2020)	Handles heterogeneous effects	did_multiplegt

★ Elenev, Quintero, Rebucci & Simeonova (2024) Management Science

Staggered DID studying COVID-19 stay-at-home orders with explicit attention to spillover effects and treatment heterogeneity. Event study plots with pre-trend tests and robustness to alternative estimators.

2.3 Regression Discontinuity (RD)

Research Context

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

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.*

Mathematical Framework

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

Where $D_i = 1$ if $X_i \geq c$, and τ is the Local Average Treatment Effect (LATE) at the cutoff. Sharp RD: treatment changes deterministically. Fuzzy RD: probability jumps at cutoff.

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.

RD Diagnostic Checklist

Test	Purpose	Implementation
McCrary Density	No manipulation of running variable	rddensity
Covariate Balance	Pre-determined covariates 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, p=2, p=3

★ Calvo, Cui & Serpa (2019) Management Science 65(12):5651–5675

Sharp RDD studying federal procurement oversight at the $150,000 simplified acquisition threshold. 262,857 projects, 71 agencies. Complete diagnostics: McCrary density, covariate balance, placebo tests at ±$25K, multi-bandwidth robustness. Found oversight increases delays by 6.1–13.8% and cost overruns by 1.4–1.6%. Gold standard for RD diagnostics in OM.

★ Flammer (2015) Management Science 61(11):2549–2568

Sharp RDD using CSR shareholder voting at the 50% majority threshold. Validated no manipulation via density test, confirmed covariate balance, tested bandwidth sensitivity. Found passing CSR proposals increases announcement returns by 1.77% and ROA by 0.7–0.8 percentage points.

2.4 Matching and Propensity Score Methods

Research Context

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

Intuition: Statistical Twins

> *You have apples and oranges and want to compare sweetness. But they differ in size, ripeness, and origin. The matching approach: find apples and oranges similar in size, color, and origin. By creating "statistical twins" differing only in treatment status, you approximate the counterfactual.*

Mathematical Framework

$$\textit{Propensity Score: } e(X) = P(D = 1 \mid X)$$

$$ATT = E[Y_1 - Y_0 \mid D = 1]$$

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

Balance Diagnostic Standards

Metric	Acceptable	Excellent	Notes
Standardized Mean Difference	< 0.25	< 0.10	Most common metric
Variance Ratio	0.5–2.0	0.8–1.25	Tests variance equality
Common Support	Visual inspection	Substantial overlap	Propensity score overlap

⚠ PSM Cannot Address Unobservable Confounders

The CIA assumes all confounders are observed—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.

Matching Method Variants

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

★ Yilmaz, Son, Shang & Arslan (2024) JOM 70(5):831–859

Most comprehensive methodological guide for matching and synthetic control in OM. Reviews 200+ papers (2010–2022). Provides diagnostic protocols, decision flowcharts, and Stata implementation. Essential reading before implementing any matching design.

2.5 Fixed Effects (FE)

Research Context

Research Question: Does changing managers improve employee productivity?
Core Challenge: Employees have inherently different abilities and work styles that don't change over time, confounding the manager effect.

Intuition: Each Person as Their Own Control

> *To know if coffee makes you more alert: only examine your own days when you drink coffee versus when you don't, comparing changes in your own alertness. Each person serves as their own control, eliminating all time-invariant individual differences.*

Mathematical Framework

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

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

Types of Fixed Effects

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

Fixed Effects Only Eliminate Time-Invariant Confounders

If time-varying confounders exist (e.g., motivation changes due to life events), FE is equally powerless. High-dimensional FE may over-control, absorbing variation needed for identification.

2.6 Synthetic Control Method (SCM)

Research Context

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

Intuition: Constructing a Virtual Control

> *Your friend started working out and you want to know if it's effective. Find several people with similar body types and habits, then proportionally weight them to "synthesize" a virtual friend matching your friend's pre-workout trajectory. The divergence after workout starts reveals the causal effect.*

Mathematical Framework

$$\text{Synthetic Control: } \hat{Y}_{1t} = \Sigma_j w_j Y_{jt} \text{ where } \Sigma w_j = 1,\ w_j \geq 0$$

$$\text{Treatment Effect: } \hat{\tau}_t = Y_{1t} - \hat{Y}_{1t}$$

SCM Diagnostics

Diagnostic	Purpose	What to Report
Pre-treatment RMSPE	Quality of pre-treatment fit	Small relative to outcome scale
Post/Pre RMSPE Ratio	Effect relative to fit	Large ratio suggests real effect
In-Space Placebo	Statistical inference	Permute treatment across donors
In-Time Placebo	Rule out spurious timing	Test fake treatment dates

⚠ SCM Requires Excellent Pre-Treatment Fit

Credibility depends entirely on pre-treatment fit. If synthetic control cannot track the treated unit before intervention, post-treatment divergence is uninterpretable. In-space placebo tests (permuting treatment across donors) are mandatory for inference.

★ Li & Shankar (2024) Management Science — Two-Step Synthetic Control

Methodological innovation: (1) Formal statistical test for parallel trends, replacing subjective visual inspection; (2) Allows weights to sum to values ≠ 100% when traditional SCM fails. Improves applicability when no convex combination of donors can match the treated unit.

2.7 Control Function Approach (CF)

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 like Logit or Probit.

Intuition: Extracting the Problematic Variation

> *Weighing apples, but stones got mixed in. Control function approach: first estimate the weight of stones (the endogenous part), then explicitly control for it in stage 2. The residual from stage 1 captures the correlation between endogenous variable and error term.*

Mathematical Framework: Two-Stage Residual Inclusion

$$\text{Stage 1: } X = \pi Z + \gamma W + \nu \rightarrow \text{Obtain residual } \hat{\nu}$$

$$\text{Stage 2: } Y = g(X, \beta) + \rho\hat{\nu} + \varepsilon^*$$

Including $\hat{\nu}$ in stage 2 controls for the endogenous component. The coefficient ρ provides a built-in endogeneity test: if $\rho \neq 0$, endogeneity is present.

□ 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 and provides built-in endogeneity test through t-test on residual coefficient.

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 ≥500 replications, or (2) Murphy-Topel analytical correction.

★ Petrin & Train (2010) Journal of Marketing Research 47(1):3–13

Foundational paper establishing CF standards for discrete choice. Key insight: 2SLS is inconsistent in nonlinear models; CF maintains consistency. Applied to cable TV demand: without correction, demand appears upward-sloping; with CF, properly downward-sloping. Cited 750+ times. Essential reading.

2.8 Lewbel Method (Heteroskedasticity-Based Identification)

Research Context

Research Question: How do team interactions affect open-source project performance?
Extreme Challenge: No convincing external instruments exist. The treatment is deeply embedded in the social process being studied.

Intuition: Exploiting Natural Variation Patterns

> *At a noisy party, if noise is equally loud everywhere (homoskedasticity), it's hard to hear your friend. But if noise varies—loud near speakers, quiet in corners (heteroskedasticity)—you can identify your friend's voice by comparing quiet vs loud areas. Lewbel exploits this variation pattern as an internal instrument.*

Mathematical Framework

$$\textit{Step 1: Estimate } Y_2 = \gamma' Z + \varepsilon_2 \rightarrow \textit{Obtain } \hat{\varepsilon}_2$$

$$\textit{Step 2: Construct instrument } \tilde{Z} = (Z - \bar{Z}) \cdot \hat{\varepsilon}_2$$

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

Identification Conditions:
A1: Cov(Z, $\varepsilon_1\varepsilon_2$) = 0 — Cannot be tested
A2: Cov(Z, ε_2^2) ≠ 0 — Use Breusch-Pagan test to verify heteroskedasticity

⚠ Use Only as Robustness Check

Only 1 Lewbel paper found in core OM journals (2018–2024). This reflects appropriate caution: (1) A1 cannot be tested, (2) OM typically has better quasi-experimental designs. Baum & Lewbel (2019) note: "External instruments should almost always be preferred." Use Lewbel as robustness check alongside traditional IV, not as primary identification.

Lewbel Diagnostics

Test	Purpose	Threshold
Breusch-Pagan Test	Verify heteroskedasticity (A2)	Reject at $p < 0.05$
Pagan-Hall Test	Over-identification	Fail to reject at $p > 0.10$

Test	Purpose	Threshold
First-Stage F	Instrument strength	$F > 10$
Compare External IV	Robustness check	Similar point estimates

★ Pal, Zuo & Nair (2024) JOM 70(7):1076–1099

"Collaborative Dynamics in Open Source Software"—the only Lewbel application in core OM journals (2018–2024). GitHub data: 100M+ developers. Diagnostics: Breusch-Pagan = 50,992 ($p < 0.05$); Pagan-Hall = 2.87 ($p > 0.1$); first-stage F exceeds thresholds. Exemplifies proper justification when external IVs are genuinely unavailable.

CHAPTER 3

Summary and Testing Guide

3.1 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 change
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, many controls
CF	High	Medium-High	Medium-High	Valid instruments + nonlinear
Lewbel	Medium	Low	Very High	Heteroskedasticity

3.2 Diagnostic Tests Quick Reference

Method	Essential Tests	Robustness Checks	Stata
IV	F > 10, Hansen J, K-P LM	LIML, GMM, different IVs	ivreg2
DID	Parallel trends, Event study	Placebo periods	csdid
RD	McCrary, covariate balance	Bandwidth, polynomial	rdrobust
PSM	Balance (SMD < 0.1)	Rosenbaum bounds	teffects
SCM	RMSPE, pre-fit	In-space/time placebos	synth

Method	Essential Tests	Robustness Checks	Stata
CF	First-stage F, residual t	Bootstrap SE	manual
Lewbel	Breusch-Pagan, Pagan-Hall	Compare external IV	ivreg2h

3.3 Current Opportunities

Modern DID Estimators: Callaway-Sant'Anna, Goodman-Bacon, and related estimators are almost never adopted in OM journals despite prevalence of staggered adoption designs. Researchers who adopt these methods early will differentiate their work.

Sensitivity Analysis in Matching: Rosenbaum bounds and E-values should become standard but are rarely seen. Doubly robust estimators (AIPW) are also underutilized.

Machine Learning for Causal Inference: Causal Forests, Double ML, and LASSO-based instrument selection are gaining traction in economics but remain rare in OM.

Method selection should not be about "which statistic looks better," but rather "which method's identifying assumptions are most credible in my research context."

Appendix A: Core References

Methodological Guides

- Lu, Ding, Peng & Chuang (2018). JOM 64, 53–64. □□□□□
- Ketokivi & McIntosh (2017). JOM 52, 1–14. □□□□□
- Ho, Lim, Reza & Xia (2017). MSOM 19(4), 509–525. □□□□□
- Yilmaz, Son, Shang & Arslan (2024). JOM 70(5), 831–859. □□□□□

Instrumental Variables

- Stock & Yogo (2005). Testing for Weak Instruments. □□□□□
- Angrist & Krueger (1991). QJE—Compulsory Schooling. □□□□□

Difference-in-Differences

- Goodman-Bacon (2021). Journal of Econometrics 225(2), 254–277. □□□□□
- Callaway & Sant'Anna (2021). Journal of Econometrics 225(2), 200–230. □□□□□

Regression Discontinuity

- Lee & Lemieux (2010). Journal of Economic Literature 48(2). □□□□□
- Calvo, Cui & Serpa (2019). Management Science 65(12), 5651–5675. □□□□□

Matching Methods

- Rosenbaum & Rubin (1983). Biometrika 70(1). □□□□□
- King & Nielsen (2019). Political Analysis 27(4). □□□□

Synthetic Control

- Abadie, Diamond & Hainmueller (2010). JASA. □□□□□

Control Function

- Petrin & Train (2010). JMR 47(1), 3–13. □□□□□
- Terza, Basu & Rathouz (2008). Journal of Health Economics 27(3). □□□□□

Lewbel Method

- Lewbel (2012). JBES 30(1). □□□□□
- Pal, Zuo & Nair (2024). JOM 70(7), 1076–1099. □□□□□

Appendix B: Stata Commands

Instrumental Variables

```
* Basic 2SLS
ivregress 2sls Y X1 X2 (EndogVar = IV1 IV2), first robust

* Comprehensive diagnostics
ivreg2 Y X1 X2 (EndogVar = IV1 IV2), first robust

* Post-estimation
estat firststage
estat overid
estat endogenous
```

Difference-in-Differences

```
* Traditional TWFE
reghdfe Y TreatPost X1 X2, absorb(unit_id year) cluster(unit_id)

* Callaway-Sant'Anna
csdid Y X1 X2, ivar(unit_id) time(year) gvar(first_treat)
csdid_estat event

* Goodman-Bacon decomposition
bacondecomp Y TreatPost, ddetail
```

Regression Discontinuity

```
* Basic Sharp RD
rdrobust Y RunningVar, c(0)

* McCrary density test
rddensity RunningVar, c(0)
```

Propensity Score Matching

```
* PSM with teffects
teffects psmatch (Y) (Treatment X1 X2 X3), atet nn(1)
tebalance summarize

* Alternative: psmatch2
psmatch2 Treatment X1 X2 X3, outcome(Y) neighbor(1) caliper(0.01)
pstest X1 X2 X3, both
```

Synthetic Control

```
synth Y X1 X2 Y(1990) Y(1991) Y(1992), trunit(1) trperiod(1993) fig
```

Lewbel Method

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

* Test for heteroskedasticity
```

```
regress EndogVar X1 X2
estat hettest
```

— End of Handbook —