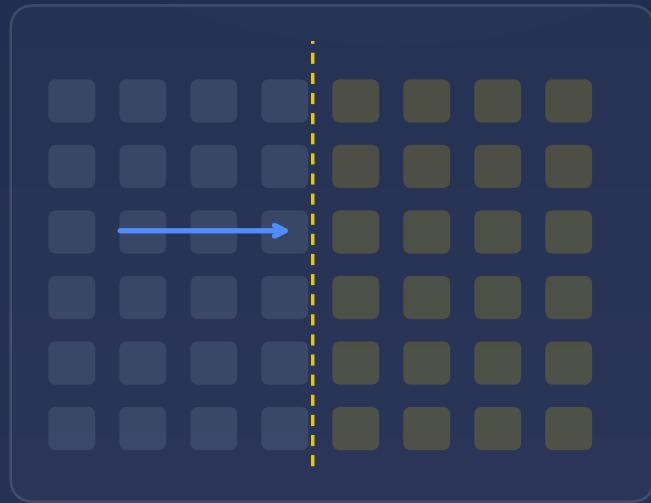


# Causal Inference in Marketing

Panel Data and Machine Learning Methods

---



**Charles Shaw**

Community Review Edition

2026

Charles Shaw

# Causal Inference in Marketing

Panel Data and Machine Learning

December 21, 2025

“We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.”

— Aristotle, *Physics II.3, 194b17–20*

“Behind every causal conclusion there must lie some causal assumption that is not testable in observational studies.”

— Judea Pearl, *Causality (2009)*



# Preface

Marketing measurement now operates under a familiar tension. Firms devote substantial resources to advertising, promotions, and customer acquisition, yet the tools used to evaluate these investments are often assembled from distinct traditions that do not share a common causal language.

On one side is a pragmatic body of practice: marketing mix models, ad-stock constructions, attribution rules, and forecasting pipelines developed to function under operational constraints. On the other is a modern identification-focused literature in causal inference and panel econometrics: difference-in-differences, synthetic control, interactive fixed effects, and related designs that emphasise explicit counterfactual reasoning.

Neither tradition, on its own, is fully adequate for the settings that dominate contemporary marketing. Purely predictive models can fit historical variation while remaining ambiguous about causal effect, especially when campaigns overlap, spillovers are plausible, and platform policies shift. Conversely, methods derived for relatively clean quasi-experiments can be fragile when panels are short, outcomes are noisy, adoption is endogenous, and the data-generating process evolves with product cycles and competitive response.

There is, of course, no shortage of rigorous foundations. Classic texts in econometrics and panel data—such as Baltagi’s *Econometric Analysis of Panel Data*, Hayashi’s *Econometrics*, and Wooldridge’s treatments of cross-sectional and panel methods—provide the theoretical and methodological backbone for much of what follows. What is less common is a practitioner-oriented synthesis that begins from the institutional realities of marketing measurement and carries the reader from design to estimation, diagnostics, and decision-relevant reporting in the settings that arise in firms.

Similarly, modern introductions to causal inference—including Angrist and Pischke’s *Mostly Harmless Econometrics* and Cunningham’s *Causal Inference: The Mixtape*—have done a great deal to make identification thinking accessible to applied researchers. Their centre of gravity, however, is not marketing measurement with short and noisy panels, platform-mediated assignment, interference, and dynamic responses. The aim here is therefore complementary: to keep the clarity of design-based reasoning, while adapting the workflow to the data structures, operational constraints, and failure modes that are characteristic of marketing practice.

This book is an attempt to make the trade-off less stark. We develop a unified framework, grounded in the potential-outcomes tradition and disciplined panel design, that is explicit about what is being estimated, under which assignment mechanism, and with what diagnostic support. The methods we emphasise

reflect a broader convergence: identification arguments remain the organising principle, but estimation must accommodate the high dimensionality and operational constraints of commercial data.

The book is written for readers who increasingly face the same practical questions from different starting points. Practitioners and data scientists need methods that are implementable and resilient to messy panels while remaining clear about what is and is not identified. Econometricians and methodologically minded analysts need a translation of familiar design logic into the language of incrementality, budget allocation, and measurement under interference and dynamics.

The exposition is deliberately design-first without being abstract. We begin with assignment mechanisms and estimands and only then introduce estimators. Assumptions are stated explicitly, diagnostics are tied to specific identification threats, and implementation choices are treated as part of the inferential argument rather than as engineering afterthoughts. Where appropriate, code and worked examples accompany formal development. What this book does not attempt is a catalogue of every industrial heuristic or ad-tech algorithm. Instead, it focuses on a core set of designs and tools that can support credible causal statements in marketing panels, and on the judgement required to recognise when the data do not.

Charles Shaw  
London, 2025

# Contents

Preface	vii
<b>Part I Foundations</b>	<b>1</b>
<b>1 Why Panel Data for Marketing? Motivation, Methods, and the Causal Revolution</b>	<b>3</b>
1.1 The Marketing Measurement Crisis . . . . .	4
1.2 What Are Panel Data Methods? A Conceptual Framework . . . . .	8
1.3 Panel Data in the Marketing Analytics Ecosystem . . . . .	11
1.4 Motivating Examples: Three Marketing Challenges . . . . .	16
1.5 The Two Cultures of Marketing Analytics . . . . .	20
1.6 Strategic Dynamics and Marketing Phenomena . . . . .	22
1.7 The Causal Revolution in Marketing . . . . .	24
1.8 Why Marketing Panel Data is Different . . . . .	27
1.9 Roadmap of the Book . . . . .	30
<b>2 Causal Frameworks and Panel Notation</b>	<b>31</b>
2.1 Potential Outcomes for Panels . . . . .	32
2.2 Panel Data Structures and Indexing . . . . .	36
2.3 Core Estimands for Panel Causality . . . . .	41
2.4 Assignment Mechanisms and Identification . . . . .	46
2.5 Regression Mechanics and Inference in Panels . . . . .	53
2.6 Crosswalk to Method Chapters . . . . .	59
2.7 Running Examples: Motivating Vignettes . . . . .	62
2.8 Notation and Assumptions Reference . . . . .	66
<b>3 Design-Based Thinking for Panels</b>	<b>69</b>
3.1 Experimental and Quasi-Experimental Regimes . . . . .	70
3.2 Assignment Mechanisms in Panels and Target Estimands . . . . .	73
3.3 Geo-Experiments and Clustered Designs . . . . .	79
3.4 Switchbacks and Platform Experiments . . . . .	83
3.5 Phased Rollouts and Staggered Adoption . . . . .	86

3.6	Ex Ante Diagnostics and Placebos . . . . .	89
3.7	Power, Minimum Detectable Effects, and Serial Dependence . . . . .	93
3.8	Threats to Validity and Design Adaptation . . . . .	97
3.9	Reporting Standards and Pre-Analysis Plans . . . . .	101
3.10	Method Selection Map . . . . .	104
3.11	Templates and Checklists . . . . .	108
3.12	Interlude: Design-First and Structural IO . . . . .	113
<b>Part II Differences-in-Differences and Event Studies</b>		<b>117</b>
<b>4</b>	<b>Difference-in-Differences: From Canonical to Staggered</b>	<b>119</b>
4.1	Canonical 2x2 DiD and Parallel Trends . . . . .	121
4.2	Causal Estimands for Staggered Adoption . . . . .	127
4.3	Identification with Staggered Timing . . . . .	134
4.4	Two-Way Fixed Effects and Its Pitfalls . . . . .	139
4.5	Modern Estimators for Staggered Designs . . . . .	143
4.6	Event-Study Specifications and Dynamics . . . . .	151
4.7	Inference . . . . .	156
4.8	Diagnostics and Design Considerations . . . . .	161
4.9	Marketing Applications and Patterns . . . . .	165
4.10	Workflow Checklist . . . . .	169
<b>5</b>	<b>Event-Study Designs</b>	<b>173</b>
5.1	Motivation and Setup . . . . .	174
5.2	Event-Time Estimands . . . . .	177
5.3	Specification: Leads and Lags . . . . .	181
5.4	Estimation under Staggered Adoption . . . . .	185
5.5	Identification: Assumptions and Design Implications . . . . .	190
5.6	Graphical Presentation and Interpretation . . . . .	195
5.7	Inference . . . . .	202
5.8	Diagnostics . . . . .	206
5.9	Extensions and Special Cases . . . . .	210
5.10	Event-Time Metrics for Marketing Decisions . . . . .	215
5.11	Workflow Checklist . . . . .	221
<b>Part III Synthetic Controls and Hybrid Methods</b>		<b>225</b>
<b>6</b>	<b>Synthetic Control</b>	<b>227</b>
6.1	Motivation and Setup . . . . .	228
6.2	Constructing the Synthetic Control . . . . .	234

<b>CONTENTS</b>	<b>xi</b>
6.3 Identification and Assumptions . . . . .	240
6.4 Inference for Synthetic Control . . . . .	246
6.5 Diagnostics and Goodness of Fit . . . . .	252
6.6 Practical Issues in Marketing Panels . . . . .	261
6.7 Data Fusion for Cold-Start Problems . . . . .	267
6.8 Extensions and Variants . . . . .	273
6.9 Marketing Applications . . . . .	279
6.10 Workflow Checklist . . . . .	285
<b>7 Generalised and Augmented Synthetic Control</b>	<b>289</b>
7.1 Motivation and Setup . . . . .	290
7.2 Augmented Synthetic Control (ASCM) . . . . .	293
7.3 Regularised and Balancing Variants of SC . . . . .	298
7.4 Synthetic Difference-in-Differences (SDID) . . . . .	303
7.5 Triply Robust Panel (TROP) Estimators . . . . .	309
7.6 Identification and Assumptions . . . . .	315
7.7 Multiple Treated Units and Staggered Adoption . . . . .	320
7.8 Tuning, Implementation, and Donor Curation . . . . .	325
7.9 Diagnostics and Goodness of Fit . . . . .	330
7.10 Inference . . . . .	335
7.11 Marketing Applications . . . . .	340
7.12 Workflow Checklist . . . . .	344
<b>Part IV Factor Models and Matrix Methods</b>	<b>349</b>
<b>8 Interactive Fixed Effects and Matrix Completion</b>	<b>351</b>
8.1 Motivation and Setup . . . . .	352
8.2 Interactive Fixed Effects (IFE) . . . . .	356
8.3 Matrix Completion Perspective . . . . .	362
8.4 Identification, Assumptions, and Design Implications . . . . .	367
8.5 Connections to SC and SDID . . . . .	373
8.6 Tuning and Implementation . . . . .	377
8.7 Inference . . . . .	382
8.8 Diagnostics and Robustness . . . . .	389
8.9 Marketing Applications . . . . .	395
8.10 Workflow Checklist . . . . .	401
<b>9 Advanced Matrix Methods for Causal Inference</b>	<b>409</b>
9.1 Introduction: Beyond Standard Matrix Completion . . . . .	410
9.2 Tensor Completion for Multi-Way Panels . . . . .	412

9.3	Robust Matrix Completion with Outliers . . . . .	418
9.4	Matrix Completion with Side Information . . . . .	424
9.5	Time-Varying Rank and Non-Stationary Panels . . . . .	430
9.6	Bayesian Matrix Completion . . . . .	435
9.7	Computational Methods for Large-Scale Problems . . . . .	441
9.8	Connections and Comparisons . . . . .	446
9.9	Diagnostics and Validation . . . . .	451
9.10	Marketing Applications . . . . .	455
9.11	Practical Workflow and Software . . . . .	460
9.12	Conclusion and Future Directions . . . . .	464
<b>Part V Dynamics, Heterogeneity, and Spillovers</b>		<b>469</b>
<b>10</b>	<b>Dynamic Treatment Effects</b>	<b>471</b>
10.1	Motivation and Setup . . . . .	472
10.2	Potential Outcomes and Dynamic Estimands . . . . .	476
10.3	Identification . . . . .	481
10.4	Estimation Strategies . . . . .	487
10.5	Anticipation, Carryover, and Mediation . . . . .	493
10.6	Inference . . . . .	498
10.7	Diagnostics . . . . .	502
10.8	Marketing Applications . . . . .	505
10.9	Workflow Checklist . . . . .	510
<b>11</b>	<b>Interference and Spillovers</b>	<b>517</b>
11.1	The Challenge of Interference in Marketing . . . . .	518
11.2	Types of Spillovers in Marketing Panels . . . . .	521
11.3	Exposure Mappings and Potential Outcomes Under Interference . . . . .	526
11.4	Partial Interference and Cluster Designs . . . . .	530
11.5	Estimation Strategies for Direct and Spillover Effects . . . . .	535
11.6	Competition and Saturation Effects . . . . .	541
11.7	Diagnostics for Detecting Interference . . . . .	546
11.8	Marketing Applications . . . . .	550
11.9	Conclusion . . . . .	554
<b>Part VI Machine Learning and High-Dimensional Methods</b>		<b>557</b>
<b>12</b>	<b>Machine Learning for Nuisance and Heterogeneity</b>	<b>559</b>
12.1	Motivation and Setup . . . . .	560
12.2	Orthogonalisation and Neyman Orthogonality . . . . .	563

<b>CONTENTS</b>	<b>xiii</b>
12.3 Cross-Fitting under Panel Dependence . . . . .	566
12.4 Double/Debiased ML Estimators in Panels . . . . .	569
12.5 Heterogeneous Treatment Effects . . . . .	575
12.6 Policy Learning . . . . .	579
12.7 Assumptions . . . . .	582
12.8 Tuning and Implementation . . . . .	586
12.9 Dose-Response Extensions . . . . .	591
12.10 Diagnostics and Robustness . . . . .	594
12.11 Inference . . . . .	598
12.12 Marketing Applications . . . . .	602
12.13 Workflow Checklist . . . . .	606
<b>13 High-Dimensional Controls and Regularisation</b>	<b>611</b>
13.1 Motivation and Setup . . . . .	612
13.2 High-Dimensional Controls in Panels . . . . .	614
13.3 Regularisation Methods . . . . .	617
13.4 Variable Selection for Causal Targets . . . . .	622
13.5 Post-Selection and Debiased Inference . . . . .	626
13.6 Panels with DiD/Event-Study and Many Controls . . . . .	631
13.7 Assumptions . . . . .	635
13.8 Tuning and Implementation . . . . .	638
13.9 Diagnostics and Robustness . . . . .	643
13.10 Inference . . . . .	646
13.11 Marketing Applications . . . . .	649
13.12 Workflow Checklist . . . . .	653
<b>14 Continuous and Nonlinear Panel Models</b>	<b>659</b>
14.1 Motivation and Setup . . . . .	660
14.2 Dose-Response Estimands . . . . .	661
14.3 Identification . . . . .	663
14.4 Estimation Strategies for Continuous Treatments . . . . .	667
14.5 Nonlinear Panel Outcome Models . . . . .	671
14.6 Handling Excess Zeros: Hurdle and Zero-Inflated Models . . . . .	676
14.7 Duration Models for Takeoff in Panels . . . . .	683
14.8 Dynamics with Continuous Intensity . . . . .	686
14.9 Assumptions . . . . .	689
14.10 Tuning and Implementation . . . . .	692
14.11 Diagnostics and Design Considerations . . . . .	695
14.12 Inference . . . . .	697
14.13 Marketing Applications . . . . .	699

14.14 Workflow Checklist . . . . .	701
<b>Part VII Validity, Inference, and Diagnostics</b>	<b>705</b>
<b>15 Threats to Validity in Marketing Panels</b>	<b>707</b>
15.1 Motivation and Scope . . . . .	708
15.2 A Taxonomy of Bias Sources . . . . .	709
15.3 Parallel Trends Violations . . . . .	711
15.4 Omitted Variable Bias and Sensitivity . . . . .	713
15.5 Measurement Error . . . . .	716
15.6 Partial Identification . . . . .	718
15.7 Structural Breaks . . . . .	719
15.8 Spillovers and SUTVA Violations . . . . .	721
15.9 Small-Sample and Dependence Corrections . . . . .	723
15.10 Diagnostics: A Practical Playbook . . . . .	724
15.11 Sensitivity Analyses . . . . .	725
15.12 Marketing-Specific Checklists . . . . .	726
15.13 Assumptions and Failure Modes . . . . .	728
15.14 Workflow Checklist . . . . .	729
<b>16 Inference and Uncertainty Quantification</b>	<b>735</b>
16.1 Motivation and Scope . . . . .	736
16.2 Unified Variance Estimation . . . . .	737
16.3 Mosaic Permutation Tests . . . . .	739
16.4 Bootstrap and Resampling . . . . .	741
16.5 Randomisation Inference . . . . .	743
16.6 Conformal and Distribution-Free Methods . . . . .	745
16.7 Aggregated Estimands and Joint Inference . . . . .	747
16.8 Inference after Selection and ML . . . . .	750
16.9 Multiplicity and Multiple Testing . . . . .	752
16.10 Instrumental Variables in Marketing . . . . .	754
16.11 Weak Instrument Diagnostics . . . . .	759
<b>17 Design and Diagnostics</b>	<b>761</b>
17.1 Motivation and Scope . . . . .	762
17.2 Control Selection and Donor Curation . . . . .	763
17.3 Overlap and Balance Diagnostics . . . . .	765
17.4 Pre-Trends and Placebo Designs . . . . .	767
17.5 Support, Exposure, and Contamination . . . . .	769
17.6 Influence and Stability . . . . .	770

17.7 Specification Curves and the Multiverse . . . . .	772
17.8 Sensitivity Analyses . . . . .	774
17.9 Implementation Details and Tuning . . . . .	776
17.10 Inference for Diagnostics . . . . .	779
17.11 Marketing Applications . . . . .	782
17.12 Assumptions and Stability . . . . .	786
17.13 Workflow Checklist . . . . .	787
<b>Part VIII Applications and Future Directions</b>	<b>791</b>
<b>18 Applications in Marketing</b>	<b>793</b>
18.1 Marketing Causal Problems Taxonomy . . . . .	794
18.2 The Application Protocol . . . . .	798
18.3 Advertising Incrementality: Geo-Experiments . . . . .	801
18.4 Digital Attribution and Multi-Touch . . . . .	807
18.5 Media Mix Modelling with Causal Foundations . . . . .	812
18.6 Loyalty Programme Valuation . . . . .	821
18.7 Promotion and Price Elasticity . . . . .	826
18.8 Customer Lifetime Value and Acquisition . . . . .	831
18.9 Dynamic Pricing in Transport Networks . . . . .	836
18.10 Subscription and Paywall Effects . . . . .	842
18.11 Platforms and Two-Sided Markets . . . . .	848
18.12 Ranking and Recommendation Algorithms . . . . .	855
18.13 Marketplace Seller Interventions . . . . .	861
18.14 Method Selection Summary . . . . .	867
18.15 Synthesis and Reporting . . . . .	871
18.16 Common Pitfalls and Anti-Patterns . . . . .	878
<b>19 Data, Measurement, and Platforms</b>	<b>883</b>
19.1 Motivation and Scope . . . . .	884
19.2 Panel Data Sources in Marketing . . . . .	885
19.3 Identity, Linking, and Keys . . . . .	889
19.4 Transformations and Aggregation . . . . .	893
19.5 Platform Metrics vs Econometric Estimands . . . . .	900
19.6 Privacy, Policy, and Governance . . . . .	904
19.7 Missing Data and Measurement Error . . . . .	908
19.8 Validation and Reconciliation . . . . .	913
19.9 Pipelines and Reproducibility . . . . .	917
19.10 Assumptions and Guardrails . . . . .	922
19.11 Workflow Checklist . . . . .	925

<b>20 Outlook and Open Problems</b>	<b>931</b>
20.1 Motivation and Scope . . . . .	932
20.2 Interference at Scale . . . . .	934
20.3 Structural Instability and Regime Change . . . . .	939
20.4 Adaptive Experimentation and Learning . . . . .	945
20.5 Method Selection and Design Guidance . . . . .	952
20.6 Robust and Distribution-Free Inference . . . . .	961
20.7 Continuous Treatments and Structural Response . . . . .	965
20.8 Partial Identification and Sensitivity . . . . .	969
20.9 Generative Synthetic Data: Promise and Peril . . . . .	973
20.10 ML Integration Beyond Nuisance . . . . .	976
20.11 Privacy-Preserving Measurement and Data Clean Rooms . . . . .	978
20.12 Reproducibility, Benchmarking, and Standards . . . . .	983
20.13 Practitioner Roadmap . . . . .	988
20.14 Assumptions for Future Practice . . . . .	991
20.15 Chapter Summary and Visual Guide . . . . .	993
<b>Part IX Appendices</b>	<b>999</b>
<b>A Time Series: Recap of Basic Principles</b>	<b>1001</b>
A.1 What Is a Time Series? . . . . .	1001
A.2 Time Series Versus iid Data . . . . .	1003
A.3 Fundamental Properties . . . . .	1004
<b>B Stationarity and Cointegration in Panels</b>	<b>1011</b>
<b>References</b>	<b>1017</b>
<b>Glossary</b>	<b>1029</b>
<b>Colour Key for Text Boxes</b>	<b>1035</b>
<b>Notation</b>	<b>1037</b>
<b>Index</b>	<b>1041</b>

# List of Tables

1.1	Comparison of Causal Inference Approaches for Marketing . . . . .	9
1.2	Overview of Three Motivating Examples . . . . .	26
1.3	Structural Differences Across Domains . . . . .	29
2.1	Core Estimands and Common Aggregations . . . . .	45
2.2	Crosswalk: Data Structures and Primary Methods . . . . .	61
3.1	Mapping from Assignment Mechanism to Estimands and Recommended Estimators . . . . .	111
4.1	Hypothetical Cohort-Time Treatment Effects . . . . .	128
4.2	Estimand Selection Guide . . . . .	130
4.3	Mapping from Estimands to Recommended Estimators and Assumptions . . . . .	149
4.4	Modern Staggered DID Estimators: Targets, Comparison Sets, and Weights . . . . .	149
4.5	Inference Method Decision Guide . . . . .	159
5.1	Specification Choices and Associated Bias Risks . . . . .	209
5.2	Event-Study Extensions Decision Guide . . . . .	214
5.3	Event-Time Metrics and Marketing Decisions . . . . .	220
6.1	Synthetic Control Diagnostic Checklist (Illustrative Heuristics) . . . . .	256
6.2	Design Choices: Implications for Bias and Variance . . . . .	266
6.3	SC Extensions: Key Features . . . . .	278
7.1	Hybrid Methods at a Glance (see [Abadie et al., 2010, Ben-Michael et al., 2021, Arkhangelsky et al., 2021, Athey et al., 2025b] for methodological details) . . . . .	348
8.1	Assumptions, Empirical Implications, and Diagnostics . . . . .	372
8.2	Comparison of SC, SDID, IFE, and Matrix Completion . . . . .	376
8.3	Inference Methods for Factor Models . . . . .	388
8.4	Diagnostic Checklist for Factor Models . . . . .	394
8.5	Factor Model Applications in Marketing . . . . .	399
8.6	When to Prefer SC, SDID, IFE, or Matrix Completion Given Data Features . . . . .	405
9.1	Chapter 9 roadmap: advanced matrix methods for counterfactual imputation . . . . .	411

9.2	Standard vs robust matrix completion . . . . .	421
9.3	Standard vs covariate-assisted matrix completion . . . . .	428
9.4	Standard vs time-varying matrix completion . . . . .	433
9.5	Frequentist vs Bayesian inference for matrix completion . . . . .	439
9.6	Algorithm comparison for matrix completion . . . . .	443
9.7	Illustrative benchmark results (Intel i7, 32GB RAM, synthetic sparse panel) . . . . .	445
9.8	Tensor vs matrix completion . . . . .	446
9.9	Summary of marketing applications . . . . .	455
9.10	Software packages by task and language . . . . .	462
9.11	Comparison of advanced matrix and tensor methods . . . . .	467
10.1	Chapter 10 Roadmap . . . . .	475
10.2	Dynamic Estimands . . . . .	480
10.3	Identification Assumptions for Dynamic Effects . . . . .	481
10.4	Control Group Choices . . . . .	482
10.5	Dynamic Estimation Methods . . . . .	487
10.6	Software for Dynamic Treatment Effect Estimation . . . . .	492
10.7	Dynamic Mechanisms in Marketing . . . . .	493
10.8	Mediation Channels in Marketing . . . . .	496
10.9	Inference Methods for Dynamic Effects . . . . .	498
10.10	Diagnostics for Dynamic Effects . . . . .	502
10.11	Marketing Applications of Dynamic Methods . . . . .	505
10.12	Estimand–Estimator Mapping for Dynamic Treatment Effects . . . . .	514
11.1	Chapter Roadmap: Interference and Spillovers . . . . .	520
11.2	Comparison of Spillover Types in Marketing . . . . .	521
11.3	Common Exposure Mappings . . . . .	527
11.4	Identification Strategies for Spillover Effects . . . . .	529
11.5	Estimands Under Partial Interference . . . . .	531
11.6	Estimation Methods for Spillover Effects . . . . .	535
11.7	DiD Spillover Groups . . . . .	537
11.8	Competition vs Saturation Effects . . . . .	541
11.9	Diagnostics for Interference . . . . .	546
11.10	Sensitivity Specifications for Spillover Analysis . . . . .	549
11.11	Summary of Marketing Applications . . . . .	550
11.12	Chapter 11 Roadmap . . . . .	554
12.1	Mapping from Estimands to Nuisance Components, Scores, and Aggregation . . . . .	609

13.1	Mapping of Method to Assumptions, Tuning, and Use-Cases. Lasso, elastic net, and group lasso often serve as building blocks for double selection or DML rather than as final estimators; see Sections 13.3–13.4. . . . .	657
14.1	Outcome-link decision guide and incidental-parameter cautions . . . . .	674
14.2	Model selection for zero-inflated and censored outcomes. All recommendations presuppose that identification for the causal effect of $D_{it}$ has been secured through design or unconfoundedness assumptions. Choosing a richer outcome model cannot compensate for violations of those assumptions. . . . .	682
14.3	Estimator to assumptions, tuning, and use-cases . . . . .	702
15.1	Threat to assumptions affected, diagnostics, and mitigation strategies . . . . .	733
16.1	Unified variance estimators for panel causal effects . . . . .	738
16.2	Test statistics for randomisation inference . . . . .	744
16.3	Methods for constructing uniform confidence bands . . . . .	748
16.4	Instruments in marketing: examples and exclusion concerns . . . . .	756
17.1	Map from design threat to diagnostic, mitigation, and inference choice . . . . .	790
18.1	Data environments and applicable methods . . . . .	796
18.2	Two-dimensional marketing problem taxonomy . . . . .	796
18.3	Indicative data requirements by marketing application domain . . . . .	800
18.4	Design comparison for advertising incrementality measurement . . . . .	804
18.5	Quasi-experimental identification strategies for digital attribution . . . . .	808
18.6	Instruments for price-elasticity estimation . . . . .	827
18.7	Potential instruments for channel effects on CLV . . . . .	833
18.8	Algorithmic discontinuities for pricing RDD . . . . .	838
18.9	Instruments for transport pricing . . . . .	839
18.10	Interference channels in platform markets . . . . .	849
18.11	Instruments for position effects . . . . .	857
18.12	Marketing problems and recommended causal methods . . . . .	869
18.13	Assumption violations and mitigations by marketing domain . . . . .	881
19.1	Marketing data sources and causal considerations . . . . .	888
19.2	Data assumptions and diagnostic checks . . . . .	924
19.3	Mapping from metric class to econometric estimand, design constraints, and diagnostic checks	930
20.1	Open problems, current tools, research gaps, and diagnostics or evaluation criteria . . . . .	995
A.1	Types of time series $X_t \in \mathbb{R}^k$ ( $t \in \mathcal{T}$ ) . . . . .	1002
B.1	Panel unit root tests: scope and assumptions . . . . .	1013

B.2	Panel cointegration tests . . . . .	1014
B.3	Implementation Map for Stationarity Diagnostics . . . . .	1015

# List of Figures

1.1	The Marketing Analytics Ecosystem: Experiments, Panels, MMM, and Attribution . . . . .	13
1.2	Timeline of Panel Data Methods: From TWFE to Modern Heterogeneity-Robust and ML-Integrated Approaches . . . . .	14
2.1	Potential-outcomes indexing: contemporaneous versus dynamic path. The figure contrasts contemporaneous notation $Y_{it}(d)$ , which assumes treatment effects depend only on current treatment status, with path-dependent notation $Y_{it}(d_i^t)$ , which allows outcomes to depend on the entire treatment history. Marketing applications often require the richer path-dependent framework to capture carryover, habit formation, and strategic dynamics. . . . .	44
3.1	Example Assignment Matrix for a Phased Rollout . . . . .	109
3.2	Geo-Cluster Map with Buffers and Stratification . . . . .	110
3.3	Switchback Schedule with Washout Periods . . . . .	110
4.1	Canonical $2 \times 2$ DiD: Timing and Parallel Trends . . . . .	124
4.2	Staggered Adoption: Cohort-Time Grid and Event-Time Alignment . . . . .	133
4.3	Event-Study Plot: Pre-Trends and Dynamic Effects . . . . .	155
5.1	Event-Study Plot with Omitted Category and Confidence Intervals. The implied cumulative effect $\sum_{k=0}^{10} \theta_k$ and long-run multiplier LRM (where appropriate) can be read directly from this event-time profile. . . . .	200
5.2	Support by Event Time $k$ and Cohort Composition Across $k$ . While the left panel reports observation counts, inference is typically driven by the number of clusters contributing to each $k$ . When clusters are few at extreme event times, standard errors will be particularly wide. . . . .	201
6.1	Pre-Treatment Fit and Post-Treatment Gap Plot (Illustrative) . . . . .	258
6.2	Donor Weight Distribution and Predictor Balance (Illustrative) . . . . .	259
6.3	Placebo-Check Gap Distribution and RMSPE Ratio (Illustrative) . . . . .	260
8.1	Matrix Completion Setup for Causal Panel Analysis . . . . .	366
8.2	Factor Model Diagnostic Plots . . . . .	394
8.3	Factor Model Schematic with Loadings and Common Shocks . . . . .	406

8.4	Pre-Period Reconstruction Error vs Rank/Penalty (Validation Curve) . . . . .	407
8.5	Matrix Completion Illustration with Treated and Missing Cells . . . . .	407
9.1	Tensor completion schematic for three-way panels . . . . .	467
9.2	Robust matrix completion: separating low-rank structure from outliers . . . . .	468
10.1	Event-Time Path with Ramp-Up Pattern. The figure shows illustrative event-time effects $\hat{\theta}_k$ with 95% confidence intervals. Pre-treatment leads ( $k < -1$ ) are near zero, which is consistent with the identification assumptions under common diagnostics but not conclusive. Post-treatment effects show a gradual ramp-up, stabilising at approximately 10% after three periods. The vertical dashed line marks the reference category ( $k = -1$ ). . . . .	514
10.2	Impulse Response under Geometric Ad-Stock. The figure shows the impulse response $\hat{\beta}_s = \hat{\beta}_0 \hat{\delta}^s$ with geometric decay ( $\delta = 0.75$ ). The contemporaneous effect is $\beta_0 = 10\%$ . The half-life (time to decay to $\beta_0/2$ ) is approximately 2.4 periods. The long-run effect (shaded area) is $\beta_0/(1 - \delta) = 40\%$ . The corresponding long-run multiplier is $LRM = 1/(1 - \delta) = 4$ , which expresses the long-run effect in units of the contemporaneous effect. . . . .	515
11.1	Cluster Randomisation Design with Varying Treatment Intensities. Each cluster (metro area) is randomly assigned a treatment intensity $p \in \{0\%, 25\%, 50\%, 75\%\}$ . Within each cluster, units are randomly selected for treatment at the assigned rate. Filled circles denote treated units; hollow circles denote untreated units. Under partial interference, spillovers occur within clusters but not between clusters. . . . .	533
13.1	Lasso/elastic net coefficient paths and inclusion frequencies. The figure summarises how coefficients and selected sets change as the penalty varies. . . . .	656
13.2	Blocked cross-validation schematics for panels (by unit and by time). Use blocking to respect dependence and avoid leakage across folds. . . . .	656
13.3	Validation error across penalty values under blocked cross-validation. The figure illustrates the stability-fit trade-off when choosing $\eta$ . . . . .	657
14.1	Average dose-response function with confidence band . . . . .	702
14.2	Overlap diagnostics for the GPS by dose bins and subgroups . . . . .	703
14.3	Marginal effect of dose as a function of dose . . . . .	704
15.1	Changepoint diagnostics around policy or algorithm updates . . . . .	730
15.2	Event-time support and seasonality overlay in pre and post windows . . . . .	731
15.3	Overlap and balance before and after donor curation or GPS weighting . . . . .	732
17.1	Overlap diagnostics and common support with trimming thresholds . . . . .	788
17.2	Event-time leads with joint bands and support-by-k overlay . . . . .	789
17.3	Weight dispersion and leverage in SC/SDID and leave-one-out influence profiles . . . . .	790

18.1	Diagnostic protocol: Overlap and balance assessment . . . . .	800
18.2	Geo-experiment design schematic. The timeline shows the division into pre-treatment (baseline) and treatment periods. Geographic units are assigned to treatment (ad exposure) or control (hold-out), allowing for difference-in-differences or synthetic control estimation. . . . .	803
18.3	Synthetic-control estimation results. Top panel: gap plot showing the difference in sales between the aggregated treated unit and the synthetic counterfactual. Bottom panel: distribution of donor weights, indicating which control DMAs contribute most to the counterfactual.	806
18.4	Media transformation functions. Left panel: geometric adstock decay showing how an impulse of spend dissipates over time. Right panel: Hill saturation curve mapping adstock to incremental sales, illustrating diminishing returns. . . . .	813
18.5	Regression discontinuity design for loyalty tier valuation . . . . .	823
18.6	Network interference in two-sided markets . . . . .	849
18.7	Switchback experiment timeline . . . . .	851
18.8	Method selection decision tree . . . . .	868
18.9	Specification curve analysis . . . . .	872
18.10	Event-study estimates with joint confidence bands . . . . .	873
19.1	Data lineage from platform logs to econometric panel with leakage guardrails . . . . .	927
19.2	Exposure construction and mapping to dose with frequency caps and viewability . . . . .	928
19.3	Policy change timeline overlaid on outcomes and exposure metrics . . . . .	929
20.1	Interference-at-scale schematic with overlapping clusters and exposure mappings. <i>Panel A illustrates non-overlapping clusters, which align with partial-interference assumptions. Panel B illustrates overlapping clusters, where exposure mappings are high-dimensional and inference must account for network-induced dependence.</i> . . . . .	993
20.2	Regime-change timeline with breakpoints overlaid on outcomes and exposures. <i>Illustrative schematic of regime breaks and drifting effects. Identification still requires design assumptions about assignment and about what is stable across regimes.</i> . . . . .	994
20.3	Method-selection decision map linking threats to design families and diagnostics. <i>A decision flowchart linking data features and threats to design families, diagnostics, and inference choices. The output is a causal estimate conditional on the stated design assumptions, accompanied by diagnostics and sensitivity analysis.</i> . . . . .	995



# **Part I**

## **Foundations**



## Chapter 1

# Why Panel Data for Marketing? Motivation, Methods, and the Causal Revolution

Marketing generates vast panel datasets, but data volume does not by itself deliver credible measurement. The central problem is causal: we want to know what would have happened under a different pricing policy, a different campaign, or no intervention at all. This chapter motivates why panel data are uniquely useful for that task, and why they still require explicit research designs and assumptions. We begin with the marketing measurement crisis, then introduce the conceptual logic of panel causal designs, place them in the broader analytics ecosystem, and preview the marketing phenomena that make identification hard. The point is simple: panel methods are not a bag of regressions; they are tools for disciplined counterfactual reasoning.

## 1.1 The Marketing Measurement Crisis

Every marketer asks the same questions: Did it work? Did our pricing change boost sales, or was it just seasonal demand? Has our new digital channel cannibalised previous ones, or has it increased the pie? These are serious questions. Correctly answering them serves as the foundation for accountability and sensible investing. However, definitive answers are notoriously difficult to uncover. The world does not pause for our campaigns. Competitors react, trends shift, and customers change. The primary problem of marketing measurement is to distinguish the genuine causal influence of our actions from this background noise.

This book is about overcoming that problem. We show how to leverage panel data—observations of the same entities across time—to estimate causal effects under explicit research designs. We move beyond simplistic before-and-after comparisons, which can be misleading. Instead, we develop econometric tools tailored to marketing data and to the specific threats that undermine causal interpretation. We focus on difference-in-differences, synthetic control, factor methods, and machine learning used in service of identification. Our purpose is to provide conceptual understanding and practical advice for measuring what matters, shifting from correlation to causality and from conjecture to evidence.

Consider a modern CMO facing a familiar conundrum. Her company holds massive amounts of data, including clickstreams that track every consumer trip, impression logs that record millions of ad exposures, transaction histories that span decades, and loyalty card data that connects behaviour to demographics. However, when the board asks the most fundamental strategic question—did our recent loyalty programme work, and should we extend it?—she cannot respond firmly. The data are plentiful, but the answers are often elusive.

The problem is not a lack of sophistication. Marketing analytics teams already often employ sophisticated statistical and machine learning methodologies. These algorithms predict churn, offer customised items, and optimise bidding tactics in real-time auctions. They can do an excellent job forecasting under the current policy—they determine which customers will react to an offer, which content will get the most clicks, and which price point will maximise short-term revenue. But these tools are optimised for prediction at the association level, not for causal identification. They cannot tell us what would happen if we modified the system. As Pearl [2009] emphasises, prediction answers “what will happen?” while intervention answers “what if we act?”—a distinction we formalise in Section 1.2.

This distinction matters. Predicting that consumers who visit the account cancellation page would churn allows us to direct our retention efforts. However, it says nothing about whether extending a discount to those consumers will reduce churn. High-churn clients may be fundamentally different in ways that no discount can alter. The link between visiting the cancellation page and churning may simply reflect a common cause such as dissatisfaction, which drives both behaviours. Predictions cannot guide strategy unless the causal mechanism is known.

The core challenge is that classic approaches cannot account for how marketing truly works. Take endogeneity. Firms do not distribute marketing budgets at random. They increase their advertising when they anticipate high demand. They conduct marketing in reaction to competitor threats. They implement loyalty programmes in areas where customer lifetime value is already rising. These decisions suggest that simple correlations between marketing tactics and outcomes might lead to confusion about cause and effect. Meth-

ods that work in cross-sectional studies with observed pre-treatment covariates fail when the confounding variables change over time.

Dynamics complicate matters. Ad campaigns rarely have immediate impact. Direct response advertising can boost sales within days. Repeated exposures build brand recognition over time. Habit formation might take several weeks or months. Carryover and adstock models from the marketing mix have long recognised this truth. However, these dynamic effects complicate identification more than functional form assumptions would predict. Consumers react to expected future pricing rather than present costs. Firms alter their strategies in anticipation of competitive moves. These intertemporal links prevent evaluating a marketing intervention by comparing a single pre-treatment period to a single post-treatment period.

Finally, spillovers are ubiquitous. Most causal inference frameworks [Rubin, 1980] rely on the stable unit treatment value assumption (SUTVA), yet marketing actions frequently violate it. A consumer who participates in a loyalty programme may refer friends, resulting in beneficial spillovers to untreated units. A retailer's local promotion attracts customers from adjacent regions, resulting in geographic spillovers. When a corporation raises its advertising, competitors modify their own marketing mix, resulting in strategic spillover. The causal effect of treating one unit depends on how the other units are treated. This contradicts the independence principle required for clean cause identification in randomised experiments.

Traditional marketing measurement struggles to handle these challenges. Cross-sectional techniques focused on observable selection do not account for time-invariant or slow-moving unobservables like brand equity, store quality, or enduring consumer preferences, all of which influence marketing decisions and outcomes. Simple before-and-after comparisons fail to distinguish the influence of a marketing initiative from secular trends, seasonality, or concurrent shocks. Randomised controlled trials (RCTs) are the gold standard for identifying causal effects when feasible. In practice, however, they face constraints in commercial settings: spillovers between treated and control units can contaminate estimates; short experimental periods may miss long-term effects; ethical constraints may restrict the interventions that can be randomised; and the external validity of closely controlled trials may not extend to natural field conditions.

Industry practice typically relies on Marketing Mix Models (MMM). They approach the problem by designing flexible functional forms (sometimes with distributed lags to capture dynamics) and estimating them with aggregate time-series data. These models have proven useful for assessing the historical relationship between marketing inputs and outcomes; yet identification issues arise when many marketing factors interact and functional form assumptions are strong. Furthermore, typical aggregate marketing mix models struggle to determine the causal influence of specific, discrete actions, such as the launch of a loyalty programme or a shift in pricing strategy. Attribution models, which are becoming more common in digital marketing platforms, track customer touchpoints and assign credit to various marketing contacts; however, they are fundamentally observational exercises that rely heavily on the absence of unmeasured confounders and spillovers across channels.

Panel data methods bridge the gap between the gold standard of randomised experiments and the often fragile inferences drawn from observational data with no explicit strategy for unobserved confounders. They use repeated observations on the same units over time to difference out time-invariant unobservables via fixed effects and to exploit shared patterns or parallel evolution between treated and control units. Causal identification still hinges on design-linked assumptions such as parallel trends, unconfoundedness, or factor

structures, which we state explicitly in later chapters. They estimate dynamic impacts by examining outcomes across longer time periods, provided assumptions about lag structure and anticipation hold. They allow for more variable patterns of treatment adoption than two-group, two-period designs. The modern panel data toolset combines traditional econometric insights with recent developments in difference-in-differences, synthetic control approaches, factor models, and machine learning to deliver robust causal estimates in observational marketing scenarios.

Our approach is design-based, with a focus on reliable measurement of causal effects from specific interventions. This is in contrast to the structural tradition, widespread in economics and marketing, in which a theoretical model of behaviour is built first before the deep parameters are estimated. Later, in an interlude on design-first and structural IO (Section 3.12), we show how design-based estimates can discipline structural models rather than compete with them.

The Berry et al. [1995] (BLP) demand model for differentiated items illustrates the structural approach. BLP specifies a consumer's utility function and uses market-level data to predict demand elasticities while accounting for price endogeneity. The goal is to retrieve the underlying parameters of consumer preference. Once estimated, these characteristics allow for extensive counterfactual simulations, such as estimating market shares after a merger or a price adjustment across all items. This power requires considerable assumptions about the utility function and market equilibrium.

Our focus is more pragmatic. We favour clear, convincing identification of an impact by applying experimental design principles to observational data. To evaluate the efficiency of a loyalty programme, we do not need a thorough consumer choice model. We do, however, require a thorough understanding of the assignment process and counterfactuals. This is not measurement without theory, but measurement with minimal theory—enough structure to clarify the causal question and identify the effect without the need for a complete consumer optimisation model.

This book methodically explores these methods, with an emphasis on the distinct characteristics of marketing applications that present both opportunities and obstacles for causal inference. Marketing generates extensive panel data, including store-level sales recorded over quarters or years, consumer purchase histories spanning many transactions, market-level advertising, and outcomes measured across regions and time periods. However, marketing faces the three basic issues described above, namely endogenous decision-making, dynamic impacts, and spillovers, in particularly severe forms. Instead of ignoring these realities, the tools we build must tackle them directly.

## What This Book Is Not

Before proceeding, we must clarify scope. This book is not an introduction to panel econometrics. We assume familiarity with fixed effects regression, basic matrix notation, and standard inference concepts at the level of a first graduate econometrics course. Readers seeking foundational econometric training should consult a standard graduate textbook before diving into the panel-specific methods we develop here. Part I provides a refresher on causal frameworks and panel notation, but it is not a substitute for prior knowledge.

This book is not a machine learning cookbook. We use machine learning tools such as random forests, lassos, and gradient boosting to identify causal relationships via nuisance function estimation or heterogeneous effect discovery. However, we do not treat ML as a replacement for research design. A causal forest without a credible identification strategy will still deliver misleading answers. Double Machine Learning (DML) [Chernozhukov et al., 2018] works due to Neyman orthogonalisation, not algorithmic sophistication. Design-based identification comes first; ML is a tool, not a solution.

Finally, this book is not a structural modelling textbook. We focus on the design-based identification of well-defined causal effects, such as the impact of a loyalty programme on sales or the effect of advertising on purchase, rather than the recovery of deep preference parameters or the estimation of equilibrium models. Structural approaches have their place, but we prefer transparent, assumption-explicit methods that value credibility over generality. We want to know “what is the effect?” not “what are the primitives of consumer utility?” When structural methods appear later in the book, they do so as complements to, not substitutes for, design-based evidence.

## 1.2 What Are Panel Data Methods? A Conceptual Framework

Panel data methods take advantage of a fundamental asymmetry: by observing the same units repeatedly over time, we can learn about causal effects by comparing changes in outcomes for treated versus control units. This within-unit, over-time variation provides us with leverage that cross-sectional or time-series analysis cannot match.

Consider a cross-sectional comparison. We observe that stores participating in a loyalty programme have higher sales than non-participating stores. This correlation may indicate a causal effect. However, it may also reflect pre-existing differences between stores that use loyalty programmes and those that do not. Even if we measure and control for observable store characteristics such as size, location, and demographics, unobservable factors such as management quality or local brand strength can throw off the comparison.

A pure time-series approach performs no better. If a single store's sales increase after launching a loyalty programme, we cannot tell whether the programme caused the increase or whether sales would have risen anyway due to seasonal trends, competitive changes, or macroeconomic conditions. Without a contemporaneous control group, we cannot distinguish between the treatment effect and the counterfactual evolution that would have occurred in the absence of the treatment.

Panel data methods combine the strengths of both approaches. By observing multiple stores over time, we compare the change in sales for stores that implement a loyalty programme to the change for stores that do not implement the programme during the same period. If the stores in the two groups are similar enough to have followed parallel trends in the absence of the programme (Assumption 6), the difference in changes identifies the programme's causal effect. This difference-in-differences logic, which compares differences across groups to changes over time, eliminates both time-invariant confounding and common time trends.

Table 1.1 summarises how panel data methods sit within the broader landscape of causal inference approaches:

We must be precise about what panel data means. Observing the same individual units over multiple time periods is what distinguishes it and gives it statistical power. This differs from repeated cross-sections, which draw different samples of a population at different points in time. Repeated cross-sections enable us to monitor group averages. Panel data enables us to account for unobserved individual characteristics.

Section 2.2 returns to this distinction by treating grouped repeated cross-sections as panels of group means, with estimands defined at the group-time level rather than the individual level.

This advantage places panel methods within a broader hierarchy of causal reasoning. Pearl [2009] describes three ascending levels. At the lowest rung, association, we observe correlations: advertising expenditure moves with sales. One step up, intervention, we consider the consequences of actions: what would happen to sales if we increased advertising expenditure by 10%? At the summit, we consider alternative histories: what would sales have been in the previous quarter if we had run a different campaign?

Marketing strategy questions often necessitate intervention or counterfactual reasoning. We want to know not only what patterns exist in historical data, but also what happens if we change our strategy in the future. Predictive models and dashboards operate at the association level, identifying patterns but not supporting causal claims. Panel methods, which make explicit assumptions about the assignment mechanism and the structure of potential outcomes, allow for interventional reasoning. Section 1.3 locates these methods within

**Table 1.1** Comparison of Causal Inference Approaches for Marketing

Method	Units	Time	Identification	Marketing Example
Cross-Section	Many independent customers or markets	Single snapshot ( $t = 1$ )	Rich covariates support unconfoundedness; overlap diagnostics critical	Survey linking satisfaction to churn with demographic controls
Time Series	Single unit observed frequently	Long horizon with $T \gg 1$	Weak/strong stationarity, no structural breaks, lag structure well-specified	Decomposition of brand-level sales into trend, seasonality, and promotion shocks
Panel	Many units tracked over time	Balanced or unbalanced panel with $N$ and $T$ large	Parallel trends or factor structure; staggered-adoption diagnostics; short- $T$ vs long- $T$ regimes, dynamic panels (Chapters 4, 5, 10)	Chain-wide sales tracking phased loyalty rollouts across 40 quarters
Multi-Way (Matrix/Tensor) Panel	Multiple cross-sectional indices (e.g., stores $\times$ products $\times$ regions)	Many periods with outcomes arranged as matrices or tensors	Low-rank factor structure, matrix/tensor completion, structural break and cross-section dependence tests	SKU-by-store panel for recovering counterfactuals via matrix completion
Spatial/Interference Panel	Many units linked by geography or networks	Many periods with possible staggered exposure	Exposure mappings $h_i(D_{-i,t})$ , partial interference, spillover and saturation models	Geo experiment with demand shifts to neighbouring markets
High-Dimensional/ML Panel	Many units with rich covariates	Many periods, often with complex dynamics	Unconfoundedness with high-dimensional controls; orthogonalised ML scores, regularisation and post-selection inference; large- $N$ , large- $T$ clustered asymptotics	Ad impressions analysed via Double Machine Learning to estimate incremental lift
RCT	Many treatment and control units	Randomised in time or across units	Designed random assignment; high compliance; restricted interference (cluster-level SUTVA); no differential attrition	Geo-level A/B test with pre-registered outcomes and clustered inference

the broader measurement ecosystem alongside attribution, MMM, and geo-experiments. In this book, panel methods primarily operate at the intervention level; under stronger assumptions—such as no dynamic effects, no spillovers, and stable outcome models across regimes—they approximate counterfactual reasoning by imputing missing untreated potential outcomes (often written as  $Y_{it}(\infty)$  in staggered-adoption settings) for treated units in treated periods. Chapter 2 formalises this logic using potential outcomes  $Y_{it}(d)$  and explicit estimands such as ATE and ATT.

Moving up Pearl’s hierarchy requires stronger assumptions. Association requires only that we observe variables and quantify relationships. Intervention requires assumptions about confounders, parallel trends,

and whether a factor structure captures unobserved heterogeneity. Counterfactuals require even stronger assumptions about stability of the data-generating process under hypothetical changes. Panel methods make these assumptions explicit, diagnosable in limited ways, and subject to sensitivity analysis—an approach we emphasise throughout the book. The potential-outcomes notation  $Y_{it}(d)$  developed in Chapter 2 provides the formal language for stating these assumptions precisely.

We develop these conceptual foundations systematically in Chapters 4–11, always grounding abstract identification arguments in the concrete marketing measurement challenges introduced in Section 1.1.

Chapter 2 turns this conceptual hierarchy into formal notation and estimands (ATE, ATT,  $\tau(g, t)$ ,  $\theta_k$ ), and Chapters 4–11 build methods on top of that foundation.

### 1.3 Panel Data in the Marketing Analytics Ecosystem

Marketing organisations measure effectiveness through four broad families of tools, each with distinct strengths and limitations: attribution modelling, marketing mix modelling (MMM), geo-experiments, and panel causal designs. Understanding where panel methods fit is critical for selecting the appropriate tool.

Attribution modelling is ubiquitous in digital marketing, but as noted in Section 1.1 it remains observational and platform-specific. It reports which touchpoints precede conversions rather than counterfactuals, and typically ignores the potential outcomes  $Y_{it}(d)$  and spillover paths  $h_i(D_{-i,t})$ . In practice, attribution is most useful for short-run optimisation and targeting, not for identifying causal effects.

Marketing mix modelling (MMM) can be estimated using aggregate time series or panels that include adstock and saturation. As discussed earlier, aggregate time-series MMM faces identification and spillover challenges: with a single aggregate unit, it leans heavily on functional form, stationarity, and exclusion restrictions rather than on explicit designs, and it struggles with discrete interventions. Panel MMMs are complementary; under a design-first identification strategy they treat adstock and saturation as outcome-model features layered on top of transparent designs such as staggered adoption or geo experiments.

The third pillar, geo-experiments, represents the gold standard of causal inference. By randomising treatment across geographic markets—designated market areas, stores, regions—firms generate clean variation in marketing actions that is, by design, uncorrelated with potential outcomes (Assumption 5 in Chapter 2). Geo-experiments are particularly valuable when spillovers are localised within markets (internal to treatment or control clusters) and when the intervention can be implemented at market level. Platforms such as Meta, Google, and Amazon now provide tools for large-scale geo-experiments, lowering adoption barriers. Nevertheless, geo-experiments face practical limitations: spillovers across geographies can contaminate estimates, small numbers of clusters limit statistical power, short durations may miss long-run effects, and the cost and complexity of randomised rollouts can be prohibitive, so design-faithful diagnostics (Chapter 17) remain essential even in nominal experiments.

Panel data methods constitute the fourth pillar. The modern approaches that are the focus of this book apply quasi-experimental identification strategies to observational data, exploiting natural variation in the timing, location, or intensity of marketing actions to estimate causal effects. Unlike attribution models, panel methods explicitly model counterfactual outcomes under control conditions while making diagnostics and sensitivity analysis central to the credibility argument. In the notation of Chapter 2, these designs invoke assumptions such as parallel trends for untreated outcomes  $Y_{it}(\infty)$  across cohorts, low-rank factor structure restrictions for untreated outcomes, or conditional independence  $D_{it} \perp Y_{it}(d) \mid X_{it}, \alpha_i, \lambda_t$ , depending on the design. Typical aggregate time-series MMM implementations struggle to isolate discrete interventions, whereas panel-based MMM under a design-first identification strategy can do so more cleanly by accommodating staggered adoption and incorporating unit and time heterogeneity using fixed effects or factor structures. Unlike geo-experiments, panel methods do not require randomisation, enabling analysis of historical data, long-run effects, and settings where randomisation is infeasible. In practice, MMM and panel designs often overlap—for example, panel MMM with adstock and saturation fitted under a design-based identification strategy.

Panel causal designs require design-specific assumptions, not a one-size-fits-all condition. Some methods — difference-in-differences, event studies, interactive fixed effects (Chapters 4, 5, 8, 9) — rely on some form of parallel evolution across treated and control units. Others — synthetic control and its variants (Chapters 6, 7) — substitute strong pre-treatment fit for parallel trends assumptions. Still others — instrumental variables, regression discontinuity, dynamic panel GMM — require entirely different identifying structures: relevance and exclusion, continuity at a threshold, or moment conditions with specific serial correlation patterns. Throughout, we signpost assumptions and diagnostics in Chapters 4–11 and 16. Section 2.8 summarises which combination of these assumptions underpins each family of estimators and how they relate to the core estimands  $\text{ATT}$ ,  $\tau(g, t)$ , and  $\theta_k$ .

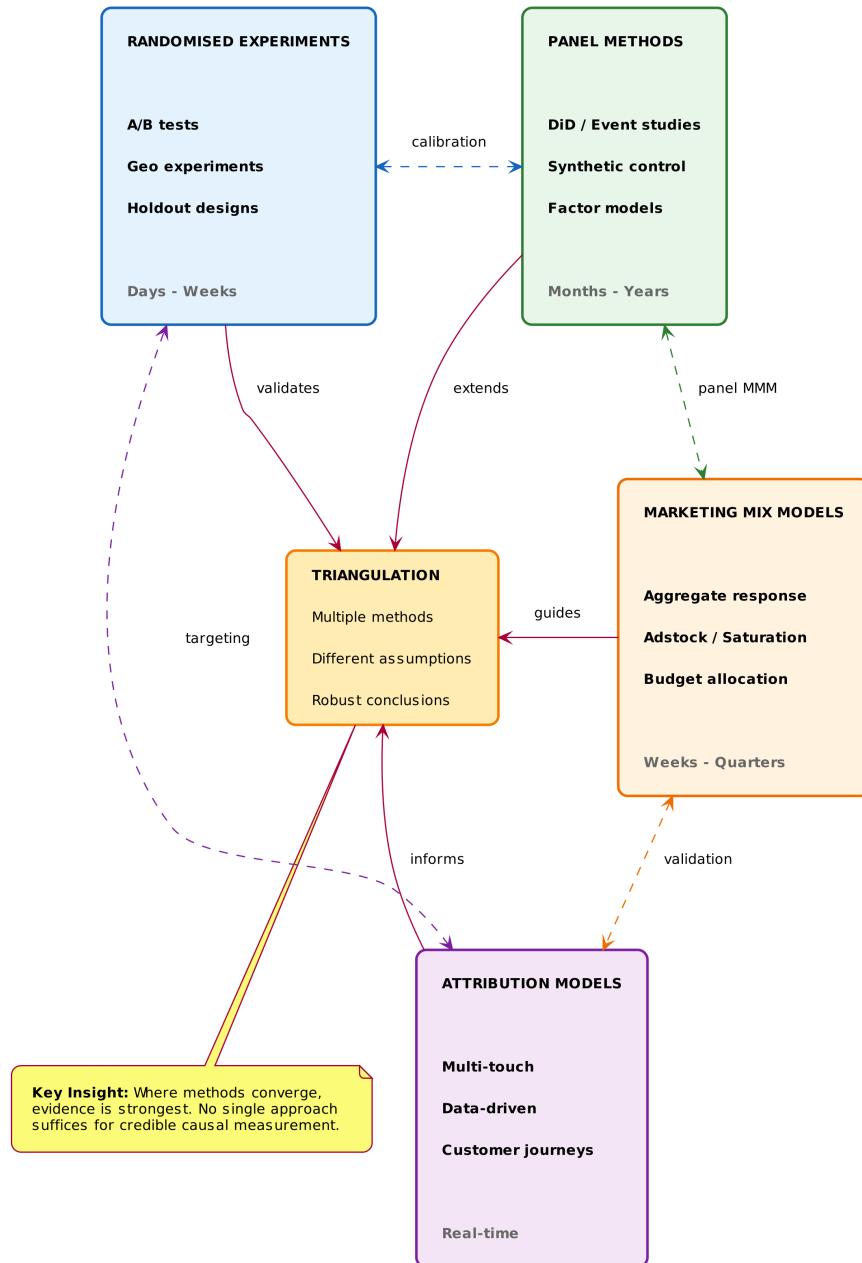
How should a practitioner decide which tool to use? The data structure drives the choice. Without repeated observations on the same units over time, panel methods are off the table. When repeated observations exist, randomisation remains the first-best option. The crosswalk in Section 2.6 and the method-selection map in Section 3.10 turn this principle into concrete design-to-method mappings. Experiments provide clean identification, though panel methods still add value by controlling for pre-treatment imbalances, quantifying heterogeneity, and extending short-term experimental results through longer observational follow-up.

When randomisation is infeasible, the structure of treatment variation guides method selection. Discrete interventions at specific points in time call for difference-in-differences or synthetic control methods (Chapters 4, 5, 6, 7) targeted at  $\text{ATT}$ , cohort–time effects  $\tau(g, t)$ , or unit-specific effects. Treatments that vary smoothly over time or across units without clear pre-post divides suit factor models and matrix completion methods (Chapter 8). Spillovers and interference demand methods that explicitly model network or geographic linkages (Chapter 11). High-dimensional covariates or heterogeneous treatment effects benefit from machine learning integration (Chapters 12, 13).

Figure 1.1 visualises these four pillars and their relationships. The overlapping circles represent not competition but complementarity. Where circles intersect, we find opportunities for triangulation: randomised experiments can validate panel-based estimates and calibrate MMM parameters; panel methods can extend short-run experimental results to longer horizons and provide causal interpretations of attribution metrics; marketing mix models can guide channel selection for deeper causal analysis; attribution models can flag segments warranting experimental scrutiny. In potential-outcomes terms, triangulation means checking whether different designs and estimators, each relying on distinct assumptions and assignment mechanisms, point to similar estimates of the same underlying effect (for example,  $\text{ATT}$  or  $\theta_k$ ) within sampling variability.

At the centre, where all four approaches converge, lies triangulation: multiple methods with different assumptions collectively build a more robust case for causality than any single method alone. The figure also highlights distinct timelines—experiments measure effects over days or weeks, panel methods track months or years, MMMs aggregate weeks or quarters, attribution provides real-time feedback. Understanding these complementarities is essential for coherent measurement strategy.

MMM and panel methods are not mutually exclusive. Many workflows combine MMM-style transforms (adstock, saturation) with design-based identification in panels (DiD, SC, IV). Panel MMMs are natural when you have multi-market or store-level data, and design-based logic strengthens identification. Factor-augmented panels (Chapters 8, 9) and instrumental variables (Chapters 2, 16) further bridge modelling and design.



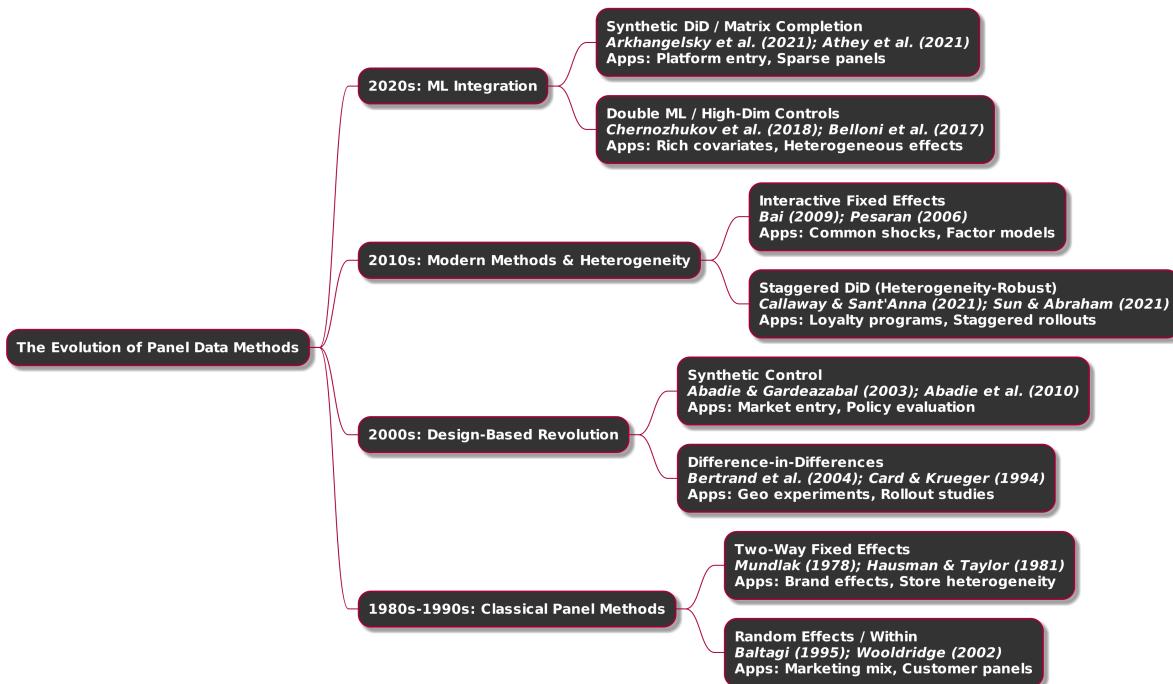
The Marketing Analytics Ecosystem

**Fig. 1.1** The Marketing Analytics Ecosystem: Experiments, Panels, MMM, and Attribution

Today's methods did not emerge fully formed. They reflect decades of intellectual development responding to new data, computational capabilities, and substantive challenges. Figure 1.2 depicts this evolution.

The 1980s and 1990s were dominated by two-way fixed effects and random effects models controlling for time-invariant unobservables. The design-based revolution of the 2000s, exemplified by Angrist and Pischke's work, brought synthetic control methods [Abadie and Gardeazabal, 2003; Abadie et al., 2010] and renewed focus on difference-in-differences with explicit parallel trends assumptions. The late 2010s saw the discovery of negative weighting in staggered TWFE designs [Goodman-Bacon, 2021; de Chaisemartin and d'Haultfoeuille, 2020], sparking heterogeneity-robust estimators that aggregate cohort-specific effects  $\tau(g, t)$  without using previously treated units as controls.

Recognition that parallel trends in levels may be too restrictive motivated synthetic control methods matching on pre-treatment trajectories. The proliferation of high-dimensional data—rich customer covariates, detailed competitor actions, granular geographic variation—prompted machine learning integration for flexible control and heterogeneous effect estimation. This timeline demonstrates the iterative, self-correcting nature of methodological progress. Today's methods will be refined or replaced as new challenges arise. The marketing applications driving this book—loyalty programmes, advertising campaigns, platform expansions—have played an important role, providing both motivation for new methods and empirical testing grounds where strengths and limitations can be identified.



**Fig. 1.2** Timeline of Panel Data Methods: From TWFE to Modern Heterogeneity-Robust and ML-Integrated Approaches

The aims of this book grew out of several converging developments rather than a single influence. A sequence of methodological breakthroughs in causal inference for panel data, together with a rapidly expanding applied literature, has made it possible to study marketing interventions with far greater rigour than traditional time-series or attribution approaches allow. At the same time, much of the marketing analytics discourse has drifted toward predictive machine learning, often with only loose connections to identification and design. This book attempts to recentre the discussion on causal structure and research design, while still drawing on modern tools where they genuinely strengthen empirical work.

Within this landscape, Arkhangelsky and Imbens [2024] provides a rigorous synthesis of modern causal panel methods. We build on similar methodological foundations but with a design-first, marketing-focused emphasis rather than a structural model-recovery focus: marketing-specific issues—attribution challenges, algorithmic confounding, applications in loyalty programmes and advertising—that receive limited attention in econometrics texts; machine learning as a supporting tool for identification rather than stand-alone prediction; and a practitioner orientation emphasising intuition, diagnostics, and method selection over formal proofs, while maintaining technical precision in assumptions and identification. Where their focus is methodological synthesis across domains, ours is to translate these methods into design playbooks, diagnostics, and method-selection workflows tailored to the persistent threats and data structures of marketing panels.

## 1.4 Motivating Examples: Three Marketing Challenges

To ground the abstract discussion of panel methods in concrete marketing problems, we consider three scenarios that motivate the methods developed in subsequent chapters. Each illustrates a distinct set of challenges and maps to a different constellation of panel data tools. While stylised, these examples reflect real measurement challenges practitioners face.

### Example 1: Evaluating a Loyalty Programme with Staggered Rollout

Consider a retail chain with 500 stores across various regions and three years of quarterly sales data. To increase customer retention and lifetime value, the chain launches a loyalty programme that rewards purchases with points redeemable for future discounts. Rather than launching the programme simultaneously, the company takes a phased approach: 100 stores in the first year, 200 more in the second, and 150 more in the third. The remaining 50 stores, mostly small and remote, never receive the programme and serve as controls.

The staggered rollout reflects both operational constraints and strategic considerations. Early adopters are typically larger stores in high-income areas where management anticipates strong programme uptake. Later adopters include stores in more competitive markets, where the company hopes the programme will help defend market share. This non-random assignment creates selection bias: programme stores differ systematically from non-programme stores in ways that affect sales even absent the programme. Simply comparing sales across store types would conflate the programme effect with pre-existing differences.

Formally, the target is a panel-level ATT over treated store-quarter cells, together with cohort-time effects  $\tau(g, t)$  and event-time effects  $\theta_k$  as defined in Section 2.3.

Additional complications arise during the evaluation. Two distinct threats deserve separate attention. First, *anticipation effects* may arise if rumours about the impending launch circulate ahead of time, causing customers to delay purchases until the programme begins—a dynamic phenomenon that shifts the timing of the treatment effect. Second, *spillovers* operate across units: customers who join may refer friends or family (positive peer effects), whereas customers from nearby non-programme stores may switch to programme stores to earn rewards (geographic interference). These are distinct identification challenges that require different modelling strategies, as we formalise in Chapters 10 and 11. A difference-in-differences design for this example therefore relies on Assumptions 6, 1, and 3, with Chapters 10 and 11 relaxing the last two when dynamic effects and interference are central. Beyond these, the programme’s effect is unlikely to be immediate or consistent—initial enrolment may be slow, habit formation takes time, and switching costs rise gradually as customers accumulate points. The effect is likely to vary by store type, with affluent, low-competition neighbourhoods responding differently than saturated urban markets where customers already have multiple loyalty options.

This scenario maps naturally to modern panel data methods: staggered difference-in-differences and event studies to estimate ATT,  $\tau(g, t)$ , and  $\theta_k$  (Chapters 4 and 5), spillover models for geographic interference (Chapter 11), and heterogeneity methods such as causal forests (Chapter 12). Chapters 17 and 16 develop

the diagnostic workflow including placebo checks, leave-one-out analyses, and sensitivity analyses. These methods rely on assumptions about parallel trends, spillover structure, and sufficient overlap, which we make explicit and diagnose in later chapters.

What might such an analysis reveal? Suppose the estimated average treatment effect were eight per cent across all programme stores and quarters. This aggregate effect would mask important dynamics: an initial effect of just two per cent in the first quarter post-launch, growing to eight per cent after four quarters and stabilising thereafter. Spillovers might be positive but modest — sales in non-programme stores within five kilometres of a programme store rising by two per cent, suggesting word-of-mouth effects. Heterogeneity analysis might reveal effects concentrated in high-income, low-competition areas (twelve per cent increase) with near-zero effects in saturated urban markets. These richer insights would guide not just whether to expand the programme, but where to expand it and how to manage expectations about the timeline for seeing results.

### **Example 2: Measuring Television Advertising Carryover in the Digital Age**

Consider a consumer packaged goods brand seeking to understand the causal effect of TV advertising on sales while accounting for carryover effects and digital channel interactions. The primary estimand is the dose-response function  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  and its dynamic profile  $\{\theta_k\}$ , together with the long-run multiplier LRM (Section 2.3). The data consist of weekly observations for 50 designated market areas over 100 weeks, including gross rating points (GRPs), online search volume, social media mentions, and sales. The advertising agency strategically varies TV spending, with higher levels during product launches, in markets with active competitors, and when previous sales trends indicate rising demand.

The first challenge is endogeneity. Markets receiving heavy TV advertising in a given week differ systematically from markets receiving little. Even with fixed effects controlling for time-invariant market characteristics, time-varying confounders—competitor actions, local economic shocks, seasonal patterns—may drive both advertising decisions and sales, inducing spurious correlation.

In practice this assumption is strong: spend often reacts mechanically to recent outcomes, so Assumption 7 must be defended using institutional knowledge and diagnostics rather than taken for granted.

The second challenge is carryover. TV advertising effects neither appear nor disappear instantly. Some viewers respond quickly, searching online or visiting stores within days. Others store the information and act weeks later. Brand awareness accumulates through repeated exposure and decays gradually in the absence of advertising. Specifying the functional form of this carryover—geometric decay, polynomial distributed lags, flexible nonparametric shapes—has occupied extensive research in the marketing mix modelling tradition.

Third, cross-channel effects complicate estimation. TV advertising may increase online searches, which in turn drive sales. Estimating the total effect of TV on sales includes both direct and search-mediated indirect effects. Whether it is appropriate to control for search depends on the estimand: controlling for a post-treatment variable targets a direct-effect object rather than the total effect, and can therefore understate incrementality when the mediator is part of the causal pathway. Understanding mediation is substantively

important but econometrically challenging. Formally identifying mediated effects requires stronger assumptions than those needed for total effects; Chapters 10 and 15 return to these mediation limits and show how to report mediated effects as sensitivity analyses rather than central estimates.

Fourth, measurement error pervades the data. Nielsen ratings derive from panels that may not represent the full population. Sales data aggregate from retail scanner panels with their own coverage gaps. Seasonal effects—holidays, weather, major events—create non-stationarity that must be distinguished from treatment effects.

Several panel methods address these challenges. Synthetic control methods (Chapters 6 and 7) create customised control groups for treated markets. Distributed lag models (Chapter 10) define carryover structure. High-dimensional control methods (Chapter 13) use data-driven variable selection to address multiple potential confounders. Chapters 16 and 17 discuss cluster-robust inference, placebo checks, and sensitivity analyses for carryover assumptions.

What might such an analysis reveal? Television advertising could boost sales by 5% during the campaign week, with a three-week half-life. Online search may mediate approximately 40% of the total effect, with TV driving search and search driving sales—though this mediation estimate relies on assumptions about the absence of unmeasured confounders of the search–sales relationship. Competitor advertising may partially offset the own-brand effect, reducing it by 20% when competitors simultaneously increase spending. According to memory and persuasion theories, emotional appeals may have a stronger carryover than informational appeals. Such insights would help guide budget allocation, creative strategy, and competitive response.

### **Example 3: Platform Market Entry and Competitive Dynamics**

Consider a food delivery platform such as DoorDash or Deliveroo that expands into 30 new cities over two years. The company tracks restaurant revenues on a monthly basis in both entry and comparison cities that do not use the platform. The data covers 50 cities over 36 months, resulting in a panel with staggered treatment timing. The firm wants to estimate the causal effect of platform entry on restaurant revenue while controlling for competitive dynamics and general equilibrium effects. The primary targets are a panel-level ATT for city–month cells, cohort–time effects  $\tau(g, t)$ , and event-time effects  $\theta_k$  for dynamics.

Several challenges complicate the analysis. Because each city is unique, exact matches between treated and control cities are impossible. The firm chooses entry cities based on market size, demographics, competitive landscape, and regulatory requirements. This creates selection bias. Entry occurs at different times in different cities, with larger, more appealing markets entering first. Once the platform enters a city, incumbent platforms (existing competitors) may respond by lowering commissions, increasing marketing, or improving service quality, thereby mitigating the treatment effect and inducing competitive spillovers. Furthermore, platform entry has an impact not only on the restaurants that join the platform, but on the entire restaurant ecosystem: consumers may dine out more frequently (category expansion), delivery drivers may shift labour supply, and restaurants that do not join the platform may see changes in foot traffic or delivery orders via third-party services.

This scenario motivates several panel methods. Staggered difference-in-differences with modern estimators (Chapter 4) provides one route; synthetic control (Chapter 6) constructs control cities from weighted averages of never-treated cities, with inference via permutation tests; synthetic difference-in-differences (Chapter 7) combines unit and time weights for staggered entry settings; and factor models (Chapter 8) handle unobserved common shocks affecting all cities. Spillover models (Chapter 11) quantify competitive responses. We develop these methods and their diagnostics in Chapters 4 through 11.

What might such an analysis reveal? Platform entry could increase restaurant revenues by fifteen per cent on average, with substantial heterogeneity. Small, independent restaurants could see twenty-five per cent gains from expanded reach, while chains with established delivery operations might see only five per cent gains. The competitive response from incumbent platforms could offset the effect by roughly thirty per cent in markets where incumbents lower commissions or increase promotions. General equilibrium effects might be positive — the category expanding as consumers order more frequently — suggesting platform entry creates value rather than merely redistributing it, with implications for regulatory policy and market structure.

These three examples illustrate the range of marketing questions panel data methods can address: selection bias, dynamics, spillovers, heterogeneity, measurement error, competitive interactions. Across these examples, we move from association (raw sales patterns) to intervention (programme launches, TV campaigns, platform entry) and, under structured assumptions, to counterfactual reasoning about alternative rollouts and budgets—ascending Pearl’s hierarchy introduced in Section 1.2 (Chapter 2 provides the formal potential-outcomes development). Each challenge will recur in subsequent chapters as we develop the technical methods and diagnostic workflows for credible causal inference.

## 1.5 The Two Cultures of Marketing Analytics

In an influential paper, and with a nod to C.P. Snow, statistician Leo Breiman distinguished two cultures in statistical modelling [Breiman, 2001]. The first, data modelling, posits explicit probabilistic models and estimates parameters under strong functional form assumptions. The second, algorithmic modelling, treats the underlying mechanism as unknown and focuses on predictive accuracy via flexible, data-adaptive algorithms. Breiman argued that statistics had overemphasised the first culture while neglecting the second, missing opportunities to leverage machine learning for prediction.

This dichotomy, while influential, has been criticised as too stark. Good statistical practice has always included elements of both cultures. Data modellers routinely validate predictive performance through cross-validation. Algorithmic modellers increasingly seek interpretability via SHAP values, partial dependence plots, and related tools. The real tension is not between prediction and understanding, but between rigid parametric assumptions and flexible, data-driven methods. Importantly, neither culture as originally conceived addresses the central challenge of marketing analytics: moving from prediction to causal inference—framed in potential outcomes  $Y_{it}(d)$  (Chapter 2) and explicit estimands such as ATE and ATT. This is the gap that design-based panel methods, introduced in Section 1.1, are meant to fill.

Modern marketing analytics reflect this tension. The rise of digital platforms has fuelled predictive modelling—churn prediction, recommendation engines, bid optimisation—that excels at forecasting outcomes within the current system. But when organisations face strategic decisions—Should we launch this loyalty programme? Will a price increase raise long-term revenue?—prediction alone is insufficient. We need causal inference. Panel data methods, grounded in quasi-experimental design but augmented with machine learning, provide a path forward.

Modern panel data methods for causal inference represent a synthesis of both cultures. They deploy the flexibility of algorithmic modelling to handle high-dimensional confounders and estimate heterogeneous treatment effects. Yet they maintain the discipline of data modelling by making causal structure explicit—parallel trends, unconfoundedness, factor structure, structured interference—and subjecting those assumptions to diagnostics. These methods primarily operate at the intervention level of Pearl’s hierarchy (Section 1.2); under stronger assumptions about dynamics, interference, and model stability they can approximate counterfactual reasoning. They map naturally onto the confounding, endogeneity, and interference threats in Section 1.1. Double machine learning exemplifies this synthesis: machine learning algorithms estimate nuisance functions with minimal parametric assumptions, while design-based identification ensures the final causal estimate is valid. This approach prioritises causal validity over pure predictive accuracy, but achieves it using tools from both traditions.

Our modelling philosophy is pragmatic. We do not seek a single true model that perfectly captures the data-generating process—no such model exists, and even if it did, we could not verify we had found it. Instead, we adopt a workflow with four components that aligns with the application protocol developed in Chapter 18.

1. **Specify the causal estimand and identification assumptions.** State the target quantity (ATT, ATE, group average treatment effects, elasticities) and the assumptions required for a causal interpretation: parallel trends, unconfoundedness, factor structure, or structured interference.
2. **Choose an estimator implementing that design.** Select difference-in-differences, synthetic control, interactive fixed effects, or a doubly robust machine learning method, matching the estimator to the identification strategy (see the crosswalk in Section 2.6).
3. **Run diagnostics.** Run diagnostics (Chapter 17): pre-trend diagnostics, placebo checks, balance assessments, leave-one-out robustness, and specification curves to assess whether assumptions hold and whether results are sensitive to modelling choices.
4. **Conduct sensitivity analysis.** Conduct sensitivity analysis for the core assumptions (parallel trends, unconfoundedness, factor structure, SUTVA), quantifying how large an assumption violation would need to be to overturn the conclusion. Assumptions are approximations rather than exact truths; sensitivity analysis makes this explicit.

This workflow embodies cultural synthesis: flexible tools deployed within a disciplined causal framework.

## 1.6 Strategic Dynamics and Marketing Phenomena

Marketing phenomena frequently exhibit dynamics and strategic interactions that panel data methods are well suited to analyse. This section previews six domains—market entry, innovation diffusion, advertising effectiveness, marketing investment returns, user-generated content, and competitive dynamics—where panel methods address long-standing strategic questions.

**Market Entry and Competitive Timing.** When should a firm enter a new market, and what advantages accrue to pioneers versus fast followers? Lieberman and Montgomery [1988] framed the first-mover advantage debate: early entrants may benefit from consumer lock-in and brand recognition, while later entrants can free-ride on pioneers’ investments in market development. Panel data on firms entering multiple markets over time enables difference-in-differences designs that compare early versus late entrants while controlling for time-invariant market characteristics and common temporal shocks, provided treated and comparison markets would have evolved similarly absent entry. When a firm expands into new cities in stages, the staggered rollout permits within-firm comparisons—though such designs require careful attention to parallel trends and heterogeneous effects, issues we address in Chapters 4 and 5. In the notation of Chapter 2, such designs rely on parallel trends for untreated outcomes  $Y_{it}(\infty)$  across entry cohorts and often adopt the cohort-time estimands  $\tau(g, t)$  and event-time effects  $\theta_k$  to describe dynamic entry impacts.

**Innovation Diffusion and Takeoff.** Innovations spread through populations at varying rates. The Bass [1969] model describes the S-shaped adoption curve, but it does not by itself identify the causal effect of marketing interventions on diffusion speed. Panel data on adoption rates across multiple markets and time periods enables models linking diffusion to market characteristics, competitive intensity, and promotional actions. Survival analysis and hazard models estimate time-to-takeoff as a function of covariates, while fixed effects or frailty terms account for unobserved market heterogeneity. The panel structure turns a descriptive diffusion curve into a setting where credible designs can be deployed, but causal statements about the effect of interventions on diffusion speed still require an identification strategy such as staggered rollouts, instruments, or other quasi-experimental variation (Chapters 4, 6). In practice this means embedding diffusion models inside designs such as staggered rollouts (to estimate how adoption hazards shift for treated versus control markets), instrumental-variable strategies that exploit exogenous shocks to exposure, or panel difference-in-differences on takeoff times.

**Advertising Effectiveness.** How effective is advertising at driving sales, and how quickly do effects accumulate and decay? Panel data with variation in advertising across markets and time, combined with distributed lag models, allows estimation of advertising elasticities, carryover parameters, and saturation thresholds [Sethuraman et al., 2011]. Synthetic control methods offer a complementary approach: constructing counterfactual markets that resemble treated markets in the pre-campaign period and comparing post-campaign trajectories. Where traditional Koyck models describe the time path of advertising response, panel methods add the identification structure needed to interpret that response causally.

**Marketing Investments and Shareholder Value.** Do marketing investments—new product launches, brand repositioning, advertising campaigns—create shareholder value? Finance event-study designs estimate abnormal stock returns around announcements while controlling for market-wide movements and industry trends [Srinivasan and Hanssens, 2009]. Under the usual identifying conditions—no other material news in the event window and rapid incorporation of information into prices—abnormal returns can be interpreted as the market’s assessment of the announcement’s value impact. Pooling many announcements across firms and time yields a panel of events, which increases power and supports heterogeneity analysis, but it does not by itself resolve confounding from endogenous announcement timing or concurrent strategic changes.

**User-Generated Content and Online Reviews.** How does user-generated content—reviews on Amazon or Yelp, social media mentions, forum discussions—affect sales and firm value? Chevalier and Mayzlin [2006] pioneered the use of review data to estimate demand effects. Panel data on product-level sales and review activity enables quasi-experimental designs, though identification requires careful argument. Plausible exogeneity in review timing is rare: popular products attract more reviews, and reviewers select strategically. Vector autoregression models applied to panel data trace dynamic feedback between content and outcomes, distinguishing whether reviews predict future sales (information revelation) or sales drive reviews (mechanical popularity effects). These models identify Granger causality—temporal precedence—not structural causality; without additional design-based variation or strong structural assumptions, they remain associational rather than causal. From the potential-outcomes perspective, VARs characterise associations among  $Y_{it}(d)$  and related series over time; turning these into causal claims requires additional structure or quasi-experimental variation.

**Competitive Dynamics.** How do competitors react to a firm’s actions? When one company cuts prices, launches a product, or increases advertising, do rivals retaliate, accommodate, or ignore? Panel data on competitor actions enables estimation of reaction functions describing how each firm’s decisions respond to rivals’ previous moves. Difference-in-differences and synthetic control methods estimate the causal effect of one firm’s action on competitors’ outcomes, quantifying spillovers and equilibrium effects that static models miss.

In each domain, traditional marketing models provide descriptive structure. Bass curves describe adoption patterns; Koyck lags describe advertising decay; VARs track joint dynamics. But description is not causation. Modern panel methods add the identification strategies—parallel trends, synthetic counterfactuals, instrumental variation, exposure mappings—that justify the transition from association to causal inference. A Bass model predicts the adoption curve under the status quo; a difference-in-differences analysis estimates how a subsidy shifts that curve. A VAR shows advertising and sales move together; a synthetic control analysis estimates whether increased advertising caused the sales increase. Panel methods do not replace classical models—they supply the design and identification machinery that turns those descriptive structures into causal estimates. We return to each of these domains in later chapters, pairing the descriptive models with explicit identification strategies (Chapters 4–11). Throughout the book we use these classical models as building blocks for estimating ATT,  $\tau(g, t)$ , and  $\theta_k$  under explicit identification assumptions rather than as stand-alone causal tools.

## 1.7 The Causal Revolution in Marketing

Marketing is shifting from metrics-focused dashboards to mechanism-focused causal analysis. For much of the digital era, organisations measured success using intermediate metrics—impressions, clicks, page views, ad recall—that serve as proxies for business outcomes. In the language of Pearl’s hierarchy (Section 1.2), these are association-level metrics: they describe patterns in the data but do not answer intervention or counterfactual questions. These metrics help optimise within a fixed strategic framework: if clicks predict sales, we can tune bidding to maximise clicks per dollar. But when used to assess strategic changes, intermediate metrics mislead. A campaign generating many clicks may not drive incremental sales if those clicks come from users who would have bought anyway. A loyalty programme that raises repeat purchase rates may not increase profits if it merely shifts purchases forward in time rather than expanding total consumption.

The incrementality problem has prompted major platforms to build causal measurement tools. Meta’s Conversion Lift uses an intent-to-treat design: users are randomly assigned to a holdout group, blocked from seeing the campaign’s ads, or to a treatment group eligible for delivery. Conversion rates are compared to estimate incremental lift. Within the treatment group, however, the platform’s targeting algorithm still determines who actually sees ads—randomisation occurs at the eligibility level, not exposure level. This distinction matters for interpretation. The experiment identifies the effect of eligibility (an intent-to-treat estimand). Translating that into an effect of exposure requires additional assumptions or instruments, because delivery acts like imperfect compliance. Google’s Geo-Experiments randomly vary advertising intensity across geographic areas to estimate aggregate regional effects. Amazon’s Brand Lift studies use similar holdout designs for awareness and intent. These designs deliver internally valid estimates of platform-specific lift for short horizons—they are genuine experiments whose internal validity is high for their defined estimands. Internal validity still hinges on correct implementation (for example, no leakage between holdout and treatment groups, stable measurement) and on the interference structure matching the design (for example, no cross-market spillovers that cross randomisation boundaries). However, they share common limitations: proprietary implementations that researchers cannot audit, aggregated outputs that obscure heterogeneity and dynamics, platform-specific designs that prevent cross-channel comparison, and narrow scope that limits external validity and transparency. Lewis and Rao [2015] quantify a related problem: even well-designed digital experiments often lack statistical power to detect economically meaningful effects, because per-user revenue variance swamps treatment effects.

When feasible, well-designed experiments remain the first-best option for causal inference; panel methods extend this logic to historical data and to designs that are impossible or impractical to randomise. Panel data methods supplement platform tools by offering transparent, reproducible, and flexible approaches to causal inference. The methods in this book apply to any panel dataset, not just those accessible via platform interfaces. They incorporate external data—competitor actions, macroeconomic conditions, offline behaviour—that platforms do not observe. They estimate long-run effects by tracking outcomes over months or years rather than the weeks typical of platform experiments. Spillovers, dynamics, and heterogeneity can be modelled explicitly and diagnosed rigorously. Assumptions are stated, scrutinised, and debated rather than hidden inside proprietary code.

This shift toward causal rigour reflects a broader intellectual movement. Angrist and Pischke [2010] labelled it the credibility revolution: economists began emphasising research designs that approximate randomised experiments—instrumental variables exploiting exogenous shocks, regression discontinuity designs using threshold-based assignment, and difference-in-differences comparing treated and control units before and after intervention. Combined with richer administrative data, these methods raised the evidential bar. Cross-sectional regressions with long covariate lists gave way to designs grounded in institutional detail, policy discontinuities, and natural experiments that generate credible counterfactuals. Section 1.1 instantiates these ideas for marketing, organising identification threats into confounding, endogeneity, and interference.

Marketing scholarship has adopted these lessons more slowly than labour economics or public finance, but the trajectory is clear. Leading journals now routinely publish quasi-experimental studies, and methodological sophistication in causal inference is increasingly expected. This book accelerates adoption by providing a rigorous yet accessible treatment tailored to marketing’s distinctive data structures and substantive questions.

The challenges ahead are formidable. Platform algorithms now decide which users see which ads, which products rank in search, and what prices appear to different customers. This algorithmic intermediation creates confounding by design: exposures depend on predicted responses, rankings on past clicks, prices on inferred willingness to pay. Causal inference must confront data-generating processes in which treatments are functions of predicted outcomes. In such environments, classical unconfoundedness and parallel-trends assumptions become harder to defend, because  $D_{it}$  is mechanically constructed from past  $Y_{it}(d)$  or its proxies.

Privacy constraints compound the difficulty. Deprecation of third-party cookies, Apple’s App Tracking Transparency, and GDPR limit cross-device and cross-platform tracking, producing missing data and measurement error. Panel methods must adapt: aggregate or differentially private data, partial identification under incomplete information, and triangulation across imperfect sources all become necessary.

Nonstationarity poses a third challenge. Consumer preferences shift, technologies emerge, competitors enter and exit, macroeconomic shocks disrupt demand. Panel methods often assume some stationarity—parallel trends, time-invariant effects, stable factor loadings—that may fail during upheaval. Methods for structural breaks, time-varying parameters, and regime changes remain active research frontiers; we address these in the dynamic-treatment chapter (Chapter 10) and the advanced matrix/tensor chapters (Chapters 8, 9) where we allow for changing factor structures. Chapters 10, 8, and 9 develop tools for time-varying factor structures and regime changes, but they still require explicit assumptions about how  $Y_{it}(d)$  evolves across regimes.

Interference grows more complex in networked platforms where users connect to many others, and in competitive markets where firms react strategically in real time. Partial interference assumptions—spillovers only within clusters, decay with geographic distance—may prove inadequate for modern marketing ecosystems. Developing tractable yet realistic interference models is an open problem.

These challenges, however, create opportunities. The same platforms that complicate inference through algorithmic targeting generate unprecedented granular data: clickstreams, location traces, social graphs, review text. Computational advances permit estimation of complex models at scale. Machine learning contributes not through prediction per se, but through flexible estimation of nuisance functions—propensity scores, outcome models, latent factors—that sharpen causal estimates. Double machine learning [Chernozhukov et al., 2018] and causal forests [Wager and Athey, 2018] exemplify this integration, using regularised learners to handle high-dimensional confounding while preserving valid inference on treatment effects. Even the

most sophisticated ML-based estimators, however, ultimately rely on the same identification assumptions—unconfoundedness, stable exposure mappings, correctly specified interference structures; they improve estimation of nuisance functions but cannot repair a fundamentally flawed design.

The methods developed in this book help readers navigate this landscape. Whether evaluating a loyalty programme, measuring advertising effectiveness, quantifying competitive dynamics, or assessing platform design changes, causal panel methods provide a disciplined framework for moving from correlation to cause. Causal claims remain provisional—grounded in approximations, not certainties. But with rigorous methods, transparent diagnostics, and systematic sensitivity analysis, we can distinguish interpretations that deserve credibility from those that do not. The rest of this book operationalises this revolution for marketing: specify clear estimands, design identification strategies, estimate with appropriate panel methods (often aided by ML), and stress-test conclusions with diagnostics and sensitivity analysis.

**Table 1.2** Overview of Three Motivating Examples

Example	Challenges	Methods	Key Insights
Loyalty Programme Rollout	Selection bias, spillovers, dynamics, heterogeneity	Staggered DiD (Ch. 4), event studies (Ch. 5), spillover models (Ch. 11), causal forests (Ch. 12)	8% avg effect, growing over time; spillovers +2%; heterogeneous (12% high-income, 0% saturated markets)
TV Advertising Carryover	Endogeneity, carryover, mediation, seasonality	Synthetic control (Ch. 6), distributed lags (Ch. 10), lasso (Ch. 13), robust inference (Ch. 16)	5% immediate effect, 3-week half-life; 40% mediated via search; competitor response -20%
Platform Entry	Unique units, staggered timing, competition, general equilibrium	SC (Ch. 6), SDID (Ch. 7), factor models (Ch. 8), spillover models (Ch. 11)	15% avg effect; 25% for small restaurants, 5% chains; competitive response -30%; category expansion (positive-sum)

## 1.8 Why Marketing Panel Data is Different

A reasonable question for any reader is: *Why do we need a specialised book on causal inference for marketing?* Excellent textbooks exist on econometrics, both theoretical [Wooldridge, 2010, Cameron and Trivedi, 2005] and practitioner-oriented [Angrist and Pischke, 2009, Cunningham, 2021]. Industrial organisation offers sophisticated models of strategic behaviour and competition. Can the marketing analyst not simply adopt these methods wholesale?

The answer is that marketing data differs structurally from datasets typically encountered in medical or financial applications, in ways that systematically violate the assumptions underpinning many standard estimators. While marketing shares some DNA with modern epidemiology—observational studies of contagion and policy interventions—it is distinct from the randomised clinical trials that dominate biostatistics. The same features that make marketing data rich—strategic behaviour, social connectivity, high-frequency measurement—introduce specific violations of traditional assumptions. Four structural differences define the challenge.

### Targeting Bias: Endogeneity by Design

In clinical trials, treatment is randomised. In labour economics, treatment such as a minimum wage increase is often exogenous to the individual worker. In marketing, treatment is strategic and endogenous. Firms actively target advertising, coupons, and sales calls to customers most likely to buy (retargeting) or most likely to churn (retention campaigns).

Standard estimators do not merely yield biased estimates in this setting; selection can even reverse the sign of a naive estimate relative to incremental lift. A naive analysis of a retargeting campaign will show that treated users purchase at vastly higher rates than controls. This is not the effect of the ad; it is the cause of the targeting. Blake et al. [2015] demonstrated this starkly: when eBay suspended branded search advertising, the naive model implied large positive returns, but the experiment revealed near-zero incremental effect because ads were reaching users who would have purchased anyway. Disentangling targeting policy from treatment effect requires methods that go beyond fixed-attribute controls: negative control outcomes, double machine learning (Chapter 12), and synthetic control methods that match on pre-treatment trajectories (Chapter 6). These tools help, but they still rely on strong identification assumptions—unconfoundedness for DML, parallel evolution for synthetic control—and careful design choices; they are not automatic fixes for targeting bias. In potential-outcomes terms, the challenge is that  $D_{it}$  is a deterministic function of signals correlated with  $Y_{it}(d)$ , so  $D_{it} \perp Y_{it}(d) \mid X_{it}$  is typically violated unless the targeting rule is explicitly modelled or instrumented.

## Interference as the Norm

The stable unit treatment value assumption—that each unit’s outcome depends only on its own treatment—is reasonable in clinical trials (my taking aspirin does not cure your headache). In marketing, SUTVA is routinely violated, much as in infectious disease epidemiology. Social interference arises when a customer’s purchase increases friends’ purchase probability through word-of-mouth or visibility [Aral and Walker, 2011]. Spatial interference occurs when a television ad in one designated market area spills into adjacent DMAs. Competitive interference emerges when one firm’s price cut triggers rival responses, altering equilibrium outcomes for all players.

Standard panel methods assuming unit independence fail to identify true effects when interference is present. Marketing applications require exposure mapping designs that model spillover structure explicitly through functions  $h_i(D_{-i,t})$  summarising neighbours’ treatments, cluster-randomised experiments that contain interference within groups, or spatial econometric methods that parameterise cross-unit dependence (Chapter 11).

## Large- $N$ , Moderate- $T$ Panels with Sparsity

Clinical and financial datasets vary in structure, but marketing data occupies a distinctive region. While biostatistics and finance both sometimes feature large  $N$ , marketing combines massive  $N$  with moderate  $T$  and high sparsity in a way that is particularly characteristic of the domain. A retailer may track ten million customers over fifty-two weeks, but most customers do not purchase in most weeks.

This structure renders traditional time-series methods (GARCH, cointegration) and traditional biostatistical methods (mixed models) computationally infeasible or statistically inefficient. It is, however, ideal for matrix completion and interactive fixed effects models (Chapters 8 and 9). These methods exploit the low-rank structure of consumer behaviour—millions of customers sharing a small number of latent preference dimensions—to construct counterfactual predictions with precision unattainable through unit-by-unit modelling. This structure—millions of units, moderate horizons, and sparse outcomes—is precisely the regime where low-rank factor models and matrix completion (Chapters 8 and 9) are most powerful.

## Complex Dynamics: Adstock and Wear-out

Financial markets are often modelled as efficient, incorporating information instantaneously. Marketing effects exhibit complex lag structures. An ad seen today may influence a purchase next week—the adstock or carry-over effect formalised by Broadbent [1984]. But seeing the same ad repeatedly may reduce its effectiveness—wear-out or saturation. Habit formation creates state dependence, where past purchases shift future purchase probability.

**Table 1.3** Structural Differences Across Domains

Feature	Clinical / Biostats	Financial Econometrics	Marketing Analytics
<b>Primary Goal</b>	Efficacy / Safety	Risk / Forecasting	<b>Incrementality / ROAS</b>
<b>Assignment</b>	Randomised (RCT)	Systemic / Exogenous	<b>Strategic (Targeting)</b>
<b>Interference</b>	Rare (except Vaccines)	Market-wide Equilibrium	<b>High (Social / Spatial)</b>
<b>Data Shape</b>	Small $N$ , Short $T$	Small $N$ , Long $T$	<b>Large <math>N</math>, Moderate <math>T</math></b>
<b>Sparsity</b>	Low (Complete records)	Low (Continuous trading)	<b>High (Infrequent purchase)</b>
<b>Key Method</b>	Survival / Mixed Models	Time Series / GARCH	<b>DiD/Event-Study, Synthetic Control, Matrix Completion</b>

Disentangling incremental lift from carryover requires distributed lag models and event study designs (Chapters 5 and 10) tuned to the stock-and-flow nature of marketing goodwill.

Table 1.3 summarises these structural distinctions. While we draw inspiration from epidemiology for interference and from finance for high-frequency dynamics, the combination of strategic targeting, network spillovers, and massive sparse panels makes marketing analytics a distinct methodological field.

## 1.9 Roadmap of the Book

This book is organised into eight Parts that build from foundational concepts through core methodologies to inference, diagnostics, and applications.

**Part I: Foundations** (Chapters 1–3) establishes the causal framework and panel notation. Chapter 2 formalises potential outcomes for panel data, distinguishing contemporaneous from path-dependent treatments, and defines key estimands. Chapter 3 contrasts randomised experiments with quasi-experimental designs including geo-experiments, switchback tests, and phased rollouts.

**Part II: Differences-in-Differences and Event Studies** (Chapters 4–5) develops canonical and staggered DiD designs. We present modern heterogeneity-robust estimators that avoid negative weighting and trace dynamic treatment effects through event-study specifications with pre-trend diagnostics.

**Part III: Synthetic Controls and Hybrid Methods** (Chapters 6–7) introduces synthetic control for settings where pre-treatment fit substitutes for parallel trends. Chapter 7 covers hybrid approaches including synthetic DiD and augmented synthetic control with doubly robust estimation.

**Part IV: Factor Models and Matrix Methods** (Chapters 8–9) exploits low-rank structure in panel data. Interactive fixed effects capture common time-varying shocks with unit-specific loadings. Matrix completion methods handle missing data and high-dimensional settings through nuclear-norm regularisation.

**Part V: Dynamics and Spillovers** (Chapters 10–11) addresses treatment path dependence through distributed lag and adstock models. Chapter 11 tackles SUTVA violations from network, geographic, and competitive spillovers using spatial econometrics and partial identification.

**Part VI: Machine Learning Integration** (Chapters 12–14) introduces double/debiased machine learning for nuisance function estimation with Neyman orthogonalisation. Causal forests estimate heterogeneous effects. High-dimensional controls and regularisation (lasso, ridge) handle settings with many potential confounders. Extensions cover continuous treatments, dose-response functions  $\mu(d)$ , and quantile effects.

**Part VII: Validity and Inference** (Chapters 15–17) catalogues marketing-specific threats including algorithmic confounding and platform metric misalignment. Inference tools (Chapter 16) include cluster-robust standard errors, bootstrap procedures, randomisation inference, and multiple testing adjustments. Diagnostic workflows (Chapter 17) encompass placebo checks, balance checks, and specification curves.

**Part VIII: Applications and Future Directions** (Chapters 18–20) synthesises methods through integrated case studies combining DiD, synthetic control, event studies, and causal forests. Chapter 19 examines scanner data, platform logs, and measurement challenges including privacy regulations. Chapter 20 identifies open problems in interference, nonstationarity, real-time optimisation, and method selection frameworks.

Each chapter follows a consistent structure: motivation, identification assumptions, estimation, inference, diagnostics, and marketing applications—the same workflow introduced in the application protocol (Section 1.5). Together, these parts move from the measurement crisis and conceptual framework in this chapter, through core panel methods and diagnostics, to applications that revisit the motivating marketing challenges with a full, design-based causal toolkit.

## Chapter 2

# Causal Frameworks and Panel Notation

This chapter develops the causal framework and panel notation that underpin all subsequent analysis. We formalise potential outcomes for panel data, including dynamic treatment paths  $Y_{it}(\underline{d}_i^t)$ , and define core estimands such as average treatment effects and event-time effects that recur throughout the book. We distinguish common data configurations — unit-level panels, grouped repeated cross-sections, and more general row–column exchangeable data — and classify assignment mechanisms that drive identification choices. We also preview regression specifications and inference issues that arise in panel settings, setting up the designs and estimators you meet in later chapters.

After working through this chapter, you will be able to formalise potential outcomes for panels and define core estimands such as ATE, ATT, and event-time effects  $\theta_k$ . You will be able to classify your data structure and assignment mechanism, and diagnose when basic regression tools are credible. You will also be able to map your data and business question to appropriate methods via the crosswalk framework introduced in Chapter 1 and developed further in Chapters 4–7.

## 2.1 Potential Outcomes for Panels

The measurement challenges outlined in Chapter 1—endogeneity, dynamics, and spillovers—demand a precise language for counterfactual comparisons. Without such a language, we cannot state clearly what we seek to learn, let alone estimate it credibly.

As introduced in Chapter 1, the potential-outcomes framework provides this foundation [Rubin, 1974, Angrist and Pischke, 2010]. In the language of Pearl’s hierarchy (Section 1.2), potential outcomes provide the formal machinery for moving from association to intervention and, under stronger assumptions, to counterfactual reasoning. Here we adapt that framework to panel data, with particular attention to how repeated observations and treatment dynamics complicate identification.

We observe data on  $N$  units indexed by  $i = 1, \dots, N$  over  $T$  periods indexed by  $t = 1, \dots, T$ . For each unit and period, we observe an outcome  $Y_{it}$  and a treatment or exposure  $D_{it}$ . In the simplest case,  $D_{it}$  is binary: it equals one if unit  $i$  is treated in period  $t$  and zero otherwise. The treatment may be an intervention applied to the unit—a store receiving a loyalty programme, a market exposed to an advertising campaign, a platform entering a city—or it may represent a continuous intensity variable such as advertising expenditure or promotional discount depth. For now, we focus on the binary case and return to continuous and multivalued treatments in Chapter 14.

In a cross-section, where each unit is observed once, we write  $Y_i(d)$  for the potential outcome of unit  $i$  under treatment level  $d$ . Panel data introduce two additional features. First, treatment may vary over time. Second, the effect of treatment in one period may depend on past or expected future treatments.

Consider a customer who joins a loyalty programme. The effect of programme membership on purchases in the current quarter may depend not only on whether the customer is a member now but also on how long the customer has been a member, how many rewards the customer has accumulated, and whether the customer anticipates remaining a member in future quarters. These intertemporal linkages mean that potential outcomes in period  $t$  may depend on the entire history of treatments up to that period, not just the current treatment  $D_{it}$ .

To handle this reality, we adopt two representations of potential outcomes for panel data, each appropriate in different settings. We will mostly use the simpler notation when we focus on designs that ignore or partially absorb dynamics, and we will return to the full path notation in Chapters 10 and 11.

The first, simpler representation assumes that potential outcomes depend only on the current treatment. We write  $Y_{it}(d)$  to denote the potential outcome for unit  $i$  in period  $t$  if its treatment in period  $t$  is set to  $d$ . This notation fits settings where treatment effects are instantaneous and history-independent—where the impact of treating a unit in period  $t$  does not depend on whether the unit was treated in earlier periods and where anticipation effects are absent. Marketing rarely offers such clean settings, so we often need a richer representation. Formally,  $Y_{it}(d)$  is a special case of the path-dependent potential outcome  $Y_{it}(\underline{d}_i^t)$  under assumptions that rule out both carryover and anticipation, so that  $Y_{it}(\underline{d}_i^t)$  depends only on the current component  $d_{it}$ .

The second, more general representation allows potential outcomes to depend on the entire treatment path. For unit  $i$ , let  $\underline{d}_i^t = (d_{i1}, d_{i2}, \dots, d_{it})$  denote the vector of treatment assignments from period one through period  $t$ , where the underline emphasises that this is a history rather than a scalar. The potential

outcome  $Y_{it}(\underline{d}_i^t)$  depends on this entire history. For example, if a store adopts a loyalty programme in quarter two and remains in the programme through quarter four, the potential outcome at  $t = 4$  corresponds to the history  $\underline{d}_i^4 = (0, 1, 1, 1)$ : no programme in quarter one and participation in quarters two through four. A different treatment path—say, adoption in quarter three,  $\underline{d}_i^4 = (0, 0, 1, 1)$ —would generally produce a different potential outcome even though the current treatment status in quarter four is the same.

The path-dependent notation is essential for settings where carryover effects, habit formation, or strategic interactions create dynamic responses. Advertising effects typically exhibit carryover: exposures in previous weeks contribute to brand awareness and purchase propensity in the current week. Loyalty programmes create switching costs that grow over time as customers accumulate points. Competitive responses unfold over multiple periods as rivals observe actions and adjust strategies. Ignoring these dynamics biases both short-run and long-run estimates. Chapter 10 develops methods explicitly designed for path-dependent potential outcomes, including distributed-lag models, dynamic panel specifications, and structural approaches that embed marketing actions in intertemporal optimisation problems.

A further complication arises because treatment assignment itself often responds to past outcomes. Firms do not randomly allocate marketing budgets across time. They increase advertising spending when demand trends upward, launch loyalty programmes in markets where early indicators suggest success, and adjust promotional intensity based on competitor actions and recent sales performance. This endogeneity means that even if we correctly specify the dynamic structure of potential outcomes, we must still address why units receive particular treatment paths at particular times. The assignment mechanism—whether treatments are allocated based on past outcomes, anticipated future trends, or exogenous factors—determines which identification strategies are credible. Formally, the assignment mechanism is the conditional distribution of the treatment matrix  $\mathbf{D} = (D_{it})$  given the collection of path-dependent potential outcomes and covariates,  $\Pr(\mathbf{D} \mid \{Y_{it}(\underline{d}_i^t) : \underline{d}_i^t \in \mathcal{D}^t\}, \mathbf{X})$ . In settings where no anticipation and no carryover hold, this simplifies to conditioning on  $\{Y_{it}(d) : d \in \mathcal{D}\}$ . Different designs — experiments, geo-experiments, phased rollouts, observational panels — correspond to different restrictions on this distribution. We return to assignment mechanisms in detail in Section 2.4.

We now formalise two assumptions that determine when we can simplify from the general framework to more tractable special cases. Panel methods frequently invoke both assumptions, though marketing contexts often violate them.

**Assumption 1 (No Anticipation)** For each unit  $i$  and period  $t$ , potential outcomes do not depend on treatment assignments in periods  $s > t$ . Formally, for any two treatment paths that agree through period  $t$ ,

$$Y_{it}(d_{i1}, \dots, d_{it}, d_{i,t+1}, \dots, d_{iT}) = Y_{it}(d_{i1}, \dots, d_{it}, d'_{i,t+1}, \dots, d'_{iT}).$$

No anticipation rules out the possibility that units respond to expected future treatments. In marketing, consumers may learn about an impending loyalty programme launch and alter their behaviour in advance. Retailers may adjust prices in anticipation of a competitor’s entry. Firms may front-load advertising expenditures ahead of a major product launch. When anticipation is plausible, event-study specifications with pre-treatment leads (Chapter 5) can diagnose anticipatory effects, and identification must rely on comparisons that are robust to such anticipation.

**Assumption 2 (No Carryover)** For each unit  $i$  and period  $t$ , potential outcomes depend only on the current-period treatment and not on past treatments. Formally, if two treatment histories agree in period  $t$ , then they generate the same period- $t$  potential outcome.

When Assumptions 1 and 2 both hold, the contemporaneous notation  $Y_{it}(d)$  is enough to describe causal effects. When either anticipation or carryover is plausible, we must instead treat  $Y_{it}(\underline{d}_i^t)$  as the primitive object and diagnose violations using pre-treatment leads in event studies (Chapter 5) and dynamic profiles (Chapter 10).

To discuss interference across units, it is helpful to start from the most general object, in which potential outcomes for unit  $i$  may depend on the entire vector of treatment histories. Let  $\underline{d}^t = (\underline{d}_1^t, \dots, \underline{d}_N^t)$  collect all histories. Then the most general potential outcome is  $Y_{it}(\underline{d}^t)$ .

**Assumption 3 (Stable Unit Treatment Value Assumption (SUTVA) for Panels)** For each unit  $i$  and period  $t$ , the potential outcome depends only on unit  $i$ 's own treatment path and not on the treatment paths of other units. Moreover, there is a single, well-defined version of treatment at each level. Formally,

$$Y_{it}(\underline{d}_1^t, \underline{d}_2^t, \dots, \underline{d}_N^t) = Y_{it}(\underline{d}_i^t).$$

Under Assumption 3, we can write potential outcomes as  $Y_{it}(\underline{d}_i^t)$  without reference to other units' histories; when combined with no anticipation and no carryover, this further simplifies to  $Y_{it}(d)$ .

As discussed in Chapter 1, SUTVA comprises two conditions. No interference requires that unit  $i$ 's potential outcomes depend only on its own treatment assignment, not on treatments received by other units. Treatment-version irrelevance requires that there be no hidden variations of treatment: receiving treatment level  $d$  has the same meaning for all units, so potential outcomes are well defined.

Marketing settings routinely challenge both components. A loyalty programme offered to customers in one store may generate word-of-mouth effects that influence purchases at nearby stores. Advertising shown to users in one city may spill over to neighbouring cities through migration or overlapping media markets. Competitive reactions create negative spillovers when one firm increases advertising and rivals respond, partially offsetting the initial effect. The definition of treatment may also be ambiguous: a loyalty programme might be implemented differently across stores, with varying enrolment incentives, rewards structures, and customer service quality.

Violations of SUTVA do not make causal inference impossible, but they require explicit modelling of the interference structure and of treatment heterogeneity. In later chapters we formalise this via exposure mappings  $h_i(D_{-i,t})$  that summarise neighbours' treatments into spillover doses, and we index potential outcomes as  $Y_{it}(d_{it}, h_i(D_{-i,t}))$  when interference matters (Chapter 11). Chapter 11 develops methods for settings with spillovers, including spatial econometric models, network-based approaches, and partial identification strategies that bound effects when the full interference structure is unknown. SUTVA is an assumption rather than an axiom. Like all assumptions in this book, it must be justified using institutional knowledge and subjected to sensitivity analysis.

Figure 2.1 contrasts the two notations visually. The contemporaneous notation  $Y_{it}(d)$  assumes that treatment effects depend only on current treatment status. The path-dependent notation  $Y_{it}(\underline{d}_i^t)$  allows outcomes

to depend on the entire treatment history. Marketing applications often require the richer path-dependent framework to capture carryover, habit formation, and strategic dynamics.

Throughout the book, we adopt whichever notation is most appropriate for the method under discussion. In chapters focused on difference-in-differences and synthetic control (Chapters 4, 5, 6, 7), we often use the simpler notation  $Y_{it}(d)$  when the focus is on comparisons of levels or changes. These comparisons may nonetheless conflate short-run and accumulated effects if dynamics are present. In chapters on dynamics and spillovers (Chapters 10, 11), we work explicitly with path-dependent potential outcomes  $Y_{it}(\underline{d}_i^t)$  and develop estimators that identify and estimate the full dynamic response. Event-study specifications (Chapter 5) trace out how treatment effects evolve over event time, effectively estimating a sequence of path-dependent effects indexed by time since treatment adoption. Distributed-lag models (Chapter 10) parameterise carryover, allowing current outcomes to depend on current and lagged treatments. The choice of notation is purely expositional and aims to serve the substantive question; it does not change the underlying causal objects. Throughout, the primitive causal object remains the path-dependent potential outcome  $Y_{it}(\underline{d}_i^t)$ ; when we write  $Y_{it}(d)$  or work with static DiD-type estimands, we are implicitly imposing restrictions on how histories enter  $Y_{it}(\underline{d}_i^t)$  and on anticipation, and later diagnostics are designed to probe those restrictions.

## 2.2 Panel Data Structures and Indexing

The shape of your panel—how many units, how many periods—determines which asymptotic approximations apply and which estimators are feasible. A method that performs well with 500 stores over 12 quarters may be unreliable with 5 markets over 100 weeks. This section classifies the data configurations you will encounter and clarifies the dimensions and sparsity patterns that drive method selection.

### 2.2.1 Proper Panels and Data Shapes

A proper panel tracks a well-defined set of units over multiple periods: each unit has a persistent identity across time, even if some periods are missing. Both balanced and unbalanced panels fit within this structure; the key requirement is that units can be followed longitudinally. We collect outcomes and treatment assignments into  $N \times T$  matrices  $\mathbf{Y}$  and  $\mathbf{D}$ :

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & \dots & Y_{1T} \\ \vdots & \ddots & \vdots \\ Y_{N1} & \dots & Y_{NT} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} D_{11} & \dots & D_{1T} \\ \vdots & \ddots & \vdots \\ D_{N1} & \dots & D_{NT} \end{pmatrix}.$$

The relative magnitude of  $N$  and  $T$  determines the shape of the data, which in turn dictates appropriate asymptotic approximations and estimation strategies.

#### Balanced and Unbalanced Panels

A **balanced panel** observes every unit in every period, so the matrices  $\mathbf{Y}$  and  $\mathbf{D}$  contain no missing entries. An **unbalanced panel** has gaps: some units enter the sample late, exit early, or have intermittent observations. We formalise this distinction with an observation indicator:

$$M_{it} = \begin{cases} 1 & \text{if } Y_{it} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

The observation matrix  $\mathbf{M}$  has the same dimensions as  $\mathbf{Y}$ , with entries indicating which cells contain data. In a balanced panel,  $M_{it} = 1$  for all  $i, t$ . In an unbalanced panel, the pattern of zeros in  $\mathbf{M}$  determines which estimators are feasible and how standard errors must be computed. When missingness is related to potential outcomes — for example, when high-churn customers exit the panel earlier — the observation process becomes an additional assignment mechanism that can bias naive estimators unless explicitly modelled. Similar observation indicators can be defined for treatment variables when  $D_{it}$  is missing independently of  $Y_{it}$ .

Marketing data are frequently unbalanced. Customers churn, stores open and close, products launch and discontinue. Scanner panels may track different store samples over time. Digital platforms observe users only when they are active. The missingness pattern itself may be informative—customers who churn differ

systematically from those who remain—raising selection concerns that Chapter 15 addresses. Chapter 15 treats informative missingness as a threat to validity and develops diagnostics and sensitivity analyses for selection on observables and unobservables. For now, we note that unbalanced panels require explicit attention to the observation process, and estimators must be adapted accordingly. For example, methods that rely on complete pre-period histories (synthetic control, some factor-based designs) may exclude units with short histories or require imputation.

### Thin Panels ( $N \gg T$ )

Thin panels are the classic setting in microeconomics, with many units observed over few periods. In marketing, think of customer purchase histories: thousands of customers tracked over a handful of quarters, or hundreds of stores observed across a dozen months. The loyalty programme example from Chapter 1—500 stores over 12 quarters—falls squarely in this regime.

$$\mathbf{Y}^{\text{thin}} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ \vdots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} \end{pmatrix} \quad (N \gg T).$$

The asymptotic regime for thin panels assumes  $N \rightarrow \infty$  with  $T$  fixed. Formally, we consider a sequence of panels  $\{(Y_{it}, D_{it}) : i = 1, \dots, N; t = 1, \dots, T\}$  where  $N$  grows but  $T$  remains constant. This fixed- $T$  asymptotic framework underpins most microeconometric panel estimators.

In this regime, we can difference out unit fixed effects  $\alpha_i$  in linear models and obtain consistent slope estimates provided the usual strict-exogeneity and regularity conditions hold, because the number of observations per unit remains bounded while the cross-sectional dimension grows. However, the incidental-parameter problem [Neyman and Scott, 1948] creates difficulties for nonlinear models. The problem arises because the number of nuisance parameters (the  $N$  fixed effects) grows with the sample size. In linear models, the fixed effects can be concentrated out, and their estimation does not distort the large-sample distribution of the treatment effect. In nonlinear models—logit, probit, Poisson—the fixed-effect estimates are inconsistent when  $T$  is small, and this inconsistency propagates to the parameters of interest. Bias-correction methods [Hahn and Newey, 2004, Fernández-Val and Weidner, 2016] or conditional-likelihood approaches can mitigate but not eliminate this problem; see Chapter 16 for further discussion.

Random-effects approaches avoid the incidental-parameter problem by treating  $\alpha_i$  as random draws from a population distribution rather than fixed parameters to estimate. This works well if the strict-exogeneity assumption holds and  $\alpha_i$  is uncorrelated with the regressors—assumptions that marketing applications frequently violate. Formally, this requires conditions such as  $\mathbb{E}[\alpha_i | X_{i1}, \dots, X_{iT}] = 0$  and  $\mathbb{E}[\varepsilon_{it} | X_{i1}, \dots, X_{iT}, \alpha_i] = 0$  for all  $t$ , assumptions that marketing applications frequently violate.

### Fat Panels ( $N \ll T$ )

Fat panels arise when we observe a few aggregate units over many periods. The TV advertising example from Chapter 1—50 DMAs tracked over 100 weeks—illustrates this shape. Brand-level sales data, where a handful of brands are tracked weekly for years, also falls here.

$$\mathbf{Y}^{\text{fat}} = \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & \dots & Y_{2T} \\ Y_{31} & Y_{32} & \dots & Y_{3T} \end{pmatrix} \quad (N \ll T).$$

This setting resembles time-series analysis more than cross-sectional microeconomics. The asymptotic regime assumes  $T \rightarrow \infty$  with  $N$  fixed. Formally, we treat the  $N$  units as a fixed collection and derive asymptotic distributions from the growing time dimension.

This regime introduces challenges absent from thin panels. Serial correlation is typically present—outcomes in adjacent periods are correlated even after conditioning on observables—so standard errors must account for temporal dependence. Heteroskedasticity-and-autocorrelation-consistent (HAC) estimators [Newey, 1987] or cluster-robust inference at the unit level address this concern (see Chapter 16 for details), though with very small  $N$  inference remains fragile. With very few units, standard HAC or cluster-robust corrections can seriously underestimate uncertainty; Chapter 16 discusses randomisation-based and wild-bootstrap methods that are better behaved in small- $N$  settings. Stationarity assumptions become relevant: if the data-generating process changes over time, long-run averages may not converge to stable population quantities (see Appendices A–B). Synthetic-control methods (Chapter 6) often operate in this regime, exploiting the long pre-treatment period to construct counterfactuals. Fat panels also tend to exhibit strong cross-sectional dependence through common shocks that affect all units. When these shocks are not fully captured by observed controls, factor-model methods (Chapters 8 and 9) can provide a more appropriate structure than unit-by-unit time-series models.

### Square Panels ( $N \approx T$ )

Square panels have comparable unit and time dimensions. The platform-entry example from Chapter 1—50 cities over 24 months—approaches this shape. Scanner data from 50 stores tracked over 52 weeks is another common instance.

$$\mathbf{Y}^{\text{square}} = \begin{pmatrix} Y_{11} & \dots & Y_{1T} \\ \vdots & \ddots & \vdots \\ Y_{N1} & \dots & Y_{NT} \end{pmatrix} \quad (N \approx T).$$

Inference here requires **joint asymptotics** where both  $N \rightarrow \infty$  and  $T \rightarrow \infty$  simultaneously. The rate at which each dimension grows matters: if  $N/T \rightarrow c$  for some constant  $c \in (0, \infty)$ , we are in a truly square regime where neither dimension dominates. Different limiting behaviours emerge depending on whether  $N/T \rightarrow 0$ ,  $N/T \rightarrow \infty$ , or  $N/T \rightarrow c$ .

Neither pure cross-sectional nor pure time-series asymptotics apply cleanly in this regime. Methods like interactive fixed effects (Chapter 8) and synthetic difference-in-differences (Chapter 7) exploit the joint struc-

ture of  $N$  and  $T$ , treating the panel as a matrix to be decomposed rather than a collection of independent units or time series. These methods typically require both  $N$  and  $T$  to be moderately large for the low-rank structure to be estimable, with convergence rates depending on  $\min(N, T)$  or  $NT$  jointly. In applications where either  $N$  or  $T$  is small, low-rank estimators can become unstable, and simpler DiD-style methods with transparent diagnostics may be preferable even if their assumptions are stronger.

### 2.2.2 Other Data Configurations

Not all marketing data arrive as proper panels. A second configuration arises when we observe different units in each period but can aggregate them into groups. Survey data, for instance, may sample different respondents each wave, but we can group respondents by demographic cell or geographic region. We then construct a panel of group-period means  $\bar{Y}_{mt}$ , where  $m \in \{1, \dots, M\}$  indexes groups. The effective sample size becomes  $M \times T$ , not the number of individual observations. Identification and inference therefore operate at the group-period level: causal estimands are defined for group averages  $\bar{Y}_{mt}(d)$ , and standard errors must account for the estimation error in  $\bar{Y}_{mt}$  arising from finite within-group samples. This grouped repeated cross-section structure requires care: within-group heterogeneity is averaged away, and inference must account for the estimation error in group means.

A third configuration arises when observations are indexed by two dimensions with no natural time ordering. Customer-product matrices are a canonical example: rows index customers, columns index products, and entries record purchase quantities or ratings.

$$\mathbf{Y}^{\text{rc}} = \begin{pmatrix} Y_{11} & \dots & Y_{1J} \\ \vdots & \ddots & \vdots \\ Y_{I1} & \dots & Y_{IJ} \end{pmatrix}.$$

This row-column exchangeable structure lacks the temporal ordering that defines a panel, so concepts like parallel trends must be reinterpreted as structural stability across dimensions. In potential-outcomes terms, we now index outcomes as  $Y_{ij}(d)$  for row  $i$  and column  $j$ , and identification hinges on assumptions about how treatment and potential outcomes vary across both dimensions rather than over time. Low-rank factor models (Chapter 8) often provide the right framework for such data, decomposing the matrix into latent customer preferences and product attributes. When an additional time index is present—for example, customer-product-time tensors—these matrices become slices of higher-order arrays, which we discuss in the advanced matrix and tensor chapters (Chapter 9).

### 2.2.3 Indexing Staggered Adoption

When treatment adoption is staggered, we need notation that distinguishes calendar time from time relative to adoption. We define  $G_i \in \{1, \dots, T\} \cup \{\infty\}$  as the adoption time for unit  $i$ —the first period in which unit  $i$

receives treatment. Units that never adopt during the sample window have  $G_i = \infty$  by convention. In simple “once treated, always treated” settings, the treatment path satisfies  $D_{it} = \mathbf{1}\{t \geq G_i\}$ . When treatment can switch on and off — for example, temporary promotions or campaigns that start and stop — this absorbing-path assumption is violated, and the event-time notation must be adapted to allow for multiple treatment spells (see Chapters 10 and 5).

Event time measures periods relative to adoption:  $k = t - G_i$ . When  $k = 0$ , the unit has just adopted. When  $k < 0$ , the unit has not yet adopted. When  $k > 0$ , the unit has been treated for  $k$  periods. This mapping transforms the calendar-time treatment matrix  $\mathbf{D}$  into an event-time structure where all adopters are aligned at their adoption moment, regardless of when that moment occurred in calendar time.

Consider the loyalty programme example. Some stores adopt in quarter 3, others in quarter 5, others in quarter 8. In calendar time, their treatment indicators turn on at different columns of the matrix. In event time, we align them so that  $k = 0$  corresponds to the adoption quarter for each store, allowing us to trace out how effects evolve in the periods before and after adoption.

We formalise event-time treatment indicators as follows. For each event-time value  $k$ , define

$$D_{it}^k = \mathbf{1}\{t - G_i = k\} = \mathbf{1}\{t = G_i + k\},$$

where  $\mathbf{1}\{\cdot\}$  is the indicator function. The variable  $D_{it}^k$  equals one if and only if unit  $i$  is exactly  $k$  periods from its adoption date at calendar time  $t$ . For never-treated units with  $G_i = \infty$ , all event-time indicators are zero ( $D_{it}^k = 0$  for all finite  $k$ ), since no finite  $t$  satisfies  $t - \infty = k$ . The original treatment indicator can then be written as  $D_{it} = \sum_{k \geq 0} D_{it}^k$  in once-treated-always-treated designs.

These event-time indicators form the building blocks of event-study regressions,

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \beta_k D_{it}^k + \varepsilon_{it},$$

where  $\beta_k$  captures the average difference in outcomes at event time  $k$  relative to the period just before adoption, for which we normalise  $\beta_{-1} = 0$ . Under the parallel-trends and no-anticipation assumptions developed in Chapter 5, the regression coefficients  $\beta_k$  can be interpreted as estimates of the event-time effects  $\theta_k$  introduced in Section 2.3, averaging cohort-time effects at event time  $k$  across contributing cohorts. The coefficients  $\{\beta_k\}_{k < 0}$  provide diagnostic checks for pre-trends (anticipation or selection), while  $\{\beta_k\}_{k \geq 0}$  trace out the dynamic response. Causal interpretation of these coefficients requires additional assumptions—in particular, parallel trends and no anticipation in event time—which we develop in Chapter 5. This event-time indexing is essential for event-study designs and for understanding dynamic treatment effects (Chapters 5 and 10). In staggered-adoption settings with heterogeneous effects, naive TWFE event-study regressions can also mix cohort-time effects with negative or opaque weights; Chapter 4 revisits event-time regressions with estimators that recover clean  $\theta_k$  profiles under weaker conditions.

## 2.3 Core Estimands for Panel Causality

What exactly do you want to know? “The effect of the loyalty programme” is not an answer. The effect for whom—all stores, or only those that adopted? In which periods—the quarter of adoption, or cumulative over two years? Aggregated how—a single number, or a dynamic path? Until you can answer these questions precisely, no method can help you.

This section defines the estimands that make these choices precise. Building on the potential outcomes framework in Section 2.1, we define average treatment effects, cohort-time effects, event-time effects, and long-run multipliers. Each estimand answers a different question, and the choice among them depends on what you need to learn.

### 2.3.1 Average Treatment Effects

Using the contemporaneous notation  $Y_{it}(d)$  from Section 2.1, the average treatment effect (ATE) is the expected difference between potential outcomes under treatment and control, averaged over units and periods:

$$\text{ATE} = \mathbb{E}_{i,t}[Y_{it}(1) - Y_{it}(0)],$$

where the expectation is taken over the joint distribution of unit-period cells  $(i, t)$  in the population. Formally, this is an average over unit-period cells  $(i, t)$ , so it weights each observed period for each unit equally; if you care about unit-level aggregates (for example, total effect per store over a year), you must define  $Y_i(d)$  and an ATE over units instead. In applications, this typically means averaging over all observed unit-period cells in your sample or over a policy-relevant subset such as treated store-quarter observations. The ATE answers: what would be the average gain if we treated all units in all periods, compared with treating none? In a randomised experiment where every customer receives an email promotion, the ATE tells us the average lift in purchases across all customers and periods under that design.

In most marketing contexts, the ATE is of limited practical interest. Firms cannot or do not treat all units: budgets constrain advertising reach, competitive dynamics preclude universal rollout, and regulation may limit intervention scope. The average treatment effect on the treated (ATT) focuses on the part of the population that actually receives treatment:

$$\text{ATT} = \mathbb{E}_{i,t}[Y_{it}(1) - Y_{it}(0) | D_{it} = 1].$$

The ATT asks: among treated observations  $(i, t)$  with  $D_{it} = 1$ , what is the average causal effect? For the loyalty programme, the ATT tells us how much sales increased, on average, for stores that adopted the programme relative to what their sales would have been without adoption. This ties directly to ROI calculations. If the ATT is £10,000 per store per quarter and the programme costs £3,000 per store per quarter, the programme pays for itself.

### 2.3.2 Cohort–Time Effects in Staggered Adoption

Modern staggered-adoption designs motivate a further refinement. When units adopt treatment at different times  $G_i \in \{1, \dots, T\} \cup \{\infty\}$ , we move from binary potential outcomes to adoption-time indexing. When treatment is absorbing—once a unit adopts, it remains treated—the contemporaneous potential outcomes  $Y_{it}(0)$  and  $Y_{it}(1)$  can be re-indexed by adoption time  $a$ . The potential outcome  $Y_{it}(a)$  denotes the outcome for unit  $i$  in period  $t$  if it had first adopted treatment at time  $a$ , corresponding to the treatment path that is zero before  $a$  and one from  $a$  onwards. Never-treated units have  $G_i = \infty$ , and  $Y_{it}(\infty)$  denotes the potential outcome under perpetual non-treatment. The realised outcome obeys  $Y_{it} = Y_{it}(G_i)$ .

This adoption-time notation is a special case of the path-dependent framework in Section 2.1:  $Y_{it}(a)$  is shorthand for  $Y_{it}(\underline{d}_i^t)$  where  $\underline{d}_i^t = (0, \dots, 0, 1, \dots, 1)$  with the switch from zero to one occurring at period  $a$ . This re-indexing implicitly uses Assumption 1 and, when we work with  $Y_{it}(a)$  rather than the full path  $Y_{it}(\underline{d}_i^t)$ , the no-dynamic-effects restriction in Assumption 4: potential outcomes are determined by the adoption time  $a$  and current calendar period  $t$ , not by finer details of the treatment history before and after  $a$ .

We define the cohort–time average treatment effect on the treated,  $\tau(g, t)$ , as the effect for cohort  $g$  in calendar period  $t$ :

$$\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g], \quad t \geq g.$$

Here  $Y_{it}(\infty)$  represents the potential outcome had the unit never adopted treatment. This estimand allows treatment effects to vary by cohort (early versus late adopters) and by calendar time (common shocks).

Aggregating  $\tau(g, t)$  produces summary measures such as

$$\tau^{\text{ATT}} = \sum_g \sum_{t \geq g} w_{gt} \tau(g, t),$$

where weights  $w_{gt} \geq 0$  satisfy  $\sum_g \sum_{t \geq g} w_{gt} = 1$ , so that  $\tau^{\text{ATT}}$  is a convex combination of cohort–time effects. When  $w_{gt}$  is proportional to the number of treated observations in cell  $(g, t)$ ,  $\tau^{\text{ATT}}$  coincides with the panel-level ATT defined above, restricted to ever-treated units and post-treatment periods. Alternative weighting schemes—such as equal weights across cohorts or restriction to specific post-treatment horizons—yield different aggregations that answer different policy questions. Common choices set  $w_{gt}$  proportional to cohort size  $n_g$  or to the number of treated observations in cell  $(g, t)$ . Chapter 4 shows that traditional two-way fixed-effects regressions do not, in general, recover this estimand when effects are heterogeneous across  $g$  and  $t$ , because they implicitly form a different linear combination of  $\tau(g, t)$ , often with negative weights on some cohort–time cells.

### 2.3.3 Event-Time Effects and Dynamics

Event-time effects trace the dynamic evolution of the treatment response relative to the adoption date. Section 2.2 defined event time as  $k = t - G_i$ , so that  $k = 0$  is the adoption period,  $k < 0$  denotes pre-treatment periods, and  $k > 0$  denotes periods since adoption.

Let  $\theta_k$  denote the average effect  $k$  periods post-treatment:

$$\theta_k = \mathbb{E}[Y_{i,G_i+k}(G_i) - Y_{i,G_i+k}(\infty) \mid G_i < \infty],$$

for event times  $k$  such that  $G_i + k$  lies within the sample window, so  $\theta_k$  averages over all ever-treated units that contribute observations at event time  $k$  within the sample window. The sequence  $\{\theta_k\}$  for  $k \geq 0$  captures dynamic treatment effects such as habit formation or wear-out. For  $k < 0$ ,  $\theta_k$  serves as a diagnostic for pre-trends or anticipation. In Chapters 4 and 5, we show how event-study regressions with event-time indicators  $D_{it}^k$  estimate these  $\theta_k$  under parallel trends and no anticipation, and why naive two-way fixed-effects implementations can deviate from  $\theta_k$  when effects are heterogeneous.

### 2.3.4 Assumption: No Dynamic Effects

The estimands above allow treatment effects to vary by cohort, calendar time, and event time. They do not, by themselves, require current outcomes to depend on past treatments. When dynamics are not the focus, we sometimes invoke the following restriction to simplify the analysis.

**Assumption 4 (No Dynamic Effects)** The potential outcome depends only on current treatment status:

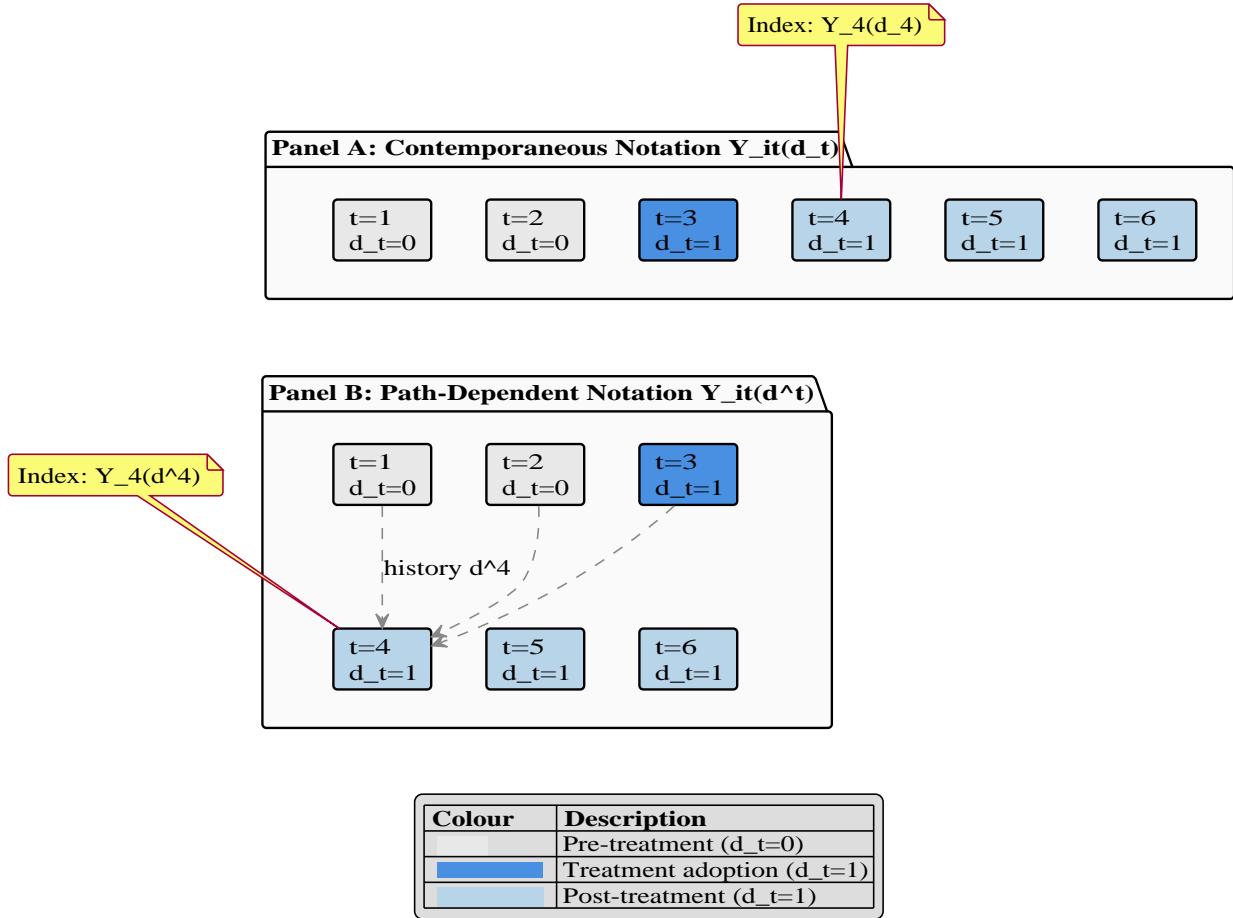
$$Y_{it}(\underline{d}_i^t) = Y_{it}(d),$$

where  $d \in \{0, 1\}$  is the current-period component of the treatment path  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$ .

This assumption rules out carryover and feedback. It says that a store's sales today depend only on whether the loyalty programme is active today, not on how long the store has been in the programme or how it expects the programme to evolve. While restrictive, this assumption simplifies identification arguments for designs such as difference-in-differences. However, when we adopt Assumption 4 we are not only simplifying notation; we are changing the underlying causal model by ruling out any dependence of  $Y_{it}(\underline{d}_i^t)$  on lagged treatments beyond their effect through  $d$ . It is a modelling simplification that is often violated in marketing contexts, not a default theoretical requirement. In this book we therefore treat Assumption 4 as a working simplification for expository purposes and for designs where short-run lifts are the sole focus, not as a default modelling choice for marketing panels. When dynamics are central—as they typically are for advertising, where exposures in previous weeks contribute to current purchases—we relax this assumption and estimate the full impulse-response function (Chapter 10).

Long-run effects matter in marketing because interventions are often intended to have persistent impacts. A loyalty programme aims to permanently increase customer retention, not just produce a temporary sales bump. Advertising seeks to build brand equity that endures beyond the campaign period. To quantify long-run effects, we aggregate event-time effects over a horizon or estimate the cumulative effect of a treatment path.

One common metric is the half-life: the event time  $k^*$  at which  $|\theta_{k^*}| = 0.5 |\theta_0|$ , indicating that half of the initial effect magnitude has dissipated. A short half-life suggests the effect wears off quickly. A long half-life



**Fig. 2.1** Potential-outcomes indexing: contemporaneous versus dynamic path. The figure contrasts contemporaneous notation  $Y_{it}(d)$ , which assumes treatment effects depend only on current treatment status, with path-dependent notation  $Y_{it}(d_i^t)$ , which allows outcomes to depend on the entire treatment history. Marketing applications often require the richer path-dependent framework to capture carryover, habit formation, and strategic dynamics.

suggests persistence. Another metric is the long-run multiplier, which compares the cumulative effect over many periods to the immediate effect:

$$\text{LRM} = \frac{\sum_{k=0}^K \theta_k}{\theta_0},$$

where  $K$  is chosen to be sufficiently long that effects have largely dissipated (often determined by examining when  $\theta_k$  becomes statistically indistinguishable from zero), subject to the constraint that  $G_i + K$  remains within the observed sample window for the cohorts contributing to  $\theta_0$ . The resulting LRM is therefore design- and horizon-specific: different treatment windows or cohorts can yield different long-run multipliers even for the same underlying mechanism.

The LRM is well defined when the immediate effect  $\theta_0$  is non-zero. If  $\text{LRM} = 1$ , the effect is purely contemporaneous with no carryover. If  $\text{LRM} > 1$ , there is positive carryover and the cumulative impact exceeds

the immediate impact. Empirical studies of advertising often find LRMs above one, indicating that campaign effects unfold over several periods. Chapter 10 develops distributed-lag models, vector autoregressions, and structural dynamic panel models that estimate these long-run responses under explicit assumptions about the lag structure and equilibrium behaviour.

These estimands—ATE, ATT, cohort-time effects, event-time effects, and long-run multipliers—provide the vocabulary for specifying what we seek to learn. Identification is not automatic: even with panel data and staggered adoption, we must invoke assumptions that link the observed distribution of  $(Y_{it}, D_{it})$  to the potential-outcomes distribution. The next section introduces those assumptions by examining assignment mechanisms and the identification strategies that connect estimands to data.

**Table 2.1** Core Estimands and Common Aggregations

Estimand	Definition	Interpretation	When Most Relevant
ATE	$\mathbb{E}[Y_{it}(1) - Y_{it}(0)]$	Average effect if all units treated	Randomised experiments; strong overlap
ATT	$\mathbb{E}[Y_{it}(1) - Y_{it}(0)   D_{it} = 1]$	Average effect on treated units	Policy evaluation; ROI calculations
$\tau(g, t)$	$\mathbb{E}[Y_{it}(g) - Y_{it}(\infty)   G_i = g]$	Effect for cohort $g$ in period $t$	Staggered rollouts; heterogeneity
$\theta_k$ (Event-time)	$\mathbb{E}[Y_{i, G_i+k}(G_i) - Y_{i, G_i+k}(\infty)   G_i < \infty]$	Effect $k$ periods post-adoption (averaged over cohorts)	Dynamic effects; pre-trends diagnostics
Long-run multiplier	$\sum_{k=0}^K \theta_k$	Cumulative vs immediate effect	Carryover; habit formation

## 2.4 Assignment Mechanisms and Identification

How did units get treated? Your answer determines which methods you can credibly use and what assumptions you must defend.

Treatment assignment in observational marketing data is rarely random. Firms choose which stores receive loyalty programmes based on expected profitability. Advertisers target campaigns to markets with high anticipated returns. Platforms enter cities based on market size and competitive conditions. These endogenous decisions mean that treated and control units differ systematically, and the observed difference in outcomes conflates the causal effect of treatment with selection bias.

Causal panel data methods address this challenge by combining substantive knowledge of the assignment mechanism with data structures—repeated observations, staggered timing, common shocks—that enable identification under weaker assumptions than cross-sectional settings require. This section classifies common assignment mechanisms and maps them to the identification assumptions that justify particular estimators.

### Randomised Assignment

In the ideal case, treatment is assigned at random. As discussed in Chapter 1, geo-experiments, switchback experiments, and A/B experiments achieve this by randomising treatment across markets, time periods, or users. In such designs, treatment status is independent of potential outcomes by construction, provided that SUTVA (Assumption 3) holds.

**Assumption 5 (Randomisation-Based Unconfoundedness)** In a randomised design, under SUTVA for panels, treatment assignment is independent of potential outcomes. Formally, the assignment mechanism satisfies

$$\Pr(\mathbf{D} \mid \{Y_{it}(\underline{d}_i^t) : \underline{d}_i^t \in \mathcal{D}^t\}_{i,t}, \mathbf{X}) = \Pr(\mathbf{D}),$$

where  $\mathcal{D}^t$  denotes the set of possible treatment histories up to time  $t$ . In stratified designs,  $\Pr(\mathbf{D} \mid \{Y_{it}(\underline{d}_i^t) : \underline{d}_i^t \in \mathcal{D}^t\}_{i,t}, \mathbf{X}) = \Pr(\mathbf{D} \mid \mathbf{X}_{\text{strata}})$ , where  $\mathbf{X}_{\text{strata}}$  collects the stratifying covariates. This implies that marginally, for each  $(i, t)$ ,  $D_{it} \perp (Y_{it}(0), Y_{it}(1))$ .

Randomised assignment is the cleanest route to causal identification. Conditional on the interference structure defined in Assumption 3 being satisfied, the difference in observed outcomes between treated and control units has a causal interpretation without further assumptions about functional forms, parallel trends, or factor structures. In practice, however, non-compliance, attrition, and measurement shifts can still undermine identification even under nominal randomisation, so design diagnostics and tracking of protocol deviations remain essential (see Chapters 17 and 15).

In marketing, randomised assignment typically takes the form of geo-experiments, where DMAs or cities are randomly assigned to treatment and control conditions. A brand might randomly assign 25 DMAs to receive a TV advertising campaign while 25 DMAs serve as controls, then compare sales across conditions.

The randomisation ensures that any difference in post-campaign sales reflects the causal effect of advertising, not pre-existing differences between markets.

Even when experiments are feasible, panel methods remain valuable. They can control for residual pre-treatment imbalances, extend short-term experimental results using observational follow-up, and estimate heterogeneous effects across units and time.

## Staggered Adoption

When randomisation is infeasible, as is often the case in marketing, we must rely on observational variation in treatment timing. Staggered adoption occurs when units adopt treatment at different times. Some units adopt in period 2 (adoption cohort  $g = 2$ ), others in period 5 (adoption cohort  $g = 5$ ), and some never adopt during the observation window ( $g = \infty$ ). Throughout, we use  $g$  to index adoption times; when we later discuss clustered experiments, we use  $c$  to index clusters. This structure is ubiquitous: a retailer rolls out a loyalty programme to batches of stores over multiple quarters, a brand launches advertising campaigns sequentially across markets, a platform enters cities in staggered fashion.

The variation in adoption timing creates opportunities for identification provided that units adopting at different times would have followed parallel trends in the absence of treatment. The parallel trends assumption asserts that differences in outcome trajectories between units with different adoption times can be attributed to treatment rather than to differential pre-existing trends.

**Assumption 6 (Parallel Trends)** For all cohorts  $g, g' \in \{1, \dots, T\} \cup \{\infty\}$  and all periods  $t < \min(g, g')$  (before either cohort is treated),

$$\mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g'],$$

where  $Y_{it}(\infty)$  denotes the potential outcome under never receiving treatment.

In the path-dependent notation of Section 2.1, this is a restriction on the evolution of  $Y_{it}(d_i^t)$  along the never-treated path, requiring that the expected increment  $Y_{it}(\infty) - Y_{i,t-1}(\infty)$  does not depend on the eventual adoption time  $G_i$ . In words, the expected change in untreated potential outcomes from one period to the next is the same across cohorts. Stating the restriction in terms of changes allows cohorts to differ in baseline levels—early adopters might have systematically higher sales than late adopters—while requiring that their growth rates would have been parallel absent treatment. This assumption does not require identical levels or slopes across cohorts, only that the evolution over time, absent treatment, would have been parallel. The restriction applies only to pre-treatment periods  $t < \min(g, g')$ ; after either cohort adopts, outcomes are allowed to diverge arbitrarily because of treatment.

Staggered adoption designs are particularly compelling when the timing of adoption is driven by factors unrelated to the anticipated magnitude of the treatment effect. If the firm rolls out a programme alphabetically by store name, or if a platform enters cities based on operational capacity constraints rather than expected profitability, then parallel trends is more plausible. Chapter 4 develops modern heterogeneity-robust

estimators that aggregate the cohort–time effects  $\tau(g, t)$  defined in Section 2.3, and Chapter 5 discusses event-study specifications that diagnose pre-trends and estimate dynamic effects. Chapter 5 shows how event-study coefficients for  $k < 0$  and placebo adoption dates in pre-treatment periods provide indirect evidence for or against this assumption.

## Single Treated Unit

In some cases, treatment variation is even more limited. When only one unit is treated while all others serve as controls—for example, a platform launching in a single pilot city while comparable cities remain untreated, or a firm implementing a major strategic change in one market—we cannot rely on variation in treatment timing.

Single-unit designs call for constructing a synthetic version of the treated unit from a weighted combination of control units, as in the synthetic control method (Chapter 6). The synthetic control is chosen such that its pre-treatment outcomes closely match the treated unit’s pre-treatment outcomes. If pre-treatment fit is sufficiently close—typically assessed by examining pre-treatment mean squared prediction error—then, under the assumption that no other shocks break this relationship, the synthetic control provides a valid counterfactual for the post-treatment period for the unit-specific effect  $Y_{it}(1) - Y_{it}(0)$  of the treated unit in each post-treatment period. This identification strategy rests on a particular view of untreated outcomes: that they can be well approximated by a weighted combination of control units, which is closely related to the factor-structure assumption discussed below. In factor notation, this corresponds to assuming that the treated unit’s untreated potential outcomes  $Y_{it}(0)$  lie approximately in the span of the control units’ factor loadings  $\{\lambda_{jr}\}$  and common shocks  $\{f_{tr}\}$ .

## Common Shocks and Differential Exposure

A related but distinct setting arises when all units experience a common event in the same period but with varying intensity or exposure. A national advertising campaign might reach different markets with different gross rating points, a regulatory change might affect firms differently based on their characteristics, or a platform algorithm update might impact sellers with different product portfolios to different degrees. These designs exploit cross-sectional variation in exposure intensity, comparing outcomes before and after the shock across units. Formally, these designs rely on either a parallel-trends condition for untreated potential outcomes across exposure levels or a factor-structure assumption for  $Y_{it}(0)$  that allows us to separate common shocks from exposure-specific effects.

Identification again relies on assumptions about untreated potential outcomes. Parallel trends requires that units with different exposure levels would have evolved similarly absent the shock. More flexible structures, such as interactive fixed effects (Chapter 8), allow for heterogeneous responses to common time-varying

factors when strict parallel trends in levels is implausible. We develop these factor-based approaches in detail in Chapter 8.

## Continuous Treatment Intensity

Treatment assignment need not be binary. In many marketing applications, treatment varies in intensity. The TV advertising example from Chapter 1 illustrates this: the brand varies GRPs across markets and weeks, and we want to know how sales respond to changes in advertising intensity. Promotional discount depth, loyalty programme reward generosity, and pricing all take continuous values.

The potential-outcomes framework extends naturally:  $Y_{it}(d)$  denotes the contemporaneous potential outcome under treatment level  $d$ . The causal effect of moving from treatment intensity  $d$  to  $d'$  is  $Y_{it}(d') - Y_{it}(d)$ . For the TV advertising example, we might ask: what is the effect of increasing GRPs from 100 to 150 in a given market-week? The dose-response function traces out how outcomes change across the full range of treatment intensities.

Identification typically relies on a conditional independence assumption.

**Assumption 7 (Unconfoundedness for Continuous Treatment)** Conditional on covariates  $X_{it}$ , unit fixed effects  $\alpha_i$ , and time fixed effects  $\lambda_t$ , the treatment intensity is independent of potential outcomes:

$$D_{it} \perp Y_{it}(d) \mid X_{it}, \alpha_i, \lambda_t \quad \text{for all } d \in \mathcal{D},$$

where  $\mathcal{D}$  denotes the support of treatment intensity. This holds for all periods  $t$  and all treatment levels  $d$  in the support  $\mathcal{D}$ . Identification also requires overlap/positivity: units must have positive probability density for treatment levels in the region of interest.

Equivalently, the conditional distribution of  $D_{it}$  given the full collection of potential outcomes  $\{Y_{it}(d) : d \in \mathcal{D}\}$  and covariates depends only on  $(X_{it}, \alpha_i, \lambda_t)$ , not on the potential outcomes themselves.

This is a strong assumption. It requires that we observe and control all confounders—all factors that jointly affect both advertising spending and sales. Stronger versions rule out feedback from current and future outcomes to current treatment intensity: after conditioning on  $X_{it}$ ,  $\alpha_i$ , and  $\lambda_t$ , there must be no omitted time-varying confounders that cause both treatment intensity and outcomes. In marketing, this assumption is often violated when spend reacts to recent sales performance, which is why instrumental-variable and calibration approaches appear in our MMM and price-elasticity chapters. When treatment reacts mechanically to recent outcomes—for example, when budgets are increased after high-sales weeks—this assumption fails even if we correctly specify the dynamic structure of  $Y_{it}(\underline{d}_i^t)$ , because  $D_{it}$  depends on past realised outcomes, which are functions of past potential outcomes.

Panel settings make this assumption more plausible by including unit and time fixed effects, which control for time-invariant unobservables and common shocks. High-dimensional controls (Chapter 13) and double machine learning (Chapter 12) provide flexible approaches to conditioning on many covariates without im-

posing restrictive functional forms. Dose-response functions, marginal effects, and elasticity estimation for continuous treatments are covered in Chapter 14.

## Combining Identification Strategies and Method Selection

In practice, you rarely rely on a single identification strategy. Difference-in-differences can incorporate high-dimensional controls; synthetic control can be followed by sensitivity analyses; factor models can be combined with staggered adoption to accommodate common shocks and timing. The key is to match the design to the variation that identifies the effect and to be explicit about the assumptions that make this variation credible.

### Parallel Trends Revisited

Parallel trends (Assumption 6) asserts that treated and control units would have evolved similarly absent treatment. It can be stated conditionally (after covariate adjustment) or unconditionally. Conditional versions are often more plausible; unconditional versions require fewer modelling choices. Although untestable directly, pre-treatment fit and placebo checks (Chapter 17) provide indirect evidence.

Difference-in-differences and event studies (Chapters 4, 5) rely on this assumption. In event-time notation, this corresponds to requiring that pre-treatment event-time effects  $\theta_k$  are zero for  $k < 0$ , up to sampling noise. Recent sensitivity analyses [Rambachan and Roth, 2023] quantify how large a deviation from parallel trends must be to overturn estimates of ATT or  $\{\theta_k\}$ , providing a disciplined way to report robustness rather than relying on informal pre-trend plots alone.

### Factor Structure

Factor structure provides an alternative when parallel trends in levels is implausible but units are subject to common time-varying shocks that affect them differentially. While parallel trends requires that treated and control units evolve similarly, factor models allow for heterogeneous responses to common shocks. Interactive fixed effects models (Chapter 8) posit that untreated potential outcomes can be decomposed as

$$Y_{it}(0) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

where  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects,  $f_{tr}$  are latent factors common to all units, and  $\lambda_{ir}$  are unit-specific loadings. This structure accommodates differential exposure to common shocks—such as seasonality, macroeconomic trends, or industry-wide demand shifts—without requiring parallel trends. Identification then requires that, after conditioning on observed covariates and the latent factors  $\{f_{tr}\}$ , the idiosyncratic component  $\varepsilon_{it}$  is uncorrelated with treatment assignment, so that the remaining variation in  $D_{it}$  is as-good-as-random with respect to  $Y_{it}(0)$ . Estimation proceeds via principal components, expectation-

maximisation algorithms, or nuclear norm regularisation. Factor models are particularly effective in marketing panels where all stores are affected by category demand shocks, all markets experience national advertising campaigns, or all platforms face common technological changes.

### Unconfoundedness with High-Dimensional Controls

Unconfoundedness with controls applies in settings where treatment intensity varies continuously or where treatment is targeted based on observable characteristics. Conditional independence assumptions can justify causal inference when we control for a sufficiently rich set of covariates—demographics, past outcomes, competitor actions, seasonal indicators—such that treatment assignment is as good as random conditional on these controls. Assumptions of this type generalise Assumption 7 beyond scalar treatments to richer treatment vectors and targeting rules.

This is a strong assumption: it requires that there are no unobserved confounders that jointly affect treatment and outcomes. Panel structures strengthen this strategy by allowing unit fixed effects (controlling for time-invariant confounders) and time fixed effects (controlling for common shocks), reducing the remaining scope for unobserved, time-varying unit-specific confounders. When the covariate set is high-dimensional, regularisation methods such as lasso (Chapter 13) can select relevant controls without overfitting, and double machine learning (Chapter 12) enables valid inference even when the selection and outcome models are estimated flexibly.

### Interference-Aware Designs

Interference-aware designs become necessary when SUTVA (Assumption 3) is violated. Identification requires either cluster randomisation or explicit modelling of spillovers. Cluster designs assign treatment to groups of units, internalising spillovers within clusters while maintaining independence across clusters. When randomisation is infeasible, spatial models and exposure mappings estimate direct and spillover effects jointly, given network or geographic structure. These designs move from potential outcomes  $Y_{it}(d)$  to  $Y_{it}(d, h_i(D_{-i,t}))$ , where  $h_i(\cdot)$  summarises neighbours' treatments, and identification hinges on assumptions about both own-treatment and spillover assignment. Chapter 11 also discusses partial identification when the spillover structure is uncertain.

### Method Selection Summary

Method choice depends on the structure of treatment variation, the plausibility of assumptions, and available diagnostics. With staggered timing and plausible parallel trends, use modern DiD (Chapters 4, 5). With one or few treated units, use synthetic control (Chapters 6, 7). With continuous treatments and rich covariates, use unconfoundedness with high-dimensional controls (Chapters 13, 12). With spillovers, use interference-aware designs (Chapter 11). When such unconfoundedness assumptions are implausible, instrumental-variable

strategies become necessary; we return to those in later chapters. Comparing multiple approaches, and showing that conclusions do not hinge on a single design, is often most persuasive.

Credible causal inference requires not just sophisticated estimators but also transparent articulation of assumptions and rigorous diagnostics to assess their plausibility. This book prioritises identification arguments and design logic, then turns to estimation mechanics once the assignment mechanism and estimand are clear.

Taken together, the estimands in Section 2.3 and the assignment mechanisms here form the backbone of all the designs we study in the rest of the book. With estimands defined and identification strategies mapped, we now turn to the mechanics: how standard panel regressions implement these ideas, and where they can go wrong.

## 2.5 Regression Mechanics and Inference in Panels

The two-way fixed effects regression is the workhorse of applied panel analysis. It is also frequently misused. This section reviews when TWFE works, when it fails, and what inference issues you must address. Many of the methods developed in later chapters build on or react against this baseline, so understanding its strengths and limitations provides essential context.

### Within Estimation and Fixed Effects

The canonical fixed effects regression for panel data takes the form

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + X'_{it} \gamma + \varepsilon_{it},$$

where  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects,  $D_{it}$  is the treatment indicator (or continuous treatment intensity),  $X_{it}$  are time-varying covariates, and  $\varepsilon_{it}$  is an error term. Consistency of the within estimator requires **strict exogeneity**:

$$\mathbb{E}[\varepsilon_{it} | D_{i1}, \dots, D_{iT}, X_{i1}, \dots, X_{iT}, \alpha_i] = 0 \quad \text{for all } t,$$

which conditions on the *entire* treatment and covariate path, not just contemporaneous values. This is a panel analogue of “no omitted time-varying confounders” and “no feedback”: after conditioning on the full treatment and covariate history and unit effects, current errors are mean independent of past, current, and future treatments and covariates. In the potential-outcomes notation of Section 2.1, this condition provides a model-based orthogonality condition that justifies treating the within-unit variation in  $D_{it}$  as as-good-as-random with respect to  $Y_{it}(d)$  once we control for  $X_{it}$ ,  $\alpha_i$ , and  $\lambda_t$ . In observational marketing panels, this condition is often violated unless a design—randomisation, staggered adoption with credible timing, or exogenous shocks—justifies it. The coefficient  $\tau$  is the parameter of interest, interpreted as the average effect of treatment on the outcome when treatment effects are constant across units and time. Under Assumptions 1, 4, and homogeneous treatment effects (Assumption 8),  $\tau$  coincides with a panel-level average treatment effect such as the ATT defined in Section 2.3, taken over the treated unit-period cells.

Unit fixed effects  $\alpha_i$  control for time-invariant differences across units. A store with persistently high sales due to a prime location, a strong manager, or loyal customer base has a large  $\alpha_i$ , and this component is absorbed by the unit fixed effect rather than confounding the estimate of  $\tau$ . Time fixed effects  $\lambda_t$  control for common shocks that affect all units in a given period: seasonal demand, macroeconomic conditions, national advertising campaigns, and industry-wide trends. With both unit and time fixed effects, the regression identifies  $\tau$  from within-unit, over-time variation in treatment that remains after removing unit means and period means.

We obtain the within estimator—also called the fixed effects estimator—by applying the two-way within transformation:

$$\ddot{Y}_{it} = Y_{it} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot t} + \bar{Y}_{\cdot\cdot},$$

where  $\bar{Y}_{i\cdot} = T^{-1} \sum_t Y_{it}$  is the unit mean,  $\bar{Y}_{\cdot t} = N^{-1} \sum_i Y_{it}$  is the time mean, and  $\bar{Y}_{\cdot\cdot} = (NT)^{-1} \sum_i \sum_t Y_{it}$  is the grand mean. The same transformation applies to  $D_{it}$  and  $X_{it}$ . OLS on the transformed data yields the within estimator. If  $\varepsilon_{it}$  is uncorrelated with the demeaned treatments and covariates, the within estimator is consistent for  $\tau$  as  $N \rightarrow \infty$  with fixed  $T$ , or as both  $N$  and  $T$  grow. The within estimator is numerically equivalent to including dummy variables for all units and all time periods (omitting one of each to avoid collinearity) and running OLS, but the demeaning formulation clarifies the identifying variation.

### Pitfalls of Two-Way Fixed Effects with Heterogeneity and Staggered Timing

Despite its simplicity and widespread use, the two-way fixed effects (TWFE) regression can produce misleading estimates when treatment effects are heterogeneous and treatment adoption is staggered. TWFE implicitly assumes homogeneous effects and, under staggered adoption, relies on parallel trends (Assumption 6).

**Assumption 8 (Homogeneous Treatment Effects)** The treatment effect is constant across units and time:

$$\tau_{it} = \tau \quad \text{for all } i, t.$$

Equivalently, the cohort-time effects  $\tau(g, t)$  from Section 2.3 are all equal to the same constant  $\tau$ .

When this assumption fails, TWFE no longer estimates the cohort-time effects  $\tau(g, t)$  defined in Section 2.3, nor any transparent convex combination of them. Instead, the TWFE coefficient is a particular weighted average of many underlying two-by-two comparisons, and some of these comparisons use already-treated units as controls for newly treated units. When treatment effects grow over time or differ across cohorts, these comparisons can receive negative weights, so the aggregate estimate can be biased toward zero, away from zero, or even have the wrong sign. Chapter 4 formally decomposes the TWFE coefficient into a weighted sum of these two-by-two comparisons and develops heterogeneity-robust alternatives.

To fix ideas, consider a setting where units adopt treatment at different times and the treatment effect grows with time since adoption. A unit that adopted early has a large effect; a unit that adopted recently has a small effect. A TWFE regression that compares a recently treated unit to an early-treated unit attributes the outcome difference to the recent treatment, but this comparison confounds the small effect of recent treatment with the large effect of earlier treatment. Formally, these problematic comparisons enter the TWFE coefficient with weights that can be negative, so that periods and cohorts with larger true effects can receive negative weight in the aggregate estimate; Chapter 4 spells out this decomposition. Modern heterogeneity-robust estimators (Chapter 4) avoid this problem by constructing comparisons that only contrast treated units with never-treated or not-yet-treated units and by aggregating  $\tau(g, t)$  with known, non-negative weights.

Despite these pitfalls, TWFE remains useful as a benchmark and in settings where treatment effects are approximately constant across units and time. It is also computationally simple and scales to large datasets. The key is to understand when TWFE can be trusted and when more sophisticated methods are required. Event-study specifications with leads and lags of treatment (Chapter 5) provide an informative

diagnostic: substantial pre-treatment coefficients or erratic post-treatment patterns signal that heterogeneity and staggered timing are confounding the TWFE estimate. In the notation of Section 2.3, these event-study coefficients aim to recover the event-time effects  $\theta_k$ ; large or systematic deviations from zero for  $k < 0$  or erratic patterns for  $k \geq 0$  signal that the constant-effect and parallel-trends assumptions underpinning the TWFE estimate are not credible.

## Random Effects and the Mundlak Device

An alternative to fixed effects is the random effects model, which treats  $\alpha_i$  as a random variable drawn from a distribution with mean zero and variance  $\sigma_\alpha^2$ . The random effects estimator uses generalised least squares (GLS) to exploit both within-unit and between-unit variation, weighting each source of variation according to the relative magnitudes of  $\sigma_\alpha^2$  and the error variance. Random effects can produce smaller standard errors than fixed effects when their assumptions hold, but they require  $\alpha_i$  to be uncorrelated with the regressors—often implausible in marketing settings where treatment assignment depends on unit characteristics. That is,  $\mathbb{E}[\alpha_i | D_{i1}, \dots, D_{iT}, X_{i1}, \dots, X_{iT}] = 0$  and  $\mathbb{E}[\varepsilon_{it} | D_{i1}, \dots, D_{iT}, X_{i1}, \dots, X_{iT}, \alpha_i] = 0$  for all  $t$ .

The Mundlak device offers a middle ground. By including unit-level means of the time-varying covariates as additional regressors,  $\bar{X}_i = T^{-1} \sum_t X_{it}$ , the Mundlak-augmented regression allows  $\alpha_i$  to be correlated with the regressors in a structured way while retaining the variance components interpretation. When unit means are included for all time-varying regressors that may be correlated with  $\alpha_i$ , the coefficient on  $D_{it}$  has a within-unit interpretation equivalent to fixed effects—it is identified from within-unit variation—but the model can accommodate random slopes and other sources of heterogeneity while remaining computationally efficient. In practice, modern marketing applications rarely rely on pure random effects, preferring the transparency of fixed effects or Mundlak-style correlated random effects, or the flexibility of methods that allow for richer forms of heterogeneity.

## Correlated Random Effects IV: Hausman–Taylor

When time-invariant regressors matter and unit effects are correlated with covariates, the Hausman–Taylor (HT) estimator provides a bridge between fixed effects and random effects. Time-varying regressors are partitioned into those plausibly exogenous with respect to unit effects and those potentially correlated; time-invariant regressors are partitioned in the same way. HT uses variation within units in the exogenous time-varying regressors as instruments for endogenous time-varying regressors, and unit means of exogenous time-varying regressors as instruments for endogenous time-invariant regressors [Hausman and Taylor, 1981]. Under these orthogonality conditions and a variance components structure, HT recovers coefficients on time-invariant regressors while allowing correlation between unit effects and a subset of regressors. In potential-outcomes terms, these orthogonality conditions assert that the designated "exogenous" regressors and their

unit means are uncorrelated with  $\alpha_i$  and with the idiosyncratic errors that remain after conditioning on  $D_{it}$  and the full regressor set.

HT is most compelling when you have credible a priori splits of regressors into exogenous and endogenous blocks, time-invariant regressors are substantively central, and the resulting internal instruments are strong. Diagnostics mirror standard IV practice: test for weak instruments, report overidentification tests, and probe sensitivity to alternative partitions. Mundlak's device remains a transparent alternative when you prefer to treat correlation with unit effects as captured by including unit means, rather than relying on internal instruments [Mundlak, 1978, Amemiya and MacCurdy, 1986]. In practice, Hausman–Taylor and related correlated random-effects IV methods are niche tools in the overall toolbox; for most of the book, fixed effects or design-based estimators are preferred because their assumptions are more transparent and more often defensible in marketing contexts.

## Serial Correlation and Clustering

Panel data violate the independence assumption that underlies classical standard error formulas, even after conditioning on covariates and fixed effects. Outcomes for the same unit observed in different periods are typically correlated because of persistent unobservables, autocorrelated shocks, or dynamic feedback. Ignoring this serial correlation leads to standard errors that are too small and hypothesis tests that reject too often.

We cluster standard errors by unit, allowing for arbitrary correlation of  $\varepsilon_{it}$  within units while maintaining independence across units—a requirement that connects to the no-interference component of SUTVA (Assumption 3). When clustering by unit, the number of clusters equals the number of units ( $G = N$ ). Formally, unit-level clustering assumes that, after conditioning on covariates and fixed effects, errors are independent across units; cross-unit correlation in  $\varepsilon_{it}$  induced by spillovers or common shocks would violate this assumption and call for cluster definitions or modelling choices that reflect the interference structure. Clustering by unit is appropriate when the primary source of correlation is persistence within units over time – the typical case in marketing panels where store characteristics, customer preferences, or market conditions create serial dependence. Two-way clustering, by both unit and time, becomes necessary when errors are also correlated across units within a period – for example, if all stores in a region are affected by a regional shock, or if all markets experience a common demand shock in a particular quarter. The cost of two-way clustering is larger standard errors and potential over-correction when cross-unit correlation is weak. Multi-way clustering can be computationally intensive, but packages implementing the method are widely available.

When the number of clusters is small—fewer than about 50—cluster-robust standard errors may be unreliable because the asymptotic approximation breaks down. In such cases, the wild cluster bootstrap [Cameron et al., 2008, Roodman et al., 2019] provides a more accurate approach to inference by resampling entire clusters; see Chapter 16 for practical guidance on wild bootstrap implementation in small- $G$  panels. Permutation tests [Young, 2019], which do not rely on asymptotic approximations, offer another route to valid inference when the number of treated units or clusters is small. With very few clusters (for example,  $G < 20$ ), con-

ventional cluster-robust standard errors can severely understate uncertainty; simulation and wild-bootstrap diagnostics are essential before drawing strong conclusions from nominal  $p$ -values.

For panels with long time series ( $T$  large), heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors that account for serial correlation up to a lag length chosen by the researcher provide an alternative to clustering. HAC estimators are particularly relevant in fat panels where the time dimension dominates and where clustering by unit would produce very few clusters—for example, brand- or country-level panels with very long  $T$  and very small  $N$ . Clustering dominates in store- and customer-level panels where  $N$  is large. Newey–West standard errors, which down-weight covariances at longer lags, are a common choice. The key trade-off is that HAC methods require choosing a lag length (often based on rules of thumb like  $T^{1/4}$ ), while clustering makes no such assumption but requires a sufficient number of clusters for asymptotic validity. Chapter 16 provides detailed guidance on choosing between clustering, HAC, and bootstrap methods depending on panel dimensions and error structure.

## Stationarity, Trends, and Spurious Regression in Panels

Panel regressions often pool many short time series. Before discussing diagnostics and remedies, we must clarify what nonstationarity means. A process is nonstationary if its joint distribution—or, under weaker definitions, its mean, variance, or autocovariance structure—depends on time. This definition covers several distinct phenomena. Deterministic trends arise when a process follows a stable linear path over time with i.i.d. errors, so that the mean changes deterministically with  $t$ . Heteroskedasticity over time occurs when the variance of the innovations changes with  $t$ . Structural breaks arise when parameters shift at known or unknown dates. Unit roots generate stochastic trends: in an  $I(1)$  process, such as  $y_t = y_{t-1} + \varepsilon_t$ , shocks accumulate over time so that the level does not revert.

The first three forms of nonstationarity are typically well behaved for inference once we adjust for the changing mean or variance. Unit roots are different: they require functional central limit theorems and specialised inference procedures. The distinction matters because different forms of nonstationarity call for different remedies. In this book we treat these time-series remedies as complements to, not substitutes for, design-based identification: robust designs and diagnostics are the first line of defence against spurious regression, with formal time-series tools used when long-run dynamics are central.

With short  $T$  (say,  $T = 12$  or  $T = 24$ ), unit root concerns are secondary. We cannot reliably test for or estimate the properties of unit roots with so few time periods. The primary concern in short panels is structural breaks and parameter drift—the data-generating process changes because the marketing environment changes (platform updates, privacy regulations, competitive entry), not because of accumulating stochastic shocks. Time fixed effects address common shocks but not unit-specific instability.

With longer  $T$  or macro-style outcomes, stochastic trends matter. Time fixed effects remove shocks and trends that are common across units, but they do not remove unit-specific stochastic trends. In a state-by-year panel, year dummies absorb the aggregate cycle, yet each state can drift as an  $I(1)$  process. Regressing

one  $I(1)$  variable on another without cointegration risks spurious relationships—high  $R^2$  and significant  $t$ -statistics even when the variables are unrelated.

Practical guidance follows a simple sequence. When  $T$  is short and identification uses within-unit, high-frequency variation (policy dummies, staggered rollouts), fixed effects with time effects are typically adequate; the main threat is structural breaks, not unit roots. When  $T$  is longer or variables plausibly follow  $I(1)$  processes (macro aggregates, prices, levels), test for unit roots with panel methods that allow heterogeneity and cross-section dependence (LLC, IPS, CIPS) and consider unit-specific trends. If variables are  $I(1)$  but cointegrated, estimate cointegrating relationships and use error-correction specifications; if not cointegrated, difference or transform to stationarity and interpret coefficients accordingly. Differencing removes unit effects and unit-specific stochastic trends, but it also removes long-run level information and does not address structural breaks; dynamic specifications in Chapter 10 discuss these trade-offs.

Appendix B summarises panel unit root and cointegration tests and offers a brief decision guide. The time-series literature emphasises how hard it is, in finite samples, to distinguish trend-stationary from difference-stationary behaviour. From a potential-outcomes perspective, nonstationarity means that the distribution of  $Y_{it}(d)$  can drift over time even along a fixed treatment path, so identification strategies that rely on stable untreated trajectories (such as parallel trends) become more fragile as structural change accumulates. These difficulties reinforce the case for using design-based identification and transparent diagnostics as the primary safeguards against spurious regression, with formal unit-root and cointegration tests as secondary tools when  $T$  is long and macro-style dynamics are central.

## 2.6 Crosswalk to Method Chapters

You have data and a causal question. Which chapter should you read next? This section pulls together the ingredients from earlier in the book and provides a crosswalk that maps your situation to the methods that can help, clarifying which estimators to deploy for different data structures and identification strategies.

When the goal is to estimate the ATT or cohort-time effects  $\tau(g, t)$  in a staggered adoption design with binary treatment, difference-in-differences estimators (Chapter 4) are the natural choice. Modern heterogeneity-robust estimators—such as those of Callaway and Sant’Anna [2021] and Sun and Abraham [2021]—avoid the negative weighting problems of TWFE and provide transparent aggregations of  $\tau(g, t)$  into summary statistics. Event-study specifications (Chapter 5) extend the logic by estimating effects as a function of event time  $k$ , enabling visualisation of pre-trends and dynamic adjustment paths. Both approaches rely on parallel trends (Assumption 6) and, for dynamic interpretations, on no anticipation (Assumption 1). These assumptions can be examined indirectly through pre-treatment diagnostics (Chapter 17) and subjected to sensitivity analysis.

When treatment variation is more limited—only one or a few units are treated—synthetic control methods (Chapter 6) construct counterfactuals by forming weighted combinations of control units that closely match pre-treatment outcomes. In terms of estimands, synthetic control targets unit-specific effects  $Y_{it}(1) - Y_{it}(0)$  for the treated unit in each post-treatment period, rather than a population-level ATT. Synthetic control does not impose parallel trends in levels; instead, it relies on the idea that if the synthetic control tracks the treated unit closely before treatment, then under a stable low-rank factor structure for untreated potential outcomes  $Y_{it}(0)$  (see Chapter 8), the synthetic control provides a valid counterfactual afterwards. Inference relies on placebo checks and permutation methods that compare the estimated effect to the distribution of placebo effects in control units. Hybrid methods (Chapter 7), particularly synthetic difference-in-differences (SDID), combine the strengths of synthetic control and difference-in-differences by weighting both units and time periods, often outperforming either method alone when their assumptions approximately hold. SDID still relies on parallel-trends-type assumptions in the reweighted sample, but its explicit weighting scheme often makes those assumptions more plausible and diagnosable.

An alternative relaxation of parallel trends comes from factor models and matrix completion methods. Interactive fixed effects models (Chapter 8) posit that outcomes are driven by a small number of latent factors, and estimation proceeds via iterative least squares. Matrix completion methods (Chapter 9) extend this logic to settings with missing data, using nuclear norm regularisation to impute counterfactual outcomes. Factor and matrix-completion methods typically aim to recover counterfactual paths  $Y_{it}(0)$  for treated units in treated periods and then aggregate these into cohort-time effects  $\tau(g, t)$  or event-time effects  $\theta_k$ . Factor methods are particularly valuable in marketing panels where category-level demand shocks, national trends, or platform-wide changes affect all units but with heterogeneous intensity.

Beyond static treatment effects, dynamics and spillovers require specialised treatment. When the No Dynamic Effects assumption (Assumption 4) fails, Chapter 10 develops distributed lag models, vector autoregressions, and structural dynamic panel methods that estimate how treatment effects unfold over time, accumulate through carryover, or decay through depreciation. When SUTVA (Assumption 3) is violated, Chapter 11 tackles interference, distinguishing direct effects from spillover effects and covering spatial econometric models, exposure mappings, and interference-robust experimental designs. In terms of Section 2.3,

these chapters move beyond static ATE and ATT to recover dynamic profiles  $\{\theta_k\}$  and spillover-aware estimands that depend on exposure mappings  $h_i(D_{-i,t})$ . Both chapters move beyond the purely static potential outcomes framework to embrace path-dependent outcomes and network interactions.

Machine learning integration is the focus of Chapters 12 and 13. Double machine learning (Chapter 12) uses flexible algorithms—random forests, boosting, neural networks, and others—to estimate nuisance functions such as propensity scores and outcome regressions, while preserving valid inference on causal parameters through Neyman orthogonalisation and cross-fitting. These methods do not relax the underlying identification assumptions (parallel trends, unconfoundedness, factor structure); they only relax functional-form assumptions for the nuisance components that enter those identification arguments. These approaches exemplify the synthesis of Breiman’s two cultures discussed in Section 1.5: flexible algorithms for nuisance estimation, deployed inside a disciplined design-based causal framework. Causal forests enable estimation of heterogeneous treatment effects, revealing which observable features moderate the treatment response. High-dimensional variable selection (Chapter 13) employs lasso and related regularisation methods to control for many potential confounders without overfitting, and post-selection inference methods provide confidence intervals that account for the data-driven selection process.

Inference and diagnostics are essential complements to estimation. Chapter 16 develops tools for uncertainty quantification, covering cluster-robust standard errors, wild bootstrap procedures, randomisation inference, conformal prediction intervals, and multiple testing adjustments when examining effects across many products, markets, or periods. This is especially important when reporting entire event-time profiles  $\{\theta_k\}$  or large collections of cohort–time effects  $\{\tau(g, t)\}$ , where naive pointwise intervals can dramatically overstate the evidence for dynamic patterns. Chapter 17 provides a practical diagnostic workflow for pre-treatment covariate balance checks, placebo checks, leave-one-out robustness analyses, sensitivity analyses for violations of parallel trends, and specification curves that aggregate results across many defensible modelling choices. Chapter 15 catalogues threats to validity specific to marketing panels, including seasonality, platform algorithm changes, and measurement error. Together, these chapters emphasise that credible causal inference requires not just sophisticated estimation but also rigorous diagnostics and transparent acknowledgment of assumptions.

Finally, continuous and multivalued treatments, nonlinear panel models, and dose–response estimation are covered in Chapter 14, extending the binary-treatment framework to settings with continuous intensity or discrete and censored outcomes.

These methods are not mutually exclusive. An analysis might employ difference-in-differences as the primary specification, synthetic control as a robustness check, factor models to relax parallel trends, and machine learning to uncover heterogeneity. Table 2.2 provides a starting point for method selection, but practitioners should consider multiple approaches when the data permit.

Taken together with the estimands in Section 2.3 and the assignment mechanisms in Section 2.4, this crosswalk completes the foundational mapping from data and design to appropriate methods. By mapping estimands to methods, this crosswalk clarifies that the choice of estimator is not arbitrary but follows logically from the structure of the data, the nature of the treatment assignment, and the assumptions you are willing to defend. The goal is not to find a single “best” method but to deploy methods that align with the identifying variation and to subject the resulting estimates to a battery of diagnostics that probe their robustness.

**Table 2.2** Crosswalk: Data Structures and Primary Methods

Data Structure	Key Feature	Primary Methods	Key Requirement
Randomised experiment	Random assignment	Geo-experiments, A/B experiments (Ch. 3)	Design-based unconfoundedness (Assump. 5); restricted interference (cluster-level SUTVA, Assump. 3); no differential attrition
Staggered adoption, binary	Variation in timing	DiD (Ch. 4), Event studies (Ch. 5)	Parallel trends (Assump. 6)
Single/few treated units	Limited variation	Synthetic control (Ch. 6), SDID (Ch. 7)	Good pre-treatment fit under a stable low-rank factor structure; treated unit is well-approximated by a convex combination of donors
Common shocks with heterogeneous exposure	Factor structure	IFE (Ch. 8), Matrix completion (Ch. 9)	Low-rank factor structure (common shocks with heterogeneous loadings)
Continuous treatment	Dose-response	DML (Ch. 12), High-dim controls (Ch. 13), Continuous treatments (Ch. 14)	Unconfoundedness (Assump. 7)
Spillovers present	SUTVA (Assump. 3) violated	Spillover models (Ch. 11)	Network/geo structure and exposure mapping for interference
Dynamic effects	No Dynamic Effects (Assump. 4) fails	Distributed lags (Ch. 10), Event studies (Ch. 5)	Lag specification; no anticipation (Assump. 1); no pre-trends; diagnostic lead coefficients near zero for $k < 0$

## 2.7 Running Examples: Motivating Vignettes

Theory without application is of limited interest to an applied practitioner. This section revisits the three examples from Chapter 1, now equipped with the precise language of estimands, assumptions, and identification strategies. These vignettes recur throughout the book as we apply each method.

Together, the loyalty programme, TV advertising, and platform entry examples span the main panel shapes and identification challenges in this book. The loyalty programme illustrates a thin panel with many stores observed over a modest number of quarters and a binary staggered treatment. The TV advertising example illustrates a fat panel with a small number of DMAs observed over many weeks and a continuous treatment intensity. The platform entry example illustrates a square panel with roughly similar numbers of cities and months and a binary staggered treatment with rich selection concerns.

### Loyalty Programme with Staggered Rollout

A retail chain operates 500 stores observed over 12 quarters. The chain launches a loyalty programme in a staggered fashion: 100 stores in quarter 3, 200 stores in quarter 5, 150 stores in quarter 8, and 50 stores never receive the programme during the sample period. The data form a balanced panel with  $N = 500$  and  $T = 12$ , a thin panel where the number of units substantially exceeds the number of periods. Outcomes  $Y_{it}$  are quarterly sales per store. Treatment  $D_{it}$  is binary, equal to one if store  $i$  has an active loyalty programme in quarter  $t$  and zero otherwise. Covariates  $X_{it}$  include store characteristics (size, location demographics, competitive intensity) and time-varying controls (local unemployment rate, competitor actions, holiday indicators).

The primary estimand is the panel-level ATT over treated store-quarter cells, as defined in Section 2.3. Given the staggered adoption structure, cohort-time effects  $\tau(g, t)$  and event-time effects  $\theta_k$  are also of interest, particularly to understand whether the programme effect grows over time as customers accumulate points and develop habits. Heterogeneity is anticipated: the programme may work well in affluent, low-competition areas but have negligible effects in saturated urban markets.

Identification relies on parallel trends (Assumption 6): in the absence of the programme, treated and control stores would have experienced similar trends in sales. This assumption can be probed and stress-tested with pre-trend diagnostics that check whether stores adopting in quarter 3 exhibit parallel trends with never-treated stores in quarters 1 and 2. Spillovers complicate identification: customers may refer friends, creating positive spillovers, or cross-shop from nearby non-programme stores, creating geographic externalities. These spillovers violate SUTVA (Assumption 3) and require explicit modelling through exposure mappings  $h_i(D_{-i,t})$  that capture neighbours' treatment status (Chapter 11). Dynamic effects mean that  $Y_{it}$  may depend on the entire treatment path  $d_i^t$ , not just  $D_{it}$ , so event-study specifications (Chapter 5) that estimate  $\theta_k$  for  $k = 0, 1, 2, \dots$  are essential. In short, a difference-in-differences design for this example relies on Assumptions 6, 1, and 3, with spillover models in Chapter 11 relaxing the last when geographic interference is important. Chapter 17 details how to implement these pre-trend checks, placebo checks, and overlap diagnostics for staggered rollouts.

Appropriate methods include staggered difference-in-differences (Chapter 4), event studies (Chapter 5), spillover models (Chapter 11), and machine-learning tools such as causal forests for treatment-effect heterogeneity (Chapter 12).

## Television Advertising Carryover Across Markets

A consumer packaged goods brand tracks sales in 50 designated market areas (DMAs) over 100 weeks. The brand varies television advertising intensity, measured in gross rating points (GRPs), across markets and weeks according to strategic considerations: higher spending during product launches, in markets with strong past performance, and in response to competitive activity. The data form a panel with  $N = 50$  and  $T = 100$ , a fat panel where the time dimension substantially exceeds the cross-sectional dimension. Treatment  $D_{it}$  is continuous (GRPs purchased), and outcomes  $Y_{it}$  are weekly sales. Covariates include online search volume, social media mentions, competitor GRPs, local economic indicators, and seasonal indicators.

The primary estimand is the dose-response function: how sales respond to changes in TV GRPs, accounting for carryover effects that allow past advertising to influence current sales. In the notation of Section 2.3, the average dose-response function  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  and its derivative with respect to  $d$  (an elasticity-style marginal effect) are the main targets. The long-run multiplier and half-life of advertising effects are functions of the distributed-lag profile  $\{\gamma_s\}$  (or an estimated impulse response), which summarises how current sales respond to GRPs purchased  $s$  weeks earlier. These dynamic measures quantify whether TV advertising has persistent effects or whether the impact dissipates quickly. Heterogeneity across markets—urban versus rural, high-income versus low-income, competitive versus concentrated—is also of interest.

Identification is challenging because TV spending is endogenous: the brand increases advertising when it anticipates high demand or when competitors are active. Parallel trends is less natural with continuous treatment, but unconfoundedness (Assumption 7)—conditional on observables  $X_{it}$  and unit and time fixed effects, TV spending is as good as random—can justify causal inference if the controls are sufficiently rich. High-dimensional controls (Chapter 13) using lasso or double machine learning (Chapter 12) can flexibly adjust for many potential confounders without overfitting. Alternatively, factor models (Chapters 8 and 9) can capture common demand shocks that affect all markets, allowing identification from deviations of individual markets’ advertising from the common trend. A design-based analysis of this example therefore rests on either Assumption 7 with rich  $X_{it}$  and fixed effects, or on a low-rank factor structure for  $Y_{it}(0)$  that justifies using deviations from common shocks for identification.

Dynamic effects are central. A distributed lag model (Chapter 10) specifies

$$Y_{it} = \alpha_i + \lambda_t + \sum_{s=0}^S \gamma_s D_{i,t-s} + X'_{it} \phi + \varepsilon_{it},$$

so that the sequence  $\{\gamma_s\}$  traces the effect of GRPs purchased  $s$  weeks ago on current sales. The cumulative effect  $\sum_{s=0}^S \gamma_s$  quantifies the total impact over  $S$  weeks, and the half-life is the lag  $s^*$  where  $\sum_{s=0}^{s^*} \gamma_s$  reaches half of the cumulative effect. Synthetic control methods (Chapter 6) provide an alternative design perspective

when a particular market experiences a discrete advertising shock, by constructing a synthetic market from control markets with low GRPs that match on pre-treatment sales trends and covariates.

Measurement issues are salient. TV ratings are based on samples, scanner sales data have coverage gaps, and attribution of sales to TV advertising is confounded by online search and social media activity. Chapter 19 discusses measurement challenges and how to bound the bias introduced by imperfect data. Inference (Chapter 16) must account for serial correlation within markets and potential cross-market correlation, pointing to two-way clustering or, given the long time series, HAC standard errors. Chapters 16 and 17 describe how to implement these serial-correlation corrections and design diagnostics in fat-panel settings.

## Platform Market Entry and Competitive Dynamics

A food delivery platform enters 30 cities over a two-year period, with entry times staggered based on market size, regulatory environment, and operational capacity. The platform observes monthly restaurant revenues in 50 cities (30 treated, 20 never-treated) over 24 months, yielding a panel with  $N = 50$  and  $T = 24$ , a square panel with roughly balanced dimensions. Treatment  $D_{it}$  is binary (platform present in city  $i$  in month  $t$ ). Outcomes  $Y_{it}$  are aggregate restaurant revenues at the city level. Covariates include city characteristics (population, income, density, number of restaurants), time-varying shocks (local economic indicators, pandemic phases, policy changes), and competitor actions (incumbent platform presence, pricing strategies).

The primary estimand is the panel-level ATT for treated city-month cells; cohort-time effects  $\tau(g, t)$  and event-time effects  $\theta_k$  help distinguish short-run disruptions from longer-run adjustments. Heterogeneity is important: small independent restaurants may benefit more than chains with established delivery operations because the platform provides access to delivery infrastructure they previously lacked. General equilibrium effects—category expansion, labour market adjustments, changes in consumer behaviour—mean that the effect on treated restaurants may differ from the aggregate effect on the restaurant sector.

Identification faces several challenges. Each city is unique, so finding comparable control cities is difficult. The platform enters larger, more attractive cities first, creating selection on observables and unobservables. Parallel trends (Assumption 6) may hold only after adjusting for covariates, or a factor structure may be needed to accommodate common shocks (macroeconomic trends, pandemic phases) that affect all cities differentially. In practice, analysts often begin with covariate-adjusted parallel-trends designs and then use factor models as robustness checks when pre-trends or residual common shocks suggest that parallel trends in levels is implausible. Competitive responses create spillovers: incumbent platforms adjust pricing and marketing in response to entry, partially offsetting the treatment effect. These spillovers can be modelled explicitly (Chapter 11) or bounded using partial identification techniques. In the potential-outcomes language, these competitive responses imply that  $Y_{it}$  depends on both own treatment  $D_{it}$  and competitors' treatments  $h_i(D_{-i,t})$ , so identification either models  $h_i(\cdot)$  explicitly (Chapter 11) or reports bounds on direct and total effects when  $h_i(\cdot)$  is uncertain.

Synthetic control methods (Chapter 6) are a natural starting point. For each treated city, construct a synthetic control from never-treated cities, weighting to match pre-entry restaurant revenues. The post-

entry difference between actual and synthetic revenues estimates the causal effect for that city. Synthetic difference-in-differences (Chapter 7) and factor models (Chapters 8, 9) combine synthetic-control logic with panel structure and latent common shocks.

Spillover models (Chapter 11) can quantify the competitive response by comparing revenues for restaurants in treated cities that do not join the platform to revenues in control cities. General equilibrium effects require comparing the total change in the restaurant sector (treated and untreated restaurants) to assess whether entry creates new demand or merely redistributes existing demand. Heterogeneous effects by restaurant type can be estimated using interactions in regression models or through stratified synthetic control analyses.

Across all three examples, the same methodological choices recur. Which assumption justifies the comparison—parallel trends, conditional independence, or factor structure? Which estimand answers the business question—ATT for ROI, event-time effects for dynamics, or dose-response functions for continuous treatments? Which diagnostics probe the conclusions—pre-trends, placebo checks, or sensitivity analyses? The methods developed in the following chapters provide the tools; these examples show how to wield them. Chapters 17 and 15 show how to turn these example-specific questions into systematic diagnostic checklists and reporting standards.

## 2.8 Notation and Assumptions Reference

This section collects the core notation and assumptions for quick reference. Return here when encountering unfamiliar symbols or when checking which assumptions underpin a particular method. The formal definitions appear in earlier sections; here we provide a consolidated summary.

### Notation

We index units (stores, customers, markets, products) by  $i = 1, \dots, N$  and periods (quarters, weeks, months, years) by  $t = 1, \dots, T$ . The observed outcome for unit  $i$  in period  $t$  is denoted  $Y_{it}$ . Potential outcomes under a contemporaneous treatment level  $d$  or a treatment path  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$  are written as  $Y_{it}(d)$  or  $Y_{it}(\underline{d}_i^t)$ . Under interference, outcomes may depend on both own treatment and others' treatments through an exposure mapping:  $Y_{it}(d, h_i(D_{-i,t}))$ , where  $h_i(\cdot)$  summarises relevant aspects of other units' treatment status. The contemporaneous form  $Y_{it}(d)$  is a shorthand for the path-dependent object  $Y_{it}(\underline{d}_i^t)$  under the no-dynamic-effects and no-anticipation restrictions from Section 2.3. For staggered adoption with absorbing treatment,  $Y_{it}(g)$  denotes the potential outcome if unit  $i$  first adopted at time  $g$ , and  $Y_{it}(\infty)$  denotes the potential outcome under never receiving treatment. The observed treatment is  $D_{it}$ , which is binary ( $D_{it} \in \{0, 1\}$ ) in most chapters but continuous ( $D_{it} \in \mathbb{R}$ ) in Chapter 14.

For staggered adoption designs,  $G_i \in \{1, \dots, T\} \cup \{\infty\}$  denotes the adoption cohort—the first period in which unit  $i$  is treated. Units never treated have  $G_i = \infty$  by convention. Event time  $k = t - G_i$  measures periods relative to adoption, where  $k = 0$  is the first treated period,  $k < 0$  are pre-treatment periods, and  $k > 0$  are post-treatment periods. Event-time indicators  $D_{it}^k = \mathbf{1}\{t - G_i = k\}$  form the building blocks of event-study regressions.

Unit fixed effects  $\alpha_i$  capture time-invariant unit heterogeneity, while time fixed effects  $\lambda_t$  capture common shocks or trends. The idiosyncratic error is denoted  $\varepsilon_{it}$ .

Using the contemporaneous potential-outcomes notation  $Y_{it}(d)$  from Section 2.1, the core estimands are the average treatment effect

$$\text{ATE} = \mathbb{E}_{i,t}[Y_{it}(1) - Y_{it}(0)]$$

and the average treatment effect on the treated

$$\text{ATT} = \mathbb{E}_{i,t}[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1].$$

For staggered adoption designs, we also define the cohort–time effect

$$\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g]$$

for  $t \geq g$ , and the event-time effect

$$\theta_k = \mathbb{E}[Y_{i,G_i+k}(G_i) - Y_{i,G_i+k}(\infty) \mid G_i < \infty],$$

which measures the average effect  $k$  periods post-adoption among ever-treated units.

## Core Assumptions

The assumptions introduced earlier in this chapter are summarised here for quick reference. See the indicated sections for formal statements and discussion.

**No Anticipation (Assumption 1, Section 2.1).** Potential outcomes in period  $t$  do not depend on treatments assigned in periods  $s > t$ . This rules out forward-looking behaviour where units respond to expected future treatments.

**SUTVA (Assumption 3, Section 2.1).** Potential outcomes for unit  $i$  depend only on unit  $i$ 's own treatment path, not on other units' treatments (no interference). There is a single, well-defined version of treatment (treatment version irrelevance).

**Parallel Trends (Assumption 6, Section 2.4).** For all cohorts  $g, g'$  and periods  $t < \min(g, g')$ , expected changes in untreated potential outcomes are the same across cohorts:

$$\mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g'].$$

**No Dynamic Effects (Assumption 4, Section 2.3).** Potential outcomes depend only on current-period treatment status:  $Y_{it}(\underline{d}_i^t) = Y_{it}(d)$ , where  $d \in \{0, 1\}$  is the current-period component of the path  $\underline{d}_i^t$ . This rules out carryover and feedback from past treatments. It is therefore a strong modelling restriction rather than a default assumption; many marketing applications in Chapters 10 and 11 explicitly relax it.

**Homogeneous Treatment Effects (Assumption 8, Section 2.5).** The treatment effect is constant across units and time:  $\tau_{it} = \tau$  for all  $i, t$ . This homogeneity condition ensures that the TWFE coefficient can be interpreted as a single common effect under staggered adoption.

**Unconfoundedness (Assumption 7, Section 2.4).** Conditional on covariates  $X_{it}$ , unit fixed effects  $\alpha_i$ , and time fixed effects  $\lambda_t$ , treatment intensity  $D_{it}$  is independent of contemporaneous potential outcomes:

$$D_{it} \perp Y_{it}(d) | X_{it}, \alpha_i, \lambda_t \quad \text{for all } d \in \mathcal{D},$$

where  $\mathcal{D}$  denotes the support of treatment intensity. This is a contemporaneous selection-on-observables condition; panel regression implementations often also require stronger assumptions about how treatment relates to past and future outcomes. In randomised experiments, a design-based version of unconfoundedness holds by construction even without conditioning on covariates (see Assumption 5 in Section 2.4).

**Factor Structure (Section 2.4).** Untreated potential outcomes follow a low-rank structure:

$$Y_{it}(\infty) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

where  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects,  $f_{tr}$  are latent factors,  $\lambda_{ir}$  are unit-specific loadings, and  $R \ll \min(N, T)$ . Identification further requires that, after conditioning on observed covariates and the latent factors  $\{f_{tr}\}$ , the idiosyncratic component  $\varepsilon_{it}$  is mean independent of treatment assignment, so that residual variation in  $D_{it}$  is as-good-as-random with respect to  $Y_{it}(\infty)$ .

These assumptions are invoked selectively depending on the method and context. Parallel trends underpins difference-in-differences and event studies. Factor structure justifies interactive fixed effects and matrix completion methods. Unconfoundedness is central to selection-on-observables approaches with high-dimensional controls. SUTVA is often violated in marketing and requires spillover modelling. No anticipation can be probed and stress-tested in event studies using pre-treatment leads. Homogeneous effects is required for a TWFE coefficient to represent a single common effect but is relaxed by modern heterogeneity-robust estimators.

## Forward References

The notation and assumptions formalised here recur throughout the book. Parallel trends and the staggered adoption structure underpin the difference-in-differences and event-study methods developed in Chapters 4 and 5. Synthetic control methods (Chapters 6 and 7) relax parallel trends in favour of pre-treatment matching or factor structure.

Factor structure assumptions are operationalised in Chapters 8 and 9 through principal components, EM algorithms, and nuclear norm regularisation. Dynamic extensions appear in Chapter 10, while Chapter 11 relaxes SUTVA to accommodate interference. Machine learning methods (Chapters 12 and 13) leverage conditional independence with flexible estimators. Chapter 16 provides tools for uncertainty quantification, and Chapter 17 develops diagnostics to assess assumption plausibility and sensitivity to violations.

Throughout the book, we return to this notation and these assumptions as the common vocabulary that links marketing questions to causal estimands, identification strategies, and estimators. In particular, the estimands ATE, ATT,  $\tau(g, t)$ , and  $\theta_k$  from Section 2.3, together with the assignment mechanisms and assumptions summarised here, form the backbone of the method-specific chapters that follow. This chapter establishes that vocabulary. The subsequent chapters show what you can do with it.

## Chapter 3

# Design-Based Thinking for Panels

This chapter frames panel studies around assignment mechanisms rather than models. You will learn how experimental and quasi-experimental regimes differ in panels, how common marketing designs (geo-experiments, switchbacks, staggered rollouts, continuous treatments) map to target estimands ( $\text{ATT}$ ,  $\tau(g, t)$ ,  $\theta_k$ , dose-response functions), and how these designs connect to the methods and diagnostics developed in later chapters.

### 3.1 Experimental and Quasi-Experimental Regimes

The central question is not which model best fits the data but how treatment was assigned. Design-based inference anchors on the assignment mechanism rather than on functional form assumptions. Angrist and Pischke’s “credibility revolution” put this transparency at the centre of applied causal work [Angrist and Pischke, 2010], and quasi-experimental methods have become central to marketing research [Goldfarb et al., 2022]. These designs exploit natural variation in treatment assignment over time and across units, using panel data to construct credible counterfactuals. Parallel trends, unconfoundedness, and exclusion restrictions are not verifiable from the observed data in the strict sense (see Chapter 17 for diagnostics and sensitivity analysis)—we cannot verify that parallel trends would hold in the post-treatment period—but pre-treatment data can provide suggestive evidence and falsification checks, and transparent design choices and rigorous diagnostics can make these assumptions more or less credible.

The key is to articulate the assignment mechanism clearly and to assess whether identification assumptions are plausible given the institutional context. Marketing applications often involve staggered adoption (loyalty programmes, product launches), geographic variation (advertising campaigns, pricing experiments), and policy changes (platform algorithm updates, privacy regulations). Each setting demands careful attention to spillovers, anticipation, and measurement alignment between platform metrics and econometric estimands.

When we know how treatment was assigned—whether by randomisation, by administrative rules, by staggered rollout schedules, or by targeting criteria—we can articulate the conditions under which observed differences in outcomes have a causal interpretation without leaning on parametric functional-form assumptions for identification.

Marketing panels can offer rich data but messy assignment. The opportunities are that repeated observations create multiple sources of identifying variation, and natural variation in timing or intensity often approximates quasi-experimental designs. The constraints are that full randomisation is rare, treatment effects unfold dynamically, and spillovers violate SUTVA (Assumption 3).

This chapter develops a design-based framework for causal inference in marketing panels, showing how assignment mechanisms map to estimands and estimators.

We distinguish between two broad regimes. Formally, let  $\mathbf{D}$  be the  $N \times T$  matrix of treatment assignments with entries  $D_{it}$ , and let  $\{Y_{it}(\underline{d}_i^t)\}_{i,t}$  denote the path-dependent potential outcomes corresponding to any assignment plan  $\mathbf{d} = (d_{it})$ , where  $\underline{d}_i^t = (d_{i1}, d_{i2}, \dots, d_{it})$  is the treatment path for unit  $i$  up to time  $t$ . The assignment mechanism is the conditional distribution of  $\mathbf{D}$  given the full schedule of potential outcomes and covariates  $\mathbf{X}$ :

$$\Pr(\mathbf{D} \mid \{Y_{it}(\underline{d}_i^t)\}_{i,t}, \mathbf{X}).$$

By *experimental* we mean designs where treatment is assigned by a known randomisation mechanism independent of potential outcomes, satisfying a design-based unconfoundedness condition such as

$$\Pr(\mathbf{D} \mid \{Y_{it}(\underline{d}_i^t)\}_{i,t}, \mathbf{X}) = \Pr(\mathbf{D} \mid \mathbf{X}),$$

possibly conditional on pre-specified strata. In cluster or unit-level randomisation, this reduces to independence between  $D_{it}$  and  $Y_{it}(d)$  within randomisation blocks.

In observational panel applications, we instead impose unit- or cell-level unconfoundedness as an *assumption*, treating the assignment as-if random after conditioning on  $X_{it}, \alpha_i, \lambda_t$ :

$$D_{it} \perp Y_{it}(d) \mid X_{it}, \alpha_i, \lambda_t \quad \text{for all } d \in \mathcal{D},$$

which is the continuous-treatment version of Assumption 7 in Chapter 2. Identification also requires overlap or positivity: for the treatment levels you want to learn about, units must have positive probability of receiving those levels given  $(X_{it}, \alpha_i, \lambda_t)$ . This is much stronger than design-based randomisation and is difficult to defend in many marketing settings where targeting is endogenous and assignment may depend on past observed outcomes through feedback loops. The key point is that, conditional on observed covariates and fixed effects, the assignment mechanism does not depend on the potential outcomes. By *quasi-experimental* we mean designs where treatment is not randomised, but institutional rules, timing, or thresholds create as-if-random variation that can be exploited given credible assumptions.

Randomised (experimental) designs provide the cleanest path to causal inference but are scarce in marketing at scale. Geo-experiments and platform A/B experiments represent the leading examples of randomised designs in marketing, and we devote considerable attention to their design and analysis. Quasi-experimental designs, where treatment is assigned by deterministic rules, strategic targeting, or staggered rollouts, dominate observational marketing data. When these assignments are driven by factors unrelated to the trajectory of untreated potential outcomes—or when we can condition on those factors—credible causal inference remains possible under explicit assumptions such as parallel trends or factor structures, which can be assessed through pre-treatment diagnostics even if they cannot be verified directly.

The notation and estimands from Chapter 2 provide the foundation. Our target estimands—ATT, cohort-time effects  $\tau(g, t)$ , and event-time effects  $\theta_k$ —summarise causal contrasts between potential outcomes. Their formal definitions appear in Section 2.3 and Table 2.1.

The assignment mechanism determines which potential outcomes are observed and which remain counterfactual, and credible inference hinges on our ability to construct valid counterfactuals for the unobserved potential outcomes. Chapter 2 develops the formal framework. This chapter focuses on practical design questions: Who randomised what? What institutional features drive treatment assignment? Which spillovers are plausible? How should we plan diagnostics and inference?

Design-based reasoning contrasts with purely model-based adjustment, which specifies parametric models for outcomes as functions of treatments and covariates and then relies on those functional form assumptions for identification. Model-based approaches require strong assumptions—linearity, additivity, correctly specified interactions—that are hard to justify a priori and can produce misleading inferences when violated. Design-based reasoning avoids leaning on those parametric functional-form assumptions for identification by anchoring the causal argument in the assignment mechanism rather than in the outcome model. In practice, modern causal work combines both: design-based arguments justify the identification strategy, while model-based adjustment (regression, propensity scores, outcome modelling) improves precision and addresses covariate imbalance. Regression remains useful even in design-based settings, but the legitimacy of the co-

efficient as a causal estimand flows from the assignment mechanism, not from the regression specification itself.

Marketing panels typically fall into four binary-treatment design categories — randomised cluster designs, staggered adoption, single treated unit, and common shock designs — plus continuous-treatment designs that we discuss separately. These archetypes often appear in combination — for example, a geo-experiment with staggered rollout or a common shock with heterogeneous continuous exposure — so the practical task is to identify which assignment features dominate the identifying variation. Randomised cluster designs, exemplified by geo-experiments and platform A/B experiments, assign treatment to groups of units through a randomisation protocol. Staggered adoption designs, common in phased rollouts of loyalty programmes or platform expansions, create variation in the timing of treatment adoption across units. Single treated unit designs arise naturally in case studies and flagship launches. Common shock designs, where all units experience a common shock at a single point in time, apply to policy changes or regulatory interventions.

Each design maps to specific estimators developed in subsequent chapters. Randomised cluster designs enable difference-in-differences comparisons, though inference must account for clustering and potential spillovers (Chapter 16). Staggered adoption designs call for modern heterogeneity-robust difference-in-differences estimators that avoid the negative weighting problems of traditional two-way fixed effects regressions [Goodman-Bacon, 2021, de Chaisemartin and d'Haultfœuille, 2020, Callaway and Sant'Anna, 2021, Sun and Abraham, 2021] (Chapters 4, 5). Single treated unit designs motivate synthetic control methods that construct weighted combinations of control units to approximate the counterfactual trajectory of the treated unit (Chapters 6, 7). Common shock designs leverage event-study specifications that trace dynamic responses before and after the shock, diagnosing anticipation effects and estimating cumulative impacts (Chapter 5).

The modern panel data literature has clarified the conditions under which each of these designs yields credible causal estimates. Parallel trends assumptions, factor structures, conditional independence, and interference-aware models each correspond to particular features of the assignment mechanism and the data-generating process. This chapter provides practical guidance on designing marketing panel studies that align with these identification strategies, choosing appropriate estimands given the assignment mechanism, and planning diagnostics and inference *ex ante* to ensure that the resulting estimates are both credible and policy-relevant.

We begin by classifying the major assignment mechanisms that define these regimes.

## 3.2 Assignment Mechanisms in Panels and Target Estimands

The assignment mechanism—how units receive treatment across periods—sits at the centre of design-based inference.

The assignment mechanism determines which estimand is appropriate, which estimator aligns with the identifying variation, and which assumptions you must defend. We classify five canonical mechanisms and map them to methods in later chapters: block designs (often randomised experiments), staggered adoption, single treated unit, common shock designs, and continuous treatment designs (typically quasi-experiments). Spillover structures, which cut across all these mechanisms, require separate treatment because they fundamentally alter the estimand and identification strategy.

### 3.2.1 Block Designs

Block designs assign treatment to a set of units at a common starting time, with all treated units remaining treated for the duration of the observation period. For example, a retailer launches a loyalty programme in 100 stores in quarter one, and those stores continue to offer the programme through quarter twelve, while 400 control stores never receive the programme. Treatment is indicated by

$$D_{it} = D_i \cdot \mathbf{1}\{t \geq t_0\},$$

where  $D_i \in \{0, 1\}$  denotes treatment-group membership. The treatment switches on at time  $t_0$  for the treated group ( $D_i = 1$ ) and remains zero for the control group ( $D_i = 0$ ).

Block designs correspond naturally to the canonical two-group, two-period difference-in-differences setting. The estimand is the Average Treatment Effect on the Treated (ATT), defined as

$$\text{ATT} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) | D_{it} = 1],$$

which averages the treatment effect over all treated unit-period cells. In block designs,  $D_{it} = 1$  if and only if  $D_i = 1$  and  $t \geq t_0$ . Identification relies on the parallel trends assumption (Assumption 6): in the absence of treatment, the treated and control groups would have experienced similar trends in outcomes. Under parallel trends, the difference-in-differences estimator — comparing the change in outcomes for treated units to the change in outcomes for control units — recovers the ATT.

Block designs are the workhorse of randomised experiments in marketing. Geo-experiments that randomise treatment across markets, store-level experiments that assign new pricing strategies to randomly selected stores, and platform A/B experiments that allocate users to treatment and control groups all produce block designs. When treatment is randomised, parallel trends holds in expectation conditional on the randomisation strata (assessed via balance diagnostics and randomisation inference, Chapter 16), though finite-sample imbalance can still occur. In finite samples with few clusters, balance can still fail by chance, so we rely on pre-treatment diagnostics and, where appropriate, randomisation-based inference to complement asymptotic

CRVE methods. Inference is straightforward provided that we account for clustering and potential spillovers. Chapter 4 develops the standard difference-in-differences estimator and its variants for block designs, and Chapter 17 provides diagnostics to assess the plausibility of parallel trends in observational block designs.

If treatment timing varies across units rather than being applied simultaneously, staggered adoption designs emerge.

### 3.2.2 Staggered Adoption Designs

Staggered adoption designs feature units adopting treatment at different times. Define  $G_i \in \{1, \dots, T\} \cup \{\infty\}$  as the adoption time (cohort) for unit  $i$ , with  $G_i = \infty$  for units that never adopt during the observation period. Some units adopt in period  $g = 2$ , others in period  $g = 5$ , others in period  $g = 8$ , and some never adopt. Treatment status is absorbing:

$$D_{it} = \mathbf{1}\{t \geq G_i\}.$$

The absorbing treatment assumption rules out treatment reversal, which is violated in some marketing settings where stores drop loyalty programmes or campaigns end. Extensions that handle non-absorbing treatment exist but require additional assumptions about treatment history effects. Staggered adoption is common in marketing. A retailer rolls out a loyalty programme in batches, with operational capacity or strategic priorities determining the timing of rollout to different store groups. A brand launches advertising campaigns sequentially across markets, with budget constraints or market readiness driving the staggered launch. A platform enters cities over time, with city size, regulatory environment, or competitive conditions influencing entry order. The variation in adoption timing creates rich opportunities for identification: units that have not yet adopted serve as controls for units that adopt early, and within-cohort comparisons over time trace dynamic effects.

The appropriate estimand in staggered adoption designs is the cohort-time effect  $\tau(g, t)$ , the average treatment effect for units in cohort  $g$  in calendar period  $t \geq g$ . As in Chapter 2,  $Y_{it}(g)$  denotes the potential outcome for unit  $i$  in period  $t$  under a once-treated-always-treated path that switches from 0 to 1 at time  $g$ , and  $Y_{it}(\infty)$  denotes the potential outcome under never treated:

$$\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g], \quad t \geq g.$$

This estimand is defined only for post-treatment periods ( $t \geq g$ ). Under the standard no-anticipation assumption (Assumption 1) we set  $Y_{it}(g) = Y_{it}(\infty)$  for  $t < g$ , so pre-treatment effects are zero by construction. Aggregating  $\tau(g, t)$  across cohorts and time produces summary measures such as the overall ATT and the event-time effects  $\theta_k$  in Section 2.3. Different estimators implement different weighting schemes over  $\tau(g, t)$ —for example, Callaway–Sant’Anna and Sun–Abraham target distinct aggregations of these cohort-time effects—so understanding  $\tau(g, t)$  is essential for interpreting what any given estimator recovers. Event-time effects  $\theta_k$  re-index these cohort-time effects by time since adoption, averaging  $\tau(g, g + k)$  over cohorts  $g$  that contribute observations at event time  $k$ . Modern heterogeneity-robust estimators [e.g., Callaway and Sant’Anna,

2021, Sun and Abraham, 2021] construct clean comparisons between treated units and not-yet-treated or never-treated control units, avoiding the negative weighting problems that plague traditional two-way fixed effects regressions [Goodman-Bacon, 2021, de Chaisemartin and d'Haultfoeuille, 2020] when treatment effects are heterogeneous (Chapter 4). Event-study specifications extend the analysis to visualise pre-trends and estimate dynamic effects (Chapter 5).

Identification in staggered adoption designs relies on a parallel trends assumption across cohorts: units adopting at different times would have followed similar trajectories in the absence of treatment, possibly after conditioning on covariates. This assumption is inherently untestable because we never observe all counterfactuals, but pre-treatment diagnostics provide indirect evidence. In practice we assess this assumption with cohort-specific event-study plots and placebo checks that compare pre-treatment trends across adoption cohorts (see Chapters 5 and 17). If pre-trends diverge, the parallel trends assumption becomes less plausible, and alternative identification strategies such as factor models (Chapters 8 and 9) or synthetic control (Chapter 6) may be required.

When only one unit receives treatment, synthetic control methods provide a natural approach.

### 3.2.3 Single Treated Unit Designs

Single treated unit designs feature one unit receiving treatment while all others serve as controls. A platform launches in a single pilot market while comparable markets remain untreated. A brand introduces a new product in one region before national rollout. A retailer implements a major store redesign in a flagship location. Finding exact matches among control units is typically impossible when the treated unit has unique characteristics.

The synthetic control method (Chapter 6) [Abadie et al., 2010] is designed for single treated unit designs. Rather than assuming that any single control unit provides a valid counterfactual, synthetic control constructs a weighted combination of control units such that the synthetic unit's pre-treatment outcomes closely match the treated unit's pre-treatment outcomes. The post-treatment difference between the actual treated unit and the synthetic control estimates the causal effect for that specific unit in each post-treatment period. Formally, the target is the unit-specific treatment effect  $Y_{i^*t}(1) - Y_{i^*t}(0)$  for the treated unit  $i^*$  in each post-treatment period  $t$ , where  $Y_{i^*t}(0)$  is approximated by the synthetic control. This is a unit-specific effect, not a population ATT: the estimand has high internal validity for the treated unit but limited external validity for generalising to other units. Inference relies on placebo checks: applying the same synthetic control procedure to each control unit (as if it had been treated) generates a distribution of placebo effects to which the actual effect is compared (Chapter 16). These placebo distributions play the role of a randomisation reference distribution under the assumption that, absent treatment, the treated unit would have been exchangeable with the control units used to build the synthetic.

Hybrid methods (Chapter 7), particularly synthetic difference-in-differences (SDID), combine the strengths of synthetic control and difference-in-differences by weighting both units and time periods to improve pre-treatment fit and leverage parallel trends where they hold. These methods are especially valuable when the

single treated unit is embedded in a panel with multiple pre-treatment and post-treatment periods, allowing both cross-sectional and time-series variation to inform the counterfactual.

### 3.2.4 Common Shock Designs

Common shock designs involve a common shock affecting all units in a single period. A regulatory change, a major advertising campaign, a pandemic, or a technological disruption creates a discrete event at time  $t_0$  that shifts outcomes for all units. Treatment is  $D_{it} = \mathbf{1}\{t \geq t_0\}$ , with no variation across units in treatment timing.

Event-study specifications (Chapter 5) are appropriate for common shock designs. Estimating separate coefficients for each period before and after the shock traces the dynamic response. Pre-treatment periods diagnose anticipation effects and provide evidence on whether the parallel trends assumption holds. Post-treatment periods estimate how effects accumulate or dissipate over time.

**Warning:** In pure common-shock settings, an event study is primarily descriptive unless you can defend an exclusion-style argument ("nothing else changed at  $t_0$ ") or bring in a comparison group/exposure gradient.

Identification in pure common shock designs, where all units receive the same binary shock at the same time, requires that the shock is exogenous and that no other concurrent changes drive the observed break. Without cross-sectional variation, identification relies entirely on temporal variation and assumes that nothing else changes coincidentally at  $t_0$ . Event studies in such settings are primarily descriptive unless you can defend an exclusion restriction or bring in cross-sectional variation in exposure intensity. When cross-sectional variation in exposure intensity or in characteristics that moderate the effect is available, difference-in-differences with continuous treatment intensity (Chapter 14) can exploit this heterogeneity for identification.

Many marketing interventions involve continuous rather than binary treatment.

### 3.2.5 Continuous Treatment Designs

In many marketing applications, treatment is not binary but varies continuously. Advertising expenditure, promotional discount depth, loyalty programme reward generosity, and pricing all take continuous values. The potential outcomes framework extends naturally:  $Y_{it}(d)$  denotes the potential outcome under treatment intensity  $d \in \mathcal{D} \subseteq \mathbb{R}$ . For a covariate value  $x$ , the conditional dose-response function

$$\mu(d | x) = \mathbb{E}[Y_{it}(d) | X_{it} = x]$$

describes how expected outcomes vary with treatment intensity conditional on covariates. The average dose-response function introduced in Chapter 2,  $\mu(d) = \mathbb{E}[Y_{it}(d)]$ , is the same object we work with throughout the book. It is obtained by averaging  $\mu(d | x)$  over the covariate distribution.

Continuous treatment designs require two key assumptions. First, conditional independence:  $Y_{it}(d) \perp D_{it} | X_{it}, \alpha_i, \lambda_t$  for all  $d$  in the support. This is the continuous-treatment version of the unconfoundedness condition

(Assumption 7) in Chapter 2. This asserts that, after conditioning on observed covariates and fixed effects, treatment intensity is independent of the potential outcomes. Second, positivity: for each covariate profile  $(X_{it}, \alpha_i, \lambda_t)$  in the target population and each treatment level  $d$  in the support of interest, the generalised propensity score  $r(d | X_{it}, \alpha_i, \lambda_t)$  is strictly positive. Positivity is often violated in practice—for example, no stores may have zero advertising spend—which limits the range of treatment intensities for which causal effects can be identified. Panel structures make conditional independence more plausible by controlling for time-invariant confounders through unit fixed effects and for common shocks through time fixed effects, reducing the set of potential unobserved confounders to time-varying unit-specific factors. High-dimensional controls (Chapter 13) and double machine learning (Chapter 12) provide flexible approaches to conditioning on many covariates without overfitting. When treatment intensity varies smoothly over time and space, factor models (Section 2.8 and Chapters 8 and 9) can capture common demand shocks, allowing identification from unit-specific deviations. Identification then hinges on the assumption that, after conditioning on observed covariates and latent factors, residual variation in treatment intensity is as good as random with respect to  $Y_{it}(d)$ . Chapter 14 formalises these assumptions and develops estimators for dose–response functions and elasticities in panel settings.

Identification also requires overlap or positivity: for the treatment levels you want to learn about, units must have positive probability of receiving those levels given  $(X_{it}, \alpha_i, \lambda_t)$ .

### 3.2.6 SUTVA and Spillover Concerns

The stable unit treatment value assumption (SUTVA), introduced in Chapter 2, asserts that potential outcomes for one unit do not depend on the treatment assignments of other units. SUTVA is routinely violated in marketing through spillovers, network effects, and competitive interactions. A loyalty programme offered to customers in one store may generate word-of-mouth that influences nearby stores. Advertising in one market may spill over to adjacent markets through media overlap or migration. One firm’s pricing decision may trigger competitive responses that alter outcomes for rival firms.

Design-based reasoning requires anticipating these violations at the design stage rather than assuming they are negligible. Cluster-randomised designs internalise spillovers within clusters while maintaining independence across clusters, if spillovers are contained within clusters. Buffer zones separate treated and control geographies to limit spatial spillovers. Exposure mappings, such as the functions  $h_i(D_{-i,t})$  introduced in Chapter 11, summarise neighbours’ treatments into scalar or vector-valued spillover doses that enter the potential outcomes. The assignment mechanism and the expected pattern of spillovers should therefore be specified together: how we randomise or select treated units and how we cluster or buffer them are design decisions that must reflect the plausible interference structure. For example, a geo-experiment that randomises individual stores but ignores advertising spillovers across neighbouring stores will mis-state its estimand: the estimated “direct” effect actually conflates own- and neighbour-treatment paths. When SUTVA fails, the estimand must be reformulated to depend on both own treatment  $D_{it}$  and exposure  $h_i(D_{-i,t})$ . Chapters 2 and 11 define these spillover-aware causal contrasts. In other words, the question shifts from “what is the ef-

fect of changing own treatment  $D_{it}$ ?" to "what is the effect of jointly changing own treatment and neighbours' treatments as encoded by  $h_i(D_{-,t})$ ?"—a different estimand with different policy implications.

These assignment mechanisms represent the building blocks of panel study designs in marketing. In practice, designs often combine elements from multiple categories—a geo-experiment with staggered rollout, a common shock with continuous treatment intensity, or a block design with spillover concerns. The key is to match the design to the substantive question, anticipate threats to validity, and select estimators that align with the identifying variation. The following section examines geo-experiments in detail, illustrating how these design principles operate in practice.

### 3.3 Geo-Experiments and Clustered Designs

Geo-experiments randomise treatment across geographic units—DMAs, stores, regions, cities—and are now the leading design for measuring aggregate market-level causal effects in marketing [Vaver and Koehler, 2011]. Major platforms offer tooling that automates randomisation and first-pass analysis, making geo-experiments accessible at scale. However, geo-experiments measure market-level effects rather than individual-level effects, require sufficient geographic variation in treatment, and may have limited power when the number of markets is small [Lewis and Rao, 2015]. They instantiate the block designs introduced in Section 3.2 in a clustered, often fully experimental regime.

Designing a credible geo-experiment requires careful choices about cluster definition, stratification, treatment window length, and inference procedures. We provide practical guidance on these design decisions, grounded in the potential outcomes framework and the design-based principles articulated above.

#### 3.3.1 Cluster Definition and Stratification

The first design choice is the definition of the cluster: the unit of randomisation. In geo-experiments, clusters are typically geographic markets—DMAs defined by Nielsen for television advertising, metropolitan areas, postal codes, or custom regions tailored to the firm’s operating footprint. Formally, we partition the  $N$  elementary units into  $G$  clusters  $\mathcal{C}_1, \dots, \mathcal{C}_G$ . Treatment is assigned at the cluster level, so  $D_{it} = D_{gt}$  for all  $i \in \mathcal{C}_g$ , where  $D_{gt}$  denotes the treatment assignment for cluster  $g$  in period  $t$ .

Define the cluster outcome as the average outcome in cluster  $g$ ,  $Y_{gt} = |\mathcal{C}_g|^{-1} \sum_{i \in \mathcal{C}_g} Y_{it}$ . Let  $Y_{gt}(1)$  and  $Y_{gt}(0)$  denote, respectively, the potential outcomes for cluster  $g$  in period  $t$  under treatment and control. Geo-experiments then target cluster-level causal contrasts such as  $\mathbb{E}[Y_{gt}(1) - Y_{gt}(0)]$  rather than individual-level effects. If you instead define  $Y_{gt}$  as total sales, the estimand implicitly weights clusters by size, which can be the right target when the business question is total incremental revenue.

The choice of cluster definition involves trade-offs. Larger clusters better internalise spillovers because consumers and competitors within a cluster interact more than consumers across clusters. Smaller clusters increase the number of clusters for a given budget, improving statistical power. The optimal cluster size depends on the strength of spillovers, the variance of outcomes within and between clusters, and the total number of units available for the experiment. In spatial settings, the choice of geographic boundaries (DMA vs postcode vs custom region) changes both the estimand and the interference structure through the Modifiable Areal Unit Problem [Openshaw, 1984]: results can differ when boundaries are redrawn, making this a causal design choice, not merely an operational convenience.

Stratification divides clusters into groups (strata) based on observable characteristics—market size, demographic composition, historical sales levels, competitive intensity—and randomises treatment within each stratum. Stratification ensures balance on the stratifying variables, reducing the variance of the treatment effect estimate and protecting against unlucky randomisations that assign treatment disproportionately to clusters with atypical characteristics. For example, if a retailer operates stores in both urban and rural mar-

kets, stratifying by urbanicity ensures that urban and rural stores are represented in both treatment and control groups in proportion to their prevalence in the population.

Over-stratification can reduce power. If strata are too fine, some strata may contain only one or two clusters, leaving little variation for within-stratum comparisons. A practical rule of thumb is to ensure at least a few treated and a few control clusters in each stratum (for example, two and two) so that within-stratum comparisons are informative.

Re-randomisation extends stratification by generating multiple candidate randomisations, checking balance on a set of covariates for each candidate, and selecting the randomisation that achieves the best balance according to a pre-specified criterion (for example, the smallest maximum standardised difference across covariates). Re-randomisation improves balance without biasing the estimate of the treatment effect, provided that the balance criterion is specified *ex ante* and the final randomisation is selected using that criterion alone [Morgan and Rubin, 2012].

Standard randomisation inference must be adjusted when re-randomisation is used. The reference distribution for the test statistic should include only those randomisations that would have passed the balance criterion, not all possible randomisations. The analysis must condition on the acceptance rule, and the rule must be fixed *ex ante* before seeing any outcomes. In practice this means drawing treatment assignments from the set of allocations that satisfy the pre-specified balance rule and recomputing the statistic under each draw. Using the unrestricted randomisation distribution after re-randomisation understates uncertainty because it ignores the selection step.

### 3.3.2 Treatment Windows and Seasonality

The treatment window is the period during which treatment is applied and outcomes are measured. Choosing an appropriate window length involves balancing statistical power (longer windows provide more data and more precise estimates) against the risk of confounding events (longer windows increase the chance that other shocks affect treated and control units differently). In marketing, seasonality is a pervasive confounder. Sales, advertising effectiveness, and competitive intensity vary systematically by week of year, month, and quarter due to holidays, weather, and consumer behaviour patterns.

Design-based solutions to seasonality include choosing treatment windows that span multiple seasons or that align treatment and control periods to the same seasonal phase. For example, if the goal is to estimate the effect of a holiday advertising campaign, the treatment window should cover the holiday season, and the control group should be observed during the same seasonal period (either in a previous year or in a concurrent, spatially separated control region). Alternatively, the analysis can include seasonal fixed effects or detrend outcomes using historical seasonal patterns, provided that seasonal patterns are stable and the adjustment is planned *ex ante*. These adjustments reintroduce functional form assumptions, but seasonal patterns are often well-understood and stable in retail and media data, which makes such adjustments relatively safe when design choices already align treatment and control on season. Identification is strongest when the design itself

### 3.3.3 Buffer Zones and Spillovers

Geographic spillovers arise when treatment in one market affects outcomes in nearby markets—a concrete violation of SUTVA (Assumption 3) that the design must anticipate and contain rather than rely on post-hoc robustness checks alone. Customers may travel across market boundaries, media markets may overlap, or advertising may reach audiences outside the target geography. Buffer zones—untreated regions surrounding treated markets that are excluded from the analysis—mitigate spillovers by creating geographic separation between treatment and control units. The width of the buffer depends on the expected strength and spatial decay of spillovers. If spillovers dissipate quickly with distance, narrow buffers suffice. If spillovers are long-range, wide buffers are needed, reducing the effective sample size.

Buffer zones address spatial spillovers but not all spillover types. Network spillovers (for example, social media word-of-mouth or supply chain effects) and competitive spillovers (for example, national pricing responses) may not decay with geographic distance. When non-spatial spillovers are plausible, buffer zones provide limited protection, and explicit spillover models with exposure mappings  $h_g(D_{-g,t})$  (Chapter 11) are required.

When spillovers are central to the research question, the design should measure outcomes in buffer zones to estimate spillover effects directly rather than ignoring them. For example, if the goal is to estimate both the direct effect of an advertising campaign on sales in treated markets and the indirect effect on sales in neighbouring markets, the analysis compares outcomes in treated markets, neighbouring buffer markets, and distant control markets. This three-group design enables identification of direct and spillover effects jointly, provided that the buffer markets are far enough from distant controls that spillovers do not reach the control group.

The three-group design is demanding. It requires correct specification of the spatial structure of spillovers: which markets are "neighbours" and which are "distant." Misspecification of this structure can bias both direct and spillover effect estimates. When the spillover structure is uncertain, sensitivity analysis across different neighbourhood definitions is essential. Chapter 11 formalises these designs and develops estimators for direct and spillover effects.

### 3.3.4 Inference with Few Clusters and Serial Correlation

Geo-experiments often feature a modest number of clusters—dozens of DMAs, tens of regions, or even fewer custom markets—and outcomes measured over multiple periods within each cluster. This structure creates two inferential challenges. First, the small number of clusters means that standard cluster-robust standard errors, which rely on large-cluster asymptotics, may be unreliable. Second, outcomes are serially correlated within

clusters over time because persistent unobservables, autocorrelated shocks, or dynamic feedback generate correlation across periods.

The independent sampling unit is the cluster, so the effective sample size is  $G$ , not the number of unit-period observations. Randomisation inference yields exact finite-sample p-values for statistics computed under the stated assignment protocol [Rubin, 1980]. This exactness holds under the sharp null hypothesis that treatment has no effect on any unit in any period. For weaker nulls (for example, an average-zero effect), randomisation inference remains useful but is no longer automatically exact without additional structure. Under the sharp null, the observed treatment assignment is just one of many possible randomisations that could have been drawn, and recomputing the test statistic across these randomisations yields its null distribution. For complex estimators (for example, regression-adjusted effects), the same logic applies: we recompute the full estimator under each admissible randomisation to obtain its null distribution. Comparing the observed statistic to this distribution gives an exact p-value that respects the randomisation protocol and automatically accounts for clustering and serial correlation without parametric assumptions.

The wild cluster bootstrap offers a complementary approach [Cameron et al., 2008, MacKinnon and Webb, 2017], resampling entire clusters with random signs on cluster-level residuals to preserve the within-cluster correlation structure. It is particularly effective when the number of clusters is small (fewer than 50) and when outcomes exhibit complex patterns of heteroskedasticity and serial correlation. Chapter 16 provides detailed algorithms and guidance on choosing between randomisation inference, wild cluster bootstrap, and asymptotic cluster-robust methods in geo-experiments.

Geo-experiments randomise across units. An alternative design randomises within units over time.

### 3.4 Switchbacks and Platform Experiments

Switchback designs randomise treatment over time within units rather than across units [Bojinov et al., 2022]. In a switchback, each unit alternates between treatment and control conditions according to a pre-specified schedule, so that the same unit serves as its own control across different time periods. Switchbacks are common in platform experiments where logistical or ethical concerns preclude withholding treatment from some users permanently, and where platforms prioritise rapid iteration and learning over long-term effect estimation. Note that platforms tend to use switchbacks in settings where they work well (short-lived effects, rapid iteration cycles), so the published literature may overrepresent success cases and underrepresent settings where carryover undermines the design. Internal post-mortems rarely get written up. Within our taxonomy, switchbacks hold the unit fixed and randomise over time blocks, contrasting with geo-experiments (across units) and phased rollouts (staggered cohorts).

For example, a ride-sharing platform testing a new surge pricing algorithm might implement a switchback where each city alternates between the new algorithm (treatment) and the existing algorithm (control) on alternating weeks. This allows the platform to measure the effect while ensuring that no city is permanently denied the potential benefits of the new algorithm.

The simplest switchback alternates treatment and control in a fixed pattern: treatment in odd periods, control in even periods. More sophisticated designs use period blocking—stratifying by day of week, time of day, or other temporal characteristics—to ensure that treatment and control observations are balanced across temporal confounders. As with stratification in geo-experiments, over-blocking can reduce power: if blocks are too fine, few observations remain within each block for comparison. Identification comes from the assignment mechanism: we randomise treatment across pre-specified time blocks (often after blocking by day-of-week or hour-of-day) so that, in expectation, treated and control blocks are comparable within a unit. The design also requires a substantive restriction on dynamics, namely no carryover (or limited carryover) across blocks, because otherwise today's treatment changes tomorrow's 'control' potential outcomes.

A natural target in a switchback is the average effect of turning treatment on for a unit during the switchback window, averaged over the time blocks that the randomisation assigns to treatment versus control. Under no carryover, and conditional on the block structure used in the randomisation, differences in mean outcomes between treated and control blocks within the same unit identify an average of  $\tau_{it} = Y_{it}(1) - Y_{it}(0)$  over the blocks included in the design. With carryover, the estimand changes because the observed contrast mixes the effect of current treatment with lingering effects from earlier treated blocks. In such cases, the analyst must decide *ex ante* whether the target is the immediate effect, the cumulative effect over a treatment cycle, or the steady-state effect under continuous treatment. Chapter 10 develops methods for decomposing dynamic responses into short-run and cumulative components.

### 3.4.1 Carryover, Anticipation, and Washout Periods

The primary threat to switchback designs is carryover: treatment effects that persist into subsequent control periods. If an advertising campaign shown in period  $t$  influences purchases in period  $t + 1$ , then the control outcome in period  $t + 1$  is contaminated by the treatment effect from period  $t$ , biasing the switchback estimate downward. Formally, the potential outcome  $Y_{it}$  depends not just on current treatment  $D_{it}$  but on the treatment history  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$ , as in the path-dependent framework of Chapter 2. A generic no-carryover assumption—a restriction on dynamics sometimes called no carryover—posits that potential outcomes depend only on current treatment,  $Y_{it}(\underline{d}_i^t) = Y_{it}(d_{it})$ , where  $d_{it} \in \{0, 1\}$  is the current-period component of the path  $\underline{d}_i^t$ . Switchback designs with rapid alternation require a relaxed version, such as order- $L$  carryover:

$$Y_{it}(\underline{d}_i^t) \approx Y_{it}(d_{it}, d_{i,t-1}, \dots, d_{i,t-L}).$$

This is a causal model restriction (finite memory) that must be defended substantively: treatments more than  $L$  periods in the past have negligible incremental effect on  $Y_{it}$  relative to more recent treatments. Sensitivity analysis should vary  $L$  to assess robustness [Athey et al., 2025a].

Anticipation effects—where users respond to expected future treatments—create a symmetric problem, contaminating pre-treatment control periods with anticipatory responses [Holland, 1986].

Washout periods—gaps between treatment and control periods during which treatment is set to baseline (outcomes may still be observed but excluded from the primary analysis)—mitigate carryover by allowing effects to dissipate before the control period begins. The length of the washout depends on the expected half-life of the treatment effect. If advertising effects decay exponentially with a half-life of one week, a two-week washout ensures that less than 25 per cent of the initial effect remains: two half-lives reduce the effect to  $(1/2)^2 = 0.25$  of its original magnitude. However, not all effects decay exponentially. Some effects have delayed peaks (for example, word-of-mouth that builds over time), non-monotonic decay (for example, habituation followed by recovery), or step-function decay (for example, effects that persist until a threshold is crossed). When the decay structure is uncertain, sensitivity analysis across different washout lengths is advisable. Chapter 10 formalises these decay concepts via impulse responses and long-run multipliers, which in turn inform washout choices in future switchback experiments.

When carryover is inevitable and washouts are impractical, the analysis must model the carryover explicitly. Distributed lag specifications (Chapter 10) include current and lagged treatments as regressors, estimating both the immediate effect and the lagged effects jointly. Event-study specifications (Chapter 5) trace the dynamic response, revealing the persistence of effects and enabling decomposition of the total effect into immediate and cumulative components.

### 3.4.2 Learning Systems and Pacing

Platform experiments often run concurrently with algorithmic learning systems that optimise targeting, bidding, or personalisation in real time. If the learning system adapts within the experimental window,

treatment and control paths can diverge not only because of the intervention itself but also because the optimiser reacts differently to their outcomes. The estimated effect then conflates the intervention with the learning dynamics. Design-based solutions include freezing the learning system during the experiment, running the experiment in a separate environment isolated from the learning system, or modelling the learning dynamics explicitly.

**External Validity Caveat** Freezing the learning system during the experiment changes the estimand. The target becomes the effect under the joint intervention (treatment plus freeze) versus the effect under production dynamics. The estimated effect is the treatment effect *conditional on a frozen learning system*, which may differ from the effect in production where the learning system is active and adapts to the treatment. If the learning system would amplify or dampen the treatment effect in production, the experimental estimate may not generalise. When the learning system remains active, credible identification typically requires assuming that, after conditioning on the experiment's randomisation and observed state variables, the learning dynamics would have evolved similarly under treatment and control in the absence of the intervention—an assumption that is often hard to defend in tightly coupled optimisation systems [Athey et al., 2025a].

Pacing—the rate at which a budget or inventory is spent over time—introduces another complication. If a platform experiment randomises ad delivery but the campaign budget runs out early in the treatment period, then treated users in later periods receive no ads, diluting the treatment effect. Specifying pacing rules in advance and ensuring that budget and inventory are sufficient to cover the full treatment window prevent this issue. Alternatively, budget exhaustion can be treated as informative: if the treatment budget runs out early, that reveals something about treatment intensity and demand, which can be modelled as treatment effect heterogeneity rather than simply excluded. Budget exhaustion is a post-treatment variable, so conditioning on it changes the estimand and can induce selection bias. Treating budget exhaustion as informative requires care because it is itself a post-treatment outcome: naïvely conditioning on it can induce selection bias unless the experiment is designed so that exhaustion patterns are part of the estimand rather than an incidental complication. These complications are primarily design problems. The experiment should be embedded in a stable learning and pacing regime rather than relying solely on post-hoc modelling adjustments.

Switchback designs offer advantages when permanent treatment assignment is infeasible or when rapid learning is prioritised over long-run effect estimation. However, they require careful attention to carryover effects and to the block structure that makes treated and control periods comparable within a unit. When these conditions are plausible and washout periods are feasible, switchbacks provide a powerful tool for causal inference in dynamic platform environments. When carryover is severe or the block structure fails to balance key time-varying confounders, alternative designs such as geo-experiments or phased rollouts may be more appropriate. The design decision is therefore not "switchbacks versus nothing" but which randomisation protocol best matches the dynamic properties of the outcome and the operational constraints of the platform. When the business question centres on long-run or steady-state effects rather than short-run lifts, slower designs such as geo-experiments or phased rollouts typically provide more credible identification than rapid switchbacks.

### 3.5 Phased Rollouts and Staggered Adoption

Phased rollouts, introduced in Section 3.2 as staggered adoption designs, assign treatment to batches of units over time. A retailer launches a loyalty programme in 100 stores in quarter one, another 200 stores in quarter three, another 150 stores in quarter five, and leaves 50 stores untreated as perpetual controls. A platform enters 30 cities over two years, with entry timing driven by operational capacity and market readiness. Phased rollouts are attractive because they spread implementation costs over time, allow mid-course corrections based on early results, and generate rich variation in treatment timing that enables credible causal inference.

There is a caveat to mid-course corrections. If the rollout schedule is adjusted based on early results—for example, accelerating rollout because early cohorts show positive effects or delaying rollout to underperforming markets—treatment timing becomes outcome-driven. Units assigned to later cohorts are then selected based on realised outcomes, violating the parallel trends assumption across cohorts. In potential-outcomes terms, the assignment mechanism now depends on realised outcomes, so adoption timing  $G_i$  becomes a function of the untreated trajectory  $\{Y_{it}(0)\}_{t=1}^T$  (even after conditioning on pre-treatment covariates). That breaks the cohort-parallel-trends logic that makes later adopters credible controls for earlier adopters. To preserve identification, the rollout schedule should be fixed in advance or adjusted only based on operational constraints unrelated to outcomes. This section focuses on practical design considerations for implementing phased rollouts effectively, while Chapter 4 provides the estimation details for modern heterogeneity-robust difference-in-differences methods.

#### 3.5.1 Cohort Mapping and Not-Yet-Treated Controls

The defining feature of phased rollouts is the cohort structure discussed in Section 3.2: units are grouped by their adoption time  $G_i$ , and cohorts are compared to one another and to never-treated controls. The key identifying variation comes from comparing units that have adopted to units that have not yet adopted. For example, units adopting in quarter three (cohort  $g = 3$ ) can be compared to units adopting later (cohorts  $g = 5$  or  $g = 7$ ) or to never-treated units (cohort  $g = \infty$ ). These comparisons use quarters one and two as pre-treatment periods for cohort 3, and quarters three and four as post-treatment for cohort 3 but pre-treatment for later cohorts.

Identification relies on a parallel trends assumption across cohorts: absent treatment, units adopting at different times would have followed similar outcome trajectories, possibly after conditioning on covariates, as discussed in Section 3.2. These comparisons also rely on no anticipation (later cohorts are unaffected before adoption) and, in the basic setup, an absorbing treatment path so that adoption time  $G_i$  is well-defined.

Modern heterogeneity-robust difference-in-differences estimators [Callaway and Sant'Anna, 2021, Sun and Abraham, 2021] (Chapter 4) formalise this logic by estimating  $\tau(g, t)$  for each cohort-time pair and aggregating them into summary measures. The Callaway–Sant’Anna estimator, for instance, compares outcomes for cohort  $g$  in period  $t$  to outcomes for not-yet-treated or never-treated units in the same period, yielding a clean estimate of the effect for that cohort at that time. Aggregating across cohorts and time produces the overall

ATT, cohort-specific effects, or event-time effects  $\theta_k$  that trace the dynamic response relative to adoption. Traditional two-way fixed effects regressions implicitly average cohort–time effects with weights that can be negative when effects are heterogeneous [Goodman-Bacon, 2021, de Chaisemartin and d’Haultfœuille, 2020], so they can even assign negative weight to cohorts where the effect is large and positive.

The primitive estimands are the cohort–time effects  $\tau(g, t)$ , while event-time effects  $\theta_k$  and overall ATT represent chosen aggregations across cohorts and time with specific weighting schemes. These aggregations are estimand choices that embody different research questions, not just different ways to plot results.

### 3.5.2 Calendar Time Versus Event Time Aggregation

Phased rollouts enable two types of aggregation: calendar time and event time. Calendar-time aggregation pools all units observed in the same period, estimating the average effect in each period  $t$  across all treated units. Event-time aggregation pools all units observed at the same time since adoption. We define event time as  $k = t - G_i$ , representing the number of periods elapsed since adoption.

As in Chapter 2, the event-time effect  $\theta_k$  averages cohort–time effects at event time  $k$  across all cohorts that contribute observations. The event-study parameter  $\theta_k$  aggregates the cohort-specific effects  $\tau(g, t)$  across all cohorts  $g$  observed at event time  $k$ :

$$\theta_k = \sum_{g: g+k \leq T} w_{gk} \tau(g, g+k),$$

where the weights  $w_{gk}$  are non-negative and sum to one over the cohorts that contribute at event time  $k$ . Different estimators implicitly choose different weighting schemes, so you should read  $\theta_k$  as an estimand defined by a chosen set of weights.

The choice depends on the substantive question. Calendar time aggregation is natural when the goal is to estimate the contemporaneous impact of the programme in a given period, accounting for macroeconomic conditions or seasonal effects specific to that period. Event time aggregation is natural when the goal is to trace how the effect evolves as a function of time since adoption. Calendar-time aggregation yields a sequence of period-specific average treatment effects  $\{\text{ATT}_t\}$ , while event-time aggregation yields the dynamic profile  $\{\theta_k\}$  that links directly to the event-study estimands in Chapter 5.

Event-study specifications (Chapter 5) facilitate event-time aggregation by estimating separate coefficients for each period relative to adoption. The coefficients  $\theta_k$  estimate the treatment effect at each event time  $k$  relative to a baseline (typically  $k = -1$ , the period immediately before adoption). Plotting  $\theta_k$  against  $k$  produces an event-study graph that visualises pre-trends and dynamic effects. Pre-treatment coefficients near zero (for  $k < 0$ ) provide supportive but not definitive evidence that parallel trends holds, while post-treatment coefficients (for  $k \geq 0$ ) trace the treatment effect trajectory. Chapter 4 shows how modern heterogeneity-robust DiD estimators implement these aggregations in practice, ensuring that overall ATT and dynamic profiles are constructed from clean cohort–time contrasts.

Plan these aggregation choices in advance to ensure the design supports the intended analysis.

### 3.5.3 Planning for Heterogeneity-Robust Estimators

When designing a phased rollout, anticipate that treatment effects may be heterogeneous across cohorts or over time. Early adopters may differ from late adopters in ways that affect the treatment response. Effects may grow over time as users adapt to a new programme or as network effects accumulate. These sources of heterogeneity mean that traditional two-way fixed effects regressions, which impose a constant treatment effect, can produce misleading estimates (Chapter 4). Plan to use heterogeneity-robust estimators that allow  $\tau(g, t)$  to vary and that aggregate these effects with transparent, non-negative weights.

Design choices that facilitate heterogeneity-robust estimation include ensuring that never-treated or not-yet-treated control units are observed in all periods, balancing cohort sizes so that no cohort dominates the aggregation, and collecting rich covariate data that can be used to explore sources of heterogeneity. Designs with aggressive rollouts leave few controls late in the window, which increases variance and makes dynamic effects fragile.

**Trade-off on Never-treated Controls** Having never-treated controls strengthens identification by providing a comparison group unaffected by treatment at any point. However, maintaining never-treated controls may be operationally difficult or ethically problematic if the treatment is beneficial. If withholding treatment permanently is unacceptable, the design can rely solely on not-yet-treated controls, which requires that some cohorts adopt late enough to serve as controls for early adopters. In that case the primary estimand becomes an average treatment effect for ever-treated units, constructed from comparisons between early and not-yet-treated cohorts rather than between treated and never-treated units. This choice should be made at design time, not improvised after seeing the data, because it shapes both ethical exposure to treatment and the set of valid identification strategies.

Specifying in advance which aggregations will be reported—overall ATT, cohort-specific effects, event-time effects, effects by subgroup (for example, high-income versus low-income markets)—disciplines the analysis and guards against cherry-picking results *ex post*. Chapter 16 complements this design discipline with multiple-testing adjustments and joint confidence bands for dynamic profiles.

Phased rollouts combine operational flexibility with credible inference. By spreading implementation over time, they reduce operational risk and allow learning from early cohorts. By creating variation in treatment timing, they enable credible causal inference through modern heterogeneity-robust methods. When designed carefully with attention to cohort balance, control availability, and pre-specified aggregations, phased rollouts provide a powerful approach to measuring treatment effects in realistic marketing settings.

## 3.6 Ex Ante Diagnostics and Placebos

Credible design-based inference requires diagnosing the plausibility of identification assumptions before treatment is assigned and \*\*post-treatment\*\* outcomes are observed. *Ex ante* diagnostics check whether the design satisfies the conditions required for causal identification and whether balance on pre-treatment covariates is adequate. In observational settings where treatment has already been assigned, these diagnostics are performed *ex post* but remain essential for assessing credibility. These diagnostics also assess whether placebo checks using pre-treatment data or alternative outcomes support the assumptions. We outline the key *ex ante* diagnostic tasks and indicate where the detailed design workflows in Chapter 17 and inference-focused diagnostics in Chapter 16 pick up.

### 3.6.1 Pre-Treatment Fit and Balance

If the design relies on parallel trends, pre-treatment fit assesses whether treated and control units followed similar trajectories before treatment. Formally, using an event-study specification with pre-treatment leads  $\theta_k$  for  $k < 0$  relative to a reference period (typically  $k = -1$ ), we assess the joint null hypothesis that all pre-treatment coefficients are zero:

$$H_0 : \theta_k = 0 \quad \forall k < 0.$$

In an event-study regression these lead coefficients are best read as falsification checks. They are not the main causal objects of interest. The causal interpretation of post-treatment  $\theta_k$  still rests on no anticipation and cohort-parallel-trends. In panels with clustered and serially correlated outcomes, this joint check should use either cluster-robust covariance matrices, wild cluster bootstrap procedures, or randomisation-based methods, as developed in Chapter 16. Plotting outcome trends for treated and control groups over pre-treatment periods provides visual evidence. If trends diverge, parallel trends becomes less plausible, and alternative identification strategies—factor models (Chapters 8, 9) or synthetic control (Chapter 6)—may be required.

This check has low power when pre-treatment periods are short or noisy. Failing to reject  $H_0$  does not prove that parallel trends holds, only that we lack evidence against it. The check provides supportive but not definitive evidence. Additionally, using pre-tests to choose specifications can distort inference for post-treatment effects, so pretest with caution [Roth, 2022].

If you want to quantify sensitivity to plausible violations of parallel trends rather than rely on pre-tests alone, use sensitivity analyses that incorporate uncertainty about pre-trends directly into inference [Rambachan and Roth, 2023].

Balance checks assess whether treated and control units are similar on observed covariates. We recommend the normalised difference (or standardised mean difference) [Austin, 2011] as a scale-invariant measure of imbalance. For a scalar covariate component  $X^{(j)}$  (which may be a baseline aggregate such as  $\bar{X}_i$  from the panel vector  $X_{it}$ ), the normalised difference is:

$$\Delta_{X^{(j)}} = \frac{\bar{X}_1^{(j)} - \bar{X}_0^{(j)}}{\sqrt{(S_1^2 + S_0^2)/2}},$$

where  $\bar{X}_d^{(j)}$  and  $S_d^2$  are the sample mean and variance in group  $d \in \{0, 1\}$ . Here group assignment depends on the design context: ever-treated unit indicator, treated cluster indicator, or cohort membership. The same baseline covariates typically enter the propensity score. When the design includes unit and time fixed effects, the conditioning set extends beyond covariates alone, but here we illustrate the idea with a covariate-only propensity score  $e(X_{it}) = \mathbb{P}(D_{it} = 1 | X_{it})$ , introduced in Chapter 2. Common thresholds are  $|\Delta_{X^{(j)}}| > 0.1$  or 0.2, though the appropriate threshold depends on how strongly the covariate predicts outcomes: a large imbalance on a covariate weakly related to outcomes may matter less than a small imbalance on a covariate that strongly predicts outcomes. Large imbalances signal that treated units differ systematically from controls, raising concerns about unobserved confounders. Good balance on observed covariates is therefore necessary but not sufficient for credible conditional independence: large imbalances are a red flag, but small imbalances do not by themselves guarantee the absence of unobserved confounding. When balance is marginal, report both unadjusted and covariate-adjusted estimates to assess sensitivity to the imbalance.

### 3.6.2 Placebo Checks and Negative Controls

Placebo checks in time assign fictitious treatment dates in pre-treatment periods and estimate the treatment effect using those placebo dates using the same estimator, covariate adjustments, and variance estimator planned for the main analysis. Placebo estimates near zero are *consistent* with parallel trends, but noisy data can hide meaningful violations. Large placebo effects are concerning, but they can also arise by chance when you run many placebos.

Placebo checks have low power: failing to reject the null of zero placebo effects does not prove that parallel trends holds. Pre-treatment data may be too noisy or the pre-treatment period too short to detect violations. A non-significant placebo check provides supportive but not definitive evidence. Placebo checks are most credible when pre-specified as part of the design rather than chosen opportunistically after seeing outcomes.

Negative controls are outcomes that should not be affected by the treatment but that may be affected by confounders. For example, if the treatment is a loyalty programme targeted at frequent shoppers in grocery stores, a negative control outcome might be sales of a product category that should not respond to the programme but shares the same local economic shocks. If the loyalty programme appears to affect the negative control outcome, this suggests confounding rather than a causal effect.

Finding good negative controls can be challenging. A good negative control must satisfy two conditions: (1) it is unaffected by the treatment, and (2) it is affected by the same confounders as the primary outcome. Finding outcomes that satisfy both conditions is difficult. If the negative control is affected by different confounders than the primary outcome, a null result on the negative control provides little reassurance. If the negative control is itself affected by treatment (even indirectly), it stops being a diagnostic and becomes a second outcome with its own estimand. Negative controls are particularly valuable in observational studies where the assignment mechanism is not fully understood and confounding is a serious concern, and like

placebo checks they are best specified *ex ante* in the analysis plan. A detected effect on a valid negative control therefore points either to direct contamination of the control outcome by the treatment or to a failure of conditional independence, both of which undermine the causal interpretation of the main effect.

### 3.6.3 Overlap and Support Plans

In designs that rely on conditional independence—treatment is as-good-as-random conditional on covariates—overlap requires that treated and control units have similar covariate distributions. If treated units have covariates that are not observed in control units (or vice versa), then extrapolation is required to estimate counterfactual outcomes, and estimates may be sensitive to functional form assumptions. *Ex ante* overlap diagnostics plot the estimated propensity scores  $\hat{e}(X_{it}) = \mathbb{P}(D_{it} = 1 | X_{it})$  for treated and control observations and check whether their supports overlap. Trimming observations with extreme propensity scores [Crump et al., 2009], using matching or weighting to reweight the sample toward regions of overlap, or explicitly modelling the outcome as a function of covariates using flexible methods (Chapter 12) can improve robustness when overlap is limited.

Trimming observations with extreme propensity scores improves overlap but changes the target population. The estimated effect is then for the trimmed population—units with moderate propensity scores—not the original population. Formally, trimming redefines the estimand from an average treatment effect over the full population to one over the overlap region, for example  $\text{ATE}_{\text{overlap}} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) | X_{it} \in \mathcal{S}_{\text{overlap}}]$ . If units with extreme propensity scores are substantively important (for example, the most loyal customers or the largest markets), trimming may exclude precisely the units of greatest interest. Document the trimming rule and report the fraction of the sample excluded.

### 3.6.4 Planning for Spillovers and Interference

When SUTVA is likely to be violated—customers refer friends, competitors respond to rivals’ actions, advertising spills over across markets—the design should include an exposure mapping that specifies how spillovers propagate. Formally, an exposure mapping is a function  $h_i(D_{-i,t})$  that summarises the treatment assignments of other units into a spillover dose for unit  $i$  in period  $t$ , indicating which other units’ treatments affect its outcomes. In a geographic spillover model, the exposure mapping might specify that unit  $i$ ’s outcome depends on treatments in units within a certain radius or in adjacent markets. In a network spillover model, the exposure mapping follows the network structure: friends, followers, or co-purchasers.

Specifying the exposure mapping in advance disciplines the analysis by making the spillover assumptions explicit. It also guides data collection: if spillovers are expected to decay with distance, outcomes should be measured in buffer zones at varying distances from treated units. If network spillovers are expected, network data should be collected.

If the exposure mapping is wrong—spillovers decay faster or slower than assumed, or propagate through different channels than modelled—estimates of both direct and spillover effects will be biased. If the true spillover channel differs from the assumed  $h_i(\cdot)$ , then both the estimated direct effect of own treatment and the spillover effect of  $h_i(D_{-i,t})$  will generally be biased, because the potential outcomes  $Y_{it}(d, h_i)$  are indexed by a mis-specified exposure. When the spillover structure is uncertain, sensitivity analysis across different exposure mappings is essential. Chapter 11 develops estimators for direct and spillover effects under various exposure mappings, but the key insight is that spillover models should be planned *ex ante* based on substantive knowledge rather than discovered *ex post* through data mining.

These *ex ante* diagnostic tasks should be completed before treatment assignment and before the post-treatment measurement window begins. By identifying potential threats to validity early, researchers can adapt the design, collect additional data, or adjust the analysis plan to ensure credible causal inference. The diagnostics outlined here complement the detailed design workflows in Chapter 17, which provide step-by-step guidance for implementing each diagnostic task. When treatment has already been assigned, the same diagnostics are necessarily conducted *ex post*, but their role shifts from design selection to credibility assessment: failing diagnostics then signal that identification relies more heavily on untestable modelling assumptions.

## 3.7 Power, Minimum Detectable Effects, and Serial Dependence

Statistical power—the probability of detecting a true effect of a given magnitude—is a central consideration in designing experiments and quasi-experiments. An underpowered study may fail to detect effects that are substantively important, leading to false negatives and wasted resources. A very large study may detect effects that are statistically significant but substantively trivial. The solution is not to reduce sample size, but to focus on effect sizes and confidence intervals rather than p-values alone. Large samples provide precise estimates, which is valuable regardless of statistical significance. We provide practical guidance on power calculations for panel data, accounting for clustering, serial correlation, and heterogeneity. In the geo-experiments, switchbacks, and phased rollouts introduced earlier in this chapter, power depends on concrete design choices such as the number of clusters, the length of pre-treatment and treatment windows, and the number and size of adoption cohorts.

### 3.7.1 Minimum Detectable Effects Under Clustering

The minimum detectable effect (MDE) is the smallest true effect that the design can detect with specified power (typically  $1 - \beta = 0.8$ ) and significance level (typically  $\alpha = 0.05$ ). For a two-group comparison with independent observations, the MDE depends on the sample size and outcome variance. Panel data complicate this because outcomes are clustered within units over time.

Clustering reduces the effective sample size. Consider a design with  $N$  units observed for  $T$  periods, assigned to treatment at the unit level. Let the error structure be  $\varepsilon_{it} = \nu_i + \eta_{it}$ , with unit-specific component variance  $\sigma_\nu^2$  and idiosyncratic variance  $\sigma_\eta^2$ . The intra-cluster correlation (ICC) is  $\rho = \sigma_\nu^2 / (\sigma_\nu^2 + \sigma_\eta^2)$ , the same quantity that inflates standard errors in the clustering discussion of Chapter 2. High  $\rho$  means that adding more periods within a unit provides diminishing returns.

For a geo-experiment with  $C$  independent clusters in total, treated share  $p$  (so  $pC$  treated and  $(1-p)C$  control), and  $T$  periods in the analysis window, a simple approximation for the MDE of a difference-in-means estimator based on time-averaged cluster outcomes is

$$\text{MDE} \approx (z_{1-\alpha/2} + z_{1-\beta}) \times \sqrt{\frac{\sigma^2}{T} [1 + (T-1)\rho] \left( \frac{1}{pC} + \frac{1}{(1-p)C} \right)},$$

where  $\sigma^2$  is the variance of the cluster-period outcome  $Y_{ct}$  at the cluster level, and  $z$  denotes standard normal quantiles. This back-of-the-envelope calculation is most defensible for estimators that average outcomes within cluster over the analysis window. For difference-in-differences or event-study estimators, the pre/post covariance structure matters, so simulation-based power is usually safer. With  $\alpha = 0.05$  (two-sided) and power 0.8, the critical value factor is approximately  $1.96 + 0.84 = 2.8$ . The factor  $\left( \frac{1}{pC} + \frac{1}{(1-p)C} \right) = \frac{1}{C} \cdot \frac{1}{p(1-p)}$  reflects the allocation: equal allocation ( $p = 1/2$ ) minimises the variance, while unequal allocation increases it by the factor  $\frac{1}{4p(1-p)}$  relative to 50/50. For example, a 60/40 allocation increases the MDE by about two per cent relative to 50/50. This formula highlights the key trade-off: doubling the number of clusters  $C$  reduces

the variance by half, while doubling  $T$  has a negligible effect when  $\rho$  is high (since  $1 + (T - 1)\rho \approx T\rho$ , which cancels the  $T$  in the denominator).

To clarify the dependence hierarchy: we work with cluster-level outcomes  $Y_{ct}$  that aggregate unit-level outcomes within cluster  $c$  at time  $t$ . The ICC  $\rho$  captures the correlation of these cluster-level outcomes over time, arising from persistent cluster-level unobservables. If the original unit-level model is  $Y_{it} = \mu + \tau D_{it} + \nu_i + \eta_{it}$  with unit-level decomposition, then aggregating to clusters creates cluster-period outcomes with their own dependence structure approximated by this ICC formulation.

The ICC formulation captures correlation arising from unit-level effects. More general forms of serial dependence further reduce the effective information in the panel.

### 3.7.2 Serial Dependence and Effective Sample Size

Serial dependence—correlation of outcomes within a unit over time—is pervasive in marketing panels [Bertrand et al., 2004]. Sales are autocorrelated because demand is persistent, seasonality repeats, and customer bases are stable. Advertising effects exhibit carryover, creating dynamic correlation. Competitive interactions produce feedback loops. Ignoring serial dependence leads to overly optimistic power calculations and overstated precision in inference.

Effective sample size corrections account for serial dependence by down-weighting the contribution of additional periods. If outcomes follow a stationary AR(1) process with autocorrelation  $\phi$ , then for moderately large  $T$  the effective number of independent observations per unit is approximately  $T \times (1 - \phi)/(1 + \phi)$ . However, this approximation assumes stationarity and is sensitive to deviations from this assumption. In particular, if the AR(1) process is near-unit-root or has strong trends, the effective sample size will be lower than this approximation suggests. When trends, seasonality, or near-unit-root persistence dominate, estimate dependence from historical data and prefer simulation. With  $\phi = 0.5$ , for example, 12 periods contribute the equivalent of only  $12 \times (0.5)/(1.5) = 4$  independent observations. High autocorrelation ( $\phi$  near one) means that each additional period contributes little new information. In practice, serial dependence may combine unit-level effects and AR(1)-type dynamics. The AR(1) effective sample size formula is a useful guide even when the true process is more complex. You can estimate  $\phi$  from historical data using the autocorrelation function (ACF) or assume a conservative value (for example,  $\phi = 0.7$ ) if historical data are unavailable. When in doubt, it is safer to assume a higher value of  $\phi$  in power calculations, which yields more conservative (larger) required sample sizes. Incorporating these corrections into power calculations ensures that sample size and duration are set realistically.

When analytical formulas are insufficient, simulation provides a flexible alternative.

### 3.7.3 Simulation-Based Power Using Historical Panels

Analytical power formulas rely on simplifying assumptions—normality, constant variance, known correlation structure—that may not hold in practice. Simulation-based power calculations relax these assumptions by using historical data to estimate the distribution of outcomes under null and alternative hypotheses, generating synthetic treatment assignments, computing test statistics, and tallying the proportion of simulations in which the null is rejected.

The procedure is as follows. First, calibrate a model to historical panel data to estimate the mean, variance, and correlation structure of outcomes. When possible, reuse the empirical residual structure from untreated periods or use block-resampling to preserve realistic dependence patterns. Second, generate synthetic datasets by drawing outcomes from the fitted model under the null hypothesis (no treatment effect) and under alternative hypotheses (treatment effects of various magnitudes). Third, randomly assign treatment according to the proposed design (for example, randomising 50 DMAs to treatment and 50 to control). Compute the test statistic (for example, the difference-in-differences estimate divided by its standard error) using the same estimator, covariate adjustments, variance estimator, and multiple-testing corrections that will be used in the actual analysis so that simulated power reflects the planned inferential procedure, and record whether the null is rejected. Fourth, repeat for thousands of simulations to estimate power as the proportion of simulations in which the null is rejected. Typically, 1,000 to 10,000 simulations are sufficient to estimate power with acceptable precision. Use more simulations when power is close to critical thresholds (for example, 0.8). This approach accommodates realistic features of the data—skewness, heteroskedasticity, complex correlation structures—that analytical formulas cannot capture.

Simulation-based power is only as good as the model fitted to historical data. If the model misspecifies the correlation structure, variance, or treatment effect heterogeneity, the power calculations will be wrong. The calibration should be viewed as a guide rather than ground truth. Varying the assumed intra-cluster correlation  $\rho$  and autocorrelation  $\phi$  in these simulations provides a direct check on how sensitive power is to the dependence structures that drive the analytical MDE and effective sample size calculations above. Sensitivity analysis across different model specifications (for example, varying the assumed ICC or autocorrelation) helps assess robustness. Simulation is particularly valuable for the complex designs described earlier in this chapter—stratified and re-randomised geo-experiments, switchbacks with carryover, and phased rollouts with heterogeneous adoption timing—where analytical approximations are least reliable. The same simulation machinery can be re-used later for randomisation-based inference and bootstrap diagnostics (Chapter 16).

Regardless of the approach, power calculations must align with the planned analysis.

### 3.7.4 Inferential Choices to Be Planned

Power calculations should align with the planned inferential procedure. If inference will use cluster-robust standard errors, power should be computed assuming those standard errors. If inference will use randomisation inference or the wild cluster bootstrap, power should be computed by simulating those procedures. If the

plan is to conduct multiple comparisons—testing effects for multiple cohorts, subgroups, or outcomes—power should account for the multiplicity adjustment (for example, Bonferroni correction or false discovery rate control, as discussed in Chapter 16). This is particularly important for dynamic event-study profiles  $\{\theta_k\}$  and cohort–time effects  $\{\tau(g, t)\}$ , where naïve pointwise tests over many  $k$  or  $(g, t)$  pairs can dramatically overstate the evidence for effects. Specifying in advance the inferential procedure and conducting power calculations consistent with that procedure ensure that the design is appropriately powered for the actual analysis.

Power calculations are essential for designing credible panel studies. By accounting for clustering, serial dependence, and the planned inferential procedure, researchers can ensure that their designs are appropriately powered to detect substantively important effects. Simulation-based approaches provide flexibility when analytical formulas are insufficient. The key is to conduct power calculations *ex ante*, using realistic assumptions about correlation structures and effect sizes, and to adjust the design if power is inadequate.

Last but not least, note that power calculations are often optimistic. Power calculations assume the effect size is known, but in practice we are uncertain about the true effect. If the true effect is smaller than assumed, actual power will be lower than calculated. Conservative practice is to power for an effect size at the lower end of the plausible range, or to report power across a range of effect sizes rather than a single point estimate. A disciplined practice is to report a power curve—power as a function of effect size—rather than a single point calculation, and to adjust the design until power is acceptable for effects at the lower end of the substantively meaningful range.

## 3.8 Threats to Validity and Design Adaptation

Even well-designed experiments and quasi-experiments face threats to validity from confounding events, measurement issues, and violations of identifying assumptions. This section catalogues the major threats encountered in marketing panels and discusses design adaptations to mitigate them. Detailed treatments of specific threats appear in Chapter 15, which develops diagnostics and sensitivity analyses in depth. Here we provide a design-stage guide for anticipating and mitigating these threats. The key point is that threats should be anticipated *ex ante* and addressed through design choices rather than diagnosed and repaired *ex post*.

### 3.8.1 Seasonality and Event Interference

Marketing outcomes exhibit strong seasonal patterns driven by holidays, weather, school calendars, and cultural events. A loyalty programme launched in December may appear highly effective because sales rise during the holiday season, but the rise may reflect seasonality rather than a causal effect. Event interference occurs when major events—sporting championships, elections, natural disasters, pandemics—affect treated and control units differently during the experiment window.

Design solutions depend on whether concurrent controls are feasible. If concurrent control is feasible (for example, geo-experiments with spatial separation), align treatment and control periods seasonally: if a geo-experiment runs from January to March, the control group should also be observed from January to March in different geographies. If concurrent control is not feasible (for example, single-unit designs), lengthen the experiment to span multiple seasons and include seasonal fixed effects or detrended outcomes in the analysis, only if seasonal structure is stable and specified *ex ante*. For phased rollouts spanning multiple quarters, cohorts can be stratified by launch quarter to ensure that effects are estimated within seasonal periods. When concurrent controls are unavailable and identification rests on aligning different calendar periods through seasonal fixed effects or detrending alone, the design leans more heavily on functional form assumptions about seasonality and trend, and pre-treatment diagnostics become even more important.

### 3.8.2 Policy and Algorithm Changes

Platforms frequently update their algorithms for ad delivery, search ranking, recommendation, and pricing. If an algorithm change coincides with a treatment assignment, the estimated effect conflates the treatment with the algorithm change, biasing the estimate. Policy changes—new privacy regulations, shifts in platform policies, macroeconomic interventions—create similar confounds.

Design adaptations include scheduling experiments during stable periods, coordinating with platform or policy calendars to avoid known changes, and using staggered adoption to estimate effects separately for cohorts that experience different algorithm or policy regimes. If a confounding change is unavoidable, the

analysis can include indicators for the change and estimate treatment-by-change interactions, though this reduces power and complicates interpretation. Credible identification from this adjustment requires a convincing exclusion-style story for the policy or algorithm change timing—that the timing was driven by factors unrelated to potential outcomes under treatment or control, conditional on observed covariates and design structure.

### 3.8.3 Measurement Shifts

Changes in measurement systems—new data sources, revised definitions, improved tracking technologies—alter the measured outcomes without reflecting true changes in underlying behaviour. For example, a switch from survey-based measurement to scanner-based measurement may increase reported sales not because actual sales increased but because scanner data have better coverage. Platform metrics such as impressions, clicks, and conversions are subject to frequent redefinitions as platforms adjust methodologies. When the definition of the outcome changes mid-study, counterfactuals become ill-posed unless you can map old and new measures onto a common scale or restrict analysis to stable periods. Formally, the object labelled  $Y_{it}$  before and after the shift no longer corresponds to the same underlying potential outcome  $Y_{it}(d)$ , so cross-period contrasts mix level shifts in measurement with genuine treatment effects. If the measurement shift affects treated and control units differentially, it becomes an interference or confounding channel and changes the estimand unless you explicitly redefine outcomes on a common scale.

Design solutions include freezing measurement systems during the experiment, collecting both old and new measurements during transition periods to quantify the measurement shift, and using alternative outcomes or negative controls that should not be affected by the measurement shift to check for spurious effects. In practice, freezing measurement systems is often infeasible—platforms change metrics without notice, and researchers rarely control data collection infrastructure. When measurement shifts are unavoidable, documenting the timing and nature of the shift and conducting sensitivity analysis around the shift date are essential. Chapter 19 discusses measurement issues in depth and provides methods for bounding the bias introduced by measurement error.

### 3.8.4 Buffers and Robustness Windows

When threats to validity are uncertain or multiple confounding factors are plausible, robustness checks provide evidence that conclusions are stable across defensible modelling choices. Buffers—periods or units excluded from the analysis—create separation between treatment and confounds. For example, excluding the first week after a treatment starts (a washout or burn-in period) mitigates contamination from anticipation or measurement lag. Excluding the last week before a treatment ends mitigates decay or announcement effects. Robustness windows vary the length of pre-treatment and post-treatment periods to check whether conclusions depend on the choice of window. In terms of estimands, these exercises check whether average

effects such as ATT or dynamic profiles  $\{\theta_k\}$  are driven by a narrow set of periods that may coincide with unmodelled shocks.

Specification curves [Simonsohn et al., 2020] aggregate estimates across many specifications—different control sets, different fixed effects, different time windows, different clustering choices—and show the distribution of estimates. If estimates are stable across specifications, conclusions are robust. If estimates vary widely, the choice of specification matters, and the analyst should report the full distribution rather than a single preferred estimate. Because you are implicitly searching across many specifications, treat this as a transparency device rather than a licence to cherry-pick. If you run many windows and control sets, report the full curve and align inference with multiplicity. Chapter 17 develops these tools systematically.

### 3.8.5 When Parallel Trends Is Implausible: Factor Designs

If pre-treatment diagnostics reveal that treated and control units are on divergent trends, the parallel trends assumption is implausible. Factor models [Bai, 2009] (Chapters 8, 9) relax parallel trends by assuming that untreated potential outcomes are generated by an interactive fixed effects structure:

$$Y_{it}(0) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

where  $f_{tr}$  denotes the  $r$ -th latent common factor at time  $t$ ,  $\lambda_{ir}$  are the corresponding factor loadings for unit  $i$ , and  $R$  is the number of factors. In exchange, they assume a stable low-rank structure for untreated outcomes and that treatment does not change the factor loadings or the factor process in ways that would have occurred under no treatment.

This structure nests the standard two-way fixed effects model as the special case with no latent factors ( $R = 0$ ), where untreated potential outcomes reduce to  $Y_{it}(0) = \alpha_i + \lambda_t + \varepsilon_{it}$ . In that special case the counterfactual is identified under the same parallel trends assumption used in standard DiD. For example, if  $f_{tr}$  represents a "tech sector downturn" factor, units with high loadings  $\lambda_{ir}$  will be more affected. Factor designs use the control units to estimate the factors  $f_{tr}$  and the pre-treatment periods of the treated unit to estimate its loadings  $\lambda_{ir}$ . The counterfactual is then imputed as  $\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\lambda}_t + \sum_{r=1}^R \hat{\lambda}_{ir} \hat{f}_{tr}$ .

Identification requires a rich set of control units—say  $N_0$  untreated units—and pre-treatment periods  $T_0$  such that  $\min(N_0, T_0) \gg R$ , where  $R$  is the number of factors. With few control units or short pre-treatment windows, the factor structure cannot be reliably estimated. If  $\min(N_0, T_0)$  is only slightly larger than  $R$ , factor estimates become unstable and small changes in model specification or sample can lead to large swings in the imputed counterfactual  $\hat{Y}_{it}(0)$ . Information criteria or cross-validation on untreated cells can guide  $R$ , but sensitivity to  $R$  is essential. Alternative approaches include generalized synthetic control methods [Xu, 2017] and matrix completion techniques [Athey et al., 2021].

Here it is important to note the bias-variance trade-off. Factor models relax parallel trends but introduce additional assumptions (low-rank structure and time-invariant factor loadings that are not affected by treatment or its anticipation) and can have higher variance than standard DiD. Violations of these assumptions

— for example, when treatment changes a unit’s sensitivity to macro shocks so that  $\lambda_{ir}$  shifts after adoption — can bias factor-based estimates even when pre-treatment fit is excellent. In simulations, factor-based estimators often have larger standard errors than DiD when parallel trends holds, but this penalty is justified when parallel trends is violated. If parallel trends is plausible, DiD is more efficient. Factor models are most valuable when pre-treatment diagnostics clearly reject parallel trends and when the panel has sufficient dimensions to estimate the factor structure reliably. For designs where assignment plausibly generates non-parallel trends—for example, when treated markets and controls face different macro shocks—factor models provide a structured alternative to abandoning identification entirely, at the cost of stronger assumptions about the latent structure.

These threats to validity are not exhaustive but represent the most common challenges in marketing panel studies. The key principle is anticipation: threats should be identified and addressed through design choices rather than discovered during analysis. When threats cannot be eliminated, robustness checks and sensitivity analyses provide evidence on the stability of conclusions. Transparent reporting of design choices and threat mitigation strategies enhances credibility.

Finally, let us note that design-based solutions have their limits. Some threats cannot be fully addressed through design. Unobserved confounders that vary over time within units, measurement error in the treatment variable itself, and model misspecification remain concerns even in well-designed studies. In the potential-outcomes language of Chapter 2, these threats correspond to failures of conditional independence  $Y_{it}(d) \perp D_{it} | X_{it}, \alpha_i, \lambda_t$  and to mis-measured versions of  $D_{it}$  and  $Y_{it}(d)$ . Design-based reasoning reduces reliance on modelling assumptions but does not eliminate the need for judgement about the plausibility of identifying assumptions. Sensitivity analysis (Chapter 17) quantifies how conclusions would change under different assumptions about the magnitude of remaining threats.

## 3.9 Reporting Standards and Pre-Analysis Plans

Transparent reporting and pre-registration of designs enhance the credibility of causal inference by reducing researcher degrees of freedom and allowing readers to judge the robustness of conclusions. These practices complement the threat mitigation strategies discussed above by making design choices explicit and verifiable. We outline reporting standards for panel designs and provide guidance for pre-analysis plans that practitioners can adapt to their settings. These components feed directly into the diagnostic workflows of Chapter 17 and the inference procedures of Chapter 16.

### 3.9.1 Design Registries and Timelines

A design registry is a timestamped record of the key features of a study, filed before data are collected or before analysts access outcome data and run the main analysis. In industry settings this registry is often internal rather than public, but it should still be auditable. The registry includes the research question, the assignment mechanism, the treatment window, the primary outcome and estimand, the planned estimator, and the inferential procedure. Where possible, the estimand should be stated using the notation of Chapter 2—for example, ATT, cohort–time effects  $\tau(g, t)$ , event-time effects  $\theta_k$ , or a dose–response function  $\mu(d)$ —so that the registry clearly links the design to the formal causal object of interest. Registering the design disciplines the analysis by committing the researcher to a pre-specified approach and guards against data-driven specification searches that inflate false positive rates [Simmons et al., 2011]. Furthermore, the registry should also include information about the data sources, data quality checks, and any data transformations or cleaning procedures that will be applied. This additional information provides a clear understanding of the data and helps to ensure that the analysis is conducted in a transparent and reproducible manner.

Timelines document when key events occurred: when treatment was assigned, when outcomes were measured, when the analysis plan was finalised, when the analysis was conducted. Timestamped records—emails, meeting notes, version-controlled analysis scripts including failed or abandoned specifications—provide evidence that the design and analysis plan preceded data access and reveal the full path of model exploration. This approach mitigates concerns about *ex post* rationalisation and p-hacking [Simmons et al., 2011]. A clear timeline also separates design decisions (assignment mechanism, estimands, power calculations) from analysis-time choices (diagnostics, robustness checks), helping readers distinguish pre-specified elements from reactive modelling.

**When Pre-registration Is Infeasible** In observational studies using existing data, pre-registration before data access is not feasible. Alternatives include timestamping analysis scripts before running them (ideally using version control such as Git, which records all attempted specifications), using holdout samples (analysing a random subset first, then confirming on the remainder), or clearly distinguishing pre-specified from exploratory analyses in the final report. These practices cannot fully recreate the evidential strength of

pre-registration before first data access, especially when analysts are deeply familiar with the dataset, but they still constrain researcher degrees of freedom and make specification search visible.

### 3.9.2 Assignment Matrices and Exposure Maps

An assignment matrix shows which units received treatment in which periods. Formally, it is the matrix  $\mathbf{D} = (D_{it})$  introduced in Chapter 2, with rows indexed by units  $i$  and columns by periods  $t$ . For continuous treatments, the same object  $\mathbf{D} = (D_{it})$  records intensities rather than 0/1 status. In a phased rollout it records adoption timing across cohorts. In a geo-experiment it records treated and control clusters, along with any stratification variables. In a switchback it records the treatment schedule over time. The assignment matrix makes the design transparent and enables replication. It should be included in supplementary materials or deposited in a public or internal repository with auditability (for example, a timestamped internal registry, Open Science Framework, or GitHub) to facilitate verification and replication.

Exposure maps document the spillover structure by specifying, for each unit  $i$ , an exposure mapping  $h_i(D_{-i,t})$  that summarises which other units' treatments affect its outcomes. For a geographic spillover model, the map shows distances or adjacencies between units. For a network spillover model, the map shows network connections. Exposure maps clarify the assumptions required for identification and guide sensitivity analyses (Chapter 11). When interference is plausible, the estimand must be defined in terms of both own treatment and exposure (for example via  $h_i(D_{-i,t})$ ), rather than as a simple contrast  $Y_{it}(1) - Y_{it}(0)$ .

### 3.9.3 Outcome Definitions Aligned to Estimands

Specifying outcome definitions in advance ensures that the measured outcomes correspond to the conceptual estimand. If the estimand is the effect on incremental sales, the outcome should measure sales attributable to the treated units, not total sales (which may include non-incremental purchases). If the estimand is the effect on customer lifetime value, the outcome should measure long-run profitability, not short-run revenue. Aligning outcomes to estimands requires careful thought about data sources, measurement windows, and adjustment for confounders. Outcome definitions must also specify the measurement window (pre/post horizon) and avoid conditioning on post-treatment variables (bad controls). If the recorded outcome does not match the conceptual estimand—for example, using total revenue when the estimand is incremental profit—then even a perfectly executed design will identify the wrong causal quantity.

Primary and secondary outcomes should be distinguished. The primary outcome is the main target of the study, the outcome for which the design is powered and for which hypothesis tests are conducted. Secondary outcomes are exploratory, intended to generate hypotheses for future studies rather than to provide definitive answers. Declaring primary and secondary outcomes *ex ante* prevents cherry-picking: the analyst cannot elevate a secondary outcome to primary status *ex post* simply because it shows a large effect. Pre-specifying

primary and secondary outcomes also defines the family of hypotheses for which multiplicity adjustments will be applied in Chapter 16.

### 3.9.4 Inference and Multiplicity Plans

The pre-analysis plan should specify how standard errors will be computed (clustered by unit, by time, two-way), whether randomisation inference or bootstrap will be used, and how multiple comparisons will be handled. If the plan is to test effects for multiple cohorts, subgroups, outcomes, or event-time coefficients  $\{\theta_k\}$ , the plan should specify the multiplicity adjustment. When testing dynamic profiles, the preferred object is often a joint statement (confidence bands, omnibus tests, or a small set of pre-specified summaries like cumulative effect or long-run multiplier), not dozens of separate pointwise claims.

Bonferroni correction is conservative but appropriate when testing a small number of pre-specified hypotheses. False discovery rate control [Benjamini and Hochberg, 1995] is more powerful for exploratory analyses with many tests. Holm–Bonferroni [Holm, 1979] provides a middle ground, controlling family-wise error rate while being less conservative than Bonferroni. Alternatively, report results without adjustment but clearly label secondary outcomes as exploratory.

Power calculations in panel designs (Section 3.7) should use the same multiplicity and clustering rules specified here to ensure the study is powered for the actual evidential standard. Declaring these inferential choices *ex ante* protects against selective reporting and ensures that readers understand the evidential standard applied. If the plan was to cluster by unit and the analysis clusters by unit, the inference is pre-specified and credible. If the plan was to cluster by unit but the analysis clusters two-way because unit-level clustering produced large standard errors, the analyst should report the deviation, explain why it was necessary, and assess how it might affect conclusions. Ideally, both the pre-specified analysis and the revised analysis should be reported, allowing readers to compare results. The original plan should remain unchanged, with deviations documented transparently rather than edited into the historical record.

## 3.10 Method Selection Map

The design of a marketing panel study determines which estimators are appropriate and which assumptions are required for causal identification. This section provides a narrative map from design features—assignment mechanism, data structure, treatment variation—to the methods developed in subsequent chapters. The goal is not to prescribe a single correct method for each setting but to clarify the menu of options and the trade-offs among them. It complements the crosswalk in Chapter 2 by highlighting how specific design choices steer you toward particular estimators.

### 3.10.1 Randomised Block Designs

If treatment is randomised across units at a single point in time and remains constant thereafter (a block design), the standard difference-in-differences estimator (Chapter 4) is appropriate. The natural target is an average treatment effect on the treated ATT, and, when interest lies in dynamics, event-time effects  $\theta_k$  that trace the response over time. The estimator compares the change in outcomes for treated units to the change in outcomes for control units. Under randomisation, the parallel trends condition holds in expectation by design. Inference can use cluster-robust standard errors, randomisation inference, or wild cluster bootstrap depending on the number of clusters and the correlation structure (Chapter 16). Event-study specifications (Chapter 5) extend the analysis to check for pre-trends and estimate dynamic effects. In randomised settings these checks are mainly diagnostics for implementation problems, spillovers, or measurement shifts, since randomisation implies balance only in expectation.

### 3.10.2 Staggered Adoption Without Heterogeneity

If treatment is adopted at different times across units and treatment effects are truly constant across cohorts and time, the two-way fixed effects (TWFE) regression recovers an unbiased estimate of ATT under the standard parallel trends and no-anticipation assumptions. The regression includes unit and time fixed effects and regresses outcomes on a treatment indicator. The coefficient on the treatment indicator estimates the constant treatment effect. TWFE is computationally straightforward and scales to large datasets, but it is valid only when treatment effects are truly constant. When effects vary, its implicit weighting scheme can generate biased estimates, including negative weights on some cohort-time effects [Goodman-Bacon, 2021, de Chaisemartin and d'Haultfœuille, 2020] (see Chapter 4). However, constant effects is a strong, typically implausible restriction in marketing settings, a clean "TWFE is safe" decision rule is not available, and heterogeneity-robust estimators are the default when staggered adoption is present. Pre-trends and event-study plots (Chapter 5) provide diagnostics. When they suggest heterogeneous effects, the modern estimators [Callaway and Sant'Anna, 2021, Sun and Abraham, 2021] in Chapter 4 described in the next subsection should replace the static TWFE specification.

### 3.10.3 Staggered Adoption With Heterogeneity

If treatment effects are heterogeneous across cohorts or over time—as is typical in marketing—the standard TWFE estimator can be biased. Modern heterogeneity-robust estimators [Callaway and Sant’Anna, 2021, Sun and Abraham, 2021] (Chapter 4) target the cohort-time effects  $\tau(g, t)$  directly.

These methods proceed in two steps. First, estimate  $\tau(g, t)$  for each cohort  $g$  and period  $t$  using appropriate control groups such as not-yet-treated units. Second, aggregate these elementary effects using user-specified weights  $\omega_{gt}$  (non-negative and summing to one over the included  $(g, t)$  cells):

$$\tau_{\text{summary}} = \sum_{g,t} \omega_{gt} \hat{\tau}(g, t).$$

These weights define the reported estimand, which represents a weighted average of the cohort-time effects  $\tau(g, t)$ , and different choices answer different questions. Choosing  $\omega_{gt}$  to aggregate by cohort size yields an overall ATT, while aggregating by event time  $k = t - g$  recovers the dynamic profile  $\theta_k$  defined in Chapter 2. Common aggregations include the simple ATT (weighted by cohort size), event-study coefficients  $\theta_k$  (aggregating by relative time  $k = t - g$ ), or calendar-time effects. This explicit aggregation avoids the opaque and potentially negative weighting inherent in the static TWFE specification.

### 3.10.4 Single Treated Unit or Few Treated Units

If only one or a few units are treated and many controls are available, synthetic control methods [Abadie et al., 2010] (Chapter 6) construct a weighted combination of control units to approximate the counterfactual trajectory of the treated unit. Synthetic control trades parallel trends for a stable factor structure or stable relationship between treated and donor pool but instead uses pre-treatment fit: if the synthetic control matches the treated unit well before treatment, it provides a credible counterfactual after treatment for the unit-specific effect  $Y_{it}(1) - Y_{it}(0)$  of the treated unit in each post-treatment period. Inference relies on placebo checks that compare the actual effect to the distribution of placebo effects obtained by applying synthetic control to each control unit.

Hybrid methods (Chapter 7), particularly synthetic difference-in-differences (SDID) [Arkhangelsky et al., 2021], combine synthetic control with difference-in-differences by weighting both units and time periods. SDID often outperforms both DiD and SC in finite samples by improving pre-treatment fit while leveraging parallel trends when they hold. SDID still relies on a parallel trends assumption in the reweighted sample, but its explicit weighting scheme often makes this assumption more plausible and diagnosable.

### 3.10.5 Common Time-Varying Shocks Without Parallel Trends

If treated and control units are subject to common shocks—macroeconomic trends, industry demand shifts, platform algorithm changes—that affect units differentially, factor models [Bai, 2009] (Chapters 8, 9) replace the parallel-trends restriction with a stable low-rank structure for untreated outcomes and the associated stability assumptions (in particular, loadings that would not have shifted absent treatment). Interactive fixed effects models posit that untreated potential outcomes are driven by a small number of latent factors, with units loading differentially on those factors. Identification then relies on the assumption that, after conditioning on observed covariates and the estimated low-rank factor structure, residual variation in treatment is as-good-as-random with respect to the untreated potential outcomes  $Y_{it}(0)$ . Matrix completion and nuclear norm methods [Xu, 2017, Athey et al., 2021] estimate the factors and loadings jointly, imputing counterfactual outcomes for treated units in treated periods.

Factor models are particularly valuable in settings where units are heterogeneous and where common shocks dominate idiosyncratic variation. Marketing panels where all stores face category-level demand shocks, all markets experience national advertising campaigns, or all platforms face technological disruptions often exhibit low-rank structure conducive to factor methods. The panel must nonetheless have enough units and pre-treatment periods for the low-rank structure to be estimable. When  $N$  or  $T$  is very small, parallel trends-based designs may be more reliable despite their stronger assumptions. With very few units or very short pre-treatment panels, factor-based counterfactuals become unstable and high-variance, so simpler parallel trends-based designs with transparent diagnostics may be preferable even though their identifying assumptions are stronger.

### 3.10.6 Dynamic Effects and Carryover

If treatment effects exhibit carryover, anticipation, or other forms of dynamic path dependence, distributed lag models (Chapter 10) parameterise the lag structure, estimating the effect of current and past treatments on current outcomes. In the potential-outcomes notation of Chapter 2, these models aim to recover path-dependent responses  $Y_{it}(\underline{d}_i^t)$  by imposing structure on how past treatments in  $\underline{d}_i^t$  affect current outcomes. Vector autoregressions, state-space models, and structural dynamic panel models extend the approach to incorporate feedback, equilibrium, and optimisation. These methods require assumptions about the lag length, the functional form of the decay, and the absence of omitted dynamics. When those assumptions are misspecified, dynamic estimates of half-lives and long-run multipliers can be severely biased even if static average effects are estimated well. However, they enable richer substantive conclusions about the time path of effects and the cumulative long-run impact.

### 3.10.7 Spillovers and Interference

If SUTVA is violated—customers influence friends, competitors respond to rivals’ actions, advertising spills across markets—estimators must model the interference structure explicitly via exposure mappings  $h_i(D_{-i,t})$  (Chapter 11). Spatial econometric models, network models, and exposure mappings enable joint estimation of direct and spillover effects. Cluster-randomised designs that internalise spillovers within clusters provide clean identification of total effects (direct plus within-cluster spillovers), though they do not separately identify direct and spillover components without additional assumptions.

Partial identification approaches bound effects when the spillover structure is uncertain, providing intervals rather than point estimates. These bounds are often wide but more honest about the limits of what can be learned without strong assumptions. These estimators should be paired with the cluster definitions and exposure maps specified at the design stage (Sections 3.2 and 3.3), rather than retrofitted after the fact.

### 3.10.8 Machine Learning for Nuisance Functions and Heterogeneity

When the goal is to uncover heterogeneous treatment effects or to flexibly control for high-dimensional confounders, machine learning methods (Chapters 12, 13) provide powerful tools. These methods still rely on the same identification conditions as their low-dimensional counterparts—conditional independence and overlap in the panel setting—but relax functional form assumptions for the nuisance components. Double machine learning (DML) [Chernozhukov et al., 2017] uses random forests, gradient boosting, or neural networks to estimate nuisance functions—propensity scores and outcome models—while preserving valid inference on causal parameters through Neyman orthogonalisation and cross-fitting. In designs that also use factor models, ML tools can help estimate nuisance components. Causal forests [Athey et al., 2019] estimate heterogeneous treatment effects as a function of covariates, revealing which features moderate the treatment response.

High-dimensional controls (Chapter 13) employ lasso, elastic net, or group lasso to select relevant confounders from a large candidate set, enabling parsimonious specifications without manual variable selection. Post-selection inference methods [Belloni et al., 2014] provide confidence intervals that account for the data-driven selection, ensuring that uncertainty is not understated. In practice these adjustments are approximate and can be sensitive to tuning choices when the selection step is aggressive, so they should be paired with robustness checks rather than treated as a complete solution. Throughout, we treat machine learning as a tool for estimating nuisance functions and heterogeneity within a design-based framework, not as a substitute for a clear assignment mechanism and estimand.

This map from design features to methods is not exhaustive, and many studies will combine multiple approaches. The key principle is to let the design guide method selection rather than imposing a preferred method on the data. When the design supports multiple methods, comparing results across approaches provides evidence on robustness. When the design clearly favours one method, that method should be used even if it is computationally complex or yields wider confidence intervals than alternatives. Credibility trumps precision.

## 3.11 Templates and Checklists

This section provides a reusable design protocol and guidance that practitioners can adapt to their marketing panel studies. The goal is to operationalise the design principles developed in this chapter, ensuring that key decisions are made *ex ante* and that reporting is transparent and reproducible.

### 3.11.1 Design Protocol Template

A design protocol documents the key features of a study in a structured format. The protocol can be adapted to various settings—geo-experiments, phased rollouts, switchbacks, observational panels—by addressing the relevant elements.

The protocol should state the research question in clear, non-technical language, specifying the intervention, the outcome of interest, the target population, and the time horizon. It should describe the assignment mechanism: whether treatment is randomised (and if so, the unit of randomisation, stratification scheme, and randomisation protocol) or observational (and if so, what drives treatment adoption). The treatment window should specify start and end dates and justify the window length in terms of power, seasonality, and threat mitigation.

The protocol should define the target estimand precisely—whether ATT, cohort–time effects  $\tau(g, t)$ , event–time effects  $\theta_k$ , or a dose–response function  $\mu(d)$ —and align the estimand to the research question and assignment mechanism. Primary and secondary outcomes should be defined (see also Section 3.9), including data sources, measurement units, and any transformations. The estimator should be specified (for example, Callaway–Sant’Anna DiD [Callaway and Sant’Anna, 2021], synthetic control [Abadie et al., 2010], SDID [Arkhangelsky et al., 2021], interactive fixed effects [Bai, 2009]), referencing the relevant chapter for details. For each estimator, the protocol should state the identification assumptions it relies on (for example, parallel trends, low-rank factor structure, conditional independence) using the language of Chapter 2.

The inferential procedure should be specified: how standard errors will be computed (clustered by unit, by time, two-way), whether bootstrap or randomisation inference will be used, and how multiplicity will be handled. *Ex ante* diagnostics should be listed (balance checks, pre-trend plots, overlap checks, placebo checks, and negative-control outcomes where plausible) with criteria for judging credibility. Main threats to validity should be identified (seasonality, algorithm changes, spillovers, measurement issues) along with design adaptations to mitigate them. Sensitivity analyses should be pre-specified to assess robustness to assumption violations.

This protocol should be timestamped and archived before data are accessed. In observational panels, interpret this as ‘before running the primary analysis on post-treatment outcomes’. In industry settings, the archive is often internal rather than public, but it should still be auditable (for example, via version control and immutable timestamps). Deviations from the protocol should be documented with justifications, ensuring transparency about *ex post* changes. The original protocol should remain archived in its original form. Revisions and addenda should be recorded as such rather than retrofitted into the initial document.

### 3.11.2 Diagnostic Checklist

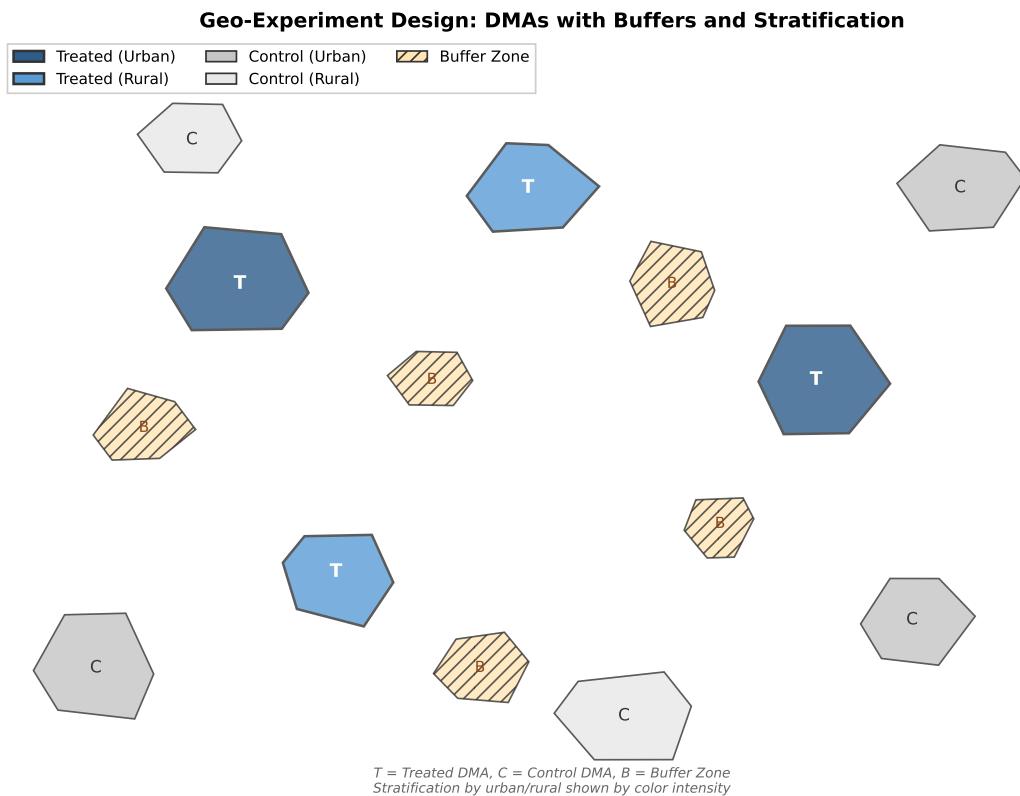
The diagnostic checklist mirrors the workflow in Chapter 17 and provides guidance for assessing the credibility of a panel design. The checklist addresses assignment transparency, covariate balance, pre-treatment trends, overlap, spillovers, seasonality and events, measurement stability, sample size and power, inference plans, and multiplicity adjustments. Each item links to a specific identification requirement: for example, transparent assignment and pre-treatment trends support parallel trends, overlap supports conditional independence, and stable measurement supports the interpretation of  $Y_{it}(d)$  across time. Each element corresponds to a specific identification requirement and should be accompanied by concrete, pre-specified questions in the analysis plan.

This checklist should be completed before data analysis begins. When diagnostics fail, the default response should be to revise the design, strengthen data collection, or narrow the estimand. When revisions are impossible, the limitations and their implications for identification should be stated explicitly in the reporting. Transparent reporting of diagnostic results builds confidence in the credibility of the final estimates.

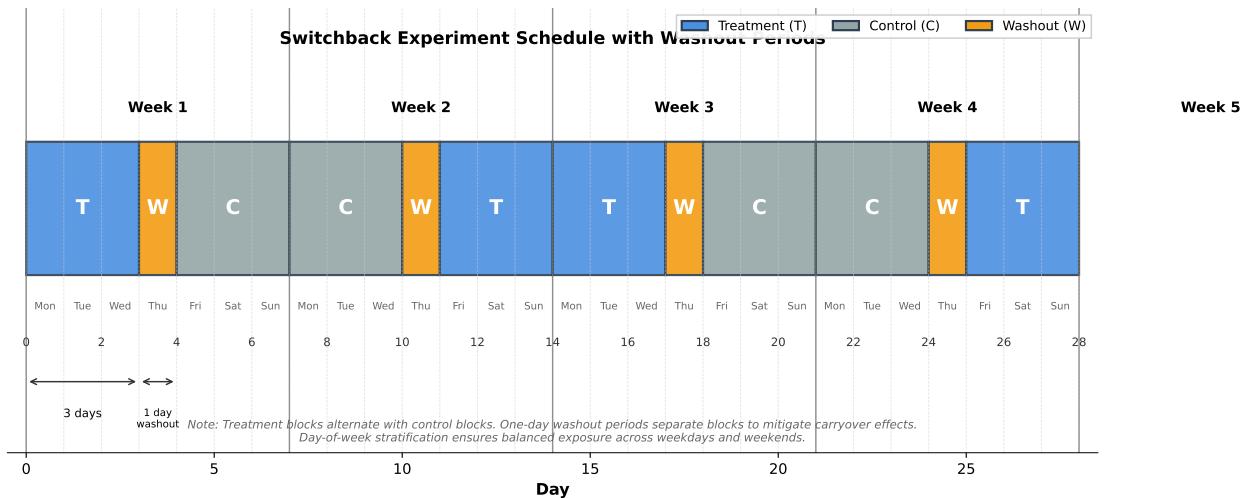
Figures 3.1, 3.2, and 3.3 provide visual templates for documenting assignment mechanisms in different design types.

Assignment Matrix: Phased Rollout with Staggered Adoption									
	T	T	T	T	T	T	T	T	
	T	T	T	T	T	T	T	T	
Cohort 1	T	T	T	T	T	T	T	T	Store 20
	T	T	T	T	T	T	T	T	Store 19
	T	T	T	T	T	T	T	T	Store 18
	T	T	T	T	T	T	T	T	Store 17
	T	T	T	T	T	T	T	T	Store 16
	T	T	T	T	T	T	T	T	Store 15
Cohort 2	C	C	T	T	T	T	T	T	Store 14
	C	C	T	T	T	T	T	T	Store 13
	C	C	T	T	T	T	T	T	Store 12
	C	C	T	T	T	T	T	T	Store 11
	C	C	T	T	T	T	T	T	Store 10
	C	C	T	T	T	T	T	T	Store 9
Cohort 3	C	C	C	C	T	T	T	T	Store 8
	C	C	C	C	T	T	T	T	Store 7
	C	C	C	C	T	T	T	T	Store 6
	C	C	C	C	T	T	T	T	Store 5
Never	C	C	C	C	C	C	C	C	Store 4
	C	C	C	C	C	C	C	C	Store 3
	C	C	C	C	C	C	C	C	Store 2
	C	C	C	C	C	C	C	C	Store 1

**Fig. 3.1** Example Assignment Matrix for a Phased Rollout



**Fig. 3.2** Geo-Cluster Map with Buffers and Stratification



**Fig. 3.3** Switchback Schedule with Washout Periods

**Table 3.1** Mapping from Assignment Mechanism to Estimands and Recommended Estimators

Assignment Mechanism	Typical Estimand	Recommended Estimators (Chapter Refs)
Randomised block (constant treatment)	ATT, event-time $\theta_k$	Standard DiD (Chapter 4), event study (Chapter 5), cluster-robust inference (Chapter 16)
Staggered adoption (heterogeneous effects)	$\tau(g, t)$ (cohort–time), event-time $\theta_k$	Callaway–Sant’Anna [Callaway and Sant’Anna, 2021], Sun–Abraham [Sun and Abraham, 2021], TWFE diagnostics (Chapter 4), event study (Chapter 5)
Single treated unit, many controls	Unit-specific effect for treated unit	Synthetic control [Abadie et al., 2010] (Chapter 6), SDID [Arkhangelsky et al., 2021] (Chapter 7), placebo inference (Chapter 16)
Common shock (single adoption time)	Event-time $\theta_k$ , cumulative effect	Event study (Chapter 5), factor models if no parallel trends (Chapters 8, 9)
Continuous intensity (observational)	Dose–response $\mu(d)$ , elasticity	Conditional DiD with high-dimensional controls (Chapter 13), DML (Chapter 12), distributed lags (Chapter 10)
Spillovers present	Total effect under cluster assignment, and direct + spillover effects under specified $h_i(D_{-i,t})$	Spatial/network models, cluster designs (Chapter 11), partial identification bounds (Chapter 11)

### 3.11.3 Pre-Analysis Plan Template for a Marketing Panel Study

A pre-analysis plan should document all critical design decisions before analysis begins. The plan should include the study title and the date the protocol was finalised. It should state the research question in one sentence and describe the assignment mechanism (whether randomisation or observational assignment). The data structure should be specified, including  $N$ ,  $T$ , panel type, and cohort structure if applicable. The treatment window should specify start and end dates along with measurement periods.

The plan should define the primary estimand (ATT,  $\tau(g, t)$ ,  $\theta_k$ , or other) with mathematical definition, and align it to the research question. The primary outcome should be defined, including data source, measurement unit, and any transformations. Secondary outcomes should be listed with clear indication of their exploratory status. The estimator should be named with reference to the relevant chapter of this book.

The identification assumption should be stated explicitly (parallel trends, factor structure, conditional independence, or other). The inference procedure should specify the clustering choice, whether bootstrap or randomisation inference will be used, the significance level, and any multiplicity adjustment. *Ex ante* diagnostics should be listed, including balance checks, pre-trend diagnostics, overlap checks, and placebo checks. Sensitivity analyses should be pre-specified to assess robustness to assumption violations.

Main threats to validity should be identified (seasonality, algorithm changes, spillovers, measurement issues) along with design adaptations to mitigate them. The reporting plan should specify which tables,

figures, and aggregations will be reported, with commitment to reporting all pre-specified analyses regardless of results. This template ensures that all critical design decisions are documented before analysis begins, reducing researcher degrees of freedom and enhancing the credibility of the study.

By following the design principles, diagnostic workflows, and reporting standards outlined in this chapter, practitioners can conduct marketing panel studies that generate credible causal estimates, withstand scrutiny, and inform strategic decisions with confidence. Design-based thinking anchors inference in the assignment mechanism, makes assumptions explicit and assessable, and prioritises transparency over sophistication.

The methods developed in subsequent chapters operationalise these principles for the estimands introduced in Chapter 2, providing the tools needed to estimate causal effects, quantify uncertainty, and diagnose threats to validity in the complex, dynamic, and strategically rich environments that characterise modern marketing. The next chapters put these principles to work in specific estimators—beginning with difference-in-differences for staggered adoption in Chapter 4—and show how to implement them in the marketing contexts introduced here.

## 3.12 Interlude: Design-First and Structural IO

This interlude puts our design-first approach next to the structural tradition in industrial organisation and marketing science. We clarify what each tradition targets, what assumptions do the identifying work, and where each approach is most useful. We then sketch hybrids that use design-based evidence to discipline structural counterfactuals.

### 3.12.1 Two Traditions, Different Targets

Design-first methods aim to estimate credible causal effects of concrete interventions under transparent assumptions with diagnostics. We emphasise identification strategies that can be stress-tested with falsification checks and sensitivity analysis, including parallel trends, pre-trend diagnostics and placebo checks, overlap, donor curation, and design-faithful inference.

Structural IO recovers behavioural primitives and market mechanisms. A canonical example is the Berry–Levinsohn–Pakes (BLP) model of demand for differentiated products Berry et al. [1995]. BLP posits a utility structure, corrects for the endogeneity of prices, and recovers own- and cross-price elasticities. Combined with a supply side, it supports counterfactual simulations of mergers, taxes, assortment changes, and pricing policies. Follow-on work extends identification and practice [Nevo, 2001, Berry et al., 2014] and offers applied guidance [Train, 2009, Einav and Levin, 2010, Aguirregabiria and Mira, 2010].

### 3.12.2 Strengths and Vulnerabilities

Design-first excels in transparency and internal validity. Diagnostics make assumptions visible, flag when they fail, and keep the link between design and estimand explicit. The resulting estimates are directly decision-ready for the specific design and time window under study. That strength is also the main limitation. The scope of inference is local. We learn little about behaviour far from observed support, about general-equilibrium feedback, or about multi-firm strategic interaction.

Structural work excels at explaining mechanisms and supporting broad counterfactuals. By embedding behaviour in an explicit model, it can simulate market-wide policy changes and recover welfare-relevant objects. This power comes with familiar risks. Utility or conduct may be misspecified, instruments or exclusion restrictions may be hard to justify, and computational burdens can crowd out robustness checks. In the potential-outcomes language of Chapter 2, these risks show up as untestable exclusion restrictions, functional-form assumptions on  $Y_{it}(d)$  far from observed support, and equilibrium-selection assumptions that are often only weakly disciplined by the data.

### 3.12.3 A Productive Synthesis

We advocate a pragmatic hybrid workflow. Begin by documenting credible effects with design-first tools such as DiD and event studies, synthetic control and SDID, and factor-based designs, using full diagnostics (Chapter 17). Then use those estimates to discipline structural models by calibrating or validating local elasticities, substitution patterns, and dynamics against design-based evidence. For example, price elasticities recovered from a BLP model can be checked against DiD-based elasticity estimates from a geo-experiment before being used in merger simulations. If the structural model cannot match credible reduced-form moments without implausible parameter values, treat the counterfactuals as sensitivity analysis, not primary evidence. Finally, deploy structure to study policy frontiers that require mechanism or equilibrium reasoning, and report sensitivity to model choices and instruments.

This sequence aligns evidence with decisions. Design-first establishes what happened under a feasible design. Structure explores why it happened and what might happen under unobserved policies.

### 3.12.4 BLP in One Paragraph

BLP specifies indirect utility for consumer  $n$  choosing product  $j$  in market  $t$  as  $u_{njt} = x'_{jt}\beta_n - \alpha p_{jt} + \xi_{jt} + \varepsilon_{njt}$ , where  $x_{jt}$  collects observed product characteristics,  $p_{jt}$  is price,  $\xi_{jt}$  is an unobserved product-market shock, and  $\varepsilon_{njt}$  is an idiosyncratic error. The random coefficients  $\beta_n$  capture heterogeneous tastes. Market shares map to mean utilities via an inversion, and instruments address the endogeneity of  $p_{jt}$  under an exclusion restriction. Estimation recovers demand elasticities and, with a supply system, markups and conduct. Counterfactuals change  $x$  or  $p$  and solve for the new equilibrium. Performance hinges on functional forms, instrument quality, and equilibrium selection. From a design-first perspective, the critical question is which price variation the instruments declare exogenous, and how far the resulting counterfactual simulations extrapolate beyond the support of observed variation.

### 3.12.5 When to Prefer Which

Use design-first when the goal is to measure realised effects of specific rollouts, when diagnostics are paramount, or when platforms induce nonstationarity and interference that complicate modelling. Use structure when the question demands market-wide simulations, welfare analysis, or strategic counterfactuals far from observed support. Combine them when decisions require both credible local evidence and mechanism-based exploration.

### 3.12.6 Reporting and Diagnostics Across Traditions

Structural work benefits from design-style diagnostics. You can run pre-trend diagnostics and placebo checks for auxiliary reduced forms, inspect weight and influence, vary instrument sets, and report first-stage strength and alternative equilibrium assumptions transparently. In practice this means treating the reduced-form implications of the structural model as design-based objects: specify clear estimands, document the sources of identifying variation, and subject them to the same placebo, balance, and influence diagnostics used elsewhere in this chapter. Design-first studies benefit from structural sanity checks. Estimated elasticities or substitution patterns should accord with economic reason and prior evidence, such as appropriate price elasticity signs and magnitudes, plausible substitution patterns between products, and pass-through bounds consistent with market structure.

### 3.12.7 Further Reading

For BLP and demand estimation, see Berry et al. [1995], Nevo [2001], Berry et al. [2014], Train [2009]. For dynamics and games, see Aguirregabiria and Mira [2010]. For field overviews, see Einav and Levin [2010].



**Part II**

**Differences-in-Differences and Event Studies**



## Chapter 4

# Difference-in-Differences: From Canonical to Staggered

Difference-in-differences is the workhorse of causal panel data analysis in marketing. Loyalty programme rollouts, advertising campaign launches, platform expansions, and pricing experiments all generate panel data where treatment is adopted at different times across units. The DiD framework exploits this variation, comparing changes in outcomes for treated units to changes for control units to isolate causal effects. The logic is elegant: if treated and control units would have followed similar trends in the absence of treatment, then the difference in their post-treatment trends identifies the causal effect of treatment.

The canonical  $2 \times 2$  DiD design—two groups, two periods, treatment switching on for one group in the second period—provides conceptual clarity and a clean identification argument grounded in parallel trends and corresponds to the block assignment mechanism introduced in Section 3.2. However, most marketing applications feature staggered adoption, where units adopt treatment at different times, and treatment effects that vary across cohorts and evolve dynamically over time. Extending DiD to these settings requires care. Traditional two-way fixed effects regressions can produce misleading estimates when treatment effects are heterogeneous, assigning negative weights to some comparisons and obscuring the true pattern of effects. Modern heterogeneity-robust estimators solve this problem by constructing clean comparisons and aggregating cohort-time effects transparently.

In this chapter, we develop the DiD framework from its canonical form through to modern methods for staggered adoption. We begin with the canonical design, establishing the parallel trends assumption and the potential outcomes framework. We then define estimands for staggered adoption—cohort-time effects  $\tau(g, t)$  and event-time effects—and explain the negative weighting problem in two-way fixed effects. Modern estimators from Callaway and Sant'Anna [2021], Sun and Abraham [2021], and others provide solutions, and we show how to implement them, conduct diagnostics, and assess robustness. In repeated cross-sections, doubly robust DiD estimators often assume that the treated and comparison groups have stable composition over time. Sant'Anna and Xu [2026] discuss estimators that remain valid under compositional change.

Throughout, we emphasise practical guidance for marketing applications: how to define estimands that align with business questions, how to diagnose assumption violations, how to choose among competing estimators, and how to report results transparently. Chapter 5 extends these methods to event-study designs that trace dynamic treatment paths and diagnose anticipation effects.

### Notation Guide for Dynamic and Heterogeneous Effects

This chapter uses the notation established in Chapter 2:

Core notation:

- $G_i \in \{1, 2, \dots, T, \infty\}$ : Adoption period for unit  $i$  ( $\infty$  = never treated)
- $D_{it} = \mathbf{1}\{t \geq G_i\}$ : Treatment indicator (1 if unit  $i$  is treated by period  $t$ )
- $\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$ : Cohort-time effect for cohort  $g$  in period  $t \geq g$
- $D_{it}^k = \mathbf{1}\{t - G_i = k\}$ : Event-time indicator
- $\theta_k = \mathbb{E}[Y_{i,G_i+k}(G_i) - Y_{i,G_i+k}(\infty) | G_i < \infty]$ : Event-time effect at relative time  $k = t - G_i$
- Equivalently,  $\theta_k = \sum_{g:g+k \leq T} w_{gk} \tau(g, g+k)$ , where  $w_{gk}$  are cohort shares among cohorts observed at event time  $k$
- LRM =  $\sum_{k=0}^K \theta_k / \theta_0$ : Long-run multiplier (cumulative effect up to horizon  $K$  relative to the immediate effect)

Comparison groups:

- *Never-treated*: Units with  $G_i = \infty$  (never adopt during sample window)
- *Not-yet-treated*: Units with  $G_i > t$  (will adopt later but currently untreated)
- Callaway and Sant'Anna [2021] use not-yet-treated or never-treated units as controls for each  $(g, t)$ . Sun and Abraham [2021] re-express DiD as event-time regressions using last-treated (or never-treated) cohorts as references.

Aggregation weights  $w_{gk}$ :

- Cohort-size weights:  $w_{gk} \propto n_g$  (number of units in cohort  $g$ )
- Equal weights:  $w_{gk} = 1/|\mathcal{G}_k|$  where  $\mathcal{G}_k$  is the set of cohorts observed at event-time  $k$
- Traditional two-way fixed effects can implicitly use weights that are *negative* when effects are heterogeneous, causing already-treated units to serve as “controls” for later-treated units

Overall ATT aggregation:

$$\text{ATT}^{\text{agg}} = \sum_{g<\infty} \sum_{t \geq g} w_{gt} \tau(g, t), \quad \text{with } \sum_{g,t} w_{gt} = 1$$

Different weighting schemes answer different questions. Cohort-size weights emphasise large cohorts, calendar-time weights emphasise recent periods, and equal weights treat all  $(g, t)$  cells symmetrically. When  $w_{gt}$  is proportional to the number of treated observations in cell  $(g, t)$ ,  $\text{ATT}^{\text{agg}}$  coincides with the panel-level ATT over ever-treated units and post-treatment periods.

Pre-treatment periods have  $k < 0$  (leads) and post-treatment periods have  $k \geq 0$  (lags). We normalise  $\theta_{-1} = 0$ . Pre-trend coefficients  $\theta_k$  for  $k < -1$  are used as diagnostics. Large systematic deviations from zero raise concern about parallel trends or anticipation.

## 4.1 Canonical 2x2 DiD and Parallel Trends

A retailer launches a loyalty programme in 50 stores and wants to know whether it lifted sales. Fifty control stores that did not receive the programme provide a benchmark. Before the launch, treated stores averaged £100K in quarterly sales while control stores averaged £80K. After the launch, treated stores averaged £130K and control stores £95K. The treated stores grew by £30K, but control stores also grew—by £15K. The difference-in-differences estimate attributes the £15K excess growth to the programme. This simple logic underpins one of the most widely used causal designs in applied economics.

The canonical  $2 \times 2$  DiD design—two groups, two periods, treatment switching on for one group in the second period—provides conceptual clarity and a clean identification argument grounded in parallel trends and corresponds to the block assignment mechanism introduced in Section 3.2. However, most marketing applications feature staggered adoption, where units adopt treatment at different times, and treatment effects that vary across cohorts and evolve dynamically over time. Extending DiD to these settings requires care. Traditional two-way fixed effects regressions can produce misleading estimates when treatment effects are heterogeneous, assigning negative weights to some comparisons and obscuring the true pattern of effects. Modern heterogeneity-robust estimators solve this problem by constructing clean comparisons and aggregating cohort-time effects transparently.

The design requires two key assumptions. Parallel Trends asserts that treated and control units would have followed the same trend absent treatment. The Stable Unit Treatment Value Assumption (SUTVA) rules out interference between units. We formalise both below.

Let  $\bar{Y}_{\text{treat,pre}}$  denote the average outcome for treated units in the pre-treatment period,  $\bar{Y}_{\text{treat,post}}$  the average for treated units in the post-treatment period, and analogously for control units. The estimator is

$$\hat{\tau}^{\text{DiD}} = (\bar{Y}_{\text{treat,post}} - \bar{Y}_{\text{treat,pre}}) - (\bar{Y}_{\text{control,post}} - \bar{Y}_{\text{control,pre}}).$$

The first term is the change in outcomes for treated units from pre to post. The second term is the change for control units. The estimator is the difference between these two changes.

Returning to our store example, the estimate is  $(130 - 100) - (95 - 80) = 30 - 15 = £15K$ . Notice that treated stores started with higher sales (£100K vs £80K), but the method does not require equal levels. It requires only that the £15K growth in control stores would have been the same for treated stores absent treatment. Level differences do not invalidate the design in principle, but they often signal trend differences in practice—a point we return to below.

Under the Parallel Trends assumption, this estimator has a causal interpretation. Parallel Trends asserts that in the absence of treatment, treated and control units would have experienced the same change in outcomes from pre to post. Formally, the assumption is

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid i \in \text{treated}] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid i \in \text{control}]$$

for the relevant pre and post periods. The notation  $Y_{it}(0)$  denotes the potential outcome under no treatment. Later, when we turn to staggered adoption, we use the adoption-time shorthand  $Y_{it}(\infty)$  for the never-treated path.

Parallel Trends does not require that treated and control units have the same levels of outcomes in the pre-treatment period. Levels can differ due to time-invariant characteristics (which are absorbed by unit fixed effects in a regression framework). Parallel Trends also does not require that treated and control units exhibited identical growth rates throughout their entire pre-treatment history. What matters is the specific counterfactual change from the final pre-treatment period to the post-treatment period. Treated and control units could have followed different trajectories in the distant past, yet still satisfy Parallel Trends if their period-to-period changes would have converged by the treatment window. Conversely, units with similar historical growth rates might violate Parallel Trends if a shock differentially affects their trajectories precisely when treatment occurs. This distinction is subtle but important: Parallel Trends is a statement about one specific counterfactual change, not about the entire historical trajectory.

When is it reasonable to believe that trends converge despite different histories? Convergence is plausible when treated and control units are exposed to the same overarching market forces, macro conditions, and seasonal patterns that dominate idiosyncratic dynamics, and implausible when the factors driving historical divergence continue into the treatment window (for example, gentrifying versus declining neighbourhoods). The burden is on the analyst to explain why convergence is credible in the specific context.

In the canonical two-period case, the general Parallel Trends assumption from Chapter 2 (Assumption 6) reduces to a statement about the change from the final pre-treatment period to the post-treatment period.

#### **Assumption 9 (Parallel Trends in the $2 \times 2$ Design)**

$$\mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{treated}] = \mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{control}].$$

This is the two-period special case of the Parallel Trends assumption introduced in Chapter 2 and of the cohort-level condition in Section 2.3, with cohorts defined by treatment status (treated vs control) and adoption time fixed at the post period for treated units.

In addition to Parallel Trends, valid inference requires the Stable Unit Treatment Value Assumption (SUTVA) from Chapter 2 (Assumption 3), which in this canonical setting reduces to the requirement that each store's potential outcomes depend only on its own treatment status, not on the assignment of other stores. Without SUTVA, the potential outcome  $Y_{it}(0)$  is not a well-defined scalar because it would depend on the entire assignment matrix  $\mathbf{D}$ , not just on unit  $i$ 's own treatment status. When interference matters, we later summarise it through an exposure mapping like  $h_i(D_{-i,t})$ . The notation we have been using presupposes SUTVA. While often plausible in isolated settings, SUTVA violations are common in marketing through spillovers, network effects, and competitive responses (see Section 4.9 and Chapter 11).

Under these assumptions, the DiD estimator consistently estimates the average treatment effect on the treated (ATT):

$$\text{ATT} = \mathbb{E}[Y_{i,\text{post}}(1) - Y_{i,\text{post}}(0) \mid i \in \text{treated}].$$

(This is the  $2 \times 2$  special case of the ATT defined in Section 2.3). The observed change for treated units is

$$\mathbb{E}[Y_{i,\text{post}} - Y_{i,\text{pre}} \mid i \in \text{treated}] = \mathbb{E}[Y_{i,\text{post}}(1) - Y_{i,\text{pre}}(0) \mid i \in \text{treated}]$$

because treated units are treated in the post period and untreated in the pre period. For control units, the observed change is

$$\mathbb{E}[Y_{i,\text{post}} - Y_{i,\text{pre}} \mid i \in \text{control}] = \mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{control}].$$

The control-group change is fully observed because control units are never treated. This is the key: we use the control group's observed change as a stand-in for the treated group's counterfactual change. Parallel Trends is the assumption that licenses this substitution.

Taking the difference, define the population DiD contrast as

$$\tau^{\text{DiD}} = \mathbb{E}[Y_{i,\text{post}}(1) - Y_{i,\text{pre}}(0) \mid i \in \text{treated}] - \mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{control}].$$

Under Parallel Trends, the second term equals  $\mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{treated}]$ , so

$$\tau^{\text{DiD}} = \mathbb{E}[Y_{i,\text{post}}(1) - Y_{i,\text{post}}(0) \mid i \in \text{treated}] = \text{ATT}.$$

The sample estimator  $\hat{\tau}^{\text{DiD}}$  based on group-period means is a consistent estimator of this quantity.

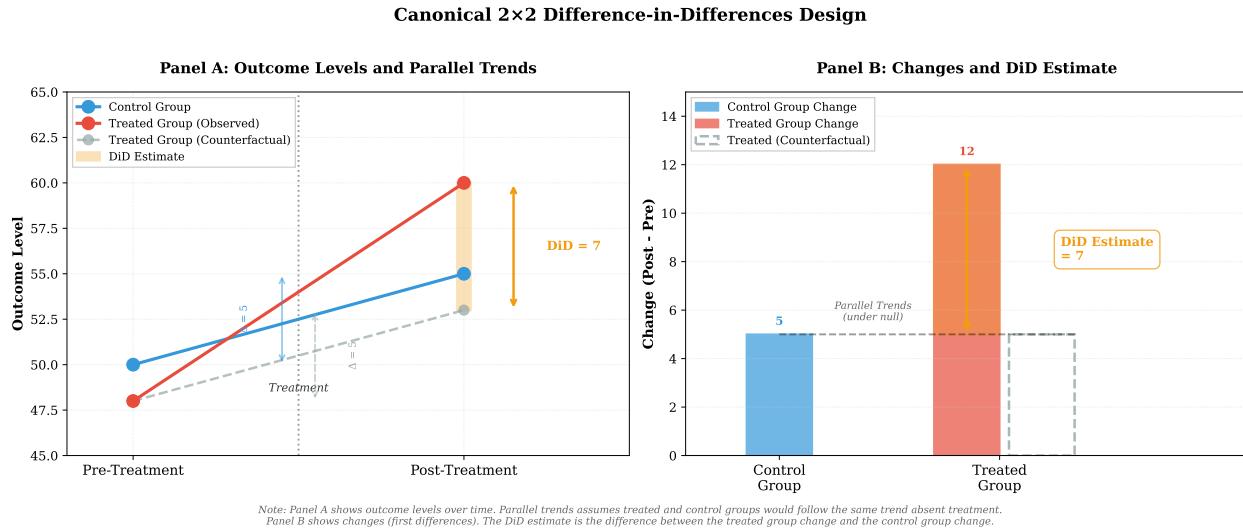
The canonical  $2 \times 2$  design is conceptually simple and provides a clear benchmark for understanding DiD logic, but it is rare in marketing applications. Most marketing interventions unfold over multiple periods, not just two, and treatment is often adopted in staggered fashion across units rather than simultaneously. Extending  $2 \times 2$  logic to panels with multiple periods and staggered adoption requires care: heterogeneous effects and comparisons with already-treated units break the simple differencing intuition. Figure 4.1 illustrates the canonical  $2 \times 2$  design and the Parallel Trends assumption visually.

In regression form for two groups and multiple periods,

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \varepsilon_{it}$$

includes unit fixed effects  $\alpha_i$  (which absorb time-invariant differences between treated and control units), time fixed effects  $\lambda_t$  (which absorb common trends or shocks affecting all units in a given period), and a treatment indicator  $D_{it}$  that switches from zero to one for treated units in the post-treatment period. The coefficient  $\tau$  on the treatment indicator is numerically identical to the  $2 \times 2$  DiD estimator when only two periods are observed. It generalises the estimator to settings with multiple pre-treatment and post-treatment periods provided that the parallel trends assumption holds in all periods and that treatment effects are constant over time (that is, Assumption 8 holds within this two-group design).

Parallel trends is fundamentally an assumption about unobservables—about what would have happened to treated units had they not been treated. We cannot verify it directly because we never observe  $Y_{it}(0)$  for treated units in treated periods. However, we can bring indirect evidence to bear. If treated and control units exhibited parallel trends in multiple pre-treatment periods, this is consistent with the idea that parallel trends might have continued into the post-treatment period. Placebo checks that apply the estimator to



**Fig. 4.1** Canonical 2 × 2 DiD: Timing and Parallel Trends

The figure illustrates the canonical difference-in-differences design with two groups and two periods. The left panel shows the treatment timing: control units (blue) remain untreated throughout, while treated units (green) switch from untreated to treated in the post-treatment period. The right panel illustrates the parallel trends assumption. Solid lines show observed outcomes: treated units start higher and grow faster. The dashed line shows the counterfactual path for treated units under parallel trends—what their outcomes would have been absent treatment. The DiD estimator measures the vertical gap between the observed treated outcome and the counterfactual at the post-treatment period. Parallel trends asserts that the dashed counterfactual line is parallel to the control group's trajectory, not that pre-treatment levels are equal.

pre-treatment periods only—treating an earlier period as if it were the post-treatment period—should yield estimates near zero when the design is behaving well. In the event-study notation of Section 2.3, these diagnostics check whether pre-treatment coefficients  $\theta_k$  for  $k < 0$  are close to zero. Chapter 5 discusses how to interpret such profiles. Divergence in pre-treatment trends is a red flag that signals parallel trends is unlikely to hold.

A word of caution: pre-trend diagnostics have well-documented limitations [Roth, 2022]. They have low power against many plausible violations, so small estimated pre-trends do not validate parallel trends. Worse, if you select specifications conditional on “passing” a pre-trend check, you can bias subsequent estimates because you are selecting for settings where pre-treatment noise happened to cancel out. Pre-trend diagnostics are useful—divergent pre-trends are a clear warning—but they cannot validate the assumption. The credibility of parallel trends ultimately rests on substantive arguments about the assignment mechanism and the data-generating process, not on a single p-value.

The plausibility of parallel trends depends on the substantive context. If treated units are selected for treatment based on anticipated future outcomes—for example, if a retailer assigns a loyalty programme to stores expected to experience rapid sales growth—then parallel trends is violated because treated units would have grown faster than controls even without treatment. Several factors can undermine parallel trends in marketing. Treated units may be selected based on growth, with high-growth units being treated first. Treatment timing may be endogenous, responding to unit-specific shocks. Competitive dynamics can also pose a challenge, as rivals may respond to treatment announcements, contaminating control outcomes. Ad-

ditionally, seasonality confounds can arise when treatment timing is correlated with seasonal patterns that differ across units.

When treatment is assigned based on past outcomes or on characteristics correlated with trends, conditioning on those characteristics can restore identification. Conditional parallel trends refines Assumption 9 by allowing treated and control units to differ in observed covariates  $X_i$ , provided that trends align within covariate strata.

**Assumption 10 (Conditional Parallel Trends)** For observed pre-treatment covariates  $X_i$ :

$$\mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{treated}, X_i] = \mathbb{E}[Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) \mid i \in \text{control}, X_i].$$

In the loyalty-programme example, such covariates  $X_i$  might include store size, local income, competitive density, and pre-treatment growth rates, chosen because they plausibly explain both treatment assignment and trends.

Conditional parallel trends is often more plausible than unconditional parallel trends in observational settings. If stores receiving a loyalty programme differ systematically from control stores in size, demographics, or competitive intensity, conditioning on these covariates through regression adjustment, propensity score weighting, or matching can make the trends assumption more credible. The cost is that identification now relies on correct specification of the conditioning set and on sufficient overlap in covariate distributions between treated and control units.

Mis-specifying the conditioning set can introduce bias rather than remove it. Conditioning on a post-treatment variable—a so-called "bad control"—can open paths through colliders or mediators and invalidate the design even when pre-treatment covariates are correctly handled. Omitting a confounder that drives both treatment assignment and trends leaves selection bias intact. Including irrelevant covariates inflates variance without improving identification. The analyst must justify each covariate on substantive grounds: why does conditioning on this variable make parallel trends more plausible? Mechanical inclusion of all available covariates is not a substitute for careful reasoning about the assignment mechanism.

**Doubly robust estimation.** When invoking conditional parallel trends, the analyst faces a choice: model the outcome (regression adjustment) or model treatment assignment (propensity score weighting). Doubly robust DiD estimators [Sant'Anna and Zhao, 2020] combine both approaches, remaining consistent if either the outcome model or the propensity score model is correctly specified (but not necessarily both). This robustness to partial misspecification is valuable in marketing applications where neither model is likely to be exactly correct. The estimator constructs inverse-probability-weighted regression-adjusted (IPWRA) comparisons that reweight control units to match the covariate distribution of treated units while also adjusting for outcome differences. Modern implementations in the `did` and `DRDID` R packages make doubly robust estimation straightforward.

Marketing applications of canonical DiD include evaluating the impact of discrete events that affect all treated units simultaneously. A regulatory change that applies to all firms in a single period, a major advertising campaign launched nationwide at a single date, or a platform algorithm update rolled out globally at once all produce canonical DiD designs. The treated group consists of units exposed to the event, and

the control group consists of comparable units not exposed (perhaps in different geographic regions, different segments, or different platforms). The identifying variation comes from the discrete timing of the event and from the availability of control units that provide a counterfactual trend.

Despite its conceptual appeal, the canonical  $2 \times 2$  DiD is ill-suited to many marketing panel datasets where treatment adoption is staggered, where treatment effects evolve dynamically over time, and where heterogeneity across units or cohorts is substantively important. The next section defines estimands for staggered adoption that accommodate heterogeneous treatment effects across cohorts and time.

## 4.2 Causal Estimands for Staggered Adoption

A retailer rolls out a loyalty programme to different regions over two years. Some regions adopt in Q1 2022, others in Q3 2022, others in Q1 2023. The firm wants to know: What is the overall effect? How does the effect evolve over time? Do early adopters benefit more than late adopters? Each question requires a different estimand, and conflating them can mislead.

Staggered adoption designs feature units adopting treatment at different times. Some units adopt in period  $g = 2$ , others in period  $g = 5$ , others in period  $g = 8$ , and some never adopt during the observation window (units with  $G_i = \infty$ ). This creates rich variation: early adopters can be compared to never-adopters and to not-yet-treated units, and within-cohort comparisons over time trace dynamic effects. But it also creates complexity. Treatment effects may differ across cohorts (early vs late adopters), and effects may evolve over time post-adoption. These sources of heterogeneity mean a single summary statistic obscures important variation and can mislead.

When no never-treated units exist—as happens when programmes eventually roll out everywhere—identification relies entirely on not-yet-treated units as controls. This raises additional concerns: anticipation effects may contaminate not-yet-treated outcomes if units change behaviour in expectation of future treatment, and selection into timing may correlate with potential outcomes. We return to these issues in Section 4.3.

Modern DiD methods address this complexity by defining estimands that respect heterogeneity and by aggregating those estimands in transparent, interpretable ways. The fundamental building block is the cohort-time effect  $\tau(g, t)$ , the average treatment effect for units in adoption cohort  $g$  in calendar period  $t$ , conditional on  $t \geq g$  (so that cohort  $g$  is treated in period  $t$ ). Following Chapter 2 and Section 2.3, we recall the cohort-time effect

$$\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g], \quad t \geq g,$$

where  $G_i$  is the adoption time for unit  $i$ . This estimand allows treatment effects to vary freely across cohorts and across calendar periods. A unit adopting in quarter two may experience a different effect in quarter three than a unit adopting in quarter four experiences in quarter five, either because the units differ in characteristics or because macroeconomic conditions, competitive landscapes, or other time-varying factors differ across periods.

### Worked Example: Computing Event-Time Effects

To make the aggregation concrete, consider a stylised loyalty programme rollout with three cohorts:

The event-time effect  $\theta_0$  (immediate impact) is a weighted average of cohort-specific effects at  $k = 0$ :

$$\theta_0 = \frac{100 \cdot 8 + 200 \cdot 6 + 150 \cdot 5}{100 + 200 + 150} = \frac{800 + 1200 + 750}{450} = \frac{2750}{450} \approx 6.1.$$

**Table 4.1** Hypothetical Cohort-Time Treatment Effects

Cohort $g$	Stores	$k = 0$	$k = 1$	$k = 2$	$k = 3$
$g = 2$	100	8	10	12	14
$g = 4$	200	6	9	11	—
$g = 6$	150	5	—	—	—

The effect one quarter post-adoption ( $\theta_1$ ) averages only cohorts  $g = 2$  and  $g = 4$ , since cohort  $g = 6$  has not yet reached  $k = 1$  by the end of the panel:

$$\theta_1 = \frac{100 \cdot 10 + 200 \cdot 9}{100 + 200} = \frac{1000 + 1800}{300} = 9.3.$$

This weighted aggregation ensures that each cohort contributes to event-time effects in proportion to its size, but only for event times it has experienced. These are point estimates. In practice, each  $\theta_k$  is an estimator of the formal parameter defined in Chapter 2, and sampling uncertainty means confidence intervals are required for any substantive dynamic claim. The apparent increase from  $\theta_0 = 6.1$  to  $\theta_1 = 9.3$  could reflect genuine effect growth, or it could be noise. Confidence intervals (discussed in Chapter 16) are essential for distinguishing real dynamics from sampling variation.

Aggregating  $\tau(g, t)$  across cohorts and time produces summary measures tailored to different policy questions. One such summary for ever-treated units is a weighted average of cohort-time effects:

$$\text{ATT}^{\text{agg}} = \sum_{g < \infty} \sum_{t \geq g} w_{gt}^{\text{ATT}} \tau(g, t),$$

where the weights  $w_{gt}^{\text{ATT}}$  reflect the relative sample sizes or other aggregation priorities (for example, weighting by treated unit-periods, by cohort size, or by importance to the business). Here  $\text{ATT}^{\text{agg}}$  is a population estimand and  $w_{gt}^{\text{ATT}}$  are population weights that sum to one over all treated  $(g, t)$  cells. Specific estimators approximate these weights in different ways. When  $w_{gt}^{\text{ATT}}$  is proportional to the treated cell share,  $\text{ATT}^{\text{agg}}$  coincides with the panel-level ATT defined in Section 2.3. Different weighting schemes can produce different overall estimates even when the underlying  $\tau(g, t)$  are the same, so transparency about weights is essential. The weights are not merely an aggregation convenience—they define the target population. Weighting by cohort size targets the effect for the average treated unit. Weighting by treated unit-periods targets the effect for the average treated unit-period. These are different estimands with different policy implications, and the choice should be driven by the business question.

Event-time aggregation pools cohort-time effects by time since adoption rather than by calendar time. Define event time  $k = t - G_i$  as the number of periods since unit  $i$  adopted treatment. The event-time effect  $\theta_k$  is the average effect  $k$  periods post-adoption, pooled across cohorts:

$$\theta_k = \sum_{g: g+k \leq T} w_{g|k}^{\text{event}} \tau(g, g+k),$$

where  $w_{g|k}^{\text{event}}$  are cohort weights (typically proportional to cohort size) normalised to sum to one over cohorts with  $g + k \leq T$ . Event-time effects are natural when the goal is to trace the dynamic evolution of the treatment response. How does the effect grow or decay over time? Does the effect take multiple periods to fully materialise, as might occur with habit formation in a loyalty programme? Does the effect dissipate quickly, as might occur with a temporary promotion? The sequence  $\{\theta_k\}_{k=0}^K$  answers these questions by estimating the effect at each event time  $k$  relative to a baseline (typically  $k = -1$ , the period immediately before adoption). As in Chapter 2, these event-time effects are interpreted relative to a pre-adoption baseline period, so  $\theta_k$  captures the incremental impact  $k$  periods after adoption compared with the immediate pre-adoption period.

**Composition Bias in Event-time Effects** Because different cohorts contribute to different event times, the composition of units changes across  $k$ . If cohort  $g = 2$  experiences systematically larger effects than cohort  $g = 6$ , and cohort  $g = 2$  contributes to all event times  $k \geq 0$  while cohort  $g = 6$  contributes only to  $k = 0$ , then the event-time profile  $\{\theta_k\}$  conflates treatment effect dynamics with cohort composition. This composition bias means that an upward-sloping event-time profile could reflect either growing effects over time or simply that early-adopting cohorts (which dominate later event times) have larger effects. To diagnose composition bias, decompose event-time effects by cohort and inspect cohort-specific event-time profiles. If cohort  $g = 2$  experiences systematically larger effects than cohort  $g = 6$ , an upward-sloping aggregate profile  $\{\theta_k\}$  can arise purely as an artefact of composition shifts toward high-effect early cohorts in later event times, even if unit-level dynamics are flat. If cohort-specific profiles have the same shape (same slopes across event times) but different intercepts, the aggregate profile reflects composition, not dynamics. If the shapes differ—some cohorts show growing effects while others show decay—heterogeneity in dynamics is present, and the aggregate profile masks important variation. Composition bias does not, by itself, violate identification if each  $\tau(g, t)$  is identified. It affects interpretation of the aggregated event-time profile and which population the profile describes.

**Balanced vs unbalanced event-time panels.** Some estimators, particularly Sun and Abraham’s interaction-weighted estimator, require restricting attention to a “balanced” event-time window where all cohorts are observed. For example, if the earliest cohort adopts in period 2 and the latest in period 8, and the panel ends in period 10, then event time  $k = 3$  is observed only for cohorts  $g \leq 7$ . Restricting to a balanced window (say,  $k \in \{-2, \dots, 2\}$ ) ensures that the same cohorts contribute to all event-time coefficients, eliminating composition effects but potentially discarding valuable long-horizon information. Crucially, balanced windows change the estimand: you average only over cohorts that survive the restriction, not all ever-treated units. Other estimators, such as Callaway and Sant’Anna’s, can work with unbalanced event-time windows but require careful interpretation: each  $\theta_k$  averages over a different set of cohorts. When reporting event-time effects, document whether the panel is balanced across event times and, if not, which cohorts contribute to each  $k$ .

We use  $\theta$  for event-time effects and  $\tau$  for calendar-time and cohort-specific effects to distinguish the different aggregation schemes (see Section 2.3 for global definitions).

Calendar-time aggregation pools cohort-time effects by calendar period rather than by event time:

$$\tau_t = \sum_{g \leq t} w_{g|t}^{\text{cal}} \tau(g, t),$$

where the sum is over cohorts that are treated in period  $t$  and  $w_{g|t}^{\text{cal}}$  are calendar-time weights (which may differ from the overall ATT weights  $w_{gt}^{\text{ATT}}$ ). Calendar-time effects are natural when the goal is to estimate the contemporaneous impact of the programme in specific periods, accounting for macroeconomic conditions, seasonal effects, or other time-specific factors. For example, if a loyalty programme is rolled out over two years and the goal is to estimate the total sales impact in each quarter, calendar-time aggregation provides the answer by summing effects across all treated units in each quarter.

Cohort-specific aggregations pool over time for a single cohort:

$$\tau_g = \sum_{t \geq g} w_{t|g}^{\text{cohort}} \tau(g, t),$$

where  $w_{t|g}^{\text{cohort}}$  are time weights within cohort  $g$  (often uniform or proportional to the number of observations). These are useful for assessing whether early adopters experience different effects than late adopters, which can inform decisions about targeting or rollout strategy. If early adopters (who may be larger, more profitable, or in more competitive markets) experience larger effects, then prioritising early rollout to similar units makes business sense. If late adopters experience larger effects (perhaps because they learn from early adopters or because market conditions improve), then patience may pay off.

## Mapping Business Questions to Estimands

Different business questions call for different estimands. The table below maps common marketing questions to the appropriate target parameter:

**Table 4.2** Estimand Selection Guide

Business Question	Estimand	Why This Estimand?
What is the overall programme effect?	Overall ATT <sup>agg</sup>	Single summary statistic for ROI calculation
How does the effect evolve over time?	Event-time effects $\theta_k$	Traces dynamics, habit formation, carryover
Do early adopters benefit more than late adopters?	Cohort-specific $\tau_g$	Reveals heterogeneity for targeting and prioritisation
What was the impact in Q4 2023?	Calendar-time $\tau_t$	Measures contemporaneous effect for specific period
What is the cumulative vs immediate effect?	Long-run multiplier LRM	Quantifies carryover for budget allocation

In practice, report multiple estimands. The overall ATT provides a single summary for executive communication, event-time effects inform forecasting and payback period calculations, and cohort-specific effects

guide rollout prioritisation. The table is a starting point, not a definitive guide: some questions admit multiple reasonable estimands, and the mapping depends on context. For example, "What is the overall programme effect?" could be answered by the overall ATT, by the cumulative effect, or by the average event-time effect—and these can differ when effects are heterogeneous or dynamic.

Long-run and cumulative effects are central to marketing, where interventions are often intended to have persistent impacts rather than transient bumps. A loyalty programme aims to increase customer lifetime value by boosting retention and visit frequency over multiple quarters or years, not just to produce a one-time sales spike. Advertising seeks to build brand equity that endures beyond the campaign period. Platform entry aims to capture market share that persists.

Estimating long-run effects requires observing treated units for many periods post-adoption, so that event-time effects  $\theta_k$  can be traced out for large  $k$ . The cumulative effect over  $K$  periods is

$$\sum_{k=0}^K \theta_k,$$

and the long-run multiplier, which compares the cumulative effect to the immediate effect, is

$$\text{LRM} = \frac{\sum_{k=0}^K \theta_k}{\theta_0}.$$

This formula assumes that effects at different event times can be summed—that the cumulative impact is the sum of period-specific impacts. This is valid for flow outcomes such as quarterly sales or monthly conversions, where each period's effect adds to the total. For stock outcomes such as market share or brand awareness, the cumulative effect is the final level, not the sum of changes. When working with stock outcomes, report the effect at the final event time  $\theta_K$  rather than the sum  $\sum_{k=0}^K \theta_k$ .

When the immediate effect  $\theta_0$  is near zero—for example, in loyalty programmes where habit formation takes time, or advertising where brand awareness accumulates gradually—the long-run multiplier is undefined or unstable. In such cases, the long-run multiplier is not a well-defined estimand for the design, and the cumulative effect  $\sum_{k=0}^K \theta_k$  is the more meaningful target. Report the cumulative effect without forming a ratio. Attempting to rescue the ratio by conditioning on periods with "measurably non-zero" effects re-introduces specification search. Report the cumulative effect instead. Never condition on a significantly non-zero  $\theta_0$  to select the estimand.

If  $\text{LRM} > 1$ , there is positive carryover, and the cumulative impact exceeds the immediate impact. If  $\text{LRM} < 1$ , the immediate effect overstates the long-run impact, perhaps because of decay or because of competitive responses that erode the effect over time. If  $\theta_k$  changes sign across event times, the cumulative effect can be small even when early effects are large. Chapter 10 develops methods explicitly designed for estimating dynamic and cumulative effects, including distributed lag models and structural dynamic panel models, but event-study specifications (discussed in Section 4.6) provide a flexible, reduced-form approach that requires fewer assumptions.

## Linking Estimands to Marketing KPIs

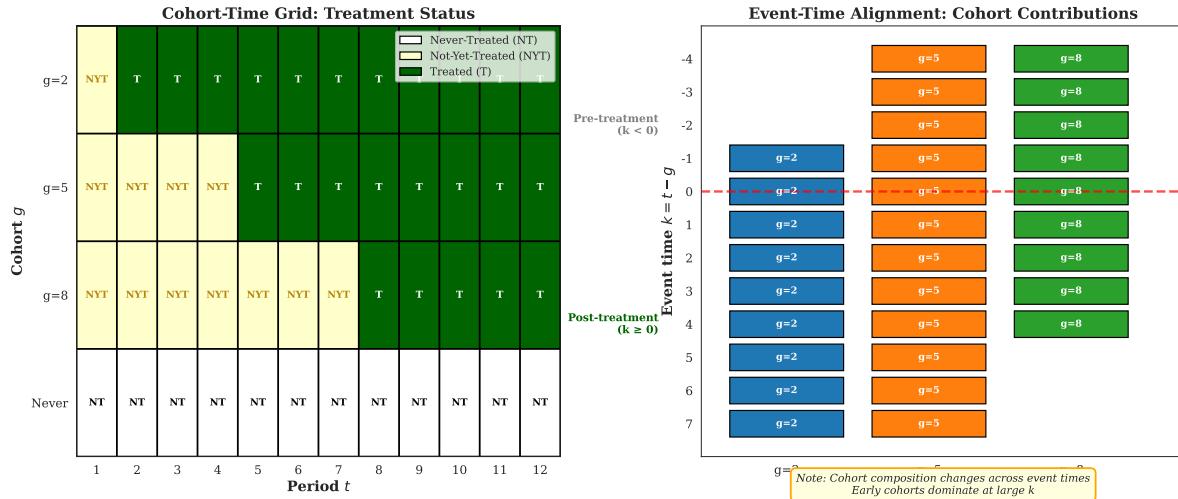
Each estimand connects to familiar marketing metrics and decision frameworks. The overall ATT translates directly to total incremental revenue or profit, the numerator in return on investment (ROI) calculations. If a programme costs \$100K and generates an average lift of \$50 per store across 500 stores over four quarters, the total incremental revenue is  $500 \times 4 \times \$50 = \$100K$ , yielding break-even ROI. Larger lifts or longer horizons push ROI positive. Smaller lifts or higher costs push it negative. The causal estimate provides the numerator. The finance team provides the denominator.

Event-time effects  $\theta_k$  inform payback period and customer lifetime value (CLV) projections. If  $\theta_0 = \$20$ ,  $\theta_1 = \$35$ ,  $\theta_2 = \$45$ , and effects plateau thereafter, the cumulative effect reaches break-even relative to a \$100 per-store cost after three quarters. CLV models incorporate the event-time trajectory to forecast long-run value, discounting future effects appropriately.

Cohort-specific effects  $\tau_g$  guide targeting and rollout prioritisation. If early-adopter cohorts (large stores, high-income areas) exhibit  $\tau_g = \$60$  while late-adopter cohorts (small stores, competitive markets) show  $\tau_g = \$10$ , the firm prioritises expansion to markets resembling early adopters and delays or forgoes rollout to low-effect segments. This heterogeneity analysis transforms a single average effect into a targeting strategy.

The long-run multiplier quantifies carryover for budget allocation across channels. If TV advertising has LRM = 2.5 (cumulative effect 2.5 times the immediate effect) while digital display has LRM = 1.2, the firm may shift budget toward TV to exploit its superior persistence, balancing immediate impact with long-run accumulation. These linkages ensure that causal estimates translate directly into actionable business insights rather than remaining abstract statistical parameters.

The choice of estimand is not neutral. Different estimands can yield different magnitudes, signs, or conclusions. This is a feature: heterogeneity is real, and methods that acknowledge it and make aggregation explicit are more credible than those that impose homogeneity. Define the estimand *ex ante*, choose an aggregation that aligns with the question, and report weights and assumptions transparently. The next section clarifies identification in staggered adoption.



**Fig. 4.2** Staggered Adoption: Cohort-Time Grid and Event-Time Alignment

The left panel displays the cohort-time grid for a staggered adoption design with three treated cohorts ( $g=2, 5, 8$ ) and never-treated units, observed over periods  $t=1$  to 12. Cells are color-coded: white for never-treated (NT), light yellow for not-yet-treated (NYT), and dark green for treated (T). The grid shows when each cohort adopts treatment and which units can serve as valid controls in each period. The right panel shows event-time alignment, illustrating which cohorts contribute observations to each event time  $k = t - g$ . Early-adopting cohorts ( $g=2$ ) contribute to a wider range of event times, while late-adopting cohorts ( $g=8$ ) contribute primarily to early event times. The red dashed line marks  $k = 0$  (treatment adoption). Cohort composition changes across event times, which can affect the interpretation of aggregated event-time effects under heterogeneity.

### 4.3 Identification with Staggered Timing

A retailer rolls out a loyalty programme to different regions over two years. What assumptions are needed to identify the causal effect? And when might those assumptions fail? This section articulates the identification assumptions for staggered adoption, discusses their plausibility in marketing settings, and clarifies when alternative strategies are required.

Identification of cohort-time effects  $\tau(g, t)$  requires assumptions about how treated and control units would have evolved in the absence of treatment and about which units can serve as valid controls for which cohorts and periods. Two types of control units are available in staggered designs: *never-treated* units ( $G_i = \infty$ ) that remain untreated throughout the observation window, and *not-yet-treated* units ( $G_i > t$ ) that will eventually be treated but have not yet adopted in period  $t$ . The distinction matters: never-treated units provide a stable control group but may be systematically different from treated units, while not-yet-treated units are more similar to treated units but may exhibit anticipation effects as their adoption date approaches. Figure 4.2 illustrates the cohort-time structure and event-time alignment in staggered designs.

#### Parallel Trends for Staggered Designs

The Parallel Trends assumption for staggered adoption asserts that units adopting at different times would have followed similar trajectories in the absence of treatment. Formally, for all cohorts  $g$  and  $g'$  and all periods  $t$  where neither cohort is yet treated ( $t < g$  and  $t < g'$ ),

$$\mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g'].$$

This is the staggered-adoption analogue of the general Parallel Trends assumption from Chapter 2 (Assumption 6): it requires that untreated period-to-period changes would have been the same across cohorts in the absence of treatment. It does not require that cohorts have the same levels or growth rates in pre-treatment periods, only that the period-to-period changes would have been the same across cohorts absent treatment. Crucially, this assumption must hold across *all* cohort pairs, not just between treated and never-treated units. If early adopters (cohort  $g = 2$ ) would have grown faster than late adopters (cohort  $g = 6$ ) absent treatment, then using cohort  $g = 6$  as a control for cohort  $g = 2$  introduces bias even if both cohorts have parallel trends with never-treated units.

A stronger version of Parallel Trends asserts that treated cohorts and never-treated units would have followed similar trajectories:

**Assumption 11 (Strong Parallel Trends)** For all  $t = 2, \dots, T$  and all cohorts  $g$ :

$$\mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = \infty].$$

This stronger version is sufficient when identification relies exclusively on never-treated units as controls. It is sufficient but not necessary when not-yet-treated units are also used. In designs that also use not-

yet-treated units as controls, modern heterogeneity-robust estimators work under the weaker condition that cohorts share common untreated trends up to the point of adoption. In marketing settings, strong parallel trends is often implausible: never-treated units may be systematically different from treated units. Stores that never receive a loyalty programme may be in declining markets, have different management, or face different competitive pressures. The assumption that they would have followed the same trajectory as treated stores requires justification. Modern heterogeneity-robust estimators can use not-yet-treated units as controls, which requires only that cohorts would have followed similar trends up to the point at which the later cohort adopts treatment—a weaker and often more credible assumption.

## Overlap and Support

Overlap and support requirements ensure that comparisons between treated and control units are valid. For identification using never-treated units as controls, we require that never-treated units exist and that they are comparable to treated units on observables. If all units eventually adopt treatment (so there are no never-treated units), identification must rely on not-yet-treated controls, which requires staggered adoption with sufficient variation in timing. If treated units are systematically different from never-treated units—for example, if the programme is rolled out first to high-performing stores and never-treated stores are persistently low-performing—parallel trends between treated and never-treated units may be implausible. Alternative approaches such as conditioning on covariates (Conditional Parallel Trends) or using factor models (Chapters 8, 9) may be required.

Assessing overlap requires examining the distribution of pre-treatment covariates across treated and control groups. Standardised mean differences (SMDs) quantify covariate imbalance: as a practical rule of thumb, values around or above 0.1–0.2 standard deviations warrant attention, with larger values (above 0.25) suggesting more serious imbalance that may threaten parallel trends. What matters is not the bare SMD, but whether imbalance on a covariate implies divergent trends in  $Y_{it}(\infty)$ . Propensity score distributions, based on  $e(X_{it})$  (or on  $e(X_{it}, \alpha_i, \lambda_t)$  when fixed effects are part of the conditioning set), reveal whether treated and control units occupy the same region of covariate space. If never-treated units cluster in a different part of the covariate distribution than treated units, comparisons between them are extrapolations rather than interpolations, and the parallel trends assumption is doing heavy lifting. Chapter 17 provides detailed guidance on balance diagnostics.

## No Anticipation

The no anticipation assumption asserts that potential outcomes in period  $t$  do not depend on treatment assignments in future periods  $s > t$ . Anticipation can arise if units learn about impending treatment and adjust behaviour in advance. For example, if customers anticipate the launch of a loyalty programme and

delay buying to qualify for rewards, then pre-treatment outcomes are contaminated by anticipation, violating No Anticipation and biasing estimates.

**Assumption 12 (No Anticipation)** For all units  $i$  and all periods  $t < G_i$ :

$$Y_{it}(g) = Y_{it}(\infty) \quad \text{for all } g > t.$$

This assumption is the staggered-adoption counterpart of the no-anticipation assumption in Chapter 2 (Assumption 1): pre-treatment outcomes for unit  $i$  are unaffected by the timing of its eventual adoption.

Anticipation is not binary: it can be partial and heterogeneous. Some units may anticipate more than others— informed insiders versus uninformed customers, large firms with market intelligence versus small firms without. Some outcomes are more susceptible to anticipation than others: purchases can be delayed, but brand awareness cannot be “saved up.” The degree of anticipation may also vary with the time horizon: units may not anticipate treatment six months in advance but may anticipate it one month in advance.

A common diagnostic for anticipation uses event-study specifications with pre-treatment leads (Section 4.6): if leads are systematically non-zero, this is suggestive evidence of anticipation or differential pre-trends. However, non-zero pre-treatment leads are observationally equivalent to differential pre-trends—both produce the same pattern in the data. Event-study leads cannot by themselves distinguish anticipation from uncontrolled pre-trends. Only institutional knowledge can. Anticipation reflects behaviour change in response to expected treatment, while differential pre-trends reflect underlying trajectory differences unrelated to treatment expectations. Distinguishing them requires institutional knowledge: if units could not have known about impending treatment, non-zero leads indicate differential pre-trends rather than anticipation. When anticipation is plausible, the analysis should estimate anticipatory effects explicitly rather than assuming them away. Modern implementations such as the `did` R package allow specifying a bounded anticipation horizon (e.g., `anticipation = 1`) to relax pure no anticipation while maintaining identification. This effectively excludes the specified number of pre-periods from the control set for each cohort.

## SUTVA and Spillovers

As introduced in Chapter 2, the stable unit treatment value assumption (SUTVA) asserts that potential outcomes for unit  $i$  do not depend on the treatment assignments of other units. SUTVA is routinely violated in marketing through spillovers, network effects, and competitive interactions. A loyalty programme offered to customers in one store may generate word-of-mouth that influences buying at nearby stores. Advertising in one market may spill over to adjacent markets through media overlap. One firm’s pricing decision may trigger competitive responses that alter outcomes for rival firms.

Spillovers complicate DiD identification in several ways. Positive spillovers from treated to control units—word-of-mouth, demonstration effects—contaminate control outcomes upward, biasing the DiD estimate toward zero and underestimating the direct effect. Negative spillovers—competitive displacement, cannibalisation—contaminate control outcomes downward, biasing the DiD estimate away from zero and

overestimating the direct effect. The direction of bias depends on the sign and magnitude of the spillover, and in practice both positive and negative spillovers may operate simultaneously. If spillovers are strong, the assumption that not-yet-treated or never-treated units provide valid counterfactuals breaks down, and estimates conflate direct effects with spillover effects.

Design-based solutions to spillovers include defining clusters that internalise spillovers (so that treatment and spillover effects occur within the same cluster and are estimated jointly) and creating buffer zones that separate treated and control units geographically or along other dimensions (so that spillovers dissipate before reaching controls). Design-based responses—such as cluster definitions and buffer zones—are discussed in detail in Chapter 3. The identification issues here should be read together with those design choices. When spillovers are central to the research question, explicit spillover models (Chapter 11) can estimate direct and spillover effects separately by specifying exposure mappings  $h_i(D_{-i,t})$  that describe how units' treatments affect other units' outcomes. These models require additional assumptions and data (for example, knowledge of the network structure or the geographic adjacency matrix). However, they avoid the biases that arise from ignoring spillovers.

## Factor Structure Relaxations

Factor structure relaxations provide an alternative when standard parallel trends is implausible but units are subject to common time-varying shocks with differential exposure. Interactive fixed effects models (Chapters 8, 9) posit that untreated potential outcomes can be decomposed as

$$Y_{it}(\infty) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

where  $R$  is the number of latent factors,  $f_{tr}$  are factors common to all units in period  $t$ , and  $\lambda_{ir}$  are unit-specific loadings that capture differential exposure to each factor. This low-rank representation can absorb unit and time fixed effects into the factors and accommodates differential trends driven by common shocks—macroeconomic conditions, industry demand shifts, platform algorithm changes—without requiring that period-to-period changes are identical across units. Identification relies on a low-rank assumption:  $R$  is small relative to the minimum of  $N$  and  $T$ , so that the factors and loadings can be estimated from the untreated observations and used to impute counterfactual outcomes for treated observations. The `gsynth` R package implements generalised synthetic control with interactive fixed effects and provides a practical entry point for these methods.

Factor models are particularly valuable when treated units are few, heterogeneous, and embedded in a panel with many control units and periods. A single treated market launching a new product, a single platform entering a city, or a single brand adopting a new advertising strategy can all be analysed using factor models if comparable control units are available and if the factor structure provides a more credible counterfactual than parallel trends.

Factor models assume that the factor structure is exogenous to treatment: the factors  $f_{tr}$  and loadings  $\lambda_{ir}$  are not affected by treatment itself. If treatment changes the factor structure—for example, if a platform entry creates a new competitive dynamic that alters how units respond to macroeconomic shocks—then the imputed counterfactuals are invalid. Factor models can also fail if the factors are correlated with treatment timing (so that units adopting early have systematically different loadings than units adopting late), or if the pre-treatment period is too short to estimate the factors reliably. The cost of factor models is the need to estimate the factors and loadings, which requires a rich pre-treatment period and may be sensitive to the choice of  $R$  or to deviations from the low-rank assumption. Factor models are not a panacea for parallel trends violations. They trade one set of assumptions for another, and do not fix violations driven by idiosyncratic shocks rather than common factors. Factor-structure relaxations do not remove the need for substantive justification. They shift the burden from parallel-trends arguments to claims about low-rank structure, stable factor loadings, and exogeneity of the factor process.

DiD assumptions should be justified using institutional knowledge, graphical and statistical diagnostics, and targeted sensitivity analyses rather than taken on faith. Pre-period evidence and diagnostic checks can be consistent with parallel trends, even though they can never validate it. Pre-treatment leads help to detect anticipation, while knowledge of likely spillover channels and attenuation patterns informs whether SUTVA is plausible. When adopting factor structures, examine pre-period fit carefully and experiment with different numbers of factors. Finally, vary control sets, windows, and estimators to check whether substantive conclusions survive reasonable perturbations. With this identification toolkit in hand, we now turn to the mechanics of two-way fixed effects and their pitfalls under heterogeneity. Sensitivity analysis tools (such as those discussed in Chapter 17) formalise how robust these conclusions are to violations of parallel trends or anticipation.

## 4.4 Two-Way Fixed Effects and Its Pitfalls

The two-way fixed effects (TWFE) regression is the traditional workhorse for difference-in-differences estimation. A retailer evaluating a loyalty programme rollout might run:

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \varepsilon_{it},$$

where  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects, and  $D_{it}$  is a treatment indicator. The coefficient  $\tau$  is interpreted as the average treatment effect. This specification is intuitive, easy to implement, and computationally fast. Under heterogeneous treatment effects it can be misleading.

### The Promise of TWFE

In the canonical  $2 \times 2$  design—two groups, two periods, treatment switching on for one group in the second period—TWFE recovers the ATT exactly, as shown in Section 4.1. The unit fixed effects absorb time-invariant differences between treated and control units, the time fixed effects absorb common shocks, and the treatment coefficient captures the causal effect under parallel trends. This logic made TWFE the default estimator for decades. Staggered adoption breaks this clean  $2 \times 2$  structure, and the same TWFE specification no longer estimates a simple average of causal effects when effects are heterogeneous.

The appeal extends to staggered adoption. With multiple cohorts adopting at different times, TWFE appears to generalise naturally: the unit fixed effects absorb cohort-specific levels, the time fixed effects absorb calendar-time shocks, and the treatment coefficient captures the average effect across all treated unit-periods. The regression pools information efficiently, and standard errors are straightforward to compute with clustering.

### The Problem: Heterogeneous Treatment Effects

The promise breaks down when treatment effects are heterogeneous—varying across cohorts, across time since adoption, or both. In staggered designs, TWFE targets a different estimand: a particular weighted average of  $2 \times 2$  comparisons that can place negative weight on some cohort–time effects. Instead, it implicitly compares:

1. newly treated units to not-yet-treated units (valid under Parallel Trends),
2. newly treated units to never-treated units (valid under Parallel Trends),
3. newly treated units to already-treated units (problematic under heterogeneity).

The third comparison is the source of the problem. When TWFE compares a newly treated unit to an already-treated unit, it uses the already-treated unit's post-treatment outcome as a counterfactual. But the already-treated unit's outcome reflects its own treatment effect, not the counterfactual of no treatment. If treatment effects differ across cohorts or evolve over time, this comparison is contaminated.

## Negative Weights and Sign Reversal

de Chaisemartin and d'Haultfoeuille [2020] and Goodman-Bacon [2021] formalised this problem by showing that TWFE can assign negative weights to some cohort-time effects. The TWFE estimator can be decomposed as:

$$\hat{\tau}^{\text{TWFE}} = \sum_g \sum_{t \geq g} w_{gt} \tau(g, t),$$

where the weights  $w_{gt}$  sum to one but are not guaranteed to be non-negative. Although the weights sum to one, they need not be non-negative, so  $\hat{\tau}^{\text{TWFE}}$  is not constrained to lie within the range of the underlying cohort-time effects. In extreme cases, TWFE can produce an estimate with the opposite sign from all underlying  $\tau(g, t)$ —a phenomenon known as sign reversal.

*Example 4.1 (Sign Reversal in a Loyalty Programme)* Consider a loyalty programme rolled out to two cohorts. Cohort A adopts in period 2 and experiences an effect of +10 in all post-treatment periods. Cohort B adopts in period 4 and experiences an effect of +5 in all post-treatment periods. Both effects are positive. However, if TWFE assigns negative weight to cohort A's later periods (because cohort A serves as a “control” for cohort B's treatment), the TWFE estimate can be attenuated toward zero or even negative, despite all true effects being positive.

The intuition is that TWFE treats already-treated units as if they were untreated when constructing comparisons for later-adopting cohorts. If early adopters have large positive effects, their elevated outcomes make later adopters look worse by comparison, biasing the estimate downward. If early adopters have negative effects, the bias goes the other way.

## When Does TWFE Fail?

TWFE is most problematic when treatment effects are heterogeneous across cohorts, so that early adopters experience different effects than late adopters and comparisons between them are contaminated. When already-treated units carry substantial weight as “controls” for later adopters, the TWFE estimate targets a different estimand because those controls embody their own treatment effects. The problem is compounded when treatment effects evolve over time: if effects grow or decay post-adoption, using already-treated units as controls conflates treatment dynamics with the counterfactual. Staggered adoption with many cohorts creates more opportunities for these forbidden comparisons, and the problem is most severe when never-treated units are scarce or absent, forcing TWFE to rely heavily on already-treated comparisons.

TWFE is less problematic when treatment effects are homogeneous—if all cohorts experience the same effect at all event times, the forbidden comparisons are not biased. Designs close to the canonical  $2 \times 2$  structure, with few cohorts and a clear pre/post distinction, approximate the simple differencing logic where TWFE works well. Abundant never-treated units also help: a large never-treated group provides valid comparisons that dominate the contribution of forbidden comparisons to the overall estimate.

## Diagnosing TWFE Problems

Before abandoning TWFE, diagnose whether the problems are severe in your application. The Bacon decomposition [Goodman-Bacon, 2021] breaks the TWFE estimator into its component  $2 \times 2$  comparisons, revealing what fraction of the estimate comes from treated vs never-treated comparisons, treated vs not-yet-treated comparisons, and treated vs already-treated comparisons. If the third category dominates, TWFE is unreliable. The `bacondecomp` command in Stata implements this decomposition directly, while the `did` R package provides similar diagnostic functionality through its aggregation routines.

Weight diagnostics from de Chaisemartin and d'Haultfœuille [2020] compute the weights  $w_{gt}$  and identify which cohort-time effects receive negative weight. If many weights are negative or if large negative weights attach to important cohort-time cells, TWFE is suspect. The `twowayfeweights` command in Stata and the `DIDmultipleg` package in R implement these diagnostics.

A practical check is to run both TWFE and a modern estimator (Callaway–Sant’Anna, Sun–Abraham, or Borusyak–Jaravel–Spiess). If the estimates are similar, TWFE may be acceptable despite its theoretical problems. If the estimates diverge substantially, the divergence reveals the magnitude of the bias from forbidden comparisons. The imputation estimator of Borusyak et al. [2024] is particularly attractive for large panels because it achieves computational efficiency comparable to TWFE while avoiding the negative weights problem.

## A Worked Example: TWFE vs Modern Estimators

Consider a stylised example with three cohorts adopting a loyalty programme in periods 2, 4, and 6, observed through period 8. True effects are heterogeneous: cohort  $g = 2$  experiences  $\tau(2, t) = 10$  for all  $t \geq 2$ . Cohort  $g = 4$  experiences  $\tau(4, t) = 5$  for all  $t \geq 4$ . Cohort  $g = 6$  experiences  $\tau(6, t) = 2$  for all  $t \geq 6$ . All effects are positive.

TWFE estimates a single coefficient  $\hat{\tau}^{\text{TWFE}}$ . Because cohort  $g = 2$  serves as a “control” for cohorts  $g = 4$  and  $g = 6$  in some comparisons, and because cohort  $g = 2$ ’s outcomes are elevated by its own treatment effect, TWFE underestimates the average effect. The Bacon decomposition reveals that a substantial fraction of the TWFE weight comes from already-treated comparisons, and the weight diagnostics show negative weights on some cohort-time cells.

A modern estimator like Callaway–Sant’Anna estimates  $\hat{\tau}(g, t)$  for each cohort-time pair using only never-treated or not-yet-treated controls. Aggregating with non-negative weights produces an overall ATT that correctly reflects the positive effects across all cohorts. The discrepancy between TWFE and CS quantifies the bias from forbidden comparisons.

## When Is TWFE Still Useful?

Despite its problems, TWFE remains useful in several contexts. As a benchmark, report TWFE alongside modern estimators to show the magnitude of the bias correction. If TWFE and modern estimators agree, readers gain confidence that heterogeneity is not severe. When effects are plausibly homogeneous across cohorts and over time, forbidden comparisons do not introduce bias, and TWFE is an efficient estimator. For computational speed with very large panels (millions of observations), TWFE is fast while some modern estimators are slow. Use TWFE for exploratory analysis, then confirm with modern estimators on subsamples or with computational optimisations. Finally, with abundant never-treated units that dominate the sample and provide most of the identifying variation, the forbidden comparisons contribute little to the TWFE estimate, and bias is small.

The key is to diagnose before deciding. Run the Bacon decomposition, check the weights, compare to modern estimators. If TWFE passes these checks, use it. If it fails, use modern estimators and report the discrepancy. The next section introduces these modern estimators in detail.

## 4.5 Modern Estimators for Staggered Designs

A retailer has rolled out a loyalty programme to 500 stores over 8 quarters, with 5 adoption cohorts. Which estimator should they use? The answer depends on the data structure, the estimand of interest, and the plausibility of identifying assumptions. This section surveys the major families of heterogeneity-robust estimators and provides practical guidance on choosing among them.

Modern difference-in-differences estimators address the pathologies of TWFE by constructing comparisons that avoid using already-treated units as controls—comparing treated units to never-treated or not-yet-treated controls—and aggregating cohort-time effects with transparent, non-negative weights.

The fundamental insight shared by all modern estimators is that identification of  $\tau(g, t)$  should not use already-treated units as controls. Instead, for each cohort  $g$  in each post-treatment period  $t \geq g$ , the estimator compares outcomes for cohort  $g$  to outcomes for a comparison group consisting of never-treated units ( $G_i = \infty$ ) or not-yet-treated units ( $G_i > t$ ). This ensures that the comparison group provides a valid counterfactual under parallel trends.

Using not-yet-treated units as controls requires an additional timing assumption: for each  $g$  and  $t < g$ , untreated potential outcomes  $Y_{it}(\infty)$  must evolve similarly for cohorts with  $G_i = g$  and  $G_i > t$  (the Staggered Parallel Trends condition in Section 4.3). If early adopters would have grown faster than late adopters even absent treatment, then not-yet-treated units do not provide a valid counterfactual. This assumption is implicit in all modern estimators that use not-yet-treated controls.

The estimators differ in how they construct these comparisons, how they aggregate them across cohorts and time, and what additional assumptions or normalisations they impose.

### Unified Setup and Notation

Throughout this section we work in the staggered-adoption setting from Section 4.2. We observe a panel with units  $i = 1, \dots, N$  and periods  $t = 1, \dots, T$ , outcomes  $Y_{it}$ , and a binary treatment indicator  $D_{it} \in \{0, 1\}$ . For each unit, let

$$G_i = \min\{t : D_{it} = 1\}$$

denote the adoption date, with  $G_i = \infty$  for never-treated units. This setup assumes treatment is absorbing: once a unit adopts, it remains treated. Non-absorbing treatments (e.g., advertising campaigns that turn on and off) require different approaches. See Chapter 10 for dynamic treatment regimes.

The panel may be balanced (all units observed in all periods) or unbalanced (units enter or exit). Callaway–Sant’Anna and Borusyak–Jaravel–Spiess accommodate unbalanced panels directly. Sun–Abraham’s regression approach may require adjustments for unbalanced data. Consult the `fixest` documentation for implementation details. Event time is  $k = t - G_i$ , and we work with adoption-time-specific potential outcomes  $Y_{it}(g)$  and  $Y_{it}(\infty)$  as in Section 4.2. The cohort-time estimand

$$\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g], \quad t \geq g$$

remains our basic building block.

For each cohort  $g$  and post-adoption period  $t \geq g$ , define the treated cohort

$$\mathcal{T}_g = \{i : G_i = g\}$$

and generic comparison sets

$$\mathcal{C}_t^{\text{NT}} = \{i : G_i = \infty\}, \quad \mathcal{C}_{g,t}^{\text{NYT}} = \{i : G_i > t\}.$$

Modern estimators construct comparisons between  $\mathcal{T}_g$  and a choice of  $\mathcal{C}_{g,t} \subseteq \mathcal{C}_t^{\text{NT}} \cup \mathcal{C}_{g,t}^{\text{NYT}}$ . Choosing between never-treated and not-yet-treated controls is a substantive decision: the former rely more heavily on strong parallel trends with potentially different units, while the latter rely on assumptions about timing being as good as random conditional on covariates. The estimators use a DiD contrast of the form

$$\hat{\tau}(g, t | \mathcal{C}_{g,t}) = \left[ \frac{1}{|\mathcal{T}_g|} \sum_{i \in \mathcal{T}_g} (Y_{it} - Y_{i,g-1}) \right] - \left[ \frac{1}{|\mathcal{C}_{g,t}|} \sum_{i \in \mathcal{C}_{g,t}} (Y_{it} - Y_{i,g-1}) \right].$$

Under the Staggered Parallel Trends assumptions in Section 4.3, and provided that  $g-1$  is strictly pre-treatment for **both** the treated cohort  $\mathcal{T}_g$  and the comparison group  $\mathcal{C}_{g,t}$  (a condition not always satisfied in arbitrary unbalanced panels), this is an unbiased estimator of  $\tau(g, t)$  for an appropriate choice of  $\mathcal{C}_{g,t}$ . In unbalanced panels, a practical workaround is to replace  $Y_{i,g-1}$  with the last observed outcome for unit  $i$  strictly before  $g$ , or with an average over available pre-treatment periods.

Aggregated estimators can then be written abstractly as

$$\hat{\tau}^{(\text{est})} = \sum_g \sum_{t \geq g} w_{g,t}^{(\text{est})} \hat{\tau}(g, t | \mathcal{C}_{g,t}), \quad \sum_g \sum_{t \geq g} w_{g,t}^{(\text{est})} = 1,$$

where the weights  $w_{g,t}^{(\text{est})}$  are estimator-specific (and not necessarily proportional to cohort size). In this unified notation, Callaway–Sant’Anna, Sun–Abraham, de Chaisemartin–d’Haultfœuille, and Borusyak–Jaravel–Spiess differ primarily in their choice of comparison sets  $\mathcal{C}_{g,t}$ , their weighting schemes  $w_{g,t}^{(\text{est})}$ , and how they incorporate covariates or factor structures.

## Callaway and Sant’Anna

Callaway and Sant’Anna [2021] (hereafter CS) estimates  $\tau(g, t)$  for each cohort-time pair using not-yet-treated or never-treated units as controls. The estimator for a specific pair  $(g, t)$  is:

$$\hat{\tau}(g, t) = \mathbb{E}_n[Y_t - Y_{g-1} | G = g] - \mathbb{E}_n[Y_t - Y_{g-1} | C],$$

where  $\mathbb{E}_n$  denotes the sample mean and  $C$  represents the chosen control group (strictly  $G = \infty$  or strictly  $G > t$ , usually not a mixture unless explicitly specified). These cohort-time estimates are then aggregated into summary measures using pre-specified weights. The aggregation can produce  $\text{ATT}^{\text{agg}}$ . When the weights are proportional to treated unit-periods,  $\text{ATT}^{\text{agg}}$  coincides with the canonical ATT. The same aggregation machinery also produces event-time effects (pooled across cohorts at each event time  $k$ ), cohort-specific effects (weighted over time for each cohort), or calendar-time effects (weighted over cohorts for each period).

CS allows for Conditional Parallel Trends by conditioning on covariates. In practice, this means assuming that expected untreated changes are comparable across cohorts after conditioning on  $X_{it}$ , not that adoption timing is independent of untreated potential outcomes. Propensity score weighting using estimated scores  $\hat{e}(X_{it})$ , inverse probability weighting, or doubly robust methods can adjust for covariate imbalances between treated and control units, making parallel trends more plausible. It also provides a rich set of aggregations and diagnostic plots, making it well suited for exploratory analysis and for settings where heterogeneity is expected and substantively interesting.

CS has limitations. Computation can be slow with many cohort-time pairs (hundreds of combinations), and standard errors are often wide because each  $\tau(g, t)$  is estimated on a subset of the data. When cohorts are small, estimates become noisy. CS also requires the researcher to specify which control group to use (never-treated, not-yet-treated, or both), and this choice can affect results when never-treated units differ systematically from eventually-treated units.

## Sun and Abraham

Sun and Abraham [2021] (hereafter SA) takes an interaction-weighted event-study approach. They estimate a dynamic specification interacting cohort indicators with event-time dummies:

$$Y_{it} = \alpha_i + \lambda_t + \sum_g \sum_{k \neq -1} \delta_{g,k} \mathbf{1}\{G_i = g\} \mathbf{1}\{t - G_i = k\} + \varepsilon_{it}.$$

The coefficients  $\delta_{g,k}$  consistently estimate the cohort-specific event-time effects  $\tau(g, g+k)$  under the Staggered Parallel Trends assumption. The key advantage is that SA avoids using already-treated units as implicit controls for later-treated cohorts. These are then aggregated using cohort-share weights:

$$\hat{\theta}_k = \sum_g \frac{N_g}{N_{\text{treated}}} \hat{\delta}_{g,k}.$$

By excluding already-treated units from the implicit comparison group and by allowing event-time effects to vary by cohort, SA avoids the negative weighting problems of TWFE under the same identifying assumptions. It produces event-study plots that are transparent and easy to interpret, making it a natural choice when the goal is to trace dynamic effects and to diagnose pre-trends.

SA has limitations. The method requires specifying a reference period (typically  $k = -1$ ), and results can be sensitive to this normalisation choice. If pre-trends are present, they will manifest as non-zero coefficients

for  $k < -1$  (or other pre-treatment  $k$ ) relative to the base period. When cohort sizes are imbalanced, the cohort-share weights can be dominated by large cohorts, obscuring effects for smaller cohorts. SA also requires sufficient variation in event-time exposure across cohorts. If all cohorts are observed for similar numbers of post-treatment periods, some event-time coefficients may be estimated imprecisely or not at all.

### **de Chaisemartin and d'Haultfœuille**

de Chaisemartin and d'Haultfœuille [2020] (hereafter dCdH) provides an alternative decomposition of the TWFE estimator, identifying which comparisons are “forbidden” (using already-treated units as controls) and which are valid. This method reports the fraction of the sample that contributes to forbidden comparisons and proposes alternative weighting schemes that exclude or downweight these comparisons. dCdH also offers extensive diagnostic tools, including diagnostics for whether the sign of the treatment effect can be inferred despite heterogeneity and tools for assessing the robustness of conclusions to different weighting choices. This makes dCdH valuable for sensitivity analysis and for understanding the sources of discrepancies between TWFE and other estimators.

dCdH is best thought of as a diagnostic layer around TWFE rather than a standalone estimator. It excels at revealing problems with TWFE. When the “corrected” estimator is used, its target is still an overall ATT-like object, but it applies weights that differ from CS or SA, so interpretation must align with the specific weighting scheme. The decomposition can be difficult to interpret when many cohorts and periods are involved, and the method assumes treatment is binary and absorbing (once treated, always treated).

### **Borusyak, Jaravel, and Spiess**

Borusyak et al. [2024] (hereafter BJS) takes an imputation-based approach. It first estimates untreated potential outcomes for all unit-period observations using only untreated observations (either pre-treatment observations for treated units or all observations for never-treated units). The estimation can use unit and time fixed effects, interactive fixed effects, or factor models. When using interactive fixed effects, identification hinges on the low-rank factor structure and on the exogeneity of factors and loadings with respect to treatment timing (Section 2.8). Treatment effects are then computed as the difference between observed outcomes and imputed untreated potential outcomes for treated observations, and aggregated using sample weights.

BJS has a key advantage over CS: it pools information across cohort-time cells rather than estimating each  $\tau(g, t)$  separately. This makes BJS more efficient when cohorts are small, producing tighter confidence intervals. BJS is particularly effective when the factor structure or the fixed effects model fits the untreated data well, and it can accommodate settings where the number of factors is large or where the panel is unbalanced.

BJS has limitations. The imputation approach relies heavily on the outcome model being correctly specified. If the factor structure is misspecified or if the number of factors  $R$  is chosen incorrectly, counterfactual

imputation will be biased, potentially exceeding the bias of simpler CS designs. BJS shifts identifying power onto stability of the outcome model for  $Y_{it}(\infty)$ . It does not repair selection on unobservables. The method requires a sufficiently long pre-treatment period to estimate factors and loadings reliably. BJS can also be computationally intensive for very large panels with interactive fixed effects, though the two-way fixed effects version scales well.

## Choosing Among Estimators

Choosing among these estimators requires matching the method to the data structure and research question. We provide decision rules rather than hedging, though these rules are conditional on the relevant identification assumptions (e.g., parallel trends, low-rank structure) being defensible in the setting at hand.

**Good Starting Point: Callaway–Sant’Anna** CS is a sensible default when you have multiple cohorts, want flexibility in aggregation, and need both ATT and event-time effects. CS handles most marketing panel settings well and provides rich diagnostics. However, CS can be computationally slow with many cohort-time pairs and may produce wide confidence intervals when cohorts are small. For very large panels or small cohorts, consider BJS as an alternative starting point.

**When to Use Sun–Abraham** Use SA when event-time effects  $\theta_k$  are the primary estimand and you want an interaction-weighted event-study plot that avoids already-treated units as controls. SA is particularly effective when cohorts are reasonably balanced in size and when the pre-treatment period is sufficient to assess pre-trends. SA integrates naturally with standard regression workflows.

**When to Use dCdH** Use dCdH for diagnostics, not as your primary estimator. Run dCdH to understand what fraction of TWFE variation comes from forbidden comparisons. As discussed in Section 4.4, a useful rule of thumb is that if more than 20% of TWFE weight comes from forbidden comparisons (already-treated vs newly treated), treat TWFE with suspicion and prefer CS or SA. This threshold is a practical heuristic rather than a formal statistical criterion. If dCdH reveals that most TWFE weight comes from valid comparisons (treated vs never-treated or not-yet-treated), TWFE may be acceptable despite its theoretical problems.

**When to Use BJS** Use BJS when simple parallel trends in raw levels is implausible but an interactive fixed effects or factor structure for untreated outcomes is credible—for example, when treated and control units are subject to common industry shocks with differential exposure. BJS still relies on that outcome model being well specified and on having enough untreated observations to estimate it. BJS requires a sufficiently long pre-treatment period (at least 5–10 periods) to estimate factors reliably. Avoid BJS with short panels.

**Decision Rules by Data Structure** If you have abundant never-treated units and a long pre-period, CS and SA will both work well. If you have no never-treated units (all units eventually adopt), you must rely on not-yet-treated comparisons, favouring CS with the not-yet-treated control option. If you have few cohorts

(2–3), SA may struggle to estimate cohort-specific effects precisely. If the pre-treatment period is short (2–3 periods), pre-trend assessment is limited and factor models (BJS) are not feasible. Rely on institutional knowledge and conditional parallel trends with covariates. If  $N$  and  $T$  are both large (thousands of units, dozens of periods), BJS with two-way fixed effects scales well. CS can become computationally slow. These rules are design-driven: they reflect which comparisons your panel can credibly support, not a preference for any specific package or estimator.

In practice, estimate multiple methods and check for agreement. If CS, SA, and BJS produce similar estimates, conclusions are robust. If estimates diverge, investigate the sources—likely differences in comparison groups, weighting, or factor structure assumptions. When divergence is large (e.g., different signs, or magnitudes differing by more than 50%), the conclusions are fragile and should be reported as such. Present results from multiple methods transparently, explain which assumptions drive the differences, and let readers assess the credibility of each approach.

## Software Implementation

Software choice should follow design and estimand choice, not the other way round. The packages below are implementations of the estimators already discussed, not independent methods. All four modern estimators have mature software implementations. In R, the `did` package implements Callaway–Sant’Anna with extensive options for control groups, covariates, and aggregations. Use `att_gt()` for cohort-time effects and `aggte()` for aggregations. The `fixest` package provides `sunab()` for Sun–Abraham event studies integrated with fast fixed effects estimation. The `did2s` package implements Gardner’s two-stage imputation, closely related to BJS. For de Chaisemartin–d’Haultfoeuille, the `DIDmultiplegt` package provides diagnostics and corrected estimators.

In Stata, `csdid` implements Callaway–Sant’Anna, `eventstudyinteract` implements Sun–Abraham, `did_multiplegt` implements dCdH, and `did_imputation` implements BJS. Python users can access these methods through `pyfixest` (Sun–Abraham style) or dedicated packages.

For practitioners, we recommend starting with the `did` package in R or `csdid` in Stata, as these provide the most complete implementation of modern methods with extensive documentation and diagnostic tools. Software implementations evolve. Check package documentation for current syntax and options. Package documentation and vignettes often contain design-specific advice and updated defaults. Consult them rather than relying solely on static code snippets. Always ground software choices in the design principles of Chapter 17 and the diagnostic checks of Chapter 16.

Marketing applications often feature rich adoption patterns, with multiple cohorts adopting at different times and with long post-treatment periods that enable tracing dynamic effects. Sample sizes vary widely: thin panels with hundreds or thousands of stores and modest  $T$  are common in retail, while fat panels with dozens of markets and many quarters or years arise in brand-level or category-level analyses. Modern estimators accommodate these structures, though computational constraints may arise with very large  $N$  and  $T$ . Pre-period length matters for assessing pre-trends and for estimating factor structures. Adoption patterns

matter for the effective sample size and for the precision of cohort-specific estimates. Interference risks shape whether SUTVA is plausible or whether spillover models (Chapter 11) are required.

Practical guidance for marketing practitioners includes starting with CS or SA, checking pre-trends using event-study plots, comparing estimates across methods to assess sensitivity, and reporting aggregated effects (overall ATT, event-time profiles, cohort-specific effects) that align with the business question. Transparency about the choice of comparison group (never-treated vs not-yet-treated), the aggregation weights, and the assumptions required for identification builds confidence in the credibility of the estimates. This enables readers to assess whether alternative choices would change conclusions.

Table 4.3 summarizes the mapping from estimands to recommended estimators and their key assumptions.

**Table 4.3** Mapping from Estimands to Recommended Estimators and Assumptions

Estimand	Recommended Estimator	Key Assumptions
ATT (all units/periods)	treated Callaway–Sant’Anna, BJS	Staggered parallel trends (unconditional or conditional), no anticipation, SUTVA
Event-time effects $\theta_k$	Sun–Abraham, Callaway–Sant’Anna	Parallel trends across cohorts, no anticipation
Cohort-specific effects $\tau_g$	Callaway–Sant’Anna cohort aggregation	Parallel trends, sufficient pre-period for each cohort
Calendar-time effects $\tau_t$	Callaway–Sant’Anna calendar-time aggregation	Parallel trends, overlap in treated cohorts per period
Sensitivity to TWFE	de Chaisemartin–d’Haultfœuille diagnostics	(Does not change identification). Assesses fraction of forbidden comparisons
Factor structure (when standard parallel trends is implausible)	BJS with interactive fixed effects	Low-rank structure, exogenous factors and loadings, sufficient pre-treatment

**Table 4.4** Modern Staggered DID Estimators: Targets, Comparison Sets, and Weights

Method	Primary estimand(s)	Comparison set $\mathcal{C}_{g,t}$	Weights $w_{g,t}^{(est)}$
Callaway–Sant’Anna (CS)	ATT, event-time effects $\theta_k$ , Never-treated and/or not-yet-treated units, possibly proportional to treated units reweighted by covariates	Non-negative, typically proportional to treated unit-periods or cohort sizes, user-chosen by aggregation	
Sun–Abraham (SA)	Event-time effects $\theta_k$	Never-treated and not-yet-treated units via cohort-pooling cohort-specific event-specific event-time regressions time profiles	Cohort-size weights when
de Chaisemartin–d’Haultfœuille (dCdH)	Decomposition and sign of overall TWFE effect	All $2 \times 2$ comparisons. Flags "forbidden" comparisons using already-treated units	Implicit TWFE weights. Can be recomputed after dropping or downweighting forbidden comparisons
Borusyak–Jaravel–Spiess (BJS)	ATT and related aggregates	Untreated observations (pre- $Y_{it}(\infty)$ )	Non-negative weights induced treatment for treated units by the imputation model and all periods for never- and sampling frequencies of treated units) used to impute treated unit-periods

These modern estimators represent a substantial advance over TWFE, providing credible estimates under heterogeneity while maintaining transparency about assumptions and aggregation choices. For a comprehensive empirical comparison of these methods, see Roth et al. [2023], who evaluate performance across a range of data-generating processes. The next section develops event-study specifications that complement these estimators by tracing dynamic treatment paths.

## 4.6 Event-Study Specifications and Dynamics

Event-study specifications extend DiD logic to estimate treatment effects as a function of event time  $k$ , the number of periods since treatment adoption. The specification includes indicators for each event time, both pre-treatment (leads,  $k < 0$ ) and post-treatment (lags,  $k \geq 0$ ), and estimates a separate coefficient for each event time. The resulting sequence of coefficients traces the dynamic evolution of the treatment effect. It provides evidence on anticipation (non-zero leads), immediate effects (the coefficient at  $k = 0$ ), and long-run effects (coefficients at large  $k$ ).

### Event-Study Specification

The event-study regression takes the form

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \beta_k \mathbf{1}\{t - G_i = k\} + \varepsilon_{it},$$

where  $\mathbf{1}\{t - G_i = k\}$  is an indicator that equals one if unit  $i$  is at event time  $k$  in period  $t$ , and the reference period is  $k = -1$  (the period immediately before treatment adoption). The coefficients  $\{\beta_k\}$  are regression coefficients that measure differences at each event time relative to the reference period, with  $\beta_{-1}$  normalised to zero by omission. Under homogeneous treatment effects and valid identification,  $\beta_k$  coincides with the event-time estimand defined in Section 4.2. Under heterogeneity,  $\beta_k$  is a TWFE regression coefficient that generally does not equal the true effect, because already-treated units implicitly enter the comparison group.

This normalisation assumes that any anticipation effects have not yet materialised by  $k = -1$ . If anticipation is already present at  $k = -1$ —for example, if customers have already begun changing behaviour in response to an announced programme—then the baseline period is contaminated. The post-treatment coefficients would then underestimate the true effect relative to a world with no anticipation. When anticipation is suspected, consider normalising to an earlier period (for example,  $k = -2$  or  $k = -3$ ) or to the average of distant pre-treatment periods.

**Caution: TWFE Event-Study Contamination** The specification above is a TWFE-style event study and inherits the negative weighting problem under heterogeneous treatment effects discussed in Section 4.4. In the notation of Section 4.4, the TWFE event-study coefficients can be written as weighted sums of cohort-time effects  $\tau(g, t)$  with weights that may be negative or attach already-treated observations as controls, so even the shape of the estimated dynamic profile can be distorted under heterogeneity. Use the TWFE-style event study only as a rough diagnostic. For effect magnitudes and formal inference, rely on Sun–Abraham or Callaway–Sant’Anna event-time aggregations, which produce event-study plots that reflect true dynamics rather than artefacts of heterogeneity.

## Pre-Trends and Anticipation

Pre-treatment leads ( $k < -1$ ) are used to diagnose pre-trends and anticipation. If Parallel Trends holds and there is No Anticipation, the pre-treatment coefficients should be close to zero. Non-zero pre-treatment coefficients signal that treated units were on a different trajectory than controls even before treatment, violating Parallel Trends.

A word of caution: pre-trend diagnostics have low power against many plausible violations [Roth, 2022], so small estimated pre-trends do not validate Parallel Trends. See Chapter 17 for a detailed discussion. Substantive arguments about the assignment mechanism remain essential.

A pattern of pre-treatment coefficients that trend toward the post-treatment effect suggests anticipation: units respond to expected future treatment by altering behaviour in advance. Anticipation can be negative or positive. Negative anticipation occurs when customers delay purchases to qualify for future rewards (producing negative pre-treatment coefficients that grow in magnitude as the launch approaches). Positive anticipation occurs when customers accelerate purchases before an expected price increase, or when firms ramp up advertising before a product launch (producing positive pre-treatment coefficients). The sign and pattern of pre-treatment coefficients, combined with institutional knowledge about what units could have anticipated, helps distinguish anticipation from differential pre-trends, complementing the No Anticipation discussion in Section 4.3. Critically, non-zero pre-treatment coefficients alone cannot distinguish anticipation from uncontrolled differential pre-trends: both generate the same pattern in the data. Only institutional knowledge about when units could plausibly have learned about treatment timing can separate the two stories.

## Dynamic Effects

Post-treatment lags ( $k \geq 0$ ) trace the dynamic response. If the effect is immediate and constant, all post-treatment coefficients should be roughly equal. If the effect grows over time—as might occur with habit formation, learning, or network effects—then post-treatment coefficients increase with  $k$ . If the effect decays—as might occur with a temporary promotion or a one-time advertising campaign—then post-treatment coefficients decrease with  $k$ . The sequence  $\{\beta_k\}_{k=0}^K$  provides a reduced-form summary of the dynamic response without imposing parametric restrictions on the lag structure.

**Worked Example: Interpreting Event-Study Coefficients** Consider a loyalty programme rollout with estimated coefficients and standard errors:  $\hat{\beta}_{-3} = 0.8$  (SE = 1.2),  $\hat{\beta}_{-2} = 0.4$  (SE = 1.0),  $\hat{\beta}_{-1} = 0$  (reference),  $\hat{\beta}_0 = 5.2$  (SE = 1.1),  $\hat{\beta}_1 = 7.1$  (SE = 1.3),  $\hat{\beta}_2 = 8.3$  (SE = 1.4),  $\hat{\beta}_3 = 8.1$  (SE = 1.5), all in £000s of quarterly sales. The pre-treatment coefficients are consistent with zero: the 95% confidence intervals for  $\hat{\beta}_{-3}$  and  $\hat{\beta}_{-2}$  include zero (approximately  $[-1.6, 3.2]$  and  $[-1.6, 2.4]$  respectively). This is consistent with parallel trends. The discrete jump at  $k = 0$  suggests an immediate effect of about £5,200 in incremental quarterly sales. The rising pattern from  $k = 0$  to  $k = 2$  ( $5.2 \rightarrow 7.1 \rightarrow 8.3$ ) suggests growing effects, consistent with habit

formation as customers accumulate points and increase visit frequency. The levelling at  $k = 3$  ( $8.1 \approx 8.3$ ) suggests the effect has reached a steady state.

## Binning and Reference Windows

Binning choices determine how event times are grouped when the number of event times is large relative to the sample size. If the panel spans many periods and if some cohorts adopt treatment early (creating many post-treatment observations), estimating a separate coefficient for each  $k$  can be impractical due to sparse data at large  $k$ . Binning groups adjacent event times into intervals—for example,  $k \in \{0\}$ ,  $k \in \{1, 2\}$ ,  $k \in \{3, 4, 5\}$ ,  $k \geq 6$ —and estimates a single coefficient for each bin. The trade-off is between resolution (finer bins provide more detail about the dynamic profile) and precision (coarser bins pool more observations and reduce standard errors). A useful rule of thumb: aim for at least 50–100 treated unit-period observations per bin, or at least 10 treated units per bin, whichever is more restrictive. When  $N$  is small, prioritise at least 10 treated units per bin over fine time resolution. Report the binning scheme transparently and check robustness to alternative binning choices.

Reference window choices determine which event time is normalised to zero. The convention is to normalise  $k = -1$ , but other choices are possible. If no anticipation is expected and if the goal is to estimate the effect relative to the average of all pre-treatment periods, the reference can be the average of  $k < 0$ . If the goal is to estimate the change from period  $k = -2$  to  $k = 0$ , the reference can be  $k = -2$ . The choice affects the interpretation of coefficients but not the differences between coefficients (which estimate treatment effect contrasts and are invariant to the reference choice). Transparency about the reference window and robustness checks using alternative references ensure that conclusions are not artefacts of the normalisation.

**Endpoint Binning Caveat** When binning distant event times into a single coefficient (for example,  $k \geq 6$ ), the estimate averages effects across different treatment durations. If effects are still evolving at large  $k$ —as occurs with habit formation, learning, or competitive adjustment—the binned coefficient obscures the dynamic trajectory. A flat binned coefficient at  $k \geq 6$  might mask continued growth from  $k = 6$  to  $k = 10$ . Report the binning cutoff explicitly, check sensitivity to alternative cutoffs (for example,  $k \geq 4$  vs  $k \geq 8$ ), and interpret binned coefficients as lower bounds on long-run effects when dynamics are likely to persist.

## Interpretation and Diagnostics

Interpretation of event-study plots requires care. A plot of  $\hat{\beta}_k$  against  $k$  visualises the estimated dynamic profile, and confidence intervals assess whether coefficients are distinguishable from zero. A pattern with flat, near-zero lead coefficients for  $k < 0$  followed by a discrete jump at  $k = 0$  is consistent with parallel trends and a causal interpretation. In practice, this idealised pattern is rare: real event-study plots often show noisy

pre-trends, gradual ramp-ups, or delayed effects. Judgement is required to assess whether deviations from the ideal are substantively meaningful or merely noise.

Pre-trend patterns that show drifts or trends for  $k < 0$  raise concerns about Parallel Trends. When pre-trends fail, several alternatives are available: condition on covariates to restore Conditional Parallel Trends, use factor models that accommodate differential trends (Chapter 8), or conduct sensitivity analysis to bound the bias from pre-trend violations. The framework of Rambachan and Roth [2023] provides formal sensitivity bounds that allow researchers to assess how robust conclusions are to violations of Parallel Trends of various magnitudes. Simply proceeding with the analysis while ignoring pre-trend violations is not acceptable.

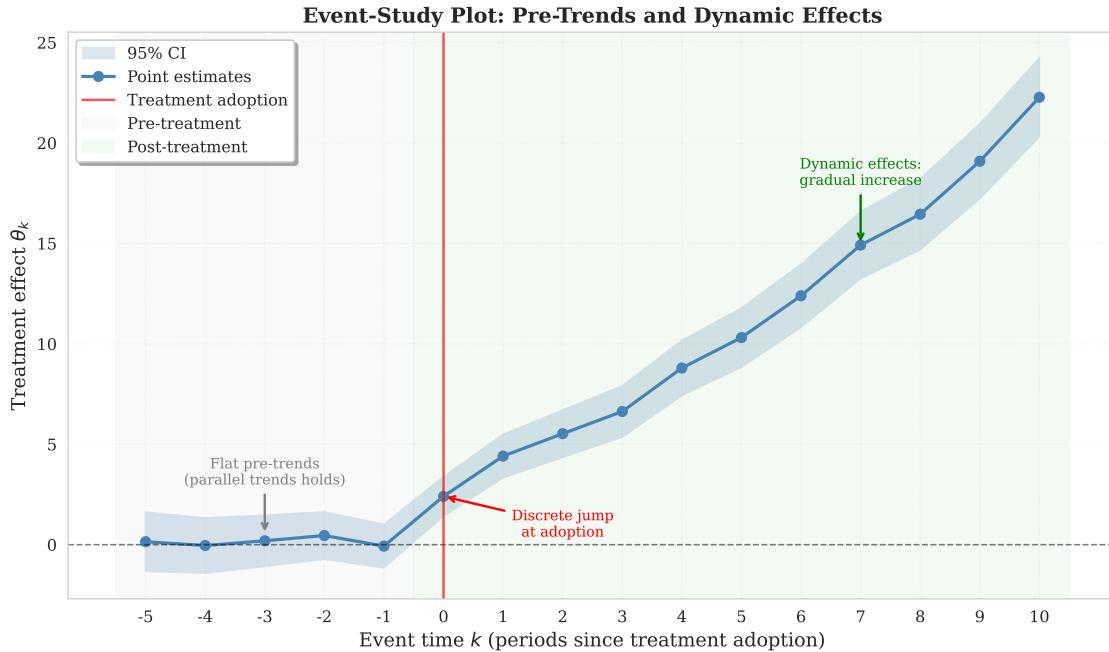
Beyond visual inspection, a joint diagnostic can complement the plot. A common choice is a Wald test of  $H_0 : \beta_{-K} = \beta_{-K+1} = \dots = \beta_{-2} = 0$  using cluster-robust standard errors (a standard F-test assumes homoskedasticity and is inappropriate when errors are clustered). Rejection is suggestive evidence of pre-trends or anticipation. Failure to reject is only weakly informative given low power, and results can be sensitive to how many leads you include and to binning choices. Report both the visual evidence and the joint test. When many event-time coefficients are examined jointly, Chapter 16 recommends controlling for multiplicity (for example, via Romano–Wolf or FDR procedures) rather than relying solely on unadjusted pointwise diagnostics.

Figure 4.3 illustrates a typical event-study plot showing pre-treatment coefficients, the treatment effect at adoption, and the dynamic evolution of effects over time.

Event-study specifications are closely related to distributed lag models (Chapter 10), which parameterise the lag structure and estimate the effect of current and lagged treatments on current outcomes. Distributed lag models impose functional form assumptions on the decay or persistence of effects, enabling extrapolation beyond the observed event times and estimation of long-run multipliers, but they require stronger assumptions than the reduced-form event study. The trade-off is between flexibility (event studies impose minimal restrictions) and efficiency and interpretability (distributed lags provide parsimonious summaries and long-run estimates).

Marketing applications of event studies abound. A loyalty programme rollout with staggered adoption across stores over multiple quarters produces an event study with many cohorts and many event times, enabling detailed tracing of how programme effects evolve as customers accumulate points and develop habits. An advertising campaign with staggered launches across markets produces an event study that reveals whether effects peak immediately (as might occur with direct response advertising) or build over time (as might occur with brand-building campaigns). A pricing policy change implemented in waves across product categories produces an event study that traces competitive responses and demand adjustments over time. In each case, the event-study plot provides transparent evidence on dynamics, anticipation, and long-run effects, complementing the summary estimates produced by modern DiD estimators.

For plotting standards and diagnostics, see Chapter 5, Sections 5.6 and 5.8. Inference for event-time paths appears in Chapter 5, Section 5.7. We now turn to inference for DiD estimators.



**Fig. 4.3** Event-Study Plot: Pre-Trends and Dynamic Effects

A well-behaved event-study plot exhibits three features: (i) flat, near-zero coefficients for  $k < 0$ , consistent with the Parallel Trends assumption. (ii) a discrete jump at  $k = 0$ , suggesting an immediate treatment effect. (iii) a clear post-treatment trajectory revealing dynamics—whether constant (flat), growing (habit formation, learning), or decaying (temporary promotion). Drifting pre-trends that slope upward or downward raise concern about Parallel Trends. Non-zero coefficients at distant leads ( $k = -2, -3, \dots$ ) may indicate anticipation or differential pre-trends. The dashed horizontal line at zero and the vertical line at  $k = 0$  provide visual reference.

## 4.7 Inference

Valid inference in DiD settings requires accounting for the correlation structure of the errors, the finite-sample properties of the estimators, and the multiplicity of hypotheses tested. This section focuses on DiD-specific implications. Chapter 16 provides general inference guidance for all designs.

### Clustering

Clustering is the standard approach to accounting for within-unit correlation over time. Outcomes for the same unit in different periods are correlated because of persistent unobservables, autocorrelated shocks, or dynamic feedback. Ignoring this correlation leads to standard errors that are too small and hypothesis tests that reject too often. Clustering standard errors by unit allows for arbitrary correlation within units while maintaining independence across units. When clustering by unit, each unit is a cluster, so the cluster count equals the number of units ( $G = N$ ). The cluster-robust variance estimator computes standard errors that are valid asymptotically as the number of clusters grows, provided that errors are uncorrelated across clusters. For a comprehensive treatment of when and how to cluster, see Abadie et al. [2023].

The independence assumption across clusters is often violated in marketing settings. Stores in the same region may be correlated through regional shocks. Brands in the same category may be correlated through category-level demand. All units may be affected by macroeconomic conditions. Two-way clustering (discussed below) addresses some of these concerns, but if cross-cluster correlation is severe and persistent, even two-way clustering may be insufficient. In such cases, consider clustering at a higher level (for example, region rather than store) or using methods that explicitly model the correlation structure.

Two-way clustering, by both unit and time, accounts for correlation within units over time and across units within periods. If all stores in a region are affected by a regional shock, errors are correlated across stores within a period. If macroeconomic conditions or industry-wide trends affect all units in a given period, errors are correlated across units. Two-way clustering captures both sources of correlation, producing standard errors that are valid under weak assumptions. The cost is larger standard errors (because more correlation is acknowledged) and greater computational complexity, though modern software implements two-way clustering efficiently. Note that two-way clustering implicitly treats cross-sectional correlation as common shocks at each time, which may not fully capture network-style dependence. For heavy interference, Chapter 11's explicit models are better.

**Common Mistake: Wrong Clustering Level** A frequent error is clustering at the wrong level. The clustering level must match the level at which treatment varies. This condition mirrors the no-interference component of SUTVA: clustering at the treatment-assignment level assumes that, after conditioning on covariates and fixed effects, residual errors are independent across clusters. When spillovers or exposure mappings  $h_i(D_{-i,t})$  operate across clusters, the effective cluster should be altered to reflect the interference structure. If treatment is assigned at the store level (each store is treated or not), cluster by store. If treatment

is assigned at the DMA level (all stores in a DMA share treatment status), cluster by DMA, not by store. Clustering at a finer level than treatment assignment (for example, store when treatment varies at DMA) produces standard errors that are too small because it ignores the correlation induced by shared treatment. Clustering at a coarser level than necessary is conservative but sacrifices power. When uncertain, cluster at the treatment assignment level or one level coarser.

**Worked Example: Impact of Clustering on Inference** Consider a loyalty programme evaluation with 500 stores over 12 quarters. Without clustering, the estimated effect is  $\hat{\tau} = 8.2$  with  $SE = 2.1$ , yielding a 95% CI of [4.1, 12.3] and  $p < 0.001$ . With unit (store) clustering, SE increases to 4.8, the CI widens to [-1.2, 17.6], and  $p = 0.09$ —no longer significant at the 5% level. The unclustered analysis dramatically overstates precision by ignoring serial correlation within stores. This example illustrates why clustering is not optional: ignoring it leads to false confidence in effects that may not be statistically distinguishable from zero.

## Small-Sample Corrections

Small- $G$  corrections address the problem that cluster-robust standard errors rely on large-cluster asymptotics, which may not provide accurate inference when the number of clusters is small [Cameron et al., 2008]. As a rough guideline—not a hard rule—with fewer than about 20 clusters, asymptotic cluster-robust SEs are often unreliable. Wild cluster bootstrap or randomisation inference is preferred. With 20–50 clusters, asymptotic SEs may be adequate but should be checked against bootstrap results, and small-sample corrections (for example, HC2 or HC3 variants) should be applied. With 50 or more clusters, asymptotic cluster-robust SEs are generally reliable. These cluster-count thresholds align with the power and MDE considerations in Chapter 3.

Here  $G$  denotes the number of clusters used for inference (for example,  $G = N$  when clustering by unit). Do not confuse this  $G$  with the unit-level adoption time  $G_i$  used elsewhere in the chapter.

These thresholds are guidelines, not hard rules. The reliability of asymptotic SEs depends not just on  $G$  but also on the balance of cluster sizes (unequal clusters degrade performance), the degree of within-cluster correlation (higher correlation requires more clusters), and the leverage of treated clusters (a few high-leverage clusters can distort inference). When in doubt, compare asymptotic SEs to bootstrap results.

In marketing panels, the number of treated cohorts or the number of DMAs in a geo-experiment may be modest, making asymptotic approximations unreliable. When only a handful of clusters are treated (for example, a few DMAs in a geo-experiment), the effective number of treated clusters, rather than total  $G$ , drives power and finite-sample behaviour. The wild cluster bootstrap provides finite-sample inference by resampling entire clusters, imposing random signs on cluster-level residuals, and computing the bootstrap distribution of the test statistic. It respects the clustering structure and accommodates heteroskedasticity and serial correlation. This provides more accurate p-values than asymptotic methods when clusters are few.

A technical note: the wild cluster bootstrap imposes the null hypothesis when constructing the bootstrap distribution, which is appropriate for hypothesis testing. Bootstrap confidence intervals constructed by inverting the test may differ slightly from intervals constructed by percentile methods. For most applications,

the difference is minor, but when precision matters, report both the bootstrap p-value and the asymptotic confidence interval, noting any discrepancies.

## Randomisation Inference

Randomisation inference and permutation tests offer design-based alternatives that do not rely on parametric assumptions about error distributions. Under the sharp null hypothesis of no effect for any unit in any period, the observed treatment assignment is just one of many possible assignments that could have been drawn from the randomisation protocol. By recomputing the test statistic for all possible assignments (or a large random sample of them), we generate the exact null distribution of the test statistic. We then compare the observed statistic to this distribution to compute an exact p-value.

This p-value is exact under the sharp null and the assumed randomisation protocol. It is not a general guarantee of correct size under misspecified assignment mechanisms or under non-sharp nulls.

Randomisation inference is particularly compelling in experimental settings where the randomisation protocol is known and where the goal is to conduct inference that respects the design. In observational staggered adoption, the “randomisation protocol” is unknown. Randomisation inference can still be applied by assuming a particular assignment mechanism—for example, that treatment timing is random conditional on covariates—but this is an assumption about the data-generating process, not a design feature. Absent known randomisation, permutation p-values are only exact under the assumed mechanism. In observational staggered adoption, any randomisation-inference procedure is only as credible as the assumed assignment mechanism. It should be justified using the institutional context, not treated as design-free.

## Multiple Testing

Multiple testing arises when estimating many coefficients across cohorts, periods, or subgroups. In DiD settings, multiple testing often arises when reporting full event-time profiles  $\{\theta_k\}$ , many cohort-specific effects  $\{\tau(g, t)\}$ , or subgroup-specific ATTs. The probability of at least one false rejection (type I error) exceeds the nominal level unless we adjust for multiplicity. Bonferroni controls family-wise error rate by dividing the significance level by the number of tests but can be conservative when tests are correlated. False discovery rate (FDR) control offers a less conservative alternative, controlling the expected proportion of false rejections. Romano–Wolf stepdown procedures exploit correlation to improve power while controlling family-wise error.

When to use each: use Bonferroni or Romano–Wolf when any false positive is costly—for example, in regulatory submissions or when decisions are irreversible. Use FDR control when some false positives are acceptable and power is a concern—for example, in exploratory subgroup analysis or when the goal is to generate hypotheses for further testing. Event-time specific multiplicity is treated in Chapter 5, Section 5.7.

## Practical Guidance

Practical guidance for marketing applications: cluster by unit as a default. Use two-way clustering when cross-unit correlation is plausible (for example, in geo-experiments or when regional shocks are present). Apply wild cluster bootstrap when the number of clusters is small. Use multiplicity adjustments when testing many coefficients or subgroup effects. When uncertain about the clustering level, report results under alternative clustering choices as a robustness check. If conclusions change substantially across clustering specifications, acknowledge the sensitivity.

Pre-specifying the primary estimand and distinguishing primary from exploratory analyses reduces the multiplicity burden. If the primary estimand is the overall ATT, control type I error for that test and treat ancillary analyses as exploratory. These priorities should be pre-specified in a pre-analysis plan (Section 3.9). Reclassifying exploratory findings as primary after seeing the data defeats the purpose of multiplicity control. As with estimands and estimators, inference choices (clustering level, bootstrap vs asymptotics, multiplicity adjustments) should be specified *ex ante* and documented in the pre-analysis plan (Section 3.9).

Modern estimators like Callaway–Sant’Anna and Sun–Abraham provide standard errors for aggregated estimands (overall ATT, event-time effects) that account for the estimation of cohort-time effects. These standard errors rely on different asymptotic arguments than those for individual coefficients. Consult the estimator documentation for details on the variance estimation approach.

For event-time inference guidance, see Chapter 5, Section 5.7. Transparency about inferential choices and robustness checks using alternative methods build confidence in conclusions.

**Table 4.5** Inference Method Decision Guide

Situation	Recommended Method	Notes
Standard panel, $G \geq 50$ clusters	Cluster-robust SEs by unit	Default choice
Cross-unit correlation (geo-experiments, regional shocks)	Two-way clustering (unit × time)	Accounts for both serial and cross-sectional correlation
Few clusters ( $G < 20$ )	Wild cluster bootstrap	Provides finite-sample valid inference
Experimental design with known randomisation	Randomisation inference	Exact p-values under the sharp null and the stated randomisation protocol
Many event-time coefficients or sub-groups	FDR control or Romano–Wolf	Adjust for multiplicity
Primary estimand + exploratory analyses	Adjust primary only	Label exploratory as such

**Software** Software choice should follow the design and estimand decisions. The commands listed below implement the inference strategies discussed above, but do not change the underlying assumptions.

- Wild cluster bootstrap: `boottest` (Stata), `fwildclusterboot` (R)
- Two-way clustering: `reghdfe` (Stata), `fixest` (R)
- Romano–Wolf: `rwolf` (Stata), `wildrwolf` (R)
- FDR control: `p.adjust(method="BH")` (R)

These inference procedures ensure that uncertainty is quantified appropriately and that hypothesis tests have correct size. The next section develops diagnostic workflows for assessing the plausibility of identifying assumptions and the robustness of conclusions.

## 4.8 Diagnostics and Design Considerations

Credible DiD analysis requires rigorous diagnostics that assess the plausibility of identifying assumptions, the sensitivity of conclusions to modelling choices, and the influence of individual observations or cohorts. This section outlines the core diagnostic workflow, specialising the general diagnostic framework from Chapter 17 to DiD applications with staggered adoption.

### Pre-Trend Assessments

Pre-trend assessments diagnose whether treated and control units followed similar trajectories before treatment. Plot outcomes for treated and control groups over pre-treatment periods and visually inspect for divergence. Estimate an event-study specification with multiple pre-treatment leads and examine whether the lead coefficients are close to zero. A joint Wald statistic for the null that all leads are zero can provide a complementary summary, but it has low power and depends on the chosen lead window. Non-zero lead coefficients raise concern about Parallel Trends. Small or noisy lead coefficients are at best consistent with Parallel Trends in the pre-period. They do not validate identification in the post-period.

**Worked Example: Joint Pre-Trend Diagnostic** Consider a loyalty programme rollout with five pre-treatment periods. The event-study specification yields  $\hat{\beta}_{-5} = 0.8$ ,  $\hat{\beta}_{-4} = -0.3$ ,  $\hat{\beta}_{-3} = 0.5$ ,  $\hat{\beta}_{-2} = 0.2$ ,  $\hat{\beta}_{-1} = 0$  (reference). A joint Wald statistic for  $H_0 : \beta_{-5} = \beta_{-4} = \beta_{-3} = \beta_{-2} = 0$  with cluster-robust standard errors yields  $\chi^2(4) = 4.92$ ,  $p = 0.29$ . This result is consistent with small pre-trends in the chosen lead window. Had  $p < 0.05$ , we would treat it as suggestive evidence of pre-trends or anticipation and investigate further. Note: use a Wald test with clustered standard errors rather than a standard  $F$ -test, which assumes homoskedasticity.

**What If Pre-Trends Are Present?** When pre-trend diagnostics suggest non-zero leads, proceed systematically. First, check whether imbalance in observables explains the pattern by conditioning on covariates and re-estimating. Second, if conditioning fails, consider factor models (Chapter 8) that accommodate differential trends via latent factors. Third, if only one treated unit or a single aggregate is available, synthetic control (Chapter 6) may construct a better counterfactual. For borderline cases ( $0.05 < p < 0.10$ ), sensitivity analysis using Rambachan–Roth bounds can quantify how much plausible pre-trend violations could shift post-treatment estimates. Rambachan–Roth bounds [Rambachan and Roth, 2023] ask: if the pre-trend pattern in  $\{\hat{\beta}_k\}_{k<0}$  continued into the post-period at the same rate observed in the pre-period, how much would the estimated overall ATT or the estimated event-time path change? This provides a range of estimates under different assumptions about the persistence of pre-trends.

## Placebo Tests

Placebo-in-time diagnostics apply the DiD logic to pre-treatment periods only, treating an earlier period as if it were the post-treatment period. For example, if treatment begins in period five and data are available from period one, a placebo analysis might treat period three as the "treatment" period and estimate the "effect" using periods one and two as pre-treatment and period three as post-treatment. If the placebo estimate is near zero, this is consistent with Parallel Trends in the pre-period. If the placebo estimate is large and statistically significant, the design may exhibit differential pre-trends. Placebo analyses cannot validate a DiD design. They only reveal whether the particular comparison between treated and control units tends to generate spurious "effects" even when no true treatment is present.

Placebo-in-units diagnostics apply the DiD logic using never-treated or not-yet-treated units as placebo-treated units. Randomly assign a subset of never-treated units to a fictitious treatment group and estimate the DiD effect comparing these placebo-treated units to the remaining never-treated units. Repeat the random assignment many times (for example, 500 or 1000 iterations) to generate a distribution of placebo estimates. The actual treatment effect estimate should not be extreme relative to this distribution. If it falls within the middle 95% of placebo estimates, the evidence for a true effect is weak. If placebo estimates are consistently non-zero even under random assignment, this suggests that the comparison between treated and never-treated units is confounded by differential trends unrelated to actual treatment.

## Covariate Balance and Overlap

Overlap and support checks assess whether treated and control units are comparable on observables. Plot covariate distributions for treated and control groups and check for imbalances. Compute standardised mean differences (SMDs) for key covariates. As a practical rule of thumb, values around or above 0.1–0.2 standard deviations warrant attention, with 0.2 often used as a simple benchmark. However, the appropriate threshold depends on context—specifically, on how strongly the covariate predicts outcomes. A 0.3 SMD on a covariate weakly related to outcomes may matter less than a 0.15 SMD on a covariate that strongly predicts outcomes. Improved balance makes Conditional Parallel Trends (Assumption 10) or unconfoundedness (Assumption 7) more plausible, but it does not guarantee them. Unobserved confounders may still remain.

**What If Covariate Balance Is Poor?** When SMDs exceed acceptable thresholds, proceed systematically. First, re-weight or match on covariates to improve balance, then re-estimate. Second, include covariates as controls in the outcome regression (Conditional Parallel Trends). Third, if balance cannot be achieved, acknowledge that treated and control units differ on observables and that Parallel Trends is less plausible. Report both unadjusted and adjusted estimates to show how covariate adjustment affects conclusions.

Covariate balance can also be assessed by regressing the treatment indicator on covariates and checking the  $R^2$ . A high  $R^2$  indicates that treatment is strongly predicted by covariates, suggesting that unobserved confounders correlated with covariates may also predict treatment. Conversely, a low  $R^2$  suggests that treat-

ment assignment is weakly related to observables. However, a low  $R^2$  does not mean treatment is "as good as random". Unobservables may still predict treatment even if observables do not. Treat the  $R^2$  heuristic as a warning sign when it is high, not as reassurance when it is low.

## Influence and Robustness

Influence and weight audits examine whether individual cohorts, periods, or observations exert undue influence on the estimates. Modern DiD estimators aggregate cohort-time effects with known weights, and these weights can be inspected to identify which comparisons contribute most to the overall estimate. If a single cohort or a single period dominates the aggregation, conclusions may be sensitive to the inclusion or exclusion of that cohort or period.

Leave-one-cohort-out analyses re-estimate the effect excluding each cohort in turn and check whether estimates are stable. A useful heuristic is that if excluding any single cohort changes the point estimate by more than roughly 25% or flips the sign, the estimate is sensitive to that cohort. If excluding a single cohort changes statistical significance (e.g., from  $p < 0.05$  to  $p > 0.10$ ), the conclusion depends on that cohort. In either case, investigate why that cohort is influential—it may be an outlier, may have different characteristics, or may have experienced the treatment differently.

Leave-one-period-out analyses check whether estimates are sensitive to the inclusion of specific periods. If a single period drives the result—for example, if excluding the first post-treatment period eliminates the estimated effect—this suggests that the effect is transient or that the period is an outlier. Robustness checks that vary the treatment window, exclude outliers, or trim the sample based on pre-treatment characteristics provide evidence on the stability of conclusions.

## Specification Curves

Specification curves aggregate estimates across many defensible modelling choices: different sets of control units (never-treated only vs never-treated and not-yet-treated), different covariate adjustments (no covariates vs rich controls), different fixed effects structures (unit and time vs unit, time, and unit-specific trends), different event-time windows (short vs long), different binning choices, and different estimators (CS vs SA vs BJS). Plot the distribution of estimates across specifications. If estimates cluster tightly, conclusions are robust to modelling choices. If estimates vary widely, the choice of specification matters.

Not all specifications are equally credible. Weight the interpretation toward specifications that are most defensible on substantive grounds—for example, specifications that use the control group most similar to treated units, that include covariates known to predict outcomes, or that use estimators appropriate for the data structure. Report the full distribution but highlight the preferred specification and explain why it is preferred. The `specr` package in R implements specification curve analysis and facilitates systematic exploration of the specification space.

Practical guidance for marketing applications: conduct pre-trend diagnostics as a matter of course. Report placebo analyses to probe whether the design generates spurious effects. Check covariate balance and adjust when necessary. Inspect weights and conduct leave-one-out analyses to assess influence. Construct specification curves to demonstrate robustness.

Transparent reporting of diagnostics builds confidence that conclusions are not artefacts of arbitrary choices and that the identifying assumptions are plausible.

#### Box 4.1: Diagnostic Workflow for DiD Studies

This box adapts the general diagnostic workflow of Chapter 17 to DiD panels, focusing on pre-trends, placebo analyses, balance, and influence in staggered adoption.

1. **Pre-treatment trends.** Plot outcomes for treated and control groups over all pre-treatment periods and check for visual divergence. Estimate an event-study specification with multiple pre-treatment leads and examine whether lead coefficients are close to zero. A joint Wald diagnostic can complement the plot. If pre-trends are evident, condition on covariates and re-estimate, or consider factor models or synthetic control (Assumption 10 and factor-structure conditions in Section 2.8).
2. **Placebo analyses.** Placebo-in-time assigns fictitious treatment in the pre-period. Placebo-in-units randomly assigns never-treated units to a fictitious treatment group and repeats the analysis many times. Placebo estimates should typically be near zero. Large systematic placebo “effects” raise concern about spurious comparisons.
3. **Balance and overlap.** Compute standardised mean differences for key covariates and inspect overlap in covariate distributions. Differences around or above 0.1–0.2 standard deviations warrant attention, but the threshold depends on how strongly covariates predict outcomes. If balance is poor, re-weight, match, or include covariates as controls.
4. **Anticipation and SUTVA.** Inspect whether lead coefficients trend toward the post-treatment period and use institutional context to judge anticipation versus differential pre-trends. If anticipation is plausible, model it explicitly or normalise to an earlier reference period. Address potential SUTVA violations (Assumption 3) by planning buffer zones, defining clusters appropriately, or using explicit spillover models (Chapter 11).
5. **Influence and robustness.** Run leave-one-cohort-out and leave-one-period-out analyses. If excluding any single cohort changes the estimate by more than 25% or flips the sign, investigate why. Vary controls, windows, and binning, and compare multiple estimators (Callaway–Sant’Anna, Sun–Abraham, Borusyak–Jaravel–Spiess) to assess sensitivity.
6. **Inference and multiplicity.** Cluster by unit at minimum. Consider two-way clustering or wild bootstrap when the number of clusters is small. Adjust for multiplicity when reporting many event times or subgroup effects. Report diagnostics transparently and disclose deviations from any pre-analysis plan.

These diagnostic procedures provide evidence on the credibility of the identifying assumptions and the robustness of conclusions. The next section illustrates how these methods apply to common marketing applications.

## 4.9 Marketing Applications and Patterns

Marketing panels feature rich adoption patterns, substantial heterogeneity, and pervasive threats to identification assumptions. This section illustrates how the DiD framework applies to common marketing applications: loyalty programme rollouts, retail pricing policy changes, and platform channel expansions. We emphasise the estimand definitions, the identification challenges, the choice of estimator, and the diagnostic and inference considerations specific to each setting.

### Loyalty Programme Rollouts with Staggered Adoption

A retail chain operates 500 stores observed over 12 quarters. The chain launches a loyalty programme in staggered fashion: 100 stores in quarter three, 200 stores in quarter five, 150 stores in quarter seven, and 50 stores never receive the programme during the sample. Outcomes are quarterly sales per store. The goal is to estimate the effect of the programme on sales and to trace how effects evolve as customers enrol, accumulate points, and develop habits.

The data structure is a balanced panel with  $N = 500$  and  $T = 12$ , a thin panel well suited to modern DiD estimators. The adoption pattern creates three treated cohorts ( $g = 3, 5, 7$ ) and a never-treated control group (stores with  $G_i = \infty$ ). The primary estimand is the overall ATT as defined in Section 4.2: the average effect on sales for treated stores and quarters relative to the counterfactual of no programme. Secondary estimands include event-time effects  $\theta_k$  (to trace dynamics) and cohort-specific effects  $\tau_g$  (to assess whether early adopters experience different effects than late adopters).

Identification relies on the Staggered Parallel Trends condition (Assumption 6) across cohorts: in the absence of the programme, treated and control stores would have experienced similar trends in sales. This assumption is plausible if the rollout timing is driven by operational capacity or by store identifiers (for example, alphabetical order) rather than by anticipated sales trends. If the programme is rolled out first to high-growth stores, parallel trends is violated. Diagnostics include plotting pre-treatment trends for each cohort, estimating an event-study specification with leads to diagnose pre-trends, and checking covariate balance (store size, demographics, competitive intensity) across cohorts.

Spillovers are a serious concern. Customers may refer friends, creating positive spillovers from treated stores to nearby control stores. Customers may cross-shop, shifting spend from non-programme stores to programme stores. Positive spillovers attenuate the estimated effect because control store outcomes are elevated by spillovers from treated stores, reducing the observed difference. Negative spillovers (e.g., treated stores cannibalising sales from control stores) would inflate the estimated effect. The direction of bias depends on the sign and magnitude of spillovers, which should be assessed using explicit spillover models or geographic discontinuities. Design-based solutions include defining clusters that group nearby stores, creating buffer zones (excluding stores within a certain radius of treated stores from the control group), or estimating explicit spillover models (Chapter 11) that quantify direct and indirect effects.

When spillovers are plausible, be explicit about what effect you are targeting: the direct effect on treated stores, spillover effects on untreated nearby stores, and the total effect (direct plus spillovers) on the relevant market.

Heterogeneity is expected. The programme may work well in affluent, low-competition areas where customers have high lifetime value and respond strongly to rewards. It may have negligible effects in saturated urban markets where customers are less loyal. Event-time effects are likely to grow over time as customers accumulate points and develop habits, so the immediate effect  $\theta_0$  understates the long-run impact. Modern DiD estimators (Callaway–Sant’Anna or Sun–Abraham) are appropriate as primary estimators, with TWFE reported only as a benchmark after diagnostics (Section 4.4).

Inference should cluster by store to account for serial correlation. With 500 store-level clusters, standard cluster-robust standard errors are generally reliable (see Section 4.7). However, because the rollout involves only a few adoption cohorts, it is good practice to complement standard errors with influence checks (for example, leave-one-cohort-out) and a wild cluster bootstrap as a robustness check. Event-time estimates require multiplicity adjustment if the goal is to treat each  $\theta_k$  as a separate primary claim, but if the event-time profile is exploratory and the primary claim concerns the overall ATT, adjustment is not required for the exploratory path.

## Retail Pricing Policy Changes with Staggered Implementation

A retailer implements a new pricing policy (for example, everyday low pricing replacing frequent promotions) across product categories in waves over six quarters. The retailer observes sales, margins, and customer visit frequency for 200 categories over 24 weeks. The policy is implemented in 50 categories in week 4, 75 categories in week 8, 50 categories in week 12, and 25 categories never adopt the policy during the sample. The goal is to estimate the effect on margins and to assess whether effects differ by category characteristics (price elasticity, brand concentration, competitive intensity).

The data structure is a panel with  $N = 200$  categories and  $T = 24$  weeks, a thin panel with moderate  $T$ . The adoption pattern creates three treated cohorts ( $g = 4, 8, 12$ ) and a never-treated group (categories with  $G_i = \infty$ ). The primary estimand is the overall ATT on margins. Secondary estimands include event-time effects (to trace dynamics) and heterogeneous effects by category characteristics (to guide targeting of future policy changes).

Identification relies on parallel trends across categories. This assumption is plausible if the rollout is driven by operational constraints (for example, staggering implementation to allow staff training and system updates) rather than by anticipated margin trends. If the policy is rolled out first to high-margin categories or to categories with declining trends, Parallel Trends is violated. Diagnostics include pre-trend diagnostics by cohort and placebo analyses using never-treated categories.

Competitive responses are a key threat. If a retailer reduces promotional frequency in one category, competitors may respond by increasing promotions in that category, partially offsetting the margin gains. Alternatively, customers may substitute across categories, shifting spend from treated categories to untreated

categories with frequent promotions. These general equilibrium effects mean that the estimated effect conflates the direct effect of the policy with indirect effects due to substitution and competition. Addressing these requires explicit modelling of cross-category substitution patterns or estimation of spillover effects (Chapter 11). SUTVA (Assumption 3) is violated if control categories are affected by substitution.

Heterogeneity by category characteristics can be explored using subgroup analyses or interactions. Estimate  $\tau(g, t)$  separately for high-elasticity and low-elasticity categories, or include interactions between treatment indicators and category characteristics in the regression. Causal forests (Chapter 12) provide a flexible approach to estimating heterogeneous effects as a function of many covariates simultaneously.

Inference should cluster by category. With 200 category-level clusters, standard cluster-robust standard errors are typically adequate, but again most of the identifying variation comes from a small number of treated cohorts. Wild cluster bootstrap is therefore recommended as a robustness check, particularly if one or two cohorts appear unusually influential. Two-way clustering (by category and week) accounts for potential cross-category correlation within weeks due to common demand shocks. Multiple testing adjustments are needed if testing heterogeneous effects across many subgroups or category characteristics.

## Platform Channel Expansion with Phased Entry

A food delivery platform enters 30 cities over two years, with entry times staggered based on market size, regulatory environment, and operational capacity. The platform observes monthly restaurant revenues in 50 cities (30 treated, 20 never-treated) over 24 months. The goal is to estimate the effect of platform entry on restaurant revenues and to assess whether effects differ by restaurant type (independent vs chain, cuisine type, price point).

The data structure is a panel with  $N = 50$  cities and  $T = 24$  months. The adoption pattern creates many cohorts (up to 30 if each city enters at a different time) and a never-treated control group of 20 cities (cities with  $G_i = \infty$ ). The primary estimand is the overall ATT on restaurant revenues. Secondary estimands include event-time effects (to trace the speed of market penetration and competitive adjustment) and heterogeneous effects by restaurant type.

Identification relies on Parallel Trends across cities. This assumption is questionable if the platform enters larger, faster-growing cities first. This creates selection on observables (market size, growth rate) and possibly on unobservables (unobserved demand shocks, entrepreneurial activity). Conditioning on city characteristics through covariate adjustment or propensity score weighting can restore Conditional Parallel Trends. Alternatively, factor models using interactive fixed effects or generalised synthetic control (Chapters 8, 6) can accommodate differential trends driven by common shocks (macroeconomic conditions, pandemic waves, national restaurant trends) without requiring parallel trends in levels. When cities differ markedly in growth trends even after covariate adjustment, factor models or generalised synthetic control may provide a more credible counterfactual than Parallel Trends in levels.

Competitive responses are central. Incumbent platforms adjust pricing and marketing in response to entry, partially offsetting the treatment effect. Restaurants respond by adjusting menus, delivery fees, and

participation decisions. These dynamic adjustments mean that the effect in the first month post-entry differs from the effect after the market reaches a new equilibrium. Event-time estimates trace these dynamics, and long-run effects (large  $k$ ) capture the equilibrium impact.

The overall ATT on restaurant revenues may mask substantial heterogeneity: platform entry may increase revenues for some restaurants (those that gain visibility and delivery capacity) while decreasing revenues for others (those that lose foot traffic or face new competition). Decomposing effects by restaurant type is essential for understanding the distributional consequences of platform entry.

Spillovers across cities are plausible if the platform's national advertising or if migration and commuting create cross-city linkages. Geographic buffer zones (excluding cities within a certain distance from treated cities) or explicit spatial models (Chapter 11) can address these spillovers. General equilibrium effects—whether entry creates new demand or merely redistributes existing demand among restaurants—require comparing total restaurant revenues (treated and untreated restaurants) to assess category expansion. In this setting, distinguish the direct effect on restaurants in treated cities, spillovers onto restaurants in untreated cities, and the total effect on the wider market.

Heterogeneous effects by restaurant type can be explored using subgroup analyses. Estimate effects separately for independent restaurants and chain restaurants, for high-priced and low-priced restaurants, and for different cuisines. If the platform is more effective for certain types of restaurants, this guides future targeting and partnership strategies.

Inference should cluster by city. With  $N = 50$  city clusters, asymptotic cluster-robust standard errors are generally adequate (per the guidance in Section 4.7), though wild cluster bootstrap provides a robustness check. Two-way clustering by city and month accounts for cross-city correlation within months due to pandemic phases, policy changes, or national trends. Multiple testing adjustments are needed if treating many restaurant-type or subgroup estimates as primary claims.

These three applications illustrate the versatility of modern DiD methods and the importance of tailoring the analysis to the substantive context. The choice of estimand, the diagnostic workflow, the estimator selection, and the inference procedure all depend on the data structure, the adoption pattern, the threats to identification, and the business question. Transparent reporting and sensitivity analyses ensure that conclusions are credible and that readers understand the assumptions required for causal interpretation. The following workflow checklist synthesizes the methods and diagnostics developed in this chapter into a practical step-by-step protocol.

## 4.10 Workflow Checklist

This section provides a compact end-to-end protocol for conducting DiD analyses in marketing panels. It synthesises the estimands, estimators, diagnostics, inference procedures, and reporting standards introduced in this chapter and in Chapter 3.

1. **Define estimands and map cohorts.** Begin by clarifying the substantive question and defining the target estimand precisely. Is the goal to estimate the overall ATT (average effect on all treated units and periods), event-time effects  $\theta_k$  (dynamic profile of effects over time since adoption), cohort-specific effects  $\tau_g$  (effects for early vs late adopters), or calendar-time effects  $\tau_t$  (effects in specific periods accounting for time-varying shocks)? Align the estimand to the business question.

Map adoption cohorts by creating a cohort-time matrix: rows for units, columns for periods, cells indicating cohort  $g$  (the first period in which each unit is treated) or never-treated status (units with  $G_i = \infty$ ). Visualise the adoption pattern to confirm that there is sufficient variation in timing and that comparison groups (never-treated or not-yet-treated units) are available for all treated cohorts and periods.

2. **Choose estimators and pre-specify aggregations.** Based on the estimand and the data structure, select one or more modern DiD estimators: Callaway–Sant’Anna for flexible aggregation and overall ATT, Sun–Abraham for event-time effects and pre-trend diagnostics, de Chaisemartin–d’Haultfœuille for diagnostics and sensitivity to TWFE, or Borusyak–Jaravel–Spiess for factor-structure settings. Estimate TWFE as a benchmark. Do not rely on it as the primary estimate unless diagnostics suggest heterogeneity is not driving the result.

Pre-specify the aggregation weights: will event-time effects be weighted by cohort size, by treated unit-periods, or uniformly? Will the overall ATT weight cohorts equally or by sample size? Document these choices in a pre-analysis plan or analysis script to ensure transparency and to prevent ex post cherry-picking of aggregations.

3. **Specify event-time windows and binning.** Decide how many pre-treatment and post-treatment event times to include. Longer pre-treatment windows provide more pre-period evidence but require cohorts to have long pre-periods. Longer post-treatment windows trace long-run dynamics but may be unavailable if some cohorts adopt late. Choose a reference period (typically  $k = -1$ ) and specify binning (if needed) to ensure adequate sample size in each event-time bin.

4. **Run pre-trend and placebo diagnostics.** Estimate an event-study specification with multiple pre-treatment leads and examine whether lead coefficients are close to zero. A joint Wald diagnostic for whether pre-treatment lead coefficients are jointly zero can complement the plot, using cluster-robust standard errors. Treat this as a falsification check. It has low power and depends on the chosen lead window and binning choices. Plot event-time coefficients with confidence intervals to visually inspect for pre-trends. Conduct placebo-in-time analyses using only pre-treatment data and check that placebo estimates are near zero. Conduct placebo-in-units analyses using never-treated units as placebo-treated and check that placebo estimates are near zero.

If pre-trends are present or placebo estimates are large, consider alternative identification strategies: conditioning on covariates (estimate propensity scores and check balance), factor models (Chapter 8), or synthetic control (Chapter 6).

- 5. Assess covariate balance and overlap.** Compute standardised mean differences (SMDs) for key covariates across treated cohorts and control units. As a practical rule of thumb, values around or above 0.1–0.2 standard deviations warrant attention, with 0.2 often used as a simple benchmark, though the appropriate threshold depends on how strongly covariates predict outcomes. If imbalances are large, adjust for covariates using regression, propensity score weighting, or matching, and re-check balance after adjustment.

Plot propensity score distributions for treated and control units. If overlap is limited, consider trimming extreme propensity scores and report how results change under reasonable trimming choices. Document any trimming rule and report the fraction of the sample excluded.

- 6. Choose clustering and inference procedures.** Cluster standard errors by unit as a default. Consider two-way clustering (by unit and time) if cross-unit correlation within periods is plausible. If the number of clusters is very small (fewer than 20), prefer wild cluster bootstrap or randomisation inference. With 20–50 clusters, compare asymptotic cluster-robust standard errors to bootstrap results, as in Section 4.7. With 50 or more clusters, asymptotic cluster-robust standard errors are generally reliable. If the design is experimental, use randomisation inference to respect the randomisation protocol.

If reporting many event-time coefficients, subgroup effects, or sensitivity analyses, pre-specify what constitutes the family of hypotheses for any multiplicity adjustment. Decide on a multiplicity adjustment: Bonferroni for conservative family-wise error rate control, FDR for less conservative control of false discoveries, or Romano–Wolf stepdown for better power while controlling family-wise error rate. Distinguish primary from exploratory analyses and apply adjustments to primary tests.

- 7. Estimate and report aggregated effects.** Estimate the chosen modern DiD estimator(s) and aggregate cohort-time effects  $\tau(g, t)$  into the pre-specified summary measures: overall ATT, event-time effects  $\{\theta_k\}$ , cohort-specific effects  $\{\tau_g\}$ , or calendar-time effects  $\{\tau_t\}$ . Report point estimates, standard errors, and confidence intervals.

Plot event-time graphs showing  $\hat{\theta}_k$  against  $k$  with confidence intervals. Highlight the reference period ( $k = -1$ ) and annotate pre-treatment and post-treatment regions.

Compare estimates across methods (CS, SA, BJS, TWFE) to assess sensitivity to estimator choice. When these methods agree, conclusions are more robust. When they diverge materially, report the spread and investigate differences in weights, comparison groups, and outcome-model assumptions.

- 8. Conduct sensitivity analyses.** Vary the set of control units (never-treated only vs never-treated and not-yet-treated), the event-time window (short vs long), the binning choices (fine vs coarse), and the covariate adjustments (no controls vs rich controls). Construct a specification curve showing the distribution of estimates across specifications. Conduct leave-one-cohort-out and leave-one-period-out analyses.

If spillovers are plausible, estimate models that exclude nearby units, define buffer zones, or estimate explicit spillover effects (Chapter 11). Compare estimates with and without spillover adjustments.

- 9. Document and report transparently.** Prepare a report that includes the research question, the data structure (panel dimensions, adoption pattern, sample characteristics), the estimand definition, and the chosen estimators and aggregation weights. Document the diagnostic results (pre-trends, placebo analyses, covariate balance), the primary estimates (overall ATT, event-time effects), and the sensitivity analyses

(specification curve, leave-one-out, spillover adjustments). Report the inference procedures (clustering, bootstrap, multiplicity) and provide substantive interpretation.

Register the analysis plan *ex ante* if possible, or timestamp the analysis script and document deviations from the initial plan. Provide replication materials: cleaned data (or simulated data if proprietary), analysis scripts, and documentation of software versions and packages used.

By following this workflow, practitioners can conduct DiD analyses that are transparent, rigorous, and aligned with modern best practices. The workflow integrates design-based reasoning, diagnostics, and sensitivity analysis, ensuring that conclusions are credible and that assumptions are articulated and assessed.

This chapter has developed the difference-in-differences framework from its canonical  $2 \times 2$  form through modern methods for staggered adoption with heterogeneous treatment effects. We have defined estimands that respect heterogeneity, explained the pitfalls of two-way fixed effects under heterogeneity, surveyed modern heterogeneity-robust estimators, developed diagnostic workflows, and illustrated applications in marketing. Chapter 5 develops event-study designs in more detail, tracing dynamic treatment paths and diagnosing anticipation effects.



## Chapter 5

# Event-Study Designs

This chapter develops event-study designs for panel data analysis. You will learn to define event-time effects  $\theta_k$  and aggregate post-treatment effects into cumulative effects and long-run multipliers (LRM). You will specify lead-lag regressions with appropriate normalisation, binning, and window selection, and estimate event-time effects under staggered adoption using the heterogeneity-robust estimators introduced in Chapter 4. You will use pre-treatment leads ( $k < 0$ ) as diagnostics for anticipation and violations of parallel trends, rather than as treatment effects. You will choose inference procedures that respect clustering and multiple testing, and extend the framework to continuous treatments and spillovers. Finally, you will learn to interpret event-study plots for marketing decisions, translating dynamic profiles into ROI and strategic metrics.

## 5.1 Motivation and Setup

Event-study designs organise panel data by event time rather than calendar time, indexing observations by their position relative to an intervention date. This chapter builds on the difference-in-differences framework of Chapter 4 and the event-study introduction in Section 4.6, focusing on event-time estimands, specifications, and diagnostics in greater depth. This simple reframing yields powerful diagnostic and substantive benefits. Event studies visualise the dynamic evolution of treatment effects, provide diagnostics for the plausibility of parallel trends and no anticipation before treatment, reveal anticipatory responses or delayed effects, and communicate causal narratives in a transparent, accessible format. In marketing applications, where interventions rarely produce immediate, constant effects, event studies are essential for understanding ramp-up periods, decay rates, carryover, and long-run impacts.

The logic of event-time analysis is straightforward. For each unit  $i$ , define the adoption time  $G_i$  as the first period in which unit  $i$  receives treatment. Event time  $k = t - G_i$  measures the number of periods since (or until) adoption:  $k = -2$  denotes two periods before adoption,  $k = 0$  denotes the adoption period,  $k = 3$  denotes three periods after adoption. By aligning units in event time rather than calendar time, event studies pool observations from units that adopted treatment at different calendar dates but are at the same point in their post-treatment trajectory. This pooling increases precision when it is plausible to assume a common dynamic profile  $\{\theta_k\}$  across cohorts (homogeneous dynamic effects). When effects are heterogeneous, pooling can introduce bias: the TWFE event-time coefficient  $\hat{\theta}_k$  becomes a weighted average of cohort-specific effects  $\tau(g, g+k)$ , with opaque regression weights. This is the event-study analogue of the TWFE weighting problem in Chapter 4. The formal event-time estimands  $\theta_k$  in Section 2.3 are aggregation-based objects that coincide with  $\hat{\theta}_k$  only under strong homogeneity conditions. Event-time pooling enables estimation of dynamic treatment effect profiles even when individual cohorts are small or when calendar-time aggregations obscure dynamics, but the analyst should assess whether homogeneity is plausible.

A related concern is composition bias at long horizons. Event-time effects at large positive  $k$  are identified only from early-adopting cohorts (late adopters have not yet reached that horizon), while effects at large negative  $k$  are identified only from late-adopting cohorts (early adopters were not yet in the panel that far before adoption). An upward-sloping profile in  $\hat{\theta}_k$  can therefore arise purely from early-adopting cohorts having larger  $\tau(g, g+k)$ , even if each cohort's effect is flat over  $k$ . If early and late adopters differ systematically, the event-time profile may reflect composition changes rather than true dynamics. Truncating the event window or reporting cohort-specific effects can diagnose this issue.

### Worked Example: Event-Time Alignment and Identification

Consider three stores adopting a loyalty programme at different calendar times: Store A (Q3), Store B (Q5), and Store C (Q7). Store D never adopts.

Observed sales data (calendar time)

Store	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
A ( $G_i = 3$ )	100	98	<b>115</b>	118	120	122	124	126
B ( $G_i = 5$ )	102	100	101	103	<b>130</b>	135	138	140
C ( $G_i = 7$ )	99	101	100	102	104	105	<b>145</b>	148
D (Never)	100	99	101	100	102	101	100	99

*Bold values indicate the adoption quarter ( $k = 0$ ).*

Event-time mapping ( $k = t - G_i$ )

Event Time	Store A	Store B	Store C	Store D	Comparison
$k = -2$	Q1	Q3	Q5	Q1/3/5	Pre-treatment baseline
$k = 0$	Q3	Q5	Q7	Q3/5/7	Adoption impact
$k = +2$	Q5	Q7	—	Q5/7	Short-run dynamics
$k = +4$	Q7	—	—	Q7	Long-run dynamics

Alignment matters for both identification and interpretation. To estimate the immediate effect  $\theta_0$ , the event-study design pools the adoption-period jump for Store A (Q3), Store B (Q5), and Store C (Q7), each compared to control units at the same calendar time. At calendar Q5, for example, Store B is treated ( $k = 0$ ) while Store C is still pre-treatment ( $k = -2$ ), so Store C can help net out common Q5 seasonality under no anticipation and staggered parallel trends (Chapter 2). The same alignment also highlights composition bias:  $k = +4$  is observed only for Store A, so  $\hat{\theta}_4$  blends treatment effects with any systematic differences between early and late adopters.

Event studies are a natural extension of the difference-in-differences framework developed in Chapter 4. The canonical two-period DiD compares treated and control units before and after treatment, implicitly estimating a single average treatment effect on the treated (ATT) that applies to all post-treatment periods. Event studies relax this restriction by estimating a separate treatment effect for each event time  $k$ , allowing the effect to vary over the post-treatment horizon and enabling pre-trend diagnostics by estimating lead coefficients for pre-treatment event times ( $k < 0$ ). The connection to DiD is direct: the overall average treatment effect on the treated (ATT) estimated by DiD is a weighted average of the event-time effects  $\theta_k$  in Section 2.3; the choice between reporting a single ATT or a full event-time profile depends on the substantive question and the data structure.

Anticipation and carryover, central concerns in marketing, motivate event-time analysis. Anticipation occurs when units respond to expected future treatment by altering behaviour in advance. Customers anticipating a loyalty programme launch may delay purchases to qualify for rewards, depressing pre-treatment outcomes relative to what they would have been absent the anticipation. This produces non-zero pre-treatment coefficients in the event study, violating the no-anticipation assumption (Assumption 1 in Chapter 2). The sign of the anticipation effect depends on the mechanism: delayed purchases produce negative coefficients,

while stockpiling in advance of a price increase produces positive coefficients. Carryover occurs when treatment effects persist or evolve over time. An advertising campaign may build brand awareness gradually, producing effects that grow over multiple quarters. A pricing change may trigger competitive responses that erode initial gains, producing effects that decay. Distributed lag models and structural dynamic panel models (Chapter 10) provide parametric approaches to modelling carryover. Event studies offer a flexible, reduced-form alternative that imposes minimal restrictions on the lag structure, though they implicitly assume a common dynamic profile  $\{\theta_k\}$  across cohorts (homogeneous dynamic effects). When this assumption fails, cohort-specific event studies or interaction models are required.

Event studies also link marketing actions to stock market reactions. Financial event studies estimate abnormal returns in tight windows around announcements and attribute them to the new information the market receives about future cash flows. A tight window is a design choice rather than a guarantee, since information leakage and overlapping announcements can confound attribution even over a few trading days. Customer satisfaction and innovation announcements are associated with positive abnormal returns, while some signals show limited or transitory effects once risk and benchmarks are accounted for [Fornell et al., 2006, Mizik and Jacobson, 2009, Jacobson and Mizik, 2009, Sood and Tellis, 2009]. Financial event studies differ methodologically from the panel event studies developed in this chapter: they use market models to define counterfactual returns rather than parallel trends, and inference typically relies on cross-sectional independence rather than clustering. We focus on panel event studies in this chapter; financial event studies are complementary for measuring capital market reactions but rely on different identification strategies and cannot substitute for panel-based designs.

Design-based thinking elevates event studies from description to identification. Pre-treatment coefficients provide diagnostics for whether parallel trends is plausible, and post-treatment coefficients trace dynamics. Plots communicate the causal narrative at a glance and help stakeholders judge credibility and relevance.

As discussed in Chapter 17 and Section 4.6, pre-trend diagnostics have low power. A non-significant result is at best supportive evidence. Rambachan–Roth sensitivity analysis (Section 5.8) formalises how conclusions change if parallel trends is violated by a bounded amount.

As discussed in Chapter 4 (Sections 4.4, 4.5, and 4.6), do not use already-treated units as controls under staggered adoption. Rely on heterogeneity-robust estimators instead. Here we focus on event-time specification, diagnostics, and interpretation. See Chapter 4 for DiD estimands and TWFE pitfalls.

We develop the event-study framework for practical use in marketing panels. We define event-time estimands and aggregations, specify lead-lag regressions (normalisation, binning, windows), and estimate under staggered adoption with heterogeneity-robust methods. We state assumptions and diagnostics, give plotting and inference guidance, and cover extensions to continuous treatment, cohort-specific event studies, and interference.

## 5.2 Event-Time Estimands

Event-time estimands define the causal quantities of interest in event-study designs. The fundamental building block is the average treatment effect at event time  $k$ , denoted  $\theta_k$  as introduced in Chapter 4 and Section 4.2, which measures the average causal effect of treatment  $k$  periods after adoption. Pre-treatment event times ( $k < 0$ ) are used as leads for diagnostics: under no anticipation and parallel trends, the corresponding lead coefficients should be close to zero. Formal definitions require care to specify the population being averaged over (all treated units, specific cohorts, or weighted combinations), the reference level relative to which effects are measured, and the aggregation scheme when treatment effects are heterogeneous across cohorts.

### Basic Event-Time Effects

The simplest definition treats  $\theta_k$  (for  $k \geq 0$ ) as the average difference between treated and untreated potential outcomes  $k$  periods after adoption, averaged over ever-treated units for whom event time  $k$  is observed:

$$\theta_k = \mathbb{E}[Y_{i, G_i+k}(G_i) - Y_{i, G_i+k}(\infty) \mid G_i < \infty, G_i + k \leq T].$$

Here  $Y_{it}(G_i)$  denotes the potential outcome under adoption at time  $G_i$  for unit  $i$ , and  $Y_{it}(\infty)$  denotes the never-treated counterfactual. The conditioning on  $G_i < \infty$  restricts to ever-treated units. Never-treated units contribute to identification as controls but not to the estimand.

To make the aggregation explicit, decompose  $\theta_k$  in terms of cohort-time effects  $\tau(g, t)$ . The event-time effect  $\theta_k$  is a weighted average of  $\tau(g, g+k)$  across cohorts:

$$\theta_k = \sum_{g: g+k \leq T} w_{gk} \tau(g, g+k), \quad \sum_g w_{gk} = 1.$$

where the weights  $w_{gk}$  reflect the relative sample sizes or importance of each cohort at event time  $k$ . Common weighting schemes include uniform weights (each cohort receives equal weight), sample-size weights (cohorts are weighted by the number of observations at event time  $k$ ), and treated-unit-period weights. Different weighting schemes can produce different  $\theta_k$  estimates even when the underlying  $\tau(g, g+k)$  are the same. As emphasised in Chapter 4, this is a design choice, not a nuisance. Transparency about weights is essential.

### Cohort-Specific Profiles

Cohort-specific event-time profiles, denoted  $\theta_{g,k}$ , estimate the effect for cohort  $g$  at event time  $k$  without aggregating across cohorts:

$$\theta_{g,k} = \mathbb{E}[Y_{i, g+k}(g) - Y_{i, g+k}(\infty) \mid G_i = g].$$

The cohort-specific event-time effect  $\theta_{g,k}$  is simply the cohort-time effect  $\tau(g, g+k)$  from Chapter 4 evaluated at calendar time  $t = g + k$ . We interpret  $\theta_{g,k}$  causally for post-treatment event times ( $k \geq 0$ ). For  $k < 0$ , the same construction yields placebo lead objects that should be close to zero under no anticipation, and we use them as diagnostics rather than as treatment effects. Estimating cohort-specific profiles is valuable when treatment effects are expected to vary substantially across cohorts—for example, if early adopters differ from late adopters in ways that moderate the treatment response, or if macroeconomic conditions or competitive environments differ across the calendar periods during which different cohorts are treated. Cohort-specific profiles also provide a diagnostic. If profiles are similar across cohorts, pooling them into a single  $\theta_k$  is defensible. If profiles diverge, aggregation obscures heterogeneity and may mislead.

Calendar-time aggregation provides an alternative to event-time aggregation when the goal is to estimate the total effect in specific calendar periods accounting for the mix of cohorts at different event times. For example, a retailer rolling out a loyalty programme in waves may want to estimate the total sales impact in quarter four, summing effects across cohorts that are at different event times in that quarter. Calendar-time aggregation weights  $\tau(g, t)$  by the prevalence of each cohort in period  $t$ . This produces an estimate of the average treatment effect in period  $t$  across all treated units observed in that period.

Event-time aggregation, by contrast, pools observations by event time  $k$  regardless of calendar period, producing an estimate of the average effect  $k$  periods post-adoption.

## Cumulative and Long-Run Effects

Long-run and cumulative effects are central to marketing applications where interventions are intended to have persistent impacts. The cumulative effect over  $K$  post-treatment periods is the sum of event-time effects:

$$\sum_{k=0}^K \theta_k.$$

This sum measures the total impact of treatment from adoption through  $K$  periods post-adoption, integrating both immediate and delayed effects. Note that the cumulative effect is only identified for  $k$  values observed in the data. If the goal is to estimate the total long-run effect but the post-treatment window is short, the analyst must either extend the observation period or extrapolate using parametric assumptions (for example, exponential decay). Extrapolation introduces model dependence and should be reported transparently. The long-run multiplier compares the cumulative effect over a chosen horizon  $K$  to the immediate effect:

$$\text{LRM} = \frac{\sum_{k=0}^K \theta_k}{\theta_0}, \quad \text{assuming } \theta_0 \neq 0.$$

The choice of  $K$  is part of the estimand and should reflect both data support and the decision horizon. If  $\text{LRM} > 1$ , the cumulative impact exceeds the immediate impact, indicating positive carryover or ramp-up. If  $\text{LRM} < 1$ , the immediate effect overstates the long-run impact, suggesting decay or erosion. Marketing interventions such as loyalty programmes, advertising campaigns, and platform launches typically exhibit

$\text{LRM} > 1$  because customer habit formation, brand awareness accumulation, and network effects generate persistence and growth. Promotional pricing or temporary discounts may exhibit  $\text{LRM} < 1$  if demand is merely shifted forward in time rather than created.

#### Edge Case: Delayed-Onset Effects

When effects ramp up gradually,  $\theta_0$  may be near zero, making LRM undefined or numerically unstable. In such cases, it is often more informative to report the cumulative effect  $\sum_{k=0}^K \theta_k$  and the full event-time profile rather than forcing a multiplier. If you do report a ratio that averages early post-treatment effects in the denominator, make the averaging rule explicit and treat the multiplier as a descriptive summary, not a primary estimand. The key is transparency: if  $\theta_0 \approx 0$ , the standard LRM is uninformative, and readers should be shown the full event-time profile rather than a single summary statistic. See Section 4.2 for the analogous issue with cohort-specific effects.

The half-life of a treatment effect—the number of periods required for the effect to decay to half its peak magnitude—provides another summary of dynamics. The half-life concept assumes that effects decay monotonically after reaching a peak. If effects ramp up before decaying, the “initial magnitude”  $\theta_0$  may not be the peak, and the half-life should be measured from the peak rather than from  $k = 0$ . If effects follow exponential decay from the peak,  $\theta_k = \theta_{\max} \exp(-\lambda(k - k_{\max}))$  for  $k \geq k_{\max}$ , where  $\lambda > 0$  is the decay rate (not to be confused with time fixed effects  $\lambda_t$ ), then the half-life is  $k_{1/2} = \log(2)/\lambda$ . Estimating the half-life requires either imposing parametric structure (as in distributed lag models, Chapter 10) or directly reading off the event-study path at the point where  $\theta_k \approx \theta_{\max}/2$ . Chapter 10 provides parametric strategies for estimating the decay rate and half-lives. Here we treat half-life primarily as a visual and descriptive feature of the event-study path. The advantage of the event-study approach is transparency. The half-life is directly visible in the plot of  $\theta_k$  against  $k$ , making it accessible to non-technical stakeholders and robust to misspecification of the functional form.

## Aggregation and Reference Choices

Aggregation schemes shape the interpretation and policy relevance of event-time estimates. Consider a loyalty programme rolled out to stores in three cohorts ( $g = 2, 5, 8$ ) over eight quarters. The programme may have different effects for each cohort because early-adopting stores differ in size, demographics, or competitive intensity. Calendar-time aggregation in quarter eight weights the cohort-two stores (at event time  $k = 6$ ), cohort-five stores (at event time  $k = 3$ ), and cohort-eight stores (at event time  $k = 0$ ) by their sample sizes. This produces an estimate of the total programme impact in quarter eight. Event-time aggregation at  $k = 3$  weights cohorts two, five, and eight at their respective  $k = 3$  observations, producing an estimate of the effect three quarters post-adoption averaged across cohorts. The choice depends on the substantive question. Calendar-time aggregation is natural for forecasting or planning decisions tied to specific periods,

while event-time aggregation is natural for understanding the dynamic profile and for generalising to future rollouts.

The reference level for event-time effects is also a choice. Effects are typically reported relative to the period immediately before adoption ( $k = -1$ ), which is normalised to zero by construction (omitting the  $k = -1$  indicator from the regression). This convention interprets  $\theta_k$  as the change in outcomes from  $k = -1$  to  $k$ . Alternative reference levels are possible: effects could be reported relative to the average of all pre-treatment periods, relative to a specific pre-treatment period further in the past, or relative to the outcome level at  $k = 0$ . The choice of reference affects the magnitude and interpretation of coefficients but not the differences between coefficients. These differences measure contrasts between event times and are invariant to the normalisation. Transparency about the reference level and robustness checks using alternative references ensure that conclusions are not artefacts of the normalisation.

There is no single “right” estimand, as emphasised in the estimand crosswalk in Chapter 2. Choose based on the decision: expansion (profiles for comparable stores), advertising ROI (cumulative effects), entry strategy (ramp-up and long-run). Estimate the full profile, then aggregate transparently into the summary that answers the question. The next section develops the regression specification.

### 5.3 Specification: Leads and Lags

Specifying an event-study regression requires choosing which event-time indicators to include (leads and lags), which event time to normalise to zero (the reference bin), how to group extreme event times (binning), and how long a window to estimate (window selection). These choices affect identification, precision, and interpretation. We provide practical guidance grounded in potential outcomes and design-based diagnostics.

#### Basic Specification

This is the TWFE-style event-study regression discussed in Chapter 4, Section 4.6. Here we focus on its specification details. The standard TWFE event-study regression takes the form

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \in \mathcal{K} \setminus \{-1\}} \beta_k^{\text{TWFE}} \mathbf{1}\{t - G_i = k\} + \varepsilon_{it},$$

where  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects, and  $\mathcal{K}$  is the set of event times included in the window. The coefficients  $\beta_k^{\text{TWFE}}$  are regression coefficients, not the causal event-time estimands  $\theta_k$  in Section 5.2. Under treatment effect heterogeneity they can differ substantially because already-treated units enter as implicit controls. Even under homogeneous treatment effects, interpreting these coefficients causally still requires the usual parallel-trends and no-anticipation assumptions from Chapter 2.  $\beta_{-1}^{\text{TWFE}}$  is normalised to zero. Under heterogeneity,  $\beta_k^{\text{TWFE}}$  is a weighted average of cohort-specific effects with potentially negative weights, distinct from the target  $\theta_k$  (Chapter 4, Sections 4.4–4.5).

To resolve this, modern estimators use interaction-weighted specifications (for example, Sun–Abraham) or aggregation of group-time effects (for example, Callaway–Sant’Anna). The interaction-weighted specification is:

$$Y_{it} = \alpha_i + \lambda_t + \sum_g \sum_{k \neq -1} \delta_{g,k} \mathbf{1}\{G_i = g\} \mathbf{1}\{t - G_i = k\} + \varepsilon_{it}.$$

This fully saturates the model with cohort-specific event-time coefficients  $\delta_{g,k}$ , identifying  $\tau(g, g+k)$  under the staggered parallel-trends and no-anticipation assumptions, using never-treated or not-yet-treated units as the comparison group. Under the staggered parallel-trends and no-anticipation assumptions (Section 4.3),  $\delta_{g,k}$  identifies the cohort-specific event-time effect  $\tau(g, g+k)$ . The target parameter  $\theta_k$  is then estimated by aggregating these clean coefficients:

$$\hat{\theta}_k = \sum_g \hat{w}_{gk}^{\text{SA}} \hat{\delta}_{g,k},$$

where the Sun–Abraham weights  $\hat{w}_{gk}^{\text{SA}}$  are user-specified (typically proportional to cohort size at event time  $k$  or uniform across cohorts). The choice of weights is a substantive decision that affects the estimand (Section 5.2). This ensures that  $\theta_k$  represents a properly weighted average of causal effects, not an opaque regression-weighted artefact of TWFE.

The omitted category should be a pre-treatment period, not a post-treatment period. Omitting a post-treatment period would interpret all coefficients as deviations from a treated baseline, obscuring the causal effect. For example, if  $k = 0$  were omitted,  $\theta_1$  would estimate the difference between  $k = 1$  and  $k = 0$ , which is the change in the effect from the first to the second post-treatment period, not the effect relative to an untreated baseline. This makes post-treatment coefficients differences from an already-treated baseline rather than from the untreated counterfactual. Omitting  $k = -1$  makes all post-treatment coefficients effects relative to a purely untreated baseline. Omitting  $k = 0$  instead makes coefficients for  $k \geq 1$  effects relative to the first treated period, which can be useful when  $k = 0$  is itself contaminated but no longer estimates the total effect relative to no treatment. The convention of omitting  $k = -1$  ensures that post-treatment coefficients  $\theta_k$  for  $k \geq 0$  measure the treatment effect relative to the immediate pre-treatment period, which aligns with the canonical DiD interpretation.

Some researchers omit  $k = 0$  (the treatment period itself) and interpret post-treatment effects relative to the immediate post-treatment baseline. This can be appropriate if the treatment is implemented part-way through the period, so that outcomes at  $k = 0$  are a mix of treated and untreated observations. It is also appropriate when anticipation contaminates the treatment period itself: if units respond to expected treatment before it formally begins,  $k = 0$  may already reflect anticipatory behaviour, making it a poor baseline. However, omitting  $k = 0$  complicates interpretation because the effect at  $k = 1$  is then the difference from  $k = 0$ , not from a purely untreated baseline. The cleanest approach is to omit  $k = -1$  and interpret  $\theta_0$  as the immediate effect,  $\theta_1$  as the effect one period post-treatment, and so on, all relative to the pre-treatment baseline.

## Binning

Binning groups extreme event times into intervals to stabilise estimates when data are sparse at the tails. If the panel spans many periods and some cohorts adopt early, observations at large event times ( $k \gg 0$  or  $k \ll 0$ ) may be few, leading to imprecise estimates and wide confidence intervals. Binning aggregates these sparse observations into a single bin. This is a bias-variance trade-off: binning reduces variance (tighter confidence intervals) but can introduce bias if effects vary within the bin. If effects at  $k = 8$  and  $k = 12$  differ substantially, pooling them into a single bin produces a coefficient that represents neither. For example, a binning scheme might define bins  $k \in \{-10, -9, -8\}$  (aggregated as  $k \leq -8$ ),  $k \in \{-7, -6, \dots, -2\}$  (separate coefficients for each),  $k = -1$  (omitted),  $k \in \{0, 1, 2, \dots, 10\}$  (separate coefficients), and  $k \in \{11, 12, 13, \dots\}$  (aggregated as  $k \geq 11$ ).

The choice of where to bin depends on the support of the data and the substantive interest in distinguishing effects at different event times. If the goal is to diagnose pre-trends, the pre-treatment window ( $k < 0$ ) should include enough separate bins to detect divergence. Bins far from treatment ( $k \ll 0$ ) can be aggregated if support is thin. If the goal is to estimate long-run effects, the post-treatment window should extend as far as the data allow. Bins at large  $k$  can be aggregated if the effect has stabilised or if observations are sparse. As a heuristic, you might aim for at least 50–100 treated unit-period observations per bin, or at least 10

treated units per bin. The right threshold depends on outcome variance and on the number of independent clusters supporting inference. This choice should be pre-specified based on the support of the data (reported in a support table or figure as described below) and should be transparent in reporting. Sensitivity analyses that vary the binning scheme provide evidence on whether conclusions are robust to aggregation choices.

#### Endpoint Binning Caveat

Endpoint bins ( $k \leq -L$  or  $k \geq K$ ) pool observations at heterogeneous event times into a single coefficient, mixing early and late dynamics. This pooling obscures information about effect evolution and can bias interpretation. For example, a coefficient for  $k \geq 8$  averages effects at  $k = 8$ ,  $k = 12$ , and  $k = 20$ , which may differ substantially if effects decay or grow over time. Interpret endpoint bins as averages over heterogeneous horizons, not as point estimates at specific event times. Report sensitivity analyses excluding endpoint bins or using alternative cutoffs to assess robustness. When the endpoint bin dominates the sample (many observations fall into  $k \geq K$ ), consider extending the window or reporting the full disaggregated profile.

## Window Selection

Window selection determines how many pre-treatment and post-treatment event times to include in the regression. Symmetric windows include equal numbers of leads and lags (for example,  $k \in \{-5, \dots, 5\}$ ), while asymmetric windows may include more lags than leads if the primary interest is in post-treatment dynamics. The window should be chosen to balance several considerations. Coverage: the window should span the period over which effects are expected to evolve. Support: the window should not extend so far that many event times have few observations. Statistical power: wider windows include more observations and improve precision, but they also increase the number of parameters estimated. Diagnostic value: the pre-treatment window should be long enough to detect pre-trends, typically at least three to five pre-treatment periods.

#### Pre-Trend Diagnostics Trade-Off

Longer pre-treatment windows provide more power to detect pre-trends, but they also increase the risk of finding spurious departures due to multiple testing. Joint inference for all pre-treatment coefficients (Section 5.8) can summarise evidence against the null that they are jointly zero while controlling the family-wise error rate. See Section 5.8 for multiplicity adjustments and diagnostic details.

## Cohort Composition and Support

Cohort composition varies across event times, and this variation affects the interpretation of  $\theta_k$ . At event time  $k = 0$ , all cohorts contribute observations (each cohort is observed at its adoption period). At event time

$k = 5$ , only cohorts that adopted at least five periods before the end of the sample contribute observations. At event time  $k = -5$ , only cohorts that adopted at least five periods after the start of the sample contribute observations. This means that estimates at extreme event times  $k$  may be identified from subsets of cohorts. If treatment effects are heterogeneous across cohorts, the composition effects can confound interpretation.

For example, suppose early adopters (cohort  $g = 2$ ) experience large effects and late adopters (cohort  $g = 8$ ) experience small effects. At event time  $k = 0$ , both cohorts contribute, and  $\theta_0$  is a weighted average of their effects. At event time  $k = 4$ , only cohort  $g = 2$  contributes (because cohort  $g = 8$  does not have four post-treatment periods in the sample). Therefore  $\theta_4$  reflects only the early-adopter effect. A plot of  $\theta_k$  might show a growing effect over event time, but this could be driven by the changing cohort composition (early adopters dominating at large  $k$ ) rather than by true dynamics within cohorts. Reporting the cohort composition at each event time (a table or figure showing which cohorts contribute to each  $k$ ) and estimating cohort-specific profiles  $\theta_{g,k}$  (linking back to the estimand definitions in Section 5.2) provide diagnostics for this issue.

Support tables report, for each event time  $k$ , the number of observations, the number of cohorts contributing, and the identities of those cohorts. Support figures plot the sample size or effective sample size (accounting for weights and clustering) as a function of  $k$ , highlighting where the data are rich and where they are sparse. These diagnostics inform binning and window selection and help readers assess the robustness of conclusions to the inclusion or exclusion of extreme event times.

The specification of leads and lags is not a one-size-fits-all decision but a design choice informed by the data structure, the substantive question, and the diagnostic priorities. Pre-specifying the reference period, binning scheme, and window in a pre-analysis plan disciplines the analysis and guards against data-driven specification searches that capitalise on chance. Reporting the support and cohort composition for each event time ensures transparency and enables readers to assess whether the estimated profile is driven by genuine dynamics or by composition effects.

## 5.4 Estimation under Staggered Adoption

Estimating event-time effects under staggered adoption with heterogeneous treatment effects requires care to avoid the biases that plague traditional two-way fixed effects (TWFE) event-study regressions. This section explains the sources of bias, presents heterogeneity-robust estimation methods, and provides practical guidance on implementation in marketing panels.

### TWFE Bias in Event Studies

TWFE event studies can suffer contamination by already-treated units and negative weights under heterogeneity. The core problem is that TWFE uses already-treated units as implicit controls when estimating effects for newly-treated units, comparing units at event time  $k = 2$  to units at  $k = 5$ . When effects evolve over event time, this produces contaminated comparisons that mix treatment effect dynamics with differential timing. Additionally, cohort-specific effects enter with opaque regression weights that need not be proportional to cohort size and can be negative on some cohort-time cells.

For formal decompositions and weight formulas, see Chapter 4, Sections 4.4 and 4.5. Note that this applies equally to the event-study coefficients, not just to a single summary effect such as ATT. Use TWFE plots as diagnostics to compare against heterogeneity-robust estimators, and rely on heterogeneity-robust estimators for primary estimation.

### Heterogeneity-Robust Estimators

Construct clean comparisons using never-treated or not-yet-treated controls. Two common choices are:

The Sun–Abraham interaction-weighted regression:

$$Y_{it} = \alpha_i + \lambda_t + \sum_g \sum_{k \neq -1} \delta_{g,k} \mathbf{1}\{G_i = g\} \mathbf{1}\{t - G_i = k\} + \varepsilon_{it}.$$

This estimates cohort-specific paths  $\delta_{g,k}$  (identifying  $\tau(g, g+k)$ ) and aggregates them into  $\hat{\theta}_k = \sum_g \hat{w}_{gk}^{\text{SA}} \hat{\delta}_{g,k}$  with user-specified cohort weights.

The Callaway–Sant’Anna aggregation from group-time effects:

$$\hat{\theta}_k = \sum_g w_{gk}^{\text{CS}} \hat{\tau}(g, g+k),$$

This aggregates group-time effects  $\hat{\tau}(g, g+k)$  into  $\hat{\theta}_k^{\text{CS}} = \sum_g w_{gk}^{\text{CS}} \hat{\tau}(g, g+k)$ . See Chapter 4, Section 4.5 for derivations and properties.

Estimator selection guide.

If your goal is...	Use...	Rationale
Event-time profile with pre-trend diagnostics	Sun–Abraham	Directly estimates $\theta_{g,k}$ for each cohort and event time, with clean pre-treatment coefficients
Flexible aggregation across Callaway–Sant’Anna cohorts		Separates estimation from aggregation and allows custom weighting schemes
Settings where an interactive Borusyak–Jaravel–Spiess fixed effects or low-rank factor structure for untreated outcomes is credible		Imputation-based: estimates counterfactuals for treated observations using untreated data, then computes effects as residuals. It requires enough untreated pre-period data and inherits the assumptions of the imputation model (misspecification bias risk). BJS relies on the outcome model (two-way fixed effects or interactive fixed effects) being correctly specified for untreated potential outcomes and on having a sufficiently long pre-treatment period to estimate factors and loadings.
Quick benchmark	TWFE	Fast but potentially biased, so always compare to robust estimators

Software: R packages `did`, `fixest` (with `sunab()`), `did2s`, and Stata commands `csdid` and `eventstudyinteract`. Compare methods for robustness.

### Estimator Choice Can Affect Results: Why and When to Worry

Different heterogeneity-robust estimators target the same population parameter only under restrictive conditions. In general, they estimate:

$$\hat{\theta}_k^{\text{SA}} = \sum_g \hat{w}_{gk}^{\text{SA}} \hat{\delta}_{g,k}, \quad \hat{\theta}_k^{\text{CS}} = \sum_g w_{gk}^{\text{CS}} \hat{\tau}(g, g+k),$$

where the weights  $w_{gk}$  depend on user choices (cohort-size, uniform, or observation-count). These coincide only when both estimators use the same control group, the same aggregation weights, and treatment effects are homogeneous across cohorts.

Different control groups (never-treated vs. not-yet-treated) identify different parameters if parallel trends holds for one comparison but not the other. Different weighting schemes (e.g.,  $w_{gk}^{\text{SA}}$  vs  $w_{gk}^{\text{CS}}$ ) amplify or dampen heterogeneous cohort-specific effects. Numerical implementation differences (trimming, bandwidth, standard error clustering) can also introduce small discrepancies.

As a rough heuristic, discrepancies much smaller than a standard error are often attributable to weighting and implementation differences, so report both and interpret them cautiously. Larger discrepancies warrant reporting cohort-specific estimates to diagnose whether a small number of cohorts drive the gap. Sign reversals or large gaps are red flags: one or both parallel trends assumptions may fail, so do not average across estimators without investigation.

Re-run both estimators with identical control groups. If discrepancies persist, plot cohort-specific  $\hat{\delta}_{g,k}$  paths. If one cohort shows a different pattern, treatment effect heterogeneity—not estimator failure—explains the gap. Report the heterogeneity transparently rather than forcing a single aggregate.

## Control Set and Weighting Choices

Control set choices—whether to use never-treated units only or to include not-yet-treated units—affect identification and precision. Using never-treated units as controls is straightforward and transparent but requires that never-treated units exist and are comparable to treated units. If all units eventually adopt treatment, never-treated controls are unavailable, and identification must rely on not-yet-treated controls. Not-yet-treated controls are units that have not yet been treated by the period in question but will be treated later. Using not-yet-treated controls increases the effective sample size and can improve precision.

### Stronger Assumption for Not-Yet-Treated Controls

Using not-yet-treated controls requires two distinct parallel trends assumptions. These conditions are the event-study restatement of the staggered parallel-trends assumptions in Section 4.3:

1. PT-Never. For all  $s < t$  and treated cohorts  $g$ ,  $\mathbb{E}[Y_{it}(\infty) - Y_{is}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{is}(\infty) | G_i = \infty]$ . Treated and never-treated units would have evolved in parallel absent treatment. (Assumption 11)
2. PT-Timing. For all  $s < t$  and cohorts  $g < g' < \infty$  with  $t < g'$ ,  $\mathbb{E}[Y_{it}(\infty) - Y_{is}(\infty) | G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{is}(\infty) | G_i = g']$ . Early and late adopters would have evolved in parallel absent treatment.

Using never-treated controls requires only PT-Never. Using not-yet-treated controls requires *both*. At event time  $k$  for cohort  $g$ , not-yet-treated controls are cohorts  $g'$  with  $g' > g + k$ . At large  $k$ , few cohorts remain not-yet-treated, so PT-Timing becomes increasingly restrictive. The precision gain from not-yet-treated controls is largest at short horizons, and the assumption risk is highest at long horizons where only the latest adopters remain as controls.

Compare pre-trends *across treated cohorts* before their respective adoption times. If early adopters exhibit steeper pre-trends than late adopters, PT-Timing is suspect. This is distinct from the standard pre-trend diagnostic (which pools all treated units) and directly probes the timing-based comparison. If late adopters anticipate treatment—adjusting inventory, pricing, or marketing before formal adoption—their pre-treatment outcomes are contaminated. They fail as controls for early adopters even if they would have been parallel absent anticipation. Anticipation thus violates PT-Timing even when PT-Never holds (making not-yet-treated controls invalid).

The Bacon decomposition [Goodman-Bacon, 2021] shows that TWFE implicitly weights three types of  $2 \times 2$  comparisons: treated vs. never-treated, early vs. late adopters, and late vs. early adopters. When PT-Timing fails, these timing-based comparisons identify different parameters, contaminating the aggregate TWFE estimate. Heterogeneity-robust estimators allow you to exclude or separately report timing-based comparisons.

Weighting conventions matter for interpretation. Cohort-size weights give each cohort weight proportional to the number of treated units in that cohort, reflecting the prevalence of the cohort in the treated population. Treated-unit-period weights give each cohort weight proportional to the number of treated unit-periods it contributes at event time  $k$ , reflecting the density of observations. Uniform weights give each cohort equal weight regardless of size. Different weighting schemes produce different  $\theta_k$  estimates when effects are heterogeneous across cohorts. The choice should be pre-specified based on the substantive question (whether large cohorts should dominate the aggregation or whether all cohorts should be weighted equally) and should be reported transparently.

## Implementation Guidance

Implementation in marketing panels requires attention to data structures and diagnostic checks. Thin panels with many units and few periods are common in retail, and event-study windows must be chosen to respect the limited time horizon. Fat panels with few units and many periods enable longer event-time windows but may have few cohorts, limiting the ability to estimate cohort-specific profiles. Unbalanced panels with entry, exit, or missing observations require care to ensure that event-time indicators are well defined and that support is adequate. Pre-specify the event-time window, binning scheme, control set, and weighting convention in a pre-analysis plan and mirror these choices in reporting templates (Section 3.9) so that readers can reconstruct the mapping from design to estimator. Reporting the support (number of units and cohorts by  $k$ ), cohort composition, and estimated weights for each event time provides transparency and enables readers to assess robustness.

The modern event-study framework clarifies that event studies are not merely a graphical tool but a principled estimation strategy grounded in the potential outcomes framework (Chapter 2). By making explicit the target estimands, the comparison groups, and the aggregation schemes, heterogeneity-robust event-study methods produce credible causal estimates that withstand scrutiny and inform strategic decisions with confidence. The next section articulates the identification assumptions that underpin these methods.

## 5.5 Identification: Assumptions and Design Implications

Causal identification of event-time effects  $\theta_k$  requires assumptions about how treated and control units would have evolved in the absence of treatment. This section articulates the identification assumptions for event studies, discusses their testable implications, and clarifies when alternative identification strategies are required.

### Parallel Trends

The parallel trends assumption for event studies asserts that, in the absence of treatment, treated units would have followed the same trajectory as control units in event time. Formally, for all event times  $k$  and for all cohorts  $g$ ,

$$\mathbb{E}[Y_{i,g+k}(\infty) - Y_{i,g-1}(\infty) \mid G_i = g] = \mathbb{E}[Y_{i,g+k}(\infty) - Y_{i,g-1}(\infty) \mid G_i \in \mathcal{C}_g],$$

where  $\mathcal{C}_g$  denotes the valid control group for cohort  $g$  (for example, never-treated units with  $G_i = \infty$  or not-yet-treated units with  $G_i > g + k$ ). Here  $Y_{it}(\infty)$  is the adoption-time shorthand from Chapter 4 for the potential outcome under never treated, consistent with the generic notation  $Y_{it}(0)$  used for binary treatments. This is the event-time restatement of the staggered parallel trends condition from Chapter 4, expressed relative to the baseline period  $g - 1$  and the omitted  $k = -1$  bin. The choice of  $g - 1$  as the baseline corresponds to the convention of omitting  $k = -1$  in the regression; alternative baseline choices (for example, the average of all pre-treatment periods) would modify the formula accordingly but leave the substantive content unchanged.

**Assumption 13 (Parallel Trends in Event Time)** In the absence of treatment, the expected change in outcomes from the baseline period  $g - 1$  to any future period  $g + k$  is the same for the treated cohort  $g$  and the control group. Formally, for all treated cohorts  $g$  and event times  $k$ :

$$\mathbb{E}[Y_{i,g+k}(\infty) - Y_{i,g-1}(\infty) \mid G_i = g] = \mathbb{E}[Y_{i,g+k}(\infty) - Y_{i,g-1}(\infty) \mid G_i \in \mathcal{C}_g],$$

where  $\mathcal{C}_g$  denotes the valid control group for cohort  $g$  (for example, never-treated or not-yet-treated units).

Under Assumption 13 and no anticipation, the regression-based event-study coefficients  $\theta_k$  identify cohort-weighted averages of the causal event-time effects defined in Section 5.2.

Conditional parallel trends weakens this assumption by allowing for covariate-dependent trends. After conditioning on covariates  $X_i$ , treated and control units follow similar event-time trajectories. Here  $X_i$  denotes (possibly high-dimensional) pre-treatment or time-invariant covariates; when we allow for time-varying controls, we return to the generic notation  $X_{it}$ . This is a weaker assumption than unconditional parallel trends and is often more plausible in observational settings where treatment assignment depends on observables. Diagnostic checks for conditional parallel trends include covariate balance assessments (Chapter 17) and propensity score overlap checks. If covariate imbalances are large, propensity score weighting, regression adjustment, or matching can restore balance. This makes conditional parallel trends more credible.

These adjustments also change the implicit estimand: the resulting  $\theta_k$  trace the average dynamic effect for a covariate-weighted population defined by the chosen weighting or matching scheme, rather than the unweighted sample.

#### Correct Specification Required: The Omitted Confounder Problem

Conditional parallel trends requires that, for all calendar periods  $t$ ,  $X_i$  blocks all backdoor paths between treatment timing  $G_i$  and untreated potential outcomes  $Y_{it}(0)$ . Formally, we need:

$$\mathbb{E}[Y_{it}(0) | X_i, G_i = g] = \mathbb{E}[Y_{it}(0) | X_i, G_i \in \mathcal{C}_g].$$

If an omitted confounder  $U_i$  affects both treatment timing and outcomes—for example, if stores with more experienced managers adopt earlier and also have stronger trend growth—the assumption fails even after conditioning on  $X_i$ . No test can verify that the conditioning set is complete.

Sensitivity analysis asks how large an omitted confounder would need to be to overturn conclusions. The coefficient stability approach [Oster, 2019] compares how  $\hat{\theta}_k$  changes when covariates are added. If adding  $X_i$  barely moves  $\hat{\theta}_k$ , remaining omitted variables are less likely to matter. The partial  $R^2$  approach [Cinelli and Hazlett, 2020] bounds the bias from omitted confounders based on their plausible correlation with treatment and outcomes. Report these bounds alongside point estimates to communicate robustness to omitted variable bias.

## No Anticipation and Pre-Trend Diagnostics

This subsection elaborates the no-anticipation assumption introduced in Chapter 2 and used in Chapter 4, now expressed in event-time language. The no anticipation assumption asserts that potential outcomes in period  $t$  do not depend on treatment assignments in future periods  $s > t$ . No anticipation means that for all treatment histories  $d_i^t$  that coincide on pre-treatment periods, the corresponding potential outcomes agree when  $t < G_i$ :  $Y_{it}(d_i^t) = Y_{it}(0)$  for all  $t < G_i$ . This implies that pre-treatment coefficients  $\theta_k$  for  $k < 0$  should be zero if treated units do not adjust behaviour in advance of treatment. Non-zero pre-treatment coefficients signal anticipation or pre-existing trends that violate parallel trends.

Diagnosing anticipation using pre-treatment leads is a central diagnostic in event studies. Estimate the event-study regression including multiple pre-treatment event times ( $k < -1$ ), and conduct a joint test that the pre-treatment coefficients are zero using a cluster-robust Wald or F-test. A naive F-test built from homoskedastic OLS standard errors is inappropriate when errors are clustered. If pre-treatment coefficients are statistically indistinguishable from zero, this is consistent with parallel trends and no anticipation and can increase their plausibility, but it does not validate either assumption. If pre-treatment coefficients are non-zero, this indicates either anticipation (units respond to expected future treatment) or pre-existing differential trends (treated and control units were diverging even before treatment).

### Anticipation vs. Pre-Trends: A Fundamental Ambiguity

Both anticipation and pre-trends produce non-zero pre-treatment coefficients, but they have different implications for identification.

Anticipation is a causal effect of expected treatment: units change behaviour because they know treatment is coming. Pre-trends indicate that treated and control units were already diverging, violating parallel trends. With anticipation, the treatment effect is real but starts before formal implementation. With pre-trends, the estimated effect is biased.

If units anticipate treatment  $\ell$  periods in advance, redefine the baseline period to  $k = -\ell - 1$  rather than  $k = -1$ . The event-study regression becomes:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k=-\ell-1} \theta_k \cdot \mathbf{1}\{t - G_i = k\} + \varepsilon_{it}.$$

The coefficients  $\theta_{-\ell}, \dots, \theta_{-1}$  now capture anticipation effects, and  $\theta_0, \theta_1, \dots$  capture post-implementation effects. The total effect of the intervention includes both.

Use institutional knowledge to interpret pre-treatment coefficients. If the rollout was publicly announced  $\ell$  periods before implementation, anticipation is likely, so model it explicitly. If treatment timing was determined internally without prior notice, pre-trends are more likely, and parallel trends may be violated.

When the source of pre-treatment coefficients is ambiguous, the Rambachan–Roth framework [Rambachan and Roth, 2023] provides honest confidence intervals under bounded trend violations. See Section 5.8 for implementation.

Pre-trend diagnostics have limited power, especially with few pre-periods [Roth, 2022]. Absence of detectable pre-trends increases plausibility but does not prove parallel trends. Combine statistical checks with institutional knowledge, balance checks, and sensitivity analyses (varying control sets, windows, and specifications).

### SUTVA and Spillovers

The stable unit treatment value assumption (SUTVA), also introduced in Chapter 2, asserts that potential outcomes for unit  $i$  do not depend on the treatment assignments of other units. SUTVA is routinely violated in marketing through spillovers, network effects, and competitive interactions. In the event-time context, spillovers can distort both pre-treatment and post-treatment coefficients. If treated units spill over to control units before treatment (for example, if an impending loyalty programme generates word-of-mouth that reaches control customers), pre-treatment coefficients may be non-zero not because of anticipation by treated units but because of contamination of control units. If spillovers occur after treatment, post-treatment coefficients conflate direct effects with spillover effects. When spillovers are present but ignored, the estimated  $\theta_k$  should

be interpreted as an equilibrium response to the overall treatment pattern in the market, not as a pure within-unit effect of  $D_{it}$ .

Design-based solutions to spillovers include defining clusters that internalise spillovers, creating buffer zones that separate treated and control units, and estimating explicit spillover models (Chapter 11). In the event-time context, you can include spatial or network leads and lags, such as terms of the form  $\sum_{j \neq i} a_{ij} \mathbf{1}\{t - G_j = k\}$ , to estimate how treatment in neighbouring units affects outcomes for unit  $i$  at different event times. These extensions are discussed in Section 5.9.

## Factor Structure Relaxations

Factor structure relaxations provide an alternative when parallel trends in levels is implausible but units face common shocks with differential exposure. Interactive fixed effects and matrix methods (Chapters 8, 9) decompose untreated outcomes into latent factors and unit-specific loadings. For never-treated potential outcomes  $Y_{it}(\infty)$  we posit

$$Y_{it}(\infty) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

and interpret  $\theta_k$  as deviations from this factor-driven trajectory. In event time, the key nuance is stability: factors and loadings must remain stable (constant loadings, no structural breaks in factor dynamics) over the event-time window. In practice, modest smooth drifts in  $\lambda_{ir}$  or  $f_{tr}$  over long horizons can be tolerated if the event window is short; what undermines identification are structural breaks or rapid shifts that coincide with treatment. If treatment coincides with a shift in factor structure,  $\theta_k$  conflates structural change with treatment response, and factor-based counterfactuals are invalid. Estimation and diagnostics live in Chapters 8 and 9; here we only flag the event-time requirement.

When using factor models, ensure event-time stability. If treatment coincides with a shift in factor structure, imputed counterfactuals are invalid and  $\theta_k$  conflates dynamics with structural change. For example, if a platform launches a loyalty programme at the same time it changes its recommendation algorithm, the factor structure governing customer behaviour may shift, and factor-based counterfactuals will be invalid. Use fit and stability checks from Chapters 8 and 9, and conduct sensitivity analysis by varying the number of factors or the estimation window.

## Diagnostic Best Practices

Limits of pre-trend diagnostics and good practice for interpretation merit emphasis. Pre-trend diagnostics have finite power, meaning that they can fail to detect pre-trends even when they exist. As a rough heuristic, pre-trend diagnostics can typically detect trends of magnitude comparable to the estimated post-treatment effects with reasonable power, but they have low power to detect smaller trends [Roth, 2022]. The actual power depends on the variance of the outcome, the sample size, the number of pre-treatment periods, and

the clustering structure; simulation-based power analysis (Section 3.7) can provide more precise guidance for a specific setting. When planning studies, use simulation-based power analysis (Section 3.7) to choose the number of pre-treatment periods and cluster sizes needed to detect economically meaningful pre-trends with high probability. This means that a non-significant pre-trend diagnostic provides reassurance only if the magnitudes of the pre-treatment coefficients are small relative to the post-treatment coefficients. If pre-treatment coefficients are large in magnitude but statistically insignificant (due to wide confidence intervals), the evidence for parallel trends is weak.

Good practice combines pre-trend diagnostics with multiple diagnostic approaches. Plot the full event-time profile to visually inspect for pre-trends. Conduct placebo tests using pre-treatment data only (for example, pretend that period  $t^* < G_i$  is the treatment date and re-estimate  $\theta_k$ ). Under parallel trends, the resulting placebo profile should fluctuate around zero. Check covariate balance. Estimate cohort-specific profiles to assess whether pre-trends are consistent across cohorts. Conduct sensitivity analyses that vary the pre-treatment window or the control group. For an approach that avoids committing to a specific identification strategy when close comparison groups exist, see Section 20.5. The goal is not to prove that parallel trends holds—which is impossible—but to build a cumulative case that parallel trends is plausible and that conclusions are robust to plausible violations.

State assumptions, show diagnostics, and report sensitivity. Use pre-treatment leads for anticipation checks, exposure models for SUTVA risks, and factor-fit diagnostics when relaxing parallel trends. The next section covers best practices for graphical presentation.

## 5.6 Graphical Presentation and Interpretation

Event-study plots are the primary vehicle for communicating causal narratives in panel data. A well-designed plot conveys the dynamic evolution of treatment effects, the credibility of the identification strategy, and the policy relevance of the findings at a glance. This section provides best practices for plotting event-time effects and for interpreting the resulting graphs in marketing applications.

### Plot Design

The basic event-study plot displays point estimates  $\hat{\theta}_k$  of the event-time effects  $\theta_k$  (defined in Section 5.2) on the vertical axis and event time  $k$  on the horizontal axis, with confidence intervals (typically 95%) around each point estimate. A vertical line at  $k = 0$  marks the treatment period, dividing the plot into pre-treatment (left of the line) and post-treatment (right of the line) regions. The omitted bin (typically  $k = -1$ ) is marked explicitly, often with a hollow marker or a different colour, to remind readers that this event time is normalised to zero by construction.

Point estimates should be connected by lines to guide the eye and to emphasise the dynamic trajectory. However, the lines should not be interpreted as interpolating the effect at non-integer event times unless the underlying estimator supports such interpolation (for example, if effects are modelled parametrically). Confidence intervals should be displayed as shaded bands or error bars, making it clear whether effects at different event times are statistically distinguishable from zero or from each other. Standard practice is to display pointwise 95% confidence intervals, which test whether each individual  $\theta_k$  differs from zero. However, pointwise intervals do not account for multiplicity: when examining many event times simultaneously, the probability of at least one false rejection exceeds 5%. For example, with 10 independent tests at the 5% level, the probability of at least one false rejection is  $1 - 0.95^{10} \approx 40\%$ . Event-time coefficients are typically positively correlated, which reduces the false rejection probability below this upper bound, but the multiplicity problem remains.

### Joint and Uniform Confidence Bands

Pointwise 95% confidence intervals test whether each individual  $\theta_k$  differs from zero, but they do not control the family-wise error rate across all event times. When testing whether *any* pre-treatment coefficient deviates from zero, or whether *any* post-treatment effect is significant, use joint inference. The sup-t (supremum-t) critical value is the  $(1 - \alpha)$  quantile of  $\max_k |t_k|$  under the joint null, accounting for correlation across event times [Montiel Olea and Plagborg-Møller, 2019]. If  $\max_{k \in \mathcal{K}} |\hat{\theta}_k / \text{SE}_k| < c_{\text{sup-t}}^{1-\alpha}$  for a chosen index set  $\mathcal{K}$  (for example, all pre-treatment  $k < 0$ ), then with probability at least  $1 - \alpha$  you would not reject  $H_0 : \theta_k = 0$  for any  $k \in \mathcal{K}$ . Sup-t bands are wider than pointwise intervals but guarantee that the probability of *any* false rejection is at most  $\alpha$ .

Bonferroni correction, Holm step-down, and bootstrap-based joint inference are available. See Chapter 16 for details on when to use each.

R packages such as `fixest` implement sup-t bands (for example, via the `iplot` method with appropriate options). In Stata, the `boottest` command can be used for bootstrap-based joint inference. For implementation details and code examples, see Chapter 16 and the software documentation of your chosen estimator.

If multiple event-time profiles are displayed (for example, cohort-specific profiles or estimates from different estimators), they should be distinguished by line style or colour, with a clear legend.

Marking  $k = 0$  with a vertical line or shaded region highlights the treatment threshold and draws attention to the immediate effect  $\theta_0$  and to the contrast between pre-treatment and post-treatment dynamics. If the treatment was phased in over multiple periods or if there was a lag between adoption and full implementation, the plot should indicate the phase-in period explicitly, and interpretation should account for the gradual ramp-up.

Showing support per event time  $k$  ensures that readers understand where the data are rich and where estimates are based on few observations. A support table or a second panel in the figure can display the number of observations, the number of cohorts contributing, or the effective sample size (accounting for weights and clustering) for each  $k$ . When using weights  $w_{it}$  or clustering at higher levels (for example, markets), raw counts by  $k$  can be misleading. Where feasible, report both the number of clusters contributing to each  $k$  and summary measures of the weights (for example, total or average  $w_{it}$ ), since inference is driven by clusters and effective weighted information, not just the number of observations. If support is thin at extreme event times, confidence intervals will be wide. Conclusions about long-run effects should then be tempered. Binning decisions (aggregating extreme event times) should be indicated in the plot, for example by using a different marker style for binned event times or by annotating the axis labels.

Overlaying cohort-specific paths  $\theta_{g,k}$  on the main event-time plot provides a diagnostic for heterogeneity. If cohort-specific paths are similar, pooling them into a single  $\theta_k$  is defensible. If paths diverge, the aggregated  $\theta_k$  obscures meaningful variation, and conclusions should acknowledge the heterogeneity. When cohort-specific paths differ sharply, the pooled  $\theta_k$  can be a non-convex or misleading average of heterogeneous effects (see Chapter 4). In such cases you should report cohort-specific profiles or alternative estimands alongside the pooled curve. Cohort-specific paths can be displayed as lighter lines or with greater transparency, with the

aggregated path displayed as a bold line, making the overall trajectory salient while retaining the diagnostic information.

## Interpretation Patterns

### Visual Inspection Can Be Misleading: Cognitive Traps and Formal Tests

Humans are prone to seeing patterns in noise. Three cognitive traps are particularly dangerous in event-study interpretation. *Confirmation bias* leads researchers expecting an effect to see “clear dynamics” in noisy post-treatment coefficients, whereas those sceptical of the identification will see “obvious pre-trends” in random pre-treatment variation. *Scale manipulation* can make flat profiles look trending or vice versa, so you should always check whether apparent slopes are economically meaningful. *Ashenfelter’s dip*—a temporary dip just before treatment (for example,  $\theta_{-1}$  below  $\theta_{-2}$ )—can be real (anticipatory delay) or noise, and misreading this inflates the apparent treatment effect.

Accompany visual inspection with:

1. *Joint pre-trend diagnostic.*  $H_0 : \theta_{-K} = \dots = \theta_{-2} = 0$ . Use a Wald test with cluster-robust variance or sup-t bands.
2. *Linear pre-trend diagnostic.* As a heuristic, you can regress the estimated pre-treatment coefficients  $\hat{\theta}_k$  on event time  $k$  for  $k < 0$  and test whether the slope differs from zero. Because this is a two-step procedure that ignores estimation error in  $\hat{\theta}_k$ , treat it as a descriptive diagnostic rather than a primary inferential test. If you want inference, define the slope restriction in a single regression with the appropriate covariance structure, or use a joint test of leads.
3. *Monotonicity or decay restriction.* For post-treatment dynamics, you can test monotonic decay or ramp-up by imposing linear inequality constraints on  $(\theta_k)_{k>0}$  in a Wald test (for example,  $\theta_1 \geq \theta_0, \theta_2 \geq \theta_1, \dots$ ). Avoid interpreting multiple pairwise t-tests as a monotonicity test.

State hypotheses about pre-trends and dynamics before seeing results. Pre-registration and a pre-defined diagnostic plan (Section 3.9) are the main antidotes to seeing patterns in noise.

Substantive interpretation for marketing requires translating the event-time profile into business insights. Ramp-up patterns, where post-treatment effects grow over event time ( $\theta_k$  increases with  $k$ ), indicate that the intervention takes time to produce its full impact. A loyalty programme may exhibit ramp-up as customers enrol, accumulate points, and develop purchasing habits. Advertising may exhibit ramp-up as brand awareness builds. Platform entry may exhibit ramp-up as network effects accumulate. Ramp-up patterns suggest that short-run evaluations understate long-run benefits and that patience is required for interventions to mature. However, apparent ramp-up can also reflect composition bias: if early adopters with large effects dominate at large  $k$  (because late adopters have not yet reached that horizon), the profile may show increasing effects even if within-cohort effects are constant. Figure 5.2 illustrates how cohort composition varies with  $k$ ; when

early adopters dominate large  $k$ , apparent ramp-up can emerge purely from who remains at risk rather than true within-unit dynamics. Cohort-specific profiles (Section 5.3) diagnose this issue.

Decay patterns, where post-treatment effects diminish over event time ( $\theta_k$  decreases with  $k$ ), indicate that the intervention produces transient effects that erode over time. Promotional pricing may exhibit decay as customers stockpile products during the promotion and reduce buying afterward. Advertising may exhibit decay if awareness fades or if competitors respond. Decay patterns suggest that sustained investment is required to maintain effects and that one-time interventions have limited long-run impact.

Persistent effects, where post-treatment effects remain at a stable level over event time ( $\theta_k$  is roughly constant for  $k \geq K$ ), indicate that the intervention produces a permanent shift in outcomes. A platform entry that captures market share may exhibit persistent effects if customers do not return to incumbents after trying the new platform. A loyalty programme that changes purchasing habits may exhibit persistent effects if habits persist even if the programme is discontinued. Persistent effect patterns suggest that the intervention has lasting effects and that the return on investment is high.

#### Identification Challenge for Persistent Effects

Persistent effects are difficult to distinguish from very slow decay. If the post-treatment window is short, what appears to be a permanent effect may actually be slow decay that would eventually return to zero.

Test whether late-horizon coefficients are jointly flat:

$$H_0 : \theta_K = \theta_{K+1} = \cdots = \theta_{K+L} \quad (\text{persistence})$$

versus  $H_1$ :  $\theta_k$  declining with  $k$ . A joint F-test on  $(\theta_K, \dots, \theta_{K+L})$  with cluster-robust variance assesses whether the profile has stabilised.

Use the following parametric model only when extrapolation beyond the observed event window is essential for the decision, and always report it side by side with the non-parametric event-study profile. For extrapolation, fit:

$$\theta_k = \theta_\infty + \delta \cdot \rho^k, \quad k > 0,$$

where  $\theta_\infty$  is the long-run persistent level,  $\delta$  is the initial transitory component, and  $\rho \in (0, 1)$  is the decay rate. The half-life of decay is  $k^* = \log(0.5)/\log(\rho)$ . If  $\hat{\theta}_\infty$  is statistically distinguishable from zero, the data are consistent with a persistent component. If not, effects are more consistent with a transitory response, although wide intervals can also reflect limited power.

At long horizons, only early adopters contribute to  $\theta_k$ . If early adopters differ systematically (for example, larger stores with stronger effects), apparent persistence may reflect composition rather than true within-unit dynamics. Check by comparing late-horizon estimates from early adopters alone against shorter-horizon estimates pooled across cohorts.

Report both the non-parametric event-time profile and the fitted decay parameters  $(\hat{\theta}_\infty, \hat{\rho}, \hat{k}^*)$ . See Section 5.10 for formulas.

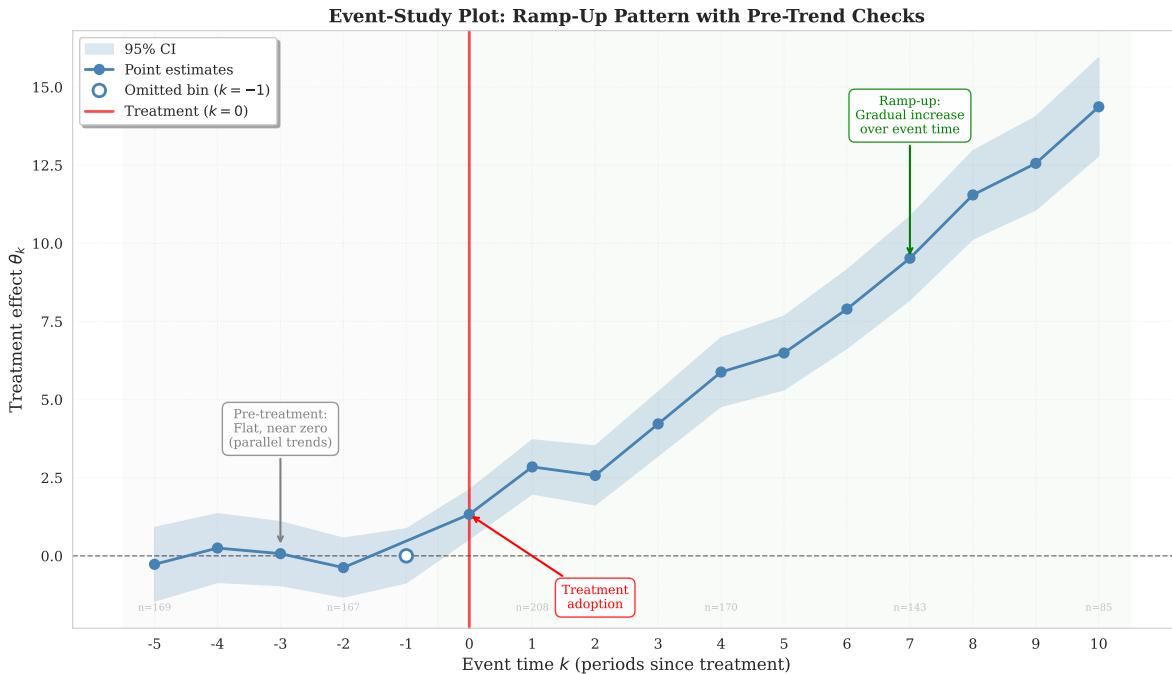
Distinguishing transitory from persistent effects is central to marketing decisions. Transitory effects, where  $\theta_k$  returns to near zero after a few periods, indicate that the intervention shifts demand forward in time but does not create new demand. Persistent effects, where  $\theta_k$  remains elevated indefinitely, indicate that the intervention changes the level or trajectory of outcomes permanently. The distinction matters for go/no-go decisions: transitory effects may not justify the cost of the intervention, while persistent effects generate long-run value that justifies upfront investment.

For quantitative metrics that translate these patterns into decision inputs—including formulas for ramp-up rate, time-to-maturity, effect multiplier, decay half-life, and cumulative effects—see Section 5.10.

## Linking Plots to Business Decisions

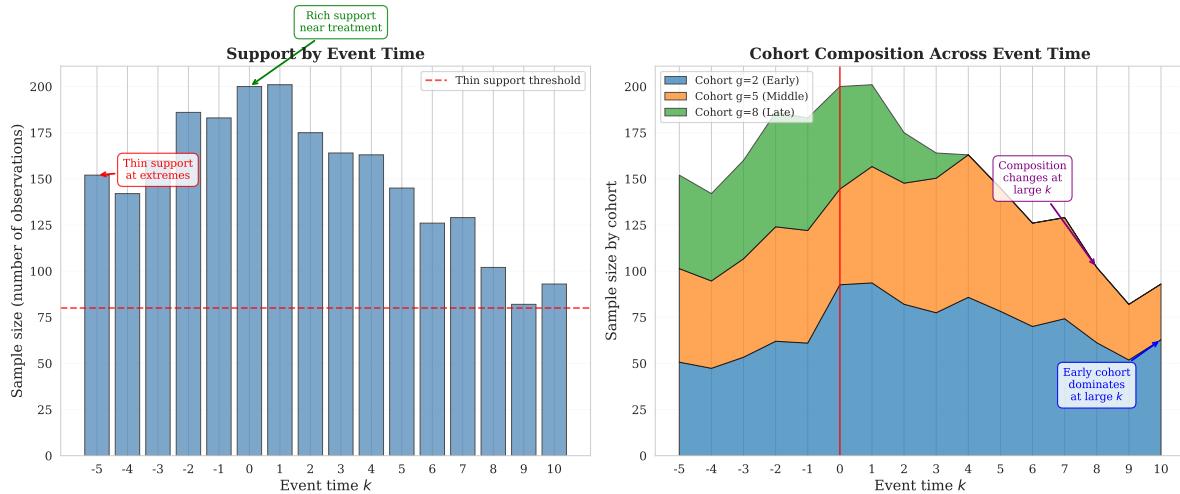
Event-study plots should be accompanied by narrative interpretation that explains the pattern, relates it to the substantive context, and draws implications for decisions. The plot provides transparent evidence on dynamics and credibility, but the narrative provides meaning and actionability. A retailer evaluating a loyalty programme uses the event-time profile to assess ramp-up speed and long-run steady-state effects. A brand estimating advertising ROI examines cumulative effects such as  $\sum_{k=0}^K \theta_k$  and, when appropriate, the long-run multiplier LRM =  $\sum_{k=0}^K \theta_k / \theta_0$  defined in Section 5.2. A platform assessing market entry tracks immediate effects and growth trajectories.

The key is to translate the visual pattern into the metrics that matter for the decision at hand. All such ROI, payback, and CLV calculations are functions of the estimated  $\theta_k$  profile and therefore inherit the identification assumptions and diagnostics discussed in Sections 5.5 and 5.6. The plot plus cumulative effect, LRM where appropriate (Section 5.2), and half-life should be mapped into ROI, payback, and CLV using the frameworks from Chapter 4. Section 5.10 provides formulas for computing these metrics directly from the estimated  $\theta_k$  profile. Together, the plot and the quantitative metrics make event studies a powerful tool for communicating causal findings to technical and non-technical audiences alike. The next section addresses inference procedures for event-time estimates.



**Fig. 5.1** Event-Study Plot with Omitted Category and Confidence Intervals. The implied cumulative effect  $\sum_{k=0}^{10} \theta_k$  and long-run multiplier LRM (where appropriate) can be read directly from this event-time profile.

The figure displays event-time treatment effects  $\theta_k$  from  $k = -5$  (five periods before treatment) to  $k = 10$  (ten periods after treatment). Pre-treatment coefficients (left of the vertical red line at  $k = 0$ ) are near zero and flat, which is consistent with parallel trends and no anticipation and increases their plausibility, although the diagnostic has limited power. The omitted reference period at  $k = -1$  is marked with a hollow circle. Post-treatment coefficients (right of the red line) show a ramp-up pattern, with effects increasing gradually over event time, consistent with habit formation or network effects. Shaded bands represent 95% confidence intervals. Sample sizes are annotated at the bottom, showing rich support near treatment and thinner support at extreme event times.



**Fig. 5.2** Support by Event Time  $k$  and Cohort Composition Across  $k$ . While the left panel reports observation counts, inference is typically driven by the number of clusters contributing to each  $k$ . When clusters are few at extreme event times, standard errors will be particularly wide.

The left panel shows sample size (number of observations) by event time  $k$ . Support is richest near the treatment period ( $k = 0$ ) and thinner at extreme event times ( $k \ll 0$  or  $k \gg 0$ ), where fewer cohorts contribute observations. Bars in coral indicate thin support (below 80 observations). The right panel displays cohort composition as a stacked area chart, showing which cohorts ( $g=2, 5, 8$ ) contribute to each event time. Early-adopting cohorts ( $g=2$ , blue) dominate at large positive  $k$ , while all cohorts contribute at  $k = 0$ . Composition changes across event time can confound interpretation if treatment effects are heterogeneous across cohorts.

## 5.7 Inference

Valid inference for event-time effects requires the same attention to clustering, small-sample corrections, and multiplicity as in Chapter 4 (Section 4.7) and Chapter 16. Here we briefly emphasise the implications that are specific to event-time paths.

### Clustering

Clustering is the default approach to accounting for within-unit correlation over time. Throughout this chapter we treat the unit  $i$  (store, customer, market) as the primary clustering dimension. When clustering by unit, the independent sampling unit is  $i$  and the number of clusters is  $N$ . Outcomes for the same unit in different periods are correlated due to persistent unobservables, autocorrelated shocks, or dynamic feedback. Clustering standard errors by unit allows for arbitrary correlation within units while maintaining independence across units. The cluster-robust variance estimator is valid asymptotically as the number of independent units grows, provided that errors are uncorrelated across units. This independence assumption is often violated in marketing settings: regional shocks affect multiple stores, competitive responses propagate across firms, and platform algorithm changes affect all users simultaneously. When cross-unit correlation is plausible, two-way clustering or other approaches are required.

Two-way clustering, by both unit and time, accounts for correlation within units over time and across units within periods. If all stores in a region are affected by a regional shock, errors are correlated across stores within a period. If macroeconomic conditions or platform algorithm changes affect all units in a given period, errors are correlated across units. Two-way clustering captures both sources of correlation, producing standard errors that are valid under weak assumptions. The cost is substantially larger standard errors and greater computational complexity, though modern software implements two-way clustering efficiently. The increase in standard errors reflects genuine uncertainty that one-way clustering ignores. It is not a defect of the method. When using two-way clustering, small-sample concerns apply to both clustering dimensions. If either the number of units or the number of time periods is small (for example, fewer than 20), treat inference with caution and consider bootstrap methods.

### Small-Sample Corrections

Small-sample concerns arise when the number of independent clusters is modest. In most event-study applications the clustering unit is  $i$  and the cluster count is  $N$ , but in some designs the clustering unit is a higher-level cluster  $c$  and the cluster count is  $G$  (for example, markets in a geo-experiment). The cluster-count thresholds from Section 4.7 apply here as rules of thumb. With fewer than about 20 clusters, use wild cluster bootstrap or randomisation inference. With 20–50 clusters, compare asymptotic to bootstrap results. With 50

or more clusters, asymptotic cluster-robust standard errors are generally reliable. See Chapter 16 for detailed algorithms and examples.

## Randomisation Inference

Randomisation inference and permutation tests offer design-based alternatives that do not rely on parametric assumptions about error distributions. Under the sharp null hypothesis of no effect for any unit at any event time, the observed treatment assignment is just one of many possible assignments that could have been drawn from the randomisation protocol. By recomputing the test statistic for all possible assignments (or a large random sample of them), we generate the exact null distribution of the test statistic under the assumed randomisation protocol. We then compare the observed statistic to this distribution to compute an exact p-value. Randomisation inference is particularly compelling in experimental settings where the randomisation protocol is known and where the goal is to conduct inference that respects the design. In event-study applications, natural randomisation-inference test statistics include (i) the cumulative effect  $\sum_{k=0}^K \hat{\theta}_k$ , (ii) the maximum absolute t-statistic across a set of event times (for example,  $\max_{k \in K} |\hat{\theta}_k|/\text{SE}_k$  for post-treatment  $k$ ), or (iii) a quadratic form in the vector of pre-treatment coefficients to diagnose pre-trends. These choices align the randomisation test with the substantive hypothesis about dynamics.

### Limitation: Randomisation Inference Requires Known Assignment

Randomisation inference requires knowledge of the randomisation protocol, which is often unavailable in observational settings. When treatment assignment is not randomised, or when the assignment mechanism is unknown or only partially specified, randomisation inference is generally not applicable. In such cases, cluster-robust standard errors or bootstrap methods remain the default, with the caveat that they rely on asymptotic approximations and modelling assumptions.

## Multiple Testing

Multiple testing arises because event studies estimate effects for many event times  $k$ , and testing whether  $\theta_k \neq 0$  for each  $k$  involves a family of related hypothesis tests. Bonferroni, FDR, and Romano–Wolf adjustments are available. Methods like Bonferroni and Romano–Wolf control the family-wise error rate, while FDR methods control the expected proportion of false discoveries among rejections. See Section 5.6 for the joint-bands discussion and Chapter 16 for implementation details.

Families of hypotheses should be defined ex ante to guide multiplicity adjustments. One natural family is the set of pre-treatment coefficients  $\{\theta_k : k < 0\}$ , which jointly test for pre-trends and anticipation. A joint Wald test with cluster-robust standard errors that all pre-treatment coefficients are zero provides a single hypothesis test at the desired significance level (a standard F-test assumes homoskedasticity and is inappropriate when errors are clustered). This avoids the need for multiplicity adjustment across individual

pre-treatment coefficients. The joint Wald test answers the question "Are all pre-treatment effects zero?" rather than "Which specific pre-treatment effect is non-zero?". Another family is the set of post-treatment coefficients  $\{\theta_k : k \geq 0\}$ , which jointly test for any treatment effect. If the goal is to test whether the intervention has any effect at any post-treatment event time, a joint test provides a single hypothesis test. If the goal is to estimate and report effects at each post-treatment event time, multiplicity adjustments (Bonferroni, FDR, Romano–Wolf) may be appropriate. Many researchers report unadjusted p-values and confidence intervals for individual event times and rely on joint pre-trend tests to establish credibility.

## Sensitivity to Parallel Trends Violations

The inference procedures above assume parallel trends holds. When this assumption is uncertain, Rambachan–Roth bounds [Rambachan and Roth, 2023] provide a sensitivity-analysis overlay on top of standard inference. The approach specifies a class of plausible violations and computes bounds on the treatment effect that are valid under all violations in the specified class. The key parameter  $\bar{M}$  controls how much the post-treatment trend can differ from the pre-treatment trend. In practice,  $\bar{M}$  is calibrated from the estimated pre-treatment coefficients  $\hat{\theta}_k$  (for example, by bounding how much steeper post-treatment trends could be, relative to pre-treatment trends, without contradicting the pre-period data). Implementation via the `honestDiD` package and design of the violation class are covered in Chapter 17, which also provides concrete calibration strategies and examples.

## Practical Guidance

For marketing applications, the default is to cluster by unit. When cross-unit correlation is plausible—for example, in geo-experiments or when regional shocks are present—two-way clustering is appropriate. Small-sample corrections should be applied when the cluster count is modest, and multiplicity adjustments are warranted when testing many event-time coefficients. As a heuristic, treat fewer than about 20 clusters as a danger zone where asymptotic approximations can be unreliable. With 20–50 clusters, compare asymptotic to bootstrap results. With 50 or more clusters, asymptotic cluster-robust standard errors are often reliable.

Pre-specifying the primary estimand (overall ATT, a cumulative effect, or specific event-time coefficients of substantive interest) and distinguishing primary from exploratory analyses reduces the multiplicity burden. If the primary estimand is the cumulative effect and individual event-time coefficients are exploratory, then only the cumulative effect test needs to control type I error at the nominal level. In many marketing applications, the primary estimand is a function of the event-time profile, such as the cumulative effect  $\sum_{k=0}^K \theta_k$  or the long-run multiplier LRM defined in Section 5.2. Declaring these functionals as primary *ex ante* clarifies which tests must be multiplicity-controlled.

Transparency about inferential choices and robustness checks using alternative methods build confidence in conclusions. Report both unadjusted and adjusted p-values or confidence intervals. Document the families of hypotheses tested. Discuss the trade-offs between controlling family-wise error rate and power.

#### Inference Choices Can Affect Conclusions

Different clustering choices, small-sample corrections, or multiplicity adjustments can lead to different conclusions about statistical significance. An effect that is significant with one-way clustering may become insignificant with two-way clustering. An effect that is significant without multiplicity adjustment may become insignificant after Bonferroni correction. Sensitivity analysis across inferential choices—reporting results under multiple specifications—provides evidence on the robustness of conclusions. If conclusions are sensitive to inferential choices, this should be acknowledged rather than hidden.

This ensures that readers understand the evidential standard applied and can assess whether alternative inferential choices would change conclusions. The next section develops diagnostic workflows for event-study analyses.

## 5.8 Diagnostics

Credible event-study analysis requires rigorous diagnostics that assess the plausibility of identification assumptions, the sensitivity of conclusions to modelling choices, and the influence of individual observations or event times. This section applies the general diagnostic framework of Chapter 17 to event-study designs, focusing on pre-trends, support by event time, and specification sensitivity specific to  $\theta_k$  profiles.

### Pre-Trend Checks

Pre-trend and anticipation checks using leads are the most important diagnostic. We assess the joint null hypothesis that all pre-treatment coefficients (except the reference) are zero:

$$H_0 : \theta_k = 0 \quad \forall k \in \{k_{\min}, \dots, -2\}.$$

Estimate the event-study regression including multiple pre-treatment event times, and assess this hypothesis using a Wald test with cluster-robust standard errors. A standard F-test assumes homoskedasticity and is inappropriate when errors are clustered. Plot the pre-treatment coefficients with confidence intervals. If we fail to reject  $H_0$  and the coefficients are visibly flat or near zero, this is consistent with parallel trends and can increase its plausibility. Rejection of  $H_0$ , or systematic trends approaching  $k = 0$ , is evidence against the identifying assumptions (anticipation or differential trends). These diagnostics have limited power, so a non-rejection does not validate the assumptions.

Pre-trend diagnostics have limited power, especially with few pre-periods [Roth, 2022]. They reliably detect violations comparable in size to post-treatment effects but can miss smaller trends. Section 5.5 and Chapter 17 provide guidance on interpreting such diagnostics. Here we focus on how to integrate them into an event-study workflow. Supplement statistical checks with visual inspection, since a plot showing flat, near-zero pre-treatment coefficients is more convincing than large but imprecisely estimated coefficients.

Placebo-in-time tests apply the event-study logic to pre-treatment periods only, treating an earlier period as if it were the treatment period. For example, if treatment begins in period five and data are available from period one, a placebo test might treat period three as the "treatment" period and estimate event-time effects using periods one and two as the baseline. If the placebo event-study shows non-zero "effects," this indicates pre-trends or other violations of parallel trends. If the placebo event-study shows near-zero effects, this is consistent with parallel trends. Like pre-trend diagnostics, placebo-in-time checks have limited power. A null result provides supportive but not definitive evidence. Formally, placebo-in-time tests examine whether an estimated path of "effects" centred on a fake treatment date fluctuates around zero. Systematic deviations indicate violations of the parallel-trends restriction in the pre-treatment window.

## Sensitivity to Parallel Trends Violations

When pre-trend diagnostics are reassuring but uncertainty about parallel trends remains, sensitivity analysis using Rambachan–Roth bounds provides a formal diagnostic for how far violations of parallel trends would have to go in order to overturn your conclusions. Rambachan–Roth bounds [Rambachan and Roth, 2023] specify a class of plausible violations of parallel trends and compute bounds on the treatment effect that are valid under all violations in the specified class. The key parameter  $\bar{M}$  controls how much the post-treatment trend can differ from the pre-treatment trend. If conclusions are robust to plausible violations (that is, if the bounds exclude zero even under the worst-case violation), the analysis is credible. If conclusions are sensitive (that is, if the bounds include zero under mild violations), the analyst should acknowledge the uncertainty.

Use these bounds as a robustness layer on top of your main estimates, not as a substitute for careful design and diagnostics. Implementation via the `honestDiD` package in R is straightforward. Chapter 17 shows how to choose the violation class and  $\bar{M}$  using the pre-treatment profile.

## Support and Sensitivity Analyses

Support and overlap checks by event time  $k$  assess whether estimates at extreme event times are based on adequate data. Plot the sample size, the number of cohorts contributing, or the effective sample size (accounting for weights and clustering). For example, if you use observation weights  $w_{it}$ , also report the number of clusters  $c$  (total  $G$ ) contributing at each  $k$ , since inference is driven by the effective information in independent clusters rather than by raw observation counts. Flag event times where support is thin or where only a single cohort contributes. There is no universal threshold for "thin" support, but as a rule of thumb, fewer than 30 observations per event time often yield unstable point estimates, and fewer than 2 contributing cohorts or clusters means the estimate reflects essentially a single group's experience. These thresholds depend on outcome variance and the clustering structure. Conclusions about effects at thin-support event times should be tempered. Sensitivity analyses that exclude extreme event times provide evidence on robustness.

Leave-one-cohort-out sensitivity analyses re-estimate the event-time profile excluding each cohort in turn and check whether the profile is stable. If excluding a single cohort changes the profile substantially, this signals that the cohort is influential and that the aggregated profile may not generalise. However, sensitivity to cohort exclusion is ambiguous: it could indicate that the excluded cohort is an outlier (suggesting the main estimate is more reliable without it), or that treatment effects are genuinely heterogeneous across cohorts (suggesting the aggregated estimate obscures meaningful variation). Cohort-specific profiles  $\theta_{g,k}$  help distinguish these cases: if profiles are similar across cohorts, the influential cohort is likely an outlier; if profiles diverge, heterogeneity is the explanation. If the profile is stable across leave-one-cohort-out specifications, this is consistent with the robustness of conclusions.

Leave-one-time-out sensitivity analyses exclude specific calendar periods and re-estimate the event-time profile. If a single period drives the result—for example, if excluding the first post-treatment period eliminates

the estimated effect—this suggests that the effect is concentrated in that period or that the period is an outlier. Robustness checks that vary the event-time window, exclude outliers, or trim the sample based on pre-treatment characteristics provide evidence on the stability of conclusions.

Weight audits examine the implicit weights assigned to cohort-time observations in the aggregation of  $\theta_k$ . Modern event-study estimators report these weights explicitly, and inspecting them reveals which cohorts and periods contribute most to each event-time coefficient. If a single cohort dominates the aggregation at certain event times due to support patterns, the estimated  $\theta_k$  reflects that cohort's effect and may not generalise. Reporting weights alongside estimates ensures transparency.

Specification curves aggregate estimates across many defensible modelling choices: different sets of control units (never-treated only vs never-treated and not-yet-treated), different covariate adjustments (no covariates vs rich controls), different event-time windows (short vs long), different binning choices, and different estimators (Sun–Abraham vs Callaway–Sant'Anna vs Borusyak–Jaravel–Spiess). Plot the distribution of estimates across specifications for each event time  $k$  or for summary measures (ATT, cumulative effect, LRM). If estimates cluster tightly, conclusions are less sensitive to modelling choices. If estimates vary widely, the choice of specification matters. Because specification curves involve many related estimates, interpret them alongside the multiple-testing considerations in Section 5.7. The goal is to assess robustness across a pre-specified grid, not to search for a specification that happens to deliver significance.

#### Multiple Testing Concern

Running many specifications and reporting all results raises multiple testing concerns. If 22 of 24 specifications yield significant results, this pattern could reflect robustness, but it can also reflect specification searching. The key distinction is whether specifications were pre-registered or chosen post hoc. Pre-specifying the specification grid in a design registry or pre-analysis plan (Section 3.9) distinguishes honest robustness analysis from specification mining. Report all pre-specified specifications regardless of results.

As an illustrative example, suppose you estimate the immediate effect  $\theta_0$  across 24 specifications (2 control sets  $\times$  2 covariate adjustments  $\times$  3 windows  $\times$  2 estimators). The resulting estimates range from 4.2 to 7.8, with median 5.8 and interquartile range [5.1, 6.4]. If 22 of 24 specifications yield positive and statistically significant estimates, this pattern is consistent with a positive effect across many defensible specifications. If estimates range from -1.2 to 8.4 with half positive and half negative, the conclusion is specification-dependent and less credible. Report the full distribution, highlight the preferred specification with justification, and discuss which modelling choices drive variation.

Practical guidance for marketing applications includes conducting pre-trend diagnostics as a matter of course, reporting placebo estimates to bolster credibility, showing support by event time, conducting leave-one-cohort-out and leave-one-time-out analyses, inspecting weights, and constructing specification curves. Transparent reporting of diagnostics builds confidence that conclusions are not artefacts of arbitrary choices and that the identifying assumptions are plausible. The goal is not to achieve perfect identification, which is impossible in observational settings, but to articulate assumptions transparently, provide evidence that they are plausible, and demonstrate that conclusions are robust to plausible deviations.

**Table 5.1** Specification Choices and Associated Bias Risks

Specification Choice	Risk if Misspecified	Diagnostic/Mitigation
Omitted category (reference bin)	Misinterpretation of levels vs differences	Clearly document the reference and check robustness to alternative references
Binning of extreme event times	Loss of resolution and aggregation bias if effects vary within bins	Report support by $k$ and check robustness to alternative binning
Event-time window (leads/lags)	Truncation bias if effects extend beyond window and loss of pre-trend evidence	Pre-specify based on data support and substantive expectations
Control set (never-treated vs not-yet-treated)	Bias if not-yet-treated are not parallel and loss of precision if never-treated are few	Test parallel trends by cohort and compare estimates across control sets
Cohort weights in aggregation	Misrepresentation of policy-relevant effect if weights do not match target population	Pre-specify weights based on the substantive question and report cohort composition
TWFE vs heterogeneity-robust estimator	Negative weights and contamination bias under heterogeneity	Estimate both, compare, and report cohort-specific profiles

### Diagnostic Quick Reference

This box summarises the key diagnostics for event-study credibility. For the full end-to-end event-study workflow with these diagnostics embedded, see Section 5.11. For non-event-specific diagnostics, see Chapter 17.

*Pre-trend checks.* Test joint significance of  $\theta_k$  for  $k < -1$ , plot pre-treatment coefficients, and conduct a placebo-in-time test. Remember that pre-trend diagnostics have limited power [Roth, 2022], so supplement them with visual inspection.

*Sensitivity to parallel trends.* Compute Rambachan–Roth bounds [Rambachan and Roth, 2023] using the `honestDiD` package to assess robustness to plausible violations.

*Support checks.* Report observations and cohorts by event time  $k$ . Flag thin support (fewer than 30 observations or fewer than 2 cohorts contributing). Bin extreme event times if needed.

*Sensitivity analyses.* Leave-one-cohort-out, leave-one-time-out, vary the control set (never-treated vs not-yet-treated), vary the window and binning, and construct a specification curve across estimators.

*Transparency.* Report weights and cohort composition, show cohort-specific profiles  $\theta_{g,k}$  if heterogeneity is material, and document deviations from the pre-specified plan.

None of these diagnostics can rescue a fundamentally flawed design. Their role is to stress-test a design that is already plausibly credible on substantive grounds.

These diagnostic procedures ensure that event-study estimates are credible and that conclusions are robust to plausible violations of identifying assumptions. The next section discusses extensions to continuous treatments, interference, and nonlinear outcomes.

## 5.9 Extensions and Special Cases

Event studies extend beyond binary treatment to settings with continuous intensities, interference across units, and nonlinear outcomes. This section provides worked examples and practical guidance for each extension, with forward references to detailed treatments in later chapters.

### Continuous and Multivalued Treatments

This subsection applies the continuous-treatment framework from Chapter 14 to event-time estimands. Continuous or multivalued treatments in event time arise when treatment intensity varies across units or over time. Advertising expenditure, promotional discount depth, loyalty programme reward generosity, and pricing all take continuous values. The event-time framework extends by defining an event-time dose-response function  $\mu_k(d) = \mathbb{E}[Y_{i,G_i+k}(d)]$  and its marginal effect  $\tau_k(d) = \partial\mu_k(d)/\partial d$  for treatment intensity  $d$  at event time  $k$ .

Intensity-based event studies regress outcomes on event-time indicators interacted with treatment intensity:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \theta_k \mathbf{1}\{t - G_i = k\} \times D_i + \varepsilon_{it},$$

where  $D_i$  is the treatment intensity for unit  $i$  (fixed at adoption). Here  $D_i$  is an aggregate of the time-varying treatment path  $D_{it}$  over a window around adoption (for example, the reward rate assigned at  $G_i$ ). When intensity itself evolves over time, we revert to the generic notation  $D_{it}$  from Chapter 14. The coefficients  $\theta_k$  estimate the effect of a unit change in treatment intensity at event time  $k$ . With a linear specification in intensity, the coefficients  $\theta_k$  estimate  $\tau_k(d)$  evaluated at the observed range of  $D_i$ , the marginal effect of a one-unit increase in intensity at event time  $k$ . When intensity varies over time within units, use  $D_{it}$  instead, but identification becomes more demanding. Event time and treatment intensity dynamics become intertwined, and the clean interpretation of  $\theta_k$  as “ $k$  periods since adoption at intensity  $d$ ” is no longer straightforward. See Chapter 10 for handling time-varying intensity. This specification assumes a linear dose-response, with proportional effects at each event time. When diminishing returns or threshold effects are expected, use flexible specifications (intensity bins, splines, or the dose-response methods in Chapter 14).

### Worked Example: Loyalty Programme with Tiered Rewards

A retailer offers three reward tiers: 2%, 5%, and 10% cashback. Stores are assigned to tiers based on customer demographics. Let  $m$  index tiers and let  $D_{im}$  indicate that store  $i$  is assigned to tier  $m$ . The event-study estimates  $\hat{\theta}_{k,m}$  for each tier over  $k = 0, \dots, 4$ :

Tier	$\hat{\theta}_{0,m}$	$\hat{\theta}_{1,m}$	$\hat{\theta}_{2,m}$	$\hat{\theta}_{3,m}$	$\hat{\theta}_{4,m}$
2% cashback	0.0	2.1	3.8	4.2	4.5
5% cashback	0.0	4.8	8.2	9.1	9.8
10% cashback	0.0	8.5	14.3	15.2	15.8

Effects are in £ thousands additional quarterly sales per store. These are point estimates, and confidence intervals (not shown) should be computed and reported in practice. Two patterns emerge: (1) effects grow over event time as customers enrol and develop habits; (2) effects scale with intensity but exhibit diminishing returns—doubling from 5% to 10% does not double the effect. This suggests saturation at higher reward levels. Chapter 14 develops dose-response methods. Chapter 10 covers distributed lag models for intensity-dependent dynamics.

### Key Identification Assumption: Continuous Treatments

A continuous-treatment event study assumes a conditional independence condition in potential-outcome terms. For all intensity levels  $d$  in the support,

$$Y_{it}(d) \perp\!\!\!\perp D_i \mid X_i,$$

Intuitively, conditional on observed covariates, intensity assignment is as-good-as-random. Fixed effects can help control for stable unobserved differences in levels, but they do not create identification if intensity is chosen in response to anticipated gains. If high-intensity stores are systematically different (for example, higher baseline sales), the estimates conflate treatment intensity effects with selection effects. This assumption is often violated in marketing: firms typically assign higher intensity to units where they expect larger effects (for example, higher discounts to price-sensitive segments or more advertising to high-potential markets). Check balance on pre-treatment characteristics across intensity levels, and consider instrumental variables or regression discontinuity designs when selection on intensity is severe.

## Interference and Spillovers

This subsection extends the exposure-mapping approach of Chapter 11 to the event-time setting. Event studies with interference estimate how treatment in one unit affects outcomes in neighbouring units at different event times. The specification includes spatial leads and lags:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \geq 0} \theta_k^{\text{direct}} \mathbf{1}\{t - G_i = k\} + \sum_{k \geq 0} \theta_k^{\text{spillover}} S_{it}(k) + \varepsilon_{it},$$

with

$$S_{it}(k) = \sum_{j \in N(i)} \mathbf{1}\{t - G_j = k\} \mathbf{1}\{G_j \leq t\},$$

noting that the  $G_j \leq t$  restriction ensures that only already-treated neighbours contribute to spillovers—not-yet-treated neighbours remain in the control comparison. This is an exposure mapping. Later chapters use  $h_i(D_{-i,t})$  to denote such mappings. Under an exposure-mapping version of SUTVA (Chapter 11) and parallel trends in event time,  $\theta_k^{\text{direct}}$  identifies the average effect of own treatment at event time  $k$  holding neighbour exposure  $S_{it}(k)$  fixed, and  $\theta_k^{\text{spillover}}$  identifies the average marginal effect of an additional treated neighbour at event time  $k$ . This specification assumes no anticipation spillovers: if neighbours anticipate treatment and adjust behaviour before their own adoption, pre-treatment coefficients for the focal unit may be contaminated. In such settings, standard pre-trend diagnostics for the treated units can fail to detect violations, because spillovers move both treated and control units in the same direction before adoption.

**Note on Exposure Mappings.** The simple neighbour count  $S_{it}(k)$  treats all neighbours equally. More sophisticated exposure mappings may weight neighbours by distance, network centrality, or competitive overlap. The choice of exposure mapping is consequential for interpretation and should be justified ex ante based on the mechanism of spillover.

#### Worked Example: Store Network Spillovers

Store A is treated in Q3. Store B is a neighbouring store that is never treated. After Store A's treatment, at  $k = 0$  (Q3), Store B experiences spillover from Store A at  $S_{B,Q3}(0) = 1$ . At  $k = 1$  (Q4), Store B experiences spillover at  $S_{B,Q4}(1) = 1$ . Store A's direct effect is  $\theta_0^{\text{direct}}$ . Store B's spillover effect is  $\theta_1^{\text{spillover}}$ .

Estimated effects might show  $\theta_0^{\text{direct}} = 8.2$ , which is the immediate lift for the treated store. They might also show  $\theta_1^{\text{spillover}} = -1.4$ , which is the effect on nearby control stores that lose customers to the treated store. Negative spillovers indicate business stealing. Positive spillovers indicate demand creation or word-of-mouth. As with all event-study estimates, these are point estimates. Confidence intervals should be computed to assess statistical significance.

#### Critical Assumption: Exposure Mapping

The exposure mapping  $N(i)$  must be specified ex ante. In marketing, “neighbours” may be geographic (stores within 10km), network-based (customers who share social connections), or competitive (stores in the same category). The choice affects interpretation. If the exposure mapping is misspecified, spillover estimates are biased. Chapter 11 develops these methods in detail.

## Nonlinear Outcomes and Log Transformations

Event studies can be estimated on transformed outcomes when the outcome distribution is non-normal. Log transformations are common for sales data (always positive, right-skewed).

**Semi-Elasticity Interpretation.** When estimating on log outcomes:

$$\log Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \theta_k \mathbf{1}\{t - G_i = k\} + \varepsilon_{it},$$

This specification requires  $Y_{it} > 0$  for all included observations. When zeros occur, either drop those observations, add a small offset, or use the inverse hyperbolic sine transformation discussed below. The coefficient  $\theta_k$  approximates the percentage change in  $Y$  at event time  $k$ . For  $\theta_k = 0.08$ , the interpretation is approximately an 8% increase in sales. For larger effects, use  $(\exp(\theta_k) - 1) \times 100\%$  for the exact percentage change. In all these cases, the identification assumptions are the same as for level outcomes. Only the scale on which  $Y_{it}(d)$  is modelled changes, and with it the interpretation of  $\theta_k$  as a percentage rather than level effect.

### Jensen's Inequality Caveat

When converting log-scale predictions back to levels,  $\exp(\mathbb{E}[\log Y]) \neq \mathbb{E}[Y]$ . This matters for forecasting aggregate outcomes. For causal effect estimation, log-specification coefficients are still interpretable as semi-elasticities. The smearing adjustment matters only when translating back into levels for forecasting. If the goal is prediction, apply a smearing adjustment such as  $\hat{Y} \approx \exp(\hat{\theta}_k) \times \frac{1}{n} \sum_i \exp(\hat{\varepsilon}_i)$ , or more generally condition the smearing factor on relevant covariates.

**When to Use Log vs Levels.** Use log transformations for sales, revenue, and expenditure—outcomes that are positive, right-skewed, and where multiplicative effects are plausible. Use levels for profit (which can be negative), customer counts (small integers where percentage changes are less meaningful), and indices (already normalised). Consider the inverse hyperbolic sine (IHS) transformation when the outcome includes zeros: IHS handles zeros gracefully and approximates the log for large values, though interpretation requires care [Bellemare and Wichman, 2020].

## Smoothing: Bias-Variance Tradeoff

Smoothing event-time profiles—by imposing polynomial, spline, or parametric structure on  $\theta_k$ —reduces variance but introduces bias if the true profile does not match the imposed structure.

**When Smoothing Is Appropriate.** Smoothing is appropriate for prediction and forecasting: if the goal is to extrapolate effects beyond the observed window, smoothing provides out-of-sample predictions that fully flexible estimates cannot. Smoothing can also aid presentation: a smoothed curve can communicate the overall trajectory to non-technical audiences, provided the underlying flexible estimates are also reported.

As with other modelling choices, smoothing decisions should be pre-specified where possible. Choosing the degree of smoothing based on the observed noise level in  $\hat{\theta}_k$  risks overfitting the visual narrative.

**When Smoothing Is Inappropriate.** Smoothing is inappropriate for causal identification: the credibility of causal claims rests on the unsmoothed estimates. Do not smooth pre-treatment coefficients when using them as diagnostics; any smoothing step should be reserved for post-treatment summarisation or forecasting and reported as secondary to the unsmoothed profile. Smoothing is also inappropriate when testing for dynamics: if the question is whether effects decay, ramp up, or exhibit non-monotonic patterns, imposing smoothness assumes the answer.

**Best Practice.** Report flexible (unsmoothed) estimates as the primary specification. If smoothing aids interpretation, overlay a smoothed fit on the event-study plot and clearly label it as such. Never report only smoothed estimates without the underlying flexible coefficients.

### Summary: When to Use Each Extension

**Table 5.2** Event-Study Extensions Decision Guide

Extension	When to Use	Key Assumption	Reference
Continuous treatment (estimand: marginal effect $\tau_k(d) = \partial\mu_k(d)/\partial d$ )	treat- Ad spend, discount depth, reward rates	Conditional independence of intensity	Ch 14
Interference/spillovers (estimand: $\theta_k^{\text{direct}}, \theta_k^{\text{spillover}}$ )	Geo-experiments, networks, platforms	Exposure mapping correctly specified	Ch 11
Log transformation	Sales, revenue (positive, right-skewed)	Multiplicative effects — plausible	—
IHS transformation	Outcomes with zeros	Large values approximate — log	—
Smoothing	Forecasting, presentation	Modelling choice: treat $\theta_k$ with causality as approximately smooth over $k$	Use with caution

Extensions to continuous treatments, interference, and nonlinear outcomes broaden the applicability of event studies beyond the canonical setting. The key is to maintain the design-based philosophy: specify estimands clearly, articulate identification assumptions transparently, estimate flexibly without imposing unnecessary structure, and report results in a way that enables readers to assess credibility and robustness.

## 5.10 Event-Time Metrics for Marketing Decisions

Chapter 4 (Section 4.9) presents detailed case studies of marketing applications for difference-in-differences and staggered adoption designs. This section focuses on what event-study analysis specifically adds: quantitative metrics that translate the event-time profile  $\theta_k$  into actionable marketing decisions.

### Diagnosing Anticipation

Anticipation occurs when units adjust behaviour before treatment begins, typically because the intervention is announced in advance. In marketing, anticipation is common: advertising campaigns are announced ahead of launch, loyalty programmes have pre-enrolment periods, and pricing changes may be leaked or inferred by customers.

Pre-treatment lead coefficients near treatment (for example,  $\theta_{-2}, \theta_{-3}, \dots$ ) can reveal anticipation. If lead coefficients are non-zero, either anticipation or pre-trends are present. In practice, use the joint lead diagnostics and uniform confidence bands from Sections 5.5 and 5.6 to assess whether the collection of pre-treatment coefficients for  $k < 0$  is jointly close to zero, rather than relying only on eyeballing a single lead coefficient.

Anticipation produces lead coefficients that are negative (customers delay purchases until the promotion begins) or positive (customers stockpile in advance of price increases), with the pattern typically appearing close to treatment and not at distant horizons. Pre-trends, by contrast, produce lead coefficients that show a systematic trend (upward or downward) extending back several periods, suggesting that treated and control units were already diverging before treatment. As noted in Section 5.5, this distinction cannot be made statistically. It requires substantive knowledge about the intervention.

Marketing examples.

Context	Pattern	Interpretation
Campaign announcement	$\theta_{-2} < 0$	Customers delay buying until campaign starts
Loyalty pre-enrolment	$\theta_{-2} > 0$	Early adopters sign up before official launch
Price increase rumours	$\theta_{-2} > 0$	Stockpiling before price rises
Seasonal promotion timing	$\theta_{-2}, \theta_{-3} \approx 0$	No anticipation, timing not predictable

If anticipation is plausible, the immediate effect  $\theta_0$  may understate the total effect (if customers pulled forward purchases) or overstate it (if customers delayed purchases until treatment). Report a windowed cumulative effect  $\sum_{k=-L}^K \theta_k$  that includes the key pre-treatment anticipation period and the post-treatment horizon of interest, with  $L$  and  $K$  chosen and justified ex ante based on the institutional setting.

## Measuring Ramp-Up Rate

Ramp-up occurs when treatment effects grow over event time, typically because customer behaviour changes gradually (enrolment, habit formation, network effects).

These metrics are deterministic functions of estimated  $\theta_k$ . For formal inference on them (confidence intervals and hypothesis tests), use the delta method or bootstrap to propagate uncertainty from the  $\theta_k$  estimates. See Chapter 16 for details.

The average per-period growth is  $(\theta_K - \theta_0)/K$ , where  $K$  is the last event time in the reporting window. It measures how much the effect grows per period on average.

The time-to-maturity is the event time  $k^*$  at which  $\theta_k$  stabilises, for example when  $|\theta_{k+1} - \theta_k| < \epsilon$  for subsequent periods. Set  $\epsilon$  ex ante and tie it to business relevance (for example, changes smaller than 5% of  $\theta_K$ ).

The effect multiplier is  $M_K = \sum_{k=0}^K \theta_k/\theta_0$ , the ratio of the cumulative effect through horizon  $K$  to the immediate effect. This coincides with the long-run multiplier LRM defined in Section 5.2 when  $K$  is chosen so that  $\theta_k$  has effectively stabilised.  $M_K$  is a horizon-dependent summary, so always report  $K$  and the support at each event time. This metric is informative when  $\theta_0$  is clearly different from zero and precisely estimated. When  $\theta_0$  is near zero or very noisy,  $M_K$  becomes unstable. In such cases, focus on the cumulative effect and the full  $\theta_k$  profile rather than a single multiplier.

Suppose a loyalty programme yields the following event-time profile (in £ thousands additional sales per store per quarter):

	Event time $k$	0	1	2	3	4	5
$\hat{\theta}_k$		3.2	5.8	8.1	9.4	10.0	10.2

The average per-period growth is  $(10.2 - 3.2)/5 = 1.4$  (£1,400 per quarter). The time-to-maturity is  $k^* \approx 4$ , since the effect stabilises between  $k = 4$  and  $k = 5$ . The effect multiplier is  $M_5 = (3.2 + 5.8 + 8.1 + 9.4 + 10.0 + 10.2)/3.2 = 46.7/3.2 \approx 14.6$ .

In practice, report confidence intervals for these metrics alongside the point estimates, not just the  $\hat{\theta}_k$  values.

If the effect multiplier is large, short-run evaluations understate long-run benefits. A manager evaluating the programme at  $k = 0$  would see only £3,200 additional sales. Waiting until maturity reveals £10,200. Set ROI evaluation timelines to  $k \geq k^*$ , not immediately post-launch.

## Estimating Decay Half-Life

Decay occurs when treatment effects diminish over event time, typically because customer attention fades, competitors respond, or promotional effects are transitory.

These definitions build on the dynamic-treatment discussion in Chapter 10. Here we focus on reading half-lives directly from the event-study path and using them as decision metrics.

*Half-life.* The event time  $k_{1/2}$  at which the effect falls to half its peak value, i.e.  $\theta_{k_{1/2}} = \theta_{\text{peak}}/2$ . Shorter half-lives indicate faster decay. When the half-life falls between observed event times, linear interpolation yields  $k_{1/2} = k_a + (k_b - k_a) \times (\theta_{k_a} - \theta_{\text{peak}}/2)/(\theta_{k_a} - \theta_{k_b})$ , where  $k_a$  and  $k_b$  bracket the half-life.

*Decay rate (optional parametric fit).* If you are willing to approximate post-treatment dynamics by an exponential curve,  $\theta_k = \theta_0 \exp(-\lambda k)$ , then  $\lambda = -\log(\theta_1/\theta_0)$  and the implied half-life is  $k_{1/2} = \log(2)/\lambda$ .

*Persistence ratio.*  $\theta_K/\theta_{\text{peak}}$ , where  $K$  is the end of the reporting window. Values near 1 indicate persistence, and values near 0 indicate transitory effects.

Suppose a TV campaign yields the following profile (in percentage points sales lift):

Event time $k$	0	1	2	3	4	5	6
$\hat{\theta}_k$	8.2	6.1	4.5	3.4	2.5	1.9	1.4

The peak effect is  $\theta_0 = 8.2$ . The half-life solves  $\theta_k = 4.1$  and occurs between  $k = 2$  and  $k = 3$ . By linear interpolation,  $k_{1/2} = 2 + (3 - 2) \times (4.5 - 4.1)/(4.5 - 3.4) \approx 2.4$  periods. The persistence ratio is  $1.4/8.2 = 0.17$ , so only 17% of the peak effect remains at  $k = 6$ .

Short half-lives indicate that sustained investment is required to maintain brand salience. If  $k_{1/2} = 2.4$  quarters, advertising effects largely dissipate within one year. Budget accordingly. One-time campaigns produce transitory effects, and sustained presence requires ongoing investment. When using such parametric fits for extrapolation, always report them alongside the non-parametric event-study profile and check that the exponential curve provides a reasonable approximation over the observed window.

## Cumulative Effects and ROI

The cumulative effect aggregates event-time coefficients over a window, providing a single summary of total impact.

The cumulative effect is

$$\text{Cumulative effect} = \sum_{k=0}^K \theta_k.$$

This sum aggregates the event-time effects over the post-treatment window. If the goal is to express the cumulative effect in outcome units (for example, total additional sales over the window), multiply by the number of units and periods as appropriate.<sup>1</sup> For ROI calculations:

$$\text{ROI} = \frac{\text{Cumulative effect} \times \text{Scale factor} - \text{Cost}}{\text{Cost}},$$

where the scale factor converts outcome units into monetary terms. In practice it is a deterministic accounting map, such as the number of treated unit-periods represented by the aggregation multiplied by a contribution margin per unit of outcome, plus any unit conversions needed to express the effect in currency. These ROI

<sup>1</sup> For long evaluation horizons, effects at distant event times should be discounted. The net present value (NPV) of the event-time profile is  $\sum_{k=0}^K \theta_k/(1+r)^k$ , where  $r$  is the per-period discount rate. For short horizons (one year or less), discounting has modest impact. For multi-year evaluations, NPV can be substantially lower than the simple cumulative sum.

calculations are functions of the estimated  $\theta_k$  profile and therefore inherit the identification assumptions, clustering choices, and diagnostic results discussed in Sections 5.5–5.8. Treat ROI as no more credible than the underlying event study on which it rests.

Two interventions with the same immediate effect  $\theta_0$  can have very different cumulative effects depending on their dynamics. A ramp-up profile produces high cumulative effects if ramp-up is fast and the long-run effect is large. A decay profile produces lower cumulative effects if decay is fast. A persistent-effect profile produces the highest cumulative effects if the effect persists indefinitely.

**Box 5.2: The Brand vs Performance Paradox**

Consider a retailer choosing between two ways to boost category revenue over the next year using the same budget per store. A performance campaign uses deep discounts that generate immediate sales spikes but little persistence. A brand-building campaign uses sustained advertising that strengthens brand equity and shifts preferences gradually.

Suppose event-time effects (in thousands of additional sales per store per quarter) from an event-study analysis are:

	$k = 0$	$k = 1$	$k = 2$	$k = 3$
Performance (discount)	8.0	2.0	0.5	-1.0
Brand-building	1.0	3.0	5.0	6.0

At equal spend per store, the discount campaign has a much larger immediate effect ( $\theta_0 = 8.0$ ) than the brand-building campaign ( $\theta_0 = 1.0$ ), but their dynamics differ sharply.

For the discount campaign, the cumulative effect through  $k = 3$  is

$$\sum_{k=0}^3 \theta_k^{\text{disc}} = 8.0 + 2.0 + 0.5 - 1.0 = 9.5,$$

with an effect multiplier of  $9.5/8.0 \approx 1.2$ . Most value arrives immediately, and negative effects from customer stockpiling or deal-seeking erode gains.

For the brand-building campaign, the cumulative effect through  $k = 3$  is

$$\sum_{k=0}^3 \theta_k^{\text{brand}} = 1.0 + 3.0 + 5.0 + 6.0 = 15.0,$$

with an effect multiplier of  $15.0/1.0 = 15$ . Evaluated only at  $k = 0$ , the brand campaign appears weak. Evaluated at maturity, it dominates on total incremental revenue and feeds into longer-run CLV and brand equity. In practice, you map the cumulative revenue profile implied by  $\theta_k$  into CLV by applying customer-level margins, churn dynamics, and discounting, as described in Chapter 4.

The *brand vs performance paradox* is that short-run ROI evaluated at  $k = 0$  favours the performance campaign, while long-run cumulative ROI and CLV favour brand-building. Event-time metrics (effect multipliers and cumulative effects) make this trade-off explicit and highlight the need to align evaluation horizons with strategic objectives.

**Table 5.3** Event-Time Metrics and Marketing Decisions

Metric	How to Compute	Marketing Decision
Anticipation	joint lead diagnostics for $\{\theta_k : k \in \{-L, \dots, -2\}\}$	Timing of announcements and interpretation of $\theta_0$
Ramp-up rate	$(\theta_K - \theta_0)/K$	Patience during rollout and evaluation timing
Time-to-maturity	$k^*$ where $ \theta_{k+1} - \theta_k  < \epsilon$	When to assess ROI and how long to support operations
Effect (equals LRM when $K$ covers the long-run horizon)	multiplier $\sum_{k=0}^K \theta_k / \theta_0$ (requires $\theta_0 \neq 0$ )	Short-run versus long-run value and stakeholder communication
Half-life	$k$ where $\theta_k = \theta_{\text{peak}}/2$	Campaign frequency and need for sustained investment
Persistence ratio	$\theta_K / \theta_{\text{peak}}$	One-time vs repeated interventions
Cumulative effect	$\sum_k \theta_k$ (for long horizons, use Total impact and ROI numerator NPV)	Total impact and ROI numerator

### Summary: Event-Time Metrics Decision Guide

These metrics transform the event-time profile from a visual diagnostic into quantitative inputs for marketing decisions. Report the profile alongside these summary metrics to communicate both the dynamic trajectory and its business implications. All of these event-time metrics are deterministic functions of the underlying causal effect path ( $\theta_k$ ). When presenting them to decision-makers, always accompany them with the original event-study plot and a brief reminder of the identification assumptions and diagnostics that support the  $\theta_k$  estimates.

## 5.11 Workflow Checklist

This section provides a numbered end-to-end protocol for conducting event-study analyses in marketing panels. It synthesises the estimands, identification assumptions, estimation strategies, diagnostics, and inference procedures developed in Chapters 4 and 5, and the general design principles of Chapter 3. Box 5.1 (Section 5.8) focuses on specification and diagnostic choices. This workflow extends to the full analysis pipeline including business outputs. Check off each step as you proceed, and treat each one as a decision point where you may conclude that the design is not credible enough to support a causal interpretation.

### Step 1: Define Estimands and Baseline

- State the substantive research question clearly
- Define target event-time effects  $\theta_k$  precisely (immediate effect  $\theta_0$ , specific horizons, or full profile?)
- Specify the baseline period (typically  $k = -1$ , or an average of several pre-periods when a single period is noisy or affected by anticipation)
- Choose aggregation weights (cohort-size, uniform, or treated-unit-period), recognising that this choice changes which units and periods the estimand emphasises
- Align estimands to business question: immediate effects for go/no-go, cumulative effects for ROI, long-run effects for strategy (for example, cumulative sums  $\sum_{k=0}^K \theta_k$  and long-run multipliers LRM as defined in Section 5.2)

### Step 2: Assess Data Support

- Construct support table: observations and cohorts by event time  $k$
- Flag thin-support regions (for example, fewer than about 30 observations, or only one cohort or cluster contributing at a given  $k$ )
- Check cohort composition across  $k$  for composition bias risk
- Choose event-time window based on support and substantive interest, truncating or heavily binning horizons where support is thin or composition changes sharply
- Pre-specify binning for extreme event times where support is thin

### Step 3: Select Estimator

- Choose heterogeneity-robust estimator: Sun–Abraham, Callaway–Sant’Anna, or Borusyak–Jaravel–Spiess

- Match estimator to design: for example, use Sun–Abraham for staggered adoption with balanced panels, Callaway–Sant’Anna for group-time average treatment effects with flexible control sets, and Borusyak–Jaravel–Spiess when an imputation-based approach suits unbalanced panels or irregular adoption
- Estimate TWFE as benchmark (but not primary specification), and interpret differences in light of its distinct weighting scheme and potential for negative weights
- Document estimator choice and rationale

#### Step 4: Run Pre-Trend and Anticipation Checks

- Estimate event study with multiple pre-treatment leads
- Assess joint significance of pre-treatment coefficients (all  $\theta_k$  with  $k < 0$  except the omitted reference period) (Wald test with cluster-robust SEs). See Sections 5.5 and 5.8 for details
- Plot pre-treatment coefficients with confidence intervals
- Conduct placebo-in-time test using only pre-treatment data
- Interpret “no significant pre-trend” cautiously (Chapter 17)
- If substantial pre-trends are detected, reconsider the design or consider alternative identification strategies (Chapter 6)

#### Step 5: Estimate and Plot Event-Time Profile

- Estimate chosen heterogeneity-robust estimator
- Construct event-time profile  $\{\theta_k\}_{k=-K_-}^{K_+}$  with confidence intervals
- Plot with vertical line at  $k = 0$  and marker for omitted bin ( $k = -1$ )
- Show support by event time in secondary panel or table
- Overlay cohort-specific profiles if heterogeneity is material

#### Step 6: Select Inference Procedure

- Cluster standard errors by unit  $i$  (default), so that the number of clusters is  $N$ , the number of units.
- Consider two-way clustering if cross-unit correlation is plausible
- If the number of clusters is small (for example, fewer than about 20), use wild cluster bootstrap or randomisation or permutation inference as the primary procedure.
- If the number of clusters is moderate (roughly 20–50), compare asymptotic cluster-robust standard errors to bootstrap-based inference and report any material differences.
- Decide multiplicity adjustment: joint tests for hypothesis families, or individual adjustments (Bonferroni, FDR, Romano–Wolf), and pre-specify which families are primary

- Ensure that clustering and resampling choices align with the dependence structure in the data (see Chapter 16 for detailed guidance on variance estimators and permutation strategies)
- Label primary vs exploratory analyses clearly

### Step 7: Conduct Sensitivity Analyses

- Vary control set (never-treated only vs including not-yet-treated)
- Vary event-time window (short vs long)
- Vary binning choices (fine vs coarse)
- Vary covariate adjustments (none vs rich controls)
- Construct specification curve for key event times ( $\theta_0, \theta_5$ , cumulative  $\sum_k \theta_k$ ) to summarise how estimates vary across reasonable specifications, as discussed further in Chapter 17
- Conduct leave-one-cohort-out and leave-one-time-out analyses
- Compare estimates across estimators (Sun–Abraham, Callaway–Sant’Anna, Borusyak–Jaravel–Spiess, TWFE)
- Treat sensitivity analyses as a way to assess robustness rather than to search for a single “best” specification

### Step 8: Compute Event-Time Metrics

This step translates the event-time profile into quantitative business outputs (see Section 5.10 for details). Because these metrics are functions of estimated event-time effects, they should be interpreted with appropriate uncertainty, for example by reporting ranges across specifications or confidence intervals where feasible.

- Anticipation.* Use joint pre-trend diagnostics and inspection of near-treatment pre-treatment coefficients (for example,  $\theta_{-2}, \theta_{-3}$ ) to detect statistically meaningful deviations from zero. Interpret these in light of institutional knowledge about announcements and predictability, and reconcile them with the pre-trend diagnostics from Step 4
- Ramp-up rate.* Compute  $(\theta_K - \theta_0)/K$ . Report average per-period effect growth, noting that the estimate can be noisy when individual  $\theta_k$  are imprecise
- Time-to-maturity.* Identify  $k^*$  where  $|\theta_{k+1} - \theta_k| < \epsilon$ . Report when programme reaches steady state, acknowledging that this depends on the chosen tolerance  $\epsilon$  and sampling variability in  $\theta_k$
- Effect multiplier.* Compute  $M_K = \sum_{k=0}^K \theta_k/\theta_0$ . Report short-run versus long-run value, and note that  $M_K$  coincides with the long-run multiplier LRM when  $K$  is chosen so that the profile has effectively stabilised.
- Half-life* (if decay profile). Identify  $k$  where  $\theta_k = \theta_{\text{peak}}/2$ . Report persistence, treating this as an approximate rather than exact calendar time

- Cumulative effect:* Compute  $\sum_{k=0}^K \theta_k$  (and, for long horizons, optionally its discounted version), and use this as the numerator in ROI and CLV calculations.

### Step 9: Document and Report

- State research question and data structure (panel dimensions, adoption pattern)
- Document estimand definition, baseline, and chosen estimator
- Report diagnostic results (pre-trends, placebo, support, composition)
- Present primary estimates: event-time profile and event-time metrics (Step 8)
- Report sensitivity analyses: specification curve, leave-one-out, estimator comparison
- Document inference procedures (clustering, bootstrap, multiplicity)
- Provide substantive interpretation linking to business decisions
- Note any deviations from pre-specified analysis plan
- Register the analysis plan or, at minimum, timestamp and version-control the analysis scripts (Section 3.9)
- Archive replication materials: scripts, data (or simulated substitute), software versions

By following this workflow, practitioners can conduct event-study analyses that are transparent, rigorous, and aligned with modern best practices. This event-study workflow parallels the DiD workflow in Section 4.10, with the key difference that estimands, diagnostics, and inference are now expressed in event time rather than calendar time.

This chapter has developed event-study designs from their foundations through practical implementation in marketing panels. We have defined event-time estimands and their aggregations, specified lead-lag regressions with appropriate normalisation and binning, presented heterogeneity-robust estimation methods, articulated identification assumptions and diagnostic workflows, and introduced quantitative metrics—ramp-up rate, effect multiplier / long-run multiplier, decay half-life, and cumulative effects—that translate event-time profiles into actionable business insights. Event studies provide a flexible, transparent framework for estimating dynamic treatment effects, diagnosing anticipation and pre-trends, and communicating causal narratives to technical and non-technical audiences. Chapter 6 extends panel methods to synthetic control designs, which provide an alternative identification strategy when parallel trends is implausible and when a single treated unit or a small number of treated units can be compared to a weighted combination of control units.

**Part III**

**Synthetic Controls and Hybrid Methods**



## Chapter 6

# Synthetic Control

In this chapter you learn how to use synthetic control when a single unit, or a small number of units, receives treatment and many others remain untreated. Our estimand is the unit-time effect for the treated unit,  $\tau_{1t} = Y_{1t}(1) - Y_{1t}(0)$  for  $t > T_0$ . By the end of the chapter you will be able to construct a synthetic control from a donor pool using convex weights and a predictor set, state the identification conditions in potential-outcomes and factor-model language, and assess credibility using pre-treatment fit diagnostics and in-space and in-time placebo checks. We then cover permutation-based inference and sensitivity analysis, and show how to apply the method to realistic marketing settings such as a city-level branding campaign or a flagship store redesign when only one market is treated. Synthetic control sits alongside the difference-in-differences and event-study methods from Chapters 4 and 5. It targets similar causal questions but replaces parallel trends for a large control group with a carefully constructed weighted control that matches the treated unit's pre-treatment path.

## 6.1 Motivation and Setup

Synthetic control methods provide a transparent, design-aligned approach to causal inference when a single unit, or a small number of units, receive treatment and a larger set of untreated units can serve as potential controls. Rather than relying on parallel trends averaged across all control units, synthetic control constructs a counterfactual by reweighting control units to closely match the treated unit's pre-treatment trajectory. This chapter develops the method rigorously, sets out its theoretical foundations in factor models, and provides a critical assessment of its limitations, in particular the lack of principled guidance for model selection in the early literature.

### Formal Estimand and Estimator

We begin with precise definitions, using the potential outcomes notation introduced earlier in the book. Units are indexed by  $i = 1, \dots, N$  and time by  $t = 1, \dots, T$ . Without loss of generality, let unit  $i = 1$  be the treated unit, and let  $T_0$  denote the last pre-treatment period, so that treatment begins in period  $T_0 + 1$ . The donor pool  $\mathcal{J} = \{2, \dots, N\}$  consists of  $N - 1$  units that remain untreated throughout the study period, so we observe  $Y_{jt} = Y_{jt}(0)$  for all  $j \in \mathcal{J}$  and all  $t$ . Throughout this chapter,  $N$  denotes the total number of units in the panel.

The causal effect for the treated unit at time  $t > T_0$  is

$$\tau_{1t} = Y_{1t}(1) - Y_{1t}(0), \quad (6.1)$$

where  $Y_{1t}(1)$  and  $Y_{1t}(0)$  are the potential outcomes with and without treatment. We observe  $Y_{1t} = Y_{1t}(1)$  for  $t > T_0$ , but  $Y_{1t}(0)$  is the missing counterfactual. The synthetic control method constructs an estimate  $\hat{Y}_{1t}(0)$  of this counterfactual by forming a weighted average of donor outcomes:

$$\hat{Y}_{1t}(0) = \sum_{j \in \mathcal{J}} w_j^* Y_{jt}, \quad (6.2)$$

where the weights  $\mathbf{w}^* = (w_2^*, w_3^*, \dots, w_N^*)'$  satisfy  $w_j^* \geq 0$  and  $\sum_{j \in \mathcal{J}} w_j^* = 1$ . The choice of these weights, based on pre-treatment data, is the core design decision that we analyse in this chapter. The synthetic control estimator of the time- $t$  treatment effect is

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j \in \mathcal{J}} w_j^* Y_{jt}, \quad t > T_0. \quad (6.3)$$

This difference, often called the *gap*, is visualised in gap plots that make the magnitude and persistence of the estimated effect transparent over time.

## Factor Model Foundation

The theoretical foundation of synthetic control rests on a latent factor model for untreated potential outcomes. Assume that untreated outcomes satisfy

$$Y_{it}(0) = \alpha_i + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_i + \varepsilon_{it}, \quad (6.4)$$

where  $\alpha_i$  is a unit fixed effect,  $\lambda_t$  is a time fixed effect,  $\mathbf{f}_t \in \mathbb{R}^R$  is a vector of common factors,  $\boldsymbol{\lambda}_i \in \mathbb{R}^R$  is a vector of unit-specific factor loadings, and  $\varepsilon_{it}$  is an idiosyncratic error with  $\mathbb{E}[\varepsilon_{it}] = 0$ . This representation is equivalent to the interactive fixed-effects notation  $Y_{it}(0) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it}$  from Chapter 8, with  $\mathbf{f}_t$  collecting the factors  $(f_{t1}, \dots, f_{tR})$  and  $\boldsymbol{\lambda}_i$  collecting the corresponding loadings  $(\lambda_{i1}, \dots, \lambda_{iR})$ .

This factor structure nests several familiar specifications. When  $R = 0$ , we recover additive fixed effects:  $Y_{it}(0) = \alpha_i + \lambda_t + \varepsilon_{it}$ . If we set  $R = 1$  with  $f_{t1} = t$  and  $\lambda_{i1} = \gamma_i$ , we obtain unit-specific linear time trends:  $Y_{it}(0) = \alpha_i + \lambda_t + \gamma_i t + \varepsilon_{it}$ . The general  $R$ -factor model encompasses interactive fixed effects, which Chapter 8 develops fully.

If weights  $\mathbf{w}^*$  can be found such that

$$\boldsymbol{\lambda}_1 = \sum_{j \in \mathcal{J}} w_j^* \boldsymbol{\lambda}_j, \quad (6.5)$$

then, conditional on the factors and weights, the synthetic control is approximately unbiased for the counterfactual untreated outcome. To see this, substitute equation (6.4) into the donor-weighted average:

$$\begin{aligned} \sum_{j \in \mathcal{J}} w_j^* Y_{jt}(0) &= \sum_{j \in \mathcal{J}} w_j^* (\alpha_j + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_j + \varepsilon_{jt}) \\ &= \sum_{j \in \mathcal{J}} w_j^* \alpha_j + \lambda_t + \mathbf{f}'_t \sum_{j \in \mathcal{J}} w_j^* \boldsymbol{\lambda}_j + \sum_{j \in \mathcal{J}} w_j^* \varepsilon_{jt} \\ &= \alpha_1^w + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_1 + \bar{\varepsilon}_t^w \\ &= Y_{1t}(0) + (\alpha_1^w - \alpha_1) - \varepsilon_{1t} + \bar{\varepsilon}_t^w, \end{aligned} \quad (6.6)$$

where  $\alpha_1^w = \sum_j w_j^* \alpha_j$  and  $\bar{\varepsilon}_t^w = \sum_j w_j^* \varepsilon_{jt}$ . When the weights additionally satisfy  $\alpha_1 = \sum_{j \in \mathcal{J}} w_j^* \alpha_j$  (so that  $\alpha_1^w = \alpha_1$ ) and the idiosyncratic errors have mean zero conditional on the factors, the synthetic control is unbiased for  $Y_{1t}(0)$  in expectation. Section 6.4 discusses large-sample approximations and inference procedures.

When the identification condition (6.5) fails, meaning that the weights cannot exactly match the treated unit's factor loadings, the estimator is biased. Let  $\boldsymbol{\lambda}_1^{sc} = \sum_j w_j^* \boldsymbol{\lambda}_j$  denote the implicit factor loadings of the synthetic control. Then, taking expectations over the idiosyncratic errors,

$$\mathbb{E}[\hat{\tau}_{1t} - \tau_{1t}] = \mathbf{f}'_t (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^{sc}) = \mathbf{f}'_t \boldsymbol{\Delta}_\lambda, \quad (6.7)$$

where  $\boldsymbol{\Delta}_\lambda = \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^{sc}$  is the factor loading mismatch. This bias expression has important implications. First, bias depends on both the mismatch  $\boldsymbol{\Delta}_\lambda$  and the post-treatment factor evolution  $\mathbf{f}_t$ . Second, good pre-

treatment fit does not guarantee small bias if  $\mathbf{f}_t$  changes after treatment. This is the synthetic-control analogue of the structural-break concern in factor-based event-study counterfactuals: a break in factor dynamics at treatment undermines identification even when pre-treatment fit is excellent. Third, bias can grow over time if the factors driving outcomes evolve differently for the treated unit than for the donor pool. This bias expression explains why pre-treatment fit is informative but not decisive for credibility. Section 6.5 develops diagnostics that make this link precise.

See Chapter 8 for a complete treatment of factor models in panel data and for conditions under which interactive fixed effects are a good approximation in marketing settings.

### The Convexity Constraint: Justification and Limitations

The defining feature of synthetic control is the convexity constraint. Weights are non-negative and sum to one, so the synthetic control is a convex combination of donors. This has both theoretical and pragmatic justifications.

The convexity constraint ensures that the synthetic control interpolates within the convex hull of the donor pool rather than extrapolating beyond the observed data. This addresses the extrapolation concern raised by Angrist and Pischke [2010]: regression methods that assign implicit negative weights can produce counterfactuals unsupported by any observed unit. By construction, the synthetic control is a “realistic” counterfactual that resembles a weighted average of actual control units.

From an estimation perspective, the convexity constraint also acts as a regulariser, shrinking weights toward uniform and preventing extreme values. This reduces variance at the cost of potential bias, a bias–variance tradeoff analogous to ridge regression. In this view, the non-negativity and sum-to-one restrictions on  $\mathbf{w}$  act as implicit regularisers, discouraging extreme weights in much the same way that an explicit  $\ell_2$  penalty shrinks regression coefficients toward zero. Doudchenko and Imbens [2016] formalise this connection, showing that synthetic control is equivalent to constrained regression:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \|\mathbf{Y}_1^{\text{pre}} - \mathbf{Y}_{\mathcal{J}}^{\text{pre}} \mathbf{w}\|_V^2 \quad \text{s.t.} \quad w_j \geq 0, \quad \sum_j w_j = 1, \quad (6.8)$$

where  $\mathbf{Y}_1^{\text{pre}} \in \mathbb{R}^{T_0}$  is the treated unit’s pre-treatment outcomes,  $\mathbf{Y}_{\mathcal{J}}^{\text{pre}} \in \mathbb{R}^{T_0 \times (N-1)}$  is the donor matrix, and  $\|\cdot\|_V$  is a weighted norm determined by predictor importance. The nested optimisation over  $V$  that we discuss later in the chapter governs how closely we match different components of the pre-treatment data (for example, levels vs trends, outcomes vs covariates).

The identification condition (6.5) can be satisfied only if

$$\boldsymbol{\lambda}_1 \in \text{conv}\{\boldsymbol{\lambda}_j : j \in \mathcal{J}\}, \quad (6.9)$$

that is, the treated unit’s factor loadings must lie within the convex hull of the donors’ factor loadings. When this condition fails, no convex combination of donors can match the treated unit. The synthetic control will

then have non-zero bias  $\Delta_\lambda \neq 0$ . Pre-treatment fit will be imperfect, and imperfect fit alone does not reveal the direction or magnitude of post-treatment bias.

In marketing applications, flagship markets are often systematically different from any potential donor market, raising concerns about the convex hull condition and the potential for growing bias. Section 6.5 proposes diagnostics for detecting and responding to such violations.

## A Critical Perspective on Early Synthetic Control Methods

The synthetic control method was introduced by Abadie and Gardeazabal [2003] and developed in the influential California tobacco study Abadie et al. [2010]. These papers established the method’s appeal: transparent weights, visual diagnostics, and placebo checks. However, the early synthetic control literature provides essentially no principled guidance for model selection.

The optimisation problem (6.8) requires choosing which predictors to include (pre-treatment outcomes, covariates, or both), how to weight different predictors (the matrix  $V$  in the nested optimisation), and which donor units to include in  $\mathcal{J}$ . The original Abadie, Diamond, and Hainmueller papers provide little guidance on these choices. The California tobacco study includes several pre-treatment outcome periods and covariates, but the selection is effectively ad hoc. Different predictor sets yield different weights, different pre-treatment fits, and potentially different post-treatment estimates. This creates opportunities for specification search, intentional or inadvertent, on the pre-treatment series.

When practitioners search over specifications until they achieve good pre-treatment fit, they are engaging in in-sample optimisation. This can lead to overfitting to noise, where a specification that fits pre-treatment idiosyncratic shocks will not generalise to post-treatment periods. It can also produce specification-dependent results, with different defensible specifications yielding materially different treatment effect estimates. Perhaps most insidiously, it creates illusory transparency: the reported weights are transparent, but the search process that produced them is not.

In the absence of principled guidance, several pieces of accepted wisdom have emerged that recent research does not support.

One is that “SC is robust to various implementations.” Simulation studies by Ferman and Pinto [2021] and others show that SC can be highly sensitive to donor pool composition, predictor selection, and weighting schemes. The robustness claim is overstated.

A second is that “covariates are unnecessary.” The original formulation emphasises matching pre-treatment outcomes, suggesting covariates add little. In practice, covariates can anchor the synthetic control to units with similar characteristics, reducing extrapolation bias when the outcome series is noisy or the donor pool is heterogeneous.

A third is that “pre-treatment fit guides model selection.” While good pre-treatment fit is necessary for credibility, it is not sufficient for post-treatment validity. Minimising pre-treatment RMSPE can lead to overfitting, and the specification with the best fit need not have the smallest post-treatment bias. Section 6.5

returns to this issue and recommends treating pre-treatment fit as one diagnostic among several, rather than as the sole model-selection criterion.

These findings have direct implications for practice. Practitioners should pre-specify the predictor set and donor pool before examining post-treatment outcomes, report the sensitivity of results to alternative specifications (Section 6.5), and consider augmented or regularised variants that reduce sensitivity to specification choices (Section 6.8). Most importantly, they should treat vanilla SC as a starting point rather than a complete solution.

## SC as Constrained Regression

The connection between synthetic control and regression clarifies the method’s properties. Doudchenko and Imbens [2016] show that unconstrained regression (OLS) permits negative weights and thus extrapolation. Synthetic control, with its non-negative sum-to-one constraints, ensures interpolation through convex weights. Ridge and elastic net approaches regularise weights toward zero or toward a uniform allocation, improving stability. Difference-in-differences represents a special case with equal weights  $w_j = 1/(N - 1)$  on the donors and a parallel trends assumption that justifies using the simple average of controls as the counterfactual. Standard DiD can also work with heterogeneous weights (for example, population weights), but parallel trends remains the identifying restriction regardless of weighting.

This taxonomy shows that DiD is nested within a broader class of weighted control methods. SC replaces the DiD parallel-trends assumption with a factor-structure assumption on untreated potential outcomes and a convex-hull restriction on factor loadings. In favourable designs, these conditions are more plausible than homogeneous parallel trends across all controls; in others, they can be more restrictive. The factor model foundation (Section 6.3) provides the theoretical basis for when SC improves on DiD and when it does not.

## Marketing Applications

Marketing applications motivate synthetic control with concrete use cases in which a single unit or a small number of units receive treatment. A consumer packaged goods brand might launch a television campaign in a single designated market area (DMA) while holding out other DMAs as controls and tracking category sales over several years. A retailer might implement a new store format in a flagship city and observe store traffic and revenue over two years. A digital platform entering a major metropolitan market might seek to estimate the impact on restaurant revenues relative to markets that do not experience entry.

In each case, the treated unit is unique or few in number, donors are plentiful, pre-treatment data are available, and transparency is paramount. However, the flagship market is often systematically different from controls (larger, more competitive, earlier adopting), raising the convex hull concern. Practitioners must assess whether the synthetic control can plausibly approximate the treated unit and whether post-treatment bias is likely to be small.

## Chapter Roadmap

This chapter develops synthetic control methods with attention to both rigorous theory and practical implementation. Section 6.2 explains how to construct the synthetic control, including predictor set selection, weight optimisation, and the role of the nested  $V$ -matrix. Section 6.3 articulates identification assumptions and their connection to factor models formally. Section 6.4 presents inference procedures, including in-space and in-time placebo checks, RMSPE ratios, and permutation-based inference. Section 6.5 outlines diagnostic workflows, goodness-of-fit assessments, and specification sensitivity analysis. Section 6.6 addresses practical issues including donor curation, missing data, anticipation, and spillovers. Section 6.8 discusses extensions to multiple treated units, staggered adoption, and augmented or regularised variants that address the limitations of vanilla SC. Section 6.9 illustrates applications in marketing. Section 6.10 provides a workflow checklist.

See Chapter 4 for DiD estimands and staggered-adoption methods, Chapter 5 for event-study design, and Chapters 8 and 9 for factor and matrix completion methods that provide the theoretical foundation for synthetic control identification.

## 6.2 Constructing the Synthetic Control

Constructing a synthetic control involves selecting predictors, optimising weights to minimise pre-treatment discrepancy, and evaluating the quality of the fit. This section explains each step with attention to the theoretical foundations established in Section 6.1 and to the pitfalls that can undermine credibility.

### Predictor Selection

The predictor set consists of variables used to match the treated unit to the donor pool. Let  $\mathbf{X}_1 \in \mathbb{R}^k$  denote the predictor vector for the treated unit, and let  $\mathbf{X}_0 \in \mathbb{R}^{k \times (N-1)}$  denote the predictor matrix for the donor pool, with columns corresponding to donors. Predictors typically include pre-treatment outcomes  $\mathbf{Y}_1^{\text{pre}} = (Y_{11}, Y_{12}, \dots, Y_{1T_0})'$ , which capture the dynamic trajectory before treatment, encoding trends, seasonality, and the unit's position in the outcome distribution. Pre-treatment covariates  $\mathbf{X}_1^{\text{cov}} = (X_1^{(1)}, X_1^{(2)}, \dots, X_1^{(m)})'$  capture characteristics that predict outcomes but may not be fully reflected in the outcome trajectory, such as market size, demographics, and baseline category shares. The predictor vector is typically constructed as  $\mathbf{X}_1 = (\mathbf{Y}_1^{\text{pre}'}, \mathbf{X}_1^{\text{cov}'})'$ , stacking outcomes and covariates.

The factor model foundation (Section 6.1) reveals why matching pre-treatment outcomes is central to identification. Under the factor model (equation (6.4)), pre-treatment outcomes for untreated units are driven by the factor loadings  $\boldsymbol{\lambda}_i$ . For  $t \leq T_0$ , no unit is treated, so  $Y_{it} = Y_{it}(0)$ . If weights  $\mathbf{w}^*$  match pre-treatment outcomes well, this provides indirect evidence that

$$\sum_{j \in \mathcal{J}} w_j^* \boldsymbol{\lambda}_j \approx \boldsymbol{\lambda}_1. \quad (6.10)$$

Pre-treatment outcome matching is thus a proxy for factor loading matching. This proxy is imperfect when  $T_0$  is small relative to the number of factors  $R$ , when idiosyncratic shocks  $\varepsilon_{it}$  are large relative to factor-driven variation, or when the optimisation over  $V$  (discussed below) emphasises periods dominated by noise.

Covariates can improve pre-treatment fit when pre-treatment outcomes alone are insufficient. If covariates are correlated with factor loadings, matching on covariates helps anchor the synthetic control to units with similar structural characteristics. This is particularly valuable when the pre-treatment period is short, when outcomes are noisy or volatile, or when the donor pool is heterogeneous in ways that covariates capture.

Longer pre-treatment periods improve identification by providing more information to distinguish factor-driven variation from idiosyncratic noise. Theoretical rate conditions [Abadie et al., 2010, Ferman and Pinto, 2021] require  $T_0 \rightarrow \infty$  at an appropriate rate, under fixed  $N$ , for consistency under the factor model with untreated donors and no spillovers. In finite samples, practitioners often use the heuristic that  $T_0$  should exceed the number of factors  $R$  (typically unknown) and should be at least as long as the post-treatment period. In practice,  $R$  is unknown; use the factor selection and diagnostics from Chapter 8 as a guide, and treat the  $T_0 \geq R$  rule as a heuristic rather than a hard requirement.

## The Weight Optimisation Problem

The synthetic control weights  $\mathbf{w}^* = (w_2^*, \dots, w_N^*)'$  are chosen to minimise the discrepancy between the treated unit and the synthetic control in predictors, subject to convexity constraints.

Given a positive semi-definite matrix  $V \in \mathbb{R}^{k \times k}$  that determines predictor importance, the inner optimisation is

$$\mathbf{w}^*(V) = \arg \min_{\mathbf{w}} (\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w})' V (\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}) \quad \text{s.t.} \quad w_j \geq 0, \quad \sum_j w_j = 1. \quad (6.11)$$

This is a quadratic programme with linear constraints, solvable by standard algorithms.

The original Abadie–Diamond–Hainmueller formulation does not fix  $V$  but selects it to minimise pre-treatment prediction error:

$$V^* = \arg \min_V \sum_{t=1}^{T_0} \left( Y_{1t} - \sum_{j \in \mathcal{J}} w_j^*(V) Y_{jt} \right)^2. \quad (6.12)$$

This creates a nested optimisation: the outer problem (6.12) selects  $V$ , and for each candidate  $V$ , the inner problem (6.11) computes  $\mathbf{w}^*(V)$ .

The nested optimisation is computationally more demanding and, more importantly, creates the specification search opportunity discussed in Section 6.1. Different choices of  $V$  yield different weights  $\mathbf{w}^*(V)$ , different pre-treatment fit ( $\text{RMSPE}_{\text{pre}}$ ), and potentially different post-treatment estimates  $\hat{\tau}_{1t}$ . The outer optimisation (6.12) automates the search over  $V$ , but this is in-sample optimisation on pre-treatment data. Note that  $V$  weights \*predictors\* (the components of  $\mathbf{X}_1$ ), but the outer loss function (6.12) uses pre-treatment \*outcomes\*. This couples two distinct objectives, which is why the nested procedure is problematic: from the factor-model perspective,  $V$  should weight predictors in a way that best proxies the latent factor loadings  $\boldsymbol{\lambda}_i$ , but minimising pre-treatment RMSPE is at best an indirect way to achieve this and can fail when noise dominates or when important dimensions of  $\boldsymbol{\lambda}_i$  are weakly reflected in pre-treatment outcomes. There is no guarantee that the selected  $V^*$  yields weights that minimise post-treatment bias. This is where data mining enters the synthetic control method, even when the analyst does not explicitly search over specifications by hand.

*Remark (Cross-validation is not external validation).* Some implementations suggest selecting  $V$  via cross-validation on pre-treatment periods. Cross-validation simply chooses  $V$  to minimise an average of pre-treatment RMSPE across folds. It therefore optimises the same proxy objective as equation (6.12), just with a different splitting scheme. Cross-validation can reduce overfitting to specific pre-treatment periods, but it does not provide out-of-sample validation for post-treatment counterfactuals. Pre-treatment fit is a proxy for factor loading matching, and optimising this proxy does not guarantee post-treatment validity. Cross-validation on pre-treatment data remains in-sample from the perspective of post-treatment inference.

## Connection to Factor Model Identification

The weight optimisation problem is intimately connected to the identification condition from Section 6.1. We now make this connection explicit.

**Proposition 6.1 (Pre-Treatment Matching and Factor Loading Match)** *Suppose outcomes follow the factor model (equation (6.4)) with  $R$  factors and  $\mathbb{E}[\varepsilon_{it}] = 0$ , and suppose the pre-treatment factor matrix  $\{\mathbf{f}_t : t \leq T_0\}$  has full column rank  $R$ . If the convex hull condition (equation (6.9)) holds and  $T_0 \geq R$ , then any convex weights  $\mathbf{w}^*$  that exactly match pre-treatment outcomes in expectation,*

$$\sum_{j \in \mathcal{J}} w_j^* \mathbb{E}[Y_{jt}(0)] = \mathbb{E}[Y_{1t}(0)] \quad \forall t \leq T_0,$$

imply that the identification condition holds:  $\sum_j w_j^* \boldsymbol{\lambda}_j = \boldsymbol{\lambda}_1$ .

*Sketch of argument.* Pre-treatment outcomes for untreated units are linear combinations of factor loadings:  $Y_{it}(0) = \alpha_i + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_i + \varepsilon_{it}$ . Matching  $T_0$  periods of outcomes in expectation (abstracting from idiosyncratic errors) imposes  $T_0$  linear constraints on the factor loading match. When  $T_0 \geq R$  and the pre-treatment factor matrix has full rank, these constraints imply that the weighted average of donor loadings equals the treated unit's loadings. In finite samples with non-negligible idiosyncratic errors, the match is approximate, and pre-treatment fit may be achieved in part by matching noise rather than factor loadings.

Perfect pre-treatment fit in expectation is thus sufficient for identification under the factor model. However, finite-sample idiosyncratic errors mean observed pre-treatment fit may be achieved by matching noise rather than factor loadings. Imperfect pre-treatment fit, captured by a strictly positive pre-treatment RMSPE ( $\text{RMSPE}_{\text{pre}} > 0$ ), implies a factor loading mismatch  $\Delta_\lambda \neq 0$  and hence post-treatment bias. The bias magnitude depends on both  $\Delta_\lambda$  and the post-treatment factor evolution  $\mathbf{f}_t$ .

## Convex Weights and the Interpolation Property

The convexity constraint ( $w_j \geq 0$ ,  $\sum_j w_j = 1$ ) restricts the synthetic control to a convex combination of donors. This ensures interpolation within the convex hull of the donor pool rather than extrapolation beyond it.

For intuition, consider a single latent factor so that the factor loadings are scalars. Suppose two donors have factor loadings  $\lambda_A = 10$  and  $\lambda_B = 2$ . If the treated unit has  $\lambda_1 = 7$ , the convex combination  $w_A = 0.625$ ,  $w_B = 0.375$  produces  $\lambda_1^{\text{sc}} = 0.625 \times 10 + 0.375 \times 2 = 7 = \lambda_1$ . The synthetic control interpolates successfully. If the treated unit has  $\lambda_1 = 12 > \max\{\lambda_A, \lambda_B\}$ , no convex combination can achieve  $\lambda_1^{\text{sc}} = 12$ . The best convex approximation is  $w_A = 1$ ,  $w_B = 0$ , yielding  $\lambda_1^{\text{sc}} = 10$  and bias  $\Delta_\lambda = 2$ .

In marketing applications, the convex hull condition (equation (6.9)) often fails. A flagship market may be larger, more competitive, or earlier-adopting than any control market, placing it outside the convex hull. The synthetic control will then be biased toward the boundary of the hull. Section 6.6 returns to this issue when we discuss diagnostics for convex hull violations.

### Non-Uniqueness and Penalised Synthetic Control

When the number of donors  $N - 1$  exceeds the number of predictors  $k$  (or the number of pre-treatment periods  $T_0$ ), the weight optimisation problem is underdetermined. Multiple weight vectors may achieve the same pre-treatment RMSPE while implying different factor loading matches and different post-treatment counterfactuals.

To address non-uniqueness and reduce overfitting, penalised variants add a regularisation term:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \left[ (\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w})' V (\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}) + \lambda \Omega(\mathbf{w}) \right] \quad \text{s.t. } w_j \geq 0, \sum_j w_j = 1, \quad (6.13)$$

where  $\Omega(\mathbf{w})$  is a penalty function and  $\lambda \geq 0$  is the regularisation strength.

Common penalty choices include a ridge penalty  $\Omega(\mathbf{w}) = \sum_j w_j^2$ , which shrinks weights toward the uniform allocation  $w_j = 1/(N - 1)$ . An entropy penalty  $\Omega(\mathbf{w}) = \sum_j w_j \log w_j$  encourages weight dispersion when weights are constrained to be strictly positive. Elastic net penalties combine ridge-style shrinkage with sparsity-inducing terms. In practice, implementations that use entropy penalties either restrict  $w_j > 0$  or impose a small lower bound  $w_j \geq \varepsilon$  to avoid taking logs of zero. The hyperparameter  $\lambda$  is typically selected via cross-validation on pre-treatment periods. As with  $V$ , this is an in-sample tuning procedure. It trades off pre-treatment fit against weight regularisation but does not provide an external check on post-treatment validity. Different  $\lambda$  values can shift weight mass across donors substantially, meaning the implied counterfactual can change even when  $\text{RMSPE}_{\text{pre}}$  barely changes. Penalised SC reduces sensitivity to specification choices and addresses non-uniqueness, at the cost of potentially worse pre-treatment fit. Because penalties reshape the weight distribution, they change the weight allocation and hence the estimated counterfactual trajectory.

### Pre-Treatment Fit

The standard diagnostic for synthetic control construction is the root mean squared prediction error in the pre-treatment period:

$$\text{RMSPE}_{\text{pre}} = \sqrt{\frac{1}{T_0} \sum_{t=1}^{T_0} \left( Y_{1t} - \sum_{j \in \mathcal{J}} w_j^* Y_{jt} \right)^2}. \quad (6.14)$$

A small pre-treatment RMSPE is necessary for credibility but not sufficient for identification: it shows that the synthetic control tracks the treated unit in-sample, but it does not rule out bias when the factor structure changes or when fit is achieved by matching noise. Section 6.5 provides a systematic treatment of RMSPE-based diagnostics and their limitations.

## Donor Pool Curation

The donor pool  $\mathcal{J}$  must be curated before weight optimisation. This is a consequential design choice that affects results.

Several categories of units require exclusion from the donor pool. First, any unit that received the same or similar treatment cannot serve as a control. Second, units whose outcomes may be affected by the treated unit's intervention—geographic neighbours or direct competitors—must be excluded to avoid spillover contamination. Third, units that differ fundamentally from the treated unit in ways that violate the factor model assumption (different industry, different country) should be removed. Excluding spillover-prone donors protects Assumption 15, and excluding structurally different units protects the factor-model assumption in Assumption 16. In dense platform settings where spillovers may affect many units, excluding all potentially affected donors can be difficult or impossible. In such cases, synthetic control may be targeting a different estimand than the direct unit-level effect, and alternative methods may be preferable.

The remaining challenge is determining which units are “comparable.” Options include units from the same industry or sector, units of similar size or scale (within a factor of 2–3), units with similar pre-treatment trends based on visual inspection, and units with geographic or market similarity.

Donor pool curation is a model selection choice. Different pools yield different weights and potentially different estimates. Practitioners should pre-specify the donor pool before examining post-treatment outcomes, report sensitivity of results to donor pool variations (Section 6.5), and acknowledge that donor pool choice is a consequential assumption.

## Practical Workflow

The practical workflow proceeds as follows. First, assemble data by gathering outcomes  $Y_{it}$  and covariates  $X_{it}$  for treated and donor units over the full sample period. Second, define intervention timing by setting  $T_0$  as the last pre-treatment period. Third, curate the donor pool by excluding treated, spillover-affected, and incomparable units, documenting exclusion criteria. Fourth, select predictors by choosing pre-treatment outcomes and covariates, considering whether to use all pre-treatment periods or selected periods.

Fifth, choose the estimation approach. Options include standard SC with nested optimisation over  $V$  (Abadie–Diamond–Hainmueller), fixed  $V$  with  $V = I$  (identity) or  $V$  diagonal with pre-specified weights, or penalised SC with regularisation to address overfitting and non-uniqueness. Sixth, solve the optimisation to compute weights  $\mathbf{w}^*$  using quadratic programming.

Seventh, evaluate pre-treatment fit by computing  $\text{RMSPE}_{\text{pre}}$ , plotting treated versus synthetic trajectories, and checking predictor balance. Eighth, if fit is poor, consider whether the treated unit lies outside the convex hull. Options include expanding the donor pool, relaxing the convexity constraint (augmented SC, Section 6.8), or acknowledging that SC may not be appropriate for this application. Ninth, report weights by documenting which donors contribute and with what weights, interpreting weight sparsity. Wherever possible,

pre-specify the donor pool, predictor set,  $V$  (or a small grid of candidate  $V$ ), and regularisation choices before examining post-treatment outcomes, and treat any deviations from this plan as exploratory.

The synthetic control's transparency is a key advantage: weights are reported explicitly, enabling readers to assess whether the comparison group is plausible. However, transparency in weights does not imply transparency in the specification search that produced them. Analysts should document and justify all choices (predictor set, donor pool,  $V$  selection, regularisation) to maintain credibility.

### 6.3 Identification and Assumptions

This section provides the formal identification theory for synthetic control methods. We state the identification assumptions precisely, prove the main identification result, discuss conditions for consistency when weights are estimated, and introduce the partial identification framework of Synthetic Parallel Trends. The factor model and bias decomposition from Sections 6.1 and 6.2 provide the foundation. Here we add the formal theorems.

#### Identification Assumptions

We collect the core assumptions required for identification. The notation follows Sections 6.1 and 6.2: unit 1 is treated at time  $T_0 + 1$ ,  $\mathcal{J}$  is the donor pool, and untreated potential outcomes follow the factor model (equation (6.4)).

**Assumption 14 (No Anticipation)** The treated unit's potential outcomes in the pre-treatment period are not affected by the anticipation of future treatment:

$$Y_{1t} = Y_{1t}(0) \quad \text{for all } t \leq T_0.$$

No anticipation asserts that the treated unit does not alter its behaviour in advance of treatment. If anticipation is present, pre-treatment outcomes reflect both the baseline untreated trajectory and the anticipatory response. Matching the treated unit to donors in the pre-treatment period then does not recover the baseline trajectory. In marketing, anticipation is common: a flagship market may stockpile inventory before a campaign launch, customers may delay buying in anticipation of a loyalty programme, or competitors may adjust behaviour in advance of platform entry. Diagnostics include checking whether the synthetic control fits better in early versus late pre-treatment periods.

**Assumption 15 (No Interference)** The treatment applied to the treated unit does not affect the outcomes of donor units:

$$Y_{jt} = Y_{jt}(0) \quad \text{for all } j \in \mathcal{J}, t > T_0.$$

This assumption also requires that treatment is well-defined with no hidden versions.

No interference, as part of SUTVA (Chapter 2), asserts that donor outcomes are unaffected by the treated unit's treatment, so that we observe  $Y_{jt} = Y_{jt}(0)$  for all  $j \in \mathcal{J}$  and  $t$ . If treatment spills over to donors through competitive effects, customer migration, or supply chain linkages, contaminated donors bias the counterfactual. Design-based solutions include excluding likely spillover targets such as nearby markets and direct competitors, and creating buffer zones. Spillovers can be diagnosed via in-space placebo checks for donor units (Section 6.4).

**Assumption 16 (Factor Model Structure)** Untreated potential outcomes follow the factor model (equation (6.4)),

$$Y_{it}(0) = \alpha_i + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_i + \varepsilon_{it},$$

where  $\alpha_i$  is a unit fixed effect,  $\lambda_t$  is a time fixed effect,  $\mathbf{f}_t \in \mathbb{R}^R$  is a vector of common factors,  $\boldsymbol{\lambda}_i \in \mathbb{R}^R$  is a vector of unit-specific factor loadings, and  $\varepsilon_{it}$  are idiosyncratic errors with  $\mathbb{E}[\varepsilon_{it}] = 0$  and  $\mathbb{E}[\varepsilon_{it}^2] \leq \sigma^2 < \infty$ .

The factor model nests additive fixed effects ( $R = 0$ ), linear time trends ( $R = 1$  with  $f_{t1} = t$ ), and general interactive fixed effects. Chapter 8 develops estimation and diagnostics.

**Assumption 17 (Convex Hull Condition)** The treated unit's factor loadings lie within the convex hull of the donor loadings:

$$\boldsymbol{\lambda}_1 \in \text{conv}\{\boldsymbol{\lambda}_j : j \in \mathcal{J}\}.$$

The convex hull condition ensures that there exist non-negative weights that sum to one and reproduce the treated unit's factor loadings. Formally, it guarantees the existence of weights  $\mathbf{w}^* = (w_2^*, \dots, w_N^*)'$  with  $w_j^* \geq 0$  and  $\sum_j w_j^* = 1$  such that  $\sum_j w_j^* \boldsymbol{\lambda}_j = \boldsymbol{\lambda}_1$ . When this condition fails, no convex combination of donors can match the treated unit's factor loadings, and the synthetic control is biased, as in equation (6.7).

## Main Identification Theorem

We now state the main identification result formally, under the assumptions above and the existence of weights that achieve factor loading match.

**Theorem 6.1 (Identification of Treatment Effect)** *Under Assumptions 14–17, suppose there exist weights  $\mathbf{w}^* = (w_2^*, \dots, w_N^*)'$  satisfying*

1. Convexity:  $w_j^* \geq 0$  for all  $j \in \mathcal{J}$  and  $\sum_{j \in \mathcal{J}} w_j^* = 1$
2. Factor loading match:  $\sum_{j \in \mathcal{J}} w_j^* \boldsymbol{\lambda}_j = \boldsymbol{\lambda}_1$
3. Unit fixed effect match:  $\sum_{j \in \mathcal{J}} w_j^* \alpha_j = \alpha_1$

then the synthetic control estimator  $\hat{\tau}_{1t} = Y_{1t} - \sum_{j \in \mathcal{J}} w_j^* Y_{jt}$  identifies the treatment effect in expectation:

$$\mathbb{E}[\hat{\tau}_{1t}] = \tau_{1t} \quad \text{for all } t > T_0.$$

The expectation is taken over the idiosyncratic errors  $\varepsilon_{it}$ , holding the factor structure and loadings fixed. This result is conditional on the existence of convex weights that satisfy the matching conditions. It does not yet address how to estimate such weights from finite pre-treatment data. Consistency of estimated weights is covered in Theorem 6.2.

**Proof** For  $t > T_0$ , we have  $Y_{1t} = Y_{1t}(1) = Y_{1t}(0) + \tau_{1t}$  by definition of the treatment effect. Under Assumption 15,  $Y_{jt} = Y_{jt}(0)$  for all  $j \in \mathcal{J}$ . Thus

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j \in \mathcal{J}} w_j^* Y_{jt} = \tau_{1t} + \left[ Y_{1t}(0) - \sum_{j \in \mathcal{J}} w_j^* Y_{jt}(0) \right].$$

Under Assumption 16, expand the counterfactual error:

$$\begin{aligned} Y_{1t}(0) - \sum_j w_j^* Y_{jt}(0) &= (\alpha_1 + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_1 + \varepsilon_{1t}) - \sum_j w_j^* (\alpha_j + \lambda_t + \mathbf{f}'_t \boldsymbol{\lambda}_j + \varepsilon_{jt}) \\ &= \left( \alpha_1 - \sum_j w_j^* \alpha_j \right) + \lambda_t \left( 1 - \sum_j w_j^* \right) + \mathbf{f}'_t \left( \boldsymbol{\lambda}_1 - \sum_j w_j^* \boldsymbol{\lambda}_j \right) + \varepsilon_{1t} - \sum_j w_j^* \varepsilon_{jt}. \end{aligned}$$

The time fixed effect term vanishes by the adding-up constraint. The factor-loading term vanishes by condition (2). The unit fixed effect term vanishes by condition (3). Taking expectations and using  $\mathbb{E}[\varepsilon_{it}] = 0$  yields  $\mathbb{E}[\hat{\tau}_{1t}] = \tau_{1t}$ .  $\square$

The theorem establishes unbiasedness when the factor loadings are matched exactly by some convex weights. In practice, factor loadings are unobserved and weights are estimated from pre-treatment data. We now discuss conditions under which the estimated synthetic control is consistent.

## Consistency and Rate Conditions

When weights are estimated from pre-treatment data, we need conditions ensuring that the estimated weights converge to weights that satisfy the identification conditions. Rather than derive a full set of asymptotic results in this chapter, we state a simplified version that captures the key intuition and refer the reader to the original sources for proofs.

**Theorem 6.2 (Consistency of Estimated Synthetic Control)** *Under Assumptions 14–17, suppose the factor model has  $R$  factors with  $\text{rank}(\boldsymbol{\Lambda}) = R$ , where  $\boldsymbol{\Lambda} = (\mathbf{f}_1, \dots, \mathbf{f}_{T_0})' \in \mathbb{R}^{T_0 \times R}$ . Suppose  $T_0$  increases and the donor pool size  $N-1$  is either fixed or grows slowly, along sequences where the convex hull condition continues to hold. Then, under regularity conditions on the weight-estimation procedure (for example, uniqueness or stable selection of approximately factor-matching weights as  $T_0$  grows), the synthetic control estimator with estimated weights  $\hat{\mathbf{w}}$  is consistent for  $\tau_{1t}$  for each fixed post-treatment period  $t$ . In particular, the estimation error decreases as  $T_0$  grows, because more pre-treatment periods identify the factor structure more precisely. It also decreases as the donor pool grows, because a larger donor pool provides better coverage of the factor loading space.*

This result summarises the main message from Abadie et al. [2010] and Ferman and Pinto [2021]: more pre-treatment periods improve estimation by providing more information about factor loadings, and larger donor pools improve estimation by providing better coverage of the factor loading space. A practical rule of thumb is that  $T_0$  should exceed the number of factors  $R$  and ideally be at least twice as large. When  $T_0$  is small relative to the noise level, pre-treatment fit may be achieved by matching noise rather than factor loadings.

### Bias When Identification Fails

When the convex hull condition (Assumption 17) fails or weights are imperfectly estimated, the bias decomposition from Section 6.1 applies. The bias equals  $\mathbf{f}_t' \boldsymbol{\Delta}_\lambda$ , where  $\boldsymbol{\Delta}_\lambda = \boldsymbol{\lambda}_1 - \sum_j w_j^* \boldsymbol{\lambda}_j$  is the factor loading mismatch.

If  $\|\boldsymbol{\Delta}_\lambda\| \leq \delta$  and  $\|\mathbf{f}_t\| \leq L$  for all  $t > T_0$ , then

$$|\text{Bias}(\hat{\tau}_{1t})| \leq L \cdot \delta.$$

In practice,  $\delta$  is proxied by pre-treatment fit diagnostics (for example, RMSPE<sub>pre</sub> and pre-treatment gap plots), while  $L$  reflects how volatile the common factors are over the post-treatment window. Large macro shocks or structural breaks correspond to large  $L$ . The bias is bounded by the product of the factor loading mismatch and the magnitude of the factors. This motivates diagnostics that assess pre-treatment fit quality (Section 6.5) and, when possible, external information on how different the treated unit is from the donor pool in terms of underlying market structure.

### Synthetic Parallel Trends and Partial Identification

The framework of Synthetic Parallel Trends [Liu, 2025] provides a unifying perspective that encompasses difference-in-differences, synthetic control, and related methods. When several weighting schemes achieve similar pre-treatment fit, point identification becomes fragile. Synthetic Parallel Trends reframes this fragility as partial identification over an admissible weight set.

Let  $\mathbf{Y}_{\mathcal{J}}^{\text{pre}}$  denote the  $T_0 \times (N - 1)$  matrix of donor outcomes in the pre-treatment period, and let  $\mathbf{Y}_{\mathcal{J},t}$  denote the  $(N - 1)$ -dimensional vector of donor outcomes at time  $t$ .

**Definition 6.1 (Admissible Weight Set)** For a given tolerance  $\epsilon \geq 0$  and a chosen norm  $\|\cdot\|$  on  $\mathbb{R}^{T_0}$  (for example, the Euclidean norm), the set of admissible weights is

$$\mathcal{W} = \left\{ \mathbf{w} \in \mathbb{R}^{N-1} : w_j \geq 0, \sum_{j \in \mathcal{J}} w_j = 1, \|\mathbf{Y}_1^{\text{pre}} - \mathbf{Y}_{\mathcal{J}}^{\text{pre}} \mathbf{w}\| \leq \epsilon \right\}.$$

In applications,  $\epsilon$  is typically chosen in relation to the pre-treatment RMSPE (for example, allowing any weight vector whose pre-treatment RMSPE is within a small factor of the minimum RMSPE achieved by a baseline estimator). When  $\epsilon = 0$ ,  $\mathcal{W}$  contains all convex weights that perfectly match pre-treatment outcomes. When  $\epsilon > 0$ ,  $\mathcal{W}$  includes weights that approximately match within tolerance  $\epsilon$ .

**Definition 6.2 (Identified Set)** The identified set for the treatment effect at time  $t > T_0$  is

$$\mathcal{I}_t = \left\{ \tau : \tau = Y_{1t} - \mathbf{Y}_{\mathcal{J},t}' \mathbf{w} \text{ for some } \mathbf{w} \in \mathcal{W} \right\}.$$

When  $\mathcal{W}$  is non-empty and convex, the mapping  $\mathbf{w} \mapsto Y_{1t} - \mathbf{Y}_{\mathcal{J},t}' \mathbf{w}$  is linear, so  $\mathcal{I}_t$  is a closed interval  $[\underline{\tau}_t, \bar{\tau}_t]$ .

**Theorem 6.3 (Bounds on Treatment Effect)** *If  $\mathcal{W} \neq \emptyset$ , the identified set  $\mathcal{I}_t$  is a closed interval with bounds*

$$\underline{\tau}_t = Y_{1t} - \max_{\mathbf{w} \in \mathcal{W}} \mathbf{Y}'_{\mathcal{J},t} \mathbf{w}, \quad (6.15)$$

$$\bar{\tau}_t = Y_{1t} - \min_{\mathbf{w} \in \mathcal{W}} \mathbf{Y}'_{\mathcal{J},t} \mathbf{w}. \quad (6.16)$$

*If, in addition,  $\epsilon = 0$  and  $\mathcal{W}$  contains a unique admissible weight vector  $\mathbf{w}^*$ , then  $\mathcal{I}_t$  collapses to a point given by  $Y_{1t} - \mathbf{Y}'_{\mathcal{J},t} \mathbf{w}^*$ . Linking this point to the causal estimand  $\tau_{1t}$  requires additional identification structure, such as the factor model assumptions used earlier in the section.*

The bounds are linear programmes over the convex set  $\mathcal{W}$ . The minimum and maximum of a linear function over a convex set are attained at extreme points. When the identification conditions hold and there is a unique admissible weight vector, the bounds coincide.

When different estimators such as DiD, SC, and SDID yield different treatment effect estimates, they are selecting different weights from  $\mathcal{W}$ . Formally, each estimator corresponds to a particular choice of weights  $\mathbf{w}^{\text{DiD}}$ ,  $\mathbf{w}^{\text{SC}}$ ,  $\mathbf{w}^{\text{SDID}}$  in  $\mathcal{W}$  when their pre-treatment fit falls within the tolerance  $\epsilon$ . Synthetic Parallel Trends treats these as three of many admissible weighting schemes rather than privileging any single one. Under Synthetic Parallel Trends, all weights in  $\mathcal{W}$  are equally admissible, and the analyst reports bounds  $[\underline{\tau}_t, \bar{\tau}_t]$  rather than privileging a single estimator.

Consider a retailer running a television campaign in one flagship DMA with twenty potential control DMAs. SC, which uses optimised convex weights matching pre-treatment outcomes, estimates a lift of 12 per cent with excellent pre-treatment fit. DiD, which uses uniform weights, estimates 6 per cent but has materially worse pre-treatment fit and falls outside the admissible weight set  $\mathcal{W}$  for the chosen tolerance  $\epsilon$ . Under Synthetic Parallel Trends, the analyst computes bounds over all admissible weights (those with pre-treatment fit comparable to SC) and reports the interval [8 per cent, 13 per cent]. The SC estimate lies near the middle of this range, and the analyst concludes that the campaign almost certainly increased sales between 8 and 13 per cent, with the precise magnitude depending on the weighting scheme within the admissible set.

## Threats to Identification

Several conditions can undermine the identification assumptions. Each threat corresponds directly to a failure of one or more assumptions above. Poor pre-treatment fit signals convex-hull or factor-structure problems (Assumptions 16–17). Anticipation violates Assumption 14. Spillovers violate Assumption 15. Structural breaks undermine Assumption 16.

Poor pre-treatment fit, captured by a large RMSPE in the pre-treatment period, signals that the synthetic control may not approximate the counterfactual. This can indicate that the treated unit lies outside the convex hull or that the pre-treatment period is too short to identify the factor structure. Options include

expanding the donor pool, extending the pre-treatment period, or using augmented methods that relax convexity (Section 6.8).

Anticipation violates Assumption 14. If the treated unit adjusts behaviour before  $T_0$ , pre-treatment outcomes reflect anticipatory effects. The synthetic control then matches the anticipation-contaminated trajectory, not the baseline. Diagnostics include checking whether fit is better in early versus late pre-treatment periods. Solutions include truncating the pre-treatment period to exclude anticipated periods or redefining the estimand to include anticipation.

Spillovers violate Assumption 15. If treatment affects donor outcomes, the counterfactual is contaminated. Diagnostics include in-space placebo checks for donors (Section 6.4). Solutions include excluding likely spillover targets, creating buffer zones, or modelling spillovers explicitly (Chapter 11).

Structural breaks threaten the factor model structure. If the factor structure shifts at or shortly after treatment because of macroeconomic shocks, regulatory changes, or competitive disruption, the pre-treatment weights may not apply post-treatment. Diagnostics include checking whether donor trajectories remain stable post-treatment. Solutions include restricting the post-treatment window, incorporating richer factor dynamics, or using methods that are robust to structural change.

Model selection, discussed in Sections 6.1 and 6.2, affects identification through the choice of predictors, donor pool, and estimation approach. Pre-specification and systematic sensitivity analysis mitigate specification search concerns and make the identification assumptions transparent.

## Summary

Identification in synthetic control requires no anticipation (Assumption 14), no interference (Assumption 15), factor model structure (Assumption 16), the convex hull condition (Assumption 17), and enough pre-treatment periods to identify the factor structure. Under these conditions, Theorem 6.1 establishes unbiasedness and Theorem 6.2 establishes consistency when weights are estimated from pre-treatment data. When point identification fails or different estimators select different admissible weights, Theorem 6.3 characterises the identified set for treatment effects under Synthetic Parallel Trends, rather than relying on a single point estimate. The next section addresses inference and uncertainty quantification.

## 6.4 Inference for Synthetic Control

Inference for synthetic control methods quantifies uncertainty in treatment effect estimates and assesses whether the observed post-treatment gap is unusually large. This section presents inference procedures with attention to the theoretical foundations from Sections 6.1–6.3. We cover permutation-based inference, conformal inference, analytical variance decompositions, and inference for bounds under Synthetic Parallel Trends.

### The Inference Challenge

Classical regression-based inference, built on large numbers of treated units and simple weighting schemes, does not translate directly to synthetic control. In typical SC applications there is a single treated unit, the synthetic control weights depend on the data in complex ways, and estimation error has both factor loading and idiosyncratic components. In particular, the primary estimand is a path of unit-specific effects  $\{\tau_{1t}\}_{t>T_0}$ , not an average over many treated units, so large- $N$  asymptotics in the number of treated units do not apply. Design-based methods that exploit the panel structure, and residual-based methods such as conformal inference, provide alternatives that respect these features. Analytical approximations are still useful, but they must be interpreted in light of the identification structure described in Section 6.3.

### Permutation-Based Inference

The classical approach to SC inference uses permutation-based inference built around in-space placebo checks. We test the sharp null hypothesis of no treatment effect for any unit,

$$H_0 : Y_{it}(1) = Y_{it}(0) \quad \text{for all } i, t > T_0. \quad (6.17)$$

Under  $H_0$ , and conditional on the set of units deemed eligible for treatment, treatment assignment is exchangeable among units: any eligible unit could have been treated at  $T_0 + 1$ . The observed post-treatment gap for the treated unit is then one realisation from the distribution of gaps that would occur under random assignment. This sharp null is strong: it asserts no effect for any unit in any post-treatment period. In observational settings, it should be interpreted as a device for calibrating how unusual the treated unit's post-treatment gap is relative to donor units, rather than as a literal claim about effects for all units. In observational settings this p-value is not a probability statement about treatment assignment. It is a calibration of extremeness relative to the donor pool under an exchangeability heuristic.

**In-Space Placebo Checks.** For each donor  $j \in \mathcal{J}$ , we treat donor  $j$  as if it received treatment at  $T_0 + 1$ , construct a synthetic control for  $j$  using the remaining donors, and compute the post-treatment gap  $\hat{\tau}_{jt}^{\text{placebo}} = Y_{jt} - \hat{Y}_{jt}^{\text{syn}}$  for  $t > T_0$ . The distribution of placebo-check gaps  $\{\hat{\tau}_{jt}^{\text{placebo}} : j \in \mathcal{J}\}$  provides the reference distribution.

**RMSPE Ratios.** Building on the pre-treatment RMSPE defined in equation (6.14), define the post-treatment RMSPE for unit  $i$  as

$$\text{RMSPE}_{i,\text{post}} = \sqrt{\frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{\tau}_{it}^2}, \quad (6.18)$$

and the RMSPE ratio as

$$R_i = \frac{\text{RMSPE}_{i,\text{post}}}{\text{RMSPE}_{i,\text{pre}}}. \quad (6.19)$$

The ratio normalises the post-treatment gap by the pre-treatment fit quality, making gaps comparable across units with different pre-treatment fit. We denote the treated unit's ratio by  $R_1$ . If the treated unit's pre-treatment fit is substantially better than most placebo-check units, ranks on raw post-treatment gaps are misleading. Ratios partially correct this by normalising for fit quality.

**Rank-Based P-Values.** Let  $r$  denote the rank of the treated unit's RMSPE ratio among all  $N$  units, with  $r = 1$  corresponding to the smallest ratio and  $r = N$  to the largest. For an upper-tail test (testing whether the treated unit's gap is unusually large),

$$p_{\text{upper}} = \frac{N + 1 - r}{N}. \quad (6.20)$$

For a two-sided test,

$$p_{\text{two-sided}} = \frac{2 \cdot \min(r, N + 1 - r)}{N}. \quad (6.21)$$

If the treated unit has the largest RMSPE ratio ( $r = N$ ), the upper-tail p-value is  $p = 1/N$ . With  $N - 1$  donors ( $N$  total units), the smallest achievable p-value is  $1/N$ , which can be relatively large when the donor pool is small. With small donor pools, permutation-based inference therefore has limited power: even if the treated unit's ratio is an extreme outlier, the p-value may remain well above conventional significance thresholds.

**Limitations of Permutation Inference.** Permutation inference faces several limitations in practice. Resolution is coarse with small donor pools: with  $N = 10$  total units, the smallest p-value is  $1/10 = 0.10$ . The sharp null assumes treatment could have been assigned to any unit, which may not hold in observational settings where exchangeability is questionable. Finally, if some placebo-check units have much worse pre-treatment fit than the treated unit, they are not comparable benchmarks.

*Remark 6.1 (On Excluding Poor-Fit Placebo Checks)* A common practice is to exclude placebo-check units with poor pre-treatment fit from the reference distribution. This is problematic because the permutation distribution is defined conditional on the full set of units and their pre-treatment data. Selecting placebo-check units based on  $\text{RMSPE}_{\text{pre}}$ , which is itself a function of those data, changes the distribution of the test statistic in ways that the naive p-value does not account for. A safer approach is to use RMSPE ratios, which account for differential fit, or to use conformal inference, which does not rely on exchangeability of treatment assignment across units. From an identification perspective, excluding poor-fit placebo-check units amounts to conditioning on part of the pre-treatment fit information without adjusting the reference

distribution accordingly. Using RMSPE ratios or conformal methods preserves the design-based logic while accommodating heterogeneous pre-treatment fit.

## Conformal Inference

Conformal inference [Chernozhukov et al., 2021] provides a modern alternative that does not require the sharp null of no effect for all units and can remain valid with heterogeneous placebo-check gaps. Rather than assuming that treatment could have been assigned to any unit, conformal methods use the distribution of residuals for the treated unit in the pre-treatment period as a reference for post-treatment residuals, under a symmetry or exchangeability condition on these residuals. The key assumption is that, under the null for a given post-treatment period, the distribution of residuals for that period is exchangeable with the distribution of pre-treatment residuals for the treated unit (possibly in blocks to account for serial correlation).

Define the residual for unit  $i$  at time  $t$  as

$$e_{it} = Y_{it} - \hat{Y}_{it}^{\text{syn}}. \quad (6.22)$$

In the pre-treatment period, these residuals reflect fit quality. Under the null of no treatment effect for the treated unit at time  $t$ , the post-treatment residual  $e_{1t}$  should be similar in distribution to the pre-treatment residuals  $\{e_{1s} : s \leq T_0\}$ , possibly after accounting for serial correlation via blocks.

**Conformal P-Value and Confidence Interval.** For testing  $H_0 : \tau_{1t} = \tau_0$  at time  $t$ , compute the adjusted outcome  $\tilde{Y}_{1t} = Y_{1t} - \tau_0$ , then compute the residual  $\tilde{e}_{1t} = \tilde{Y}_{1t} - \hat{Y}_{1t}^{\text{syn}}$ . Compare  $|\tilde{e}_{1t}|$  to the distribution of  $|e_{1s}|$  for  $s \leq T_0$ . The conformal p-value is the fraction of pre-treatment residuals with  $|e_{1s}| \geq |\tilde{e}_{1t}|$ .

Inverting this test yields a  $(1 - \alpha)$  confidence interval for  $\tau_{1t}$ ,

$$CI_{1-\alpha}(\tau_{1t}) = \{\tau_0 : p(\tau_0) \geq \alpha\}, \quad (6.23)$$

which contains all values of  $\tau_0$  that cannot be rejected at level  $\alpha$ . The `scpi` package [Cattaneo et al., 2021] (R, Stata, Python) developed by Cattaneo, Feng, and Titiunik implements these procedures, including block structures for serial correlation and both pointwise and uniform confidence bands. Conformal methods quantify uncertainty conditional on the synthetic control specification. They do not alter the identification assumptions from Section 6.3. If those assumptions fail, conformal intervals will be centred on a biased counterfactual.

If the pre-treatment residual distribution exhibits nonstationarity (for example, drift in the mean or variance), conformal coverage guarantees may not hold. In such cases, consider using only recent pre-treatment periods or employing weighted conformal methods that downweight distant periods.

## Analytical Variance Decomposition

Under the factor model (Assumption 16), the variance of the SC estimator can be decomposed into an idiosyncratic component and a component arising from uncertainty about the factor structure. Following Abadie et al. [2010, Proposition 2], and related work, one can write schematically

$$\text{Var}(\hat{\tau}_{1t}) \approx \sigma^2 \left( 1 + \sum_{j \in \mathcal{J}} (w_j^*)^2 \right) + \mathbf{f}'_t \text{Var}(\hat{\Delta}_\lambda) \mathbf{f}_t, \quad (6.24)$$

where  $\sigma^2 = \text{Var}(\varepsilon_{it})$  is the idiosyncratic error variance,  $\sum_j (w_j^*)^2$  is the “effective sample size” adjustment (the Herfindahl index of weights),  $\mathbf{f}_t$  is the factor vector at time  $t$ , and  $\text{Var}(\hat{\Delta}_\lambda)$  is the variance of the estimated factor loading mismatch. This decomposition is approximate and relies on linearisation and further regularity conditions. In most applications, practitioners rely on bootstrap or conformal methods implemented in software such as `scpi` rather than on closed-form variance formulas.

The first term reflects variance from idiosyncratic shocks. Sparse weights (few donors with large weights) increase this term. The second term reflects uncertainty in factor loading estimation and decreases as  $T_0$  grows and the factor structure is better identified. In practice,  $\sigma^2$  is estimated from pre-treatment residuals and the factor loading variance is approximated via bootstrap or analytical methods. See the `scpi` package and associated documentation for implementation details.

## In-Time Placebo Checks

In-time placebo checks assess stability by applying the SC method with pseudo-intervention dates in the pre-treatment period. Choose a pseudo-intervention date  $T_0^* < T_0$ , use periods  $1, \dots, T_0^*$  as the pseudo pre-treatment window, construct synthetic control weights using this window, and compute pseudo-gaps for periods  $T_0^* + 1, \dots, T_0$ .

If pseudo-gaps are near zero, this supports stability: the synthetic control continues to track the treated unit even outside the fitting window. Large pseudo-gaps suggest the synthetic control overfits the pre-treatment period or that the factor structure is unstable. In-time placebo checks are diagnostics, not formal tests. They provide evidence on whether the synthetic control is a stable counterfactual, but they cannot definitively validate the method. Consistently small pseudo-gaps across multiple pseudo-intervention dates support the assumption that the factor structure governing untreated outcomes is stable over time.

## Inference for Bounds under Synthetic Parallel Trends

Section 6.3 introduced the identified set  $\mathcal{I}_t = [\underline{\tau}_t, \bar{\tau}_t]$  under Synthetic Parallel Trends. When point identification fails, we need inference procedures for these bounds.

One approach is to construct a  $(1 - \alpha)$  confidence set for the identified set by combining point estimates of the bounds with uncertainty measures:

$$CS_{1-\alpha} = [\underline{\tau}_t - c_\alpha \hat{\sigma}_{\underline{\tau}}, \bar{\tau}_t + c_\alpha \hat{\sigma}_{\bar{\tau}}], \quad (6.25)$$

where  $c_\alpha$  is an appropriate critical value (for example, a quantile from a bootstrap distribution or, in large samples, a normal approximation) and  $\hat{\sigma}_{\underline{\tau}}, \hat{\sigma}_{\bar{\tau}}$  are estimated standard errors for the lower and upper bounds. This formula provides a heuristic confidence set for the identified set. For formal treatment of inference for partially identified parameters, see Chapter 16. The object  $CS_{1-\alpha}$  is a confidence set for the identified set itself: with probability approximately  $1 - \alpha$ , it contains the true interval  $[\underline{\tau}_t, \bar{\tau}_t]$ , not just a single point. Even when  $CS_{1-\alpha}$  is relatively tight, the treatment effect  $\tau_{1t}$  remains only partially identified if  $\underline{\tau}_t < \bar{\tau}_t$ . Reporting both the bounds and the confidence set makes this distinction transparent.

When DiD and SC give different estimates, analysts should report point estimates from each method, the identified set bounds  $[\underline{\tau}_t, \bar{\tau}_t]$ , and the associated confidence set, and then discuss which weighting schemes are more credible given the marketing context.

## Multiple Testing

Testing many post-treatment periods or outcomes raises multiple testing concerns. Pointwise inference tests each period separately at level  $\alpha$ , which is simple but inflates family-wise error. Uniform inference controls error across all periods simultaneously, using either Bonferroni correction (conservative) or sup- $t$  bands (tighter). See Chapter 16 for the general multiplicity framework. For dynamic treatment paths, see also the event-study multiplicity discussion in Chapter 5, which applies directly when using SC-based event-time effects.

For SC with many post-treatment periods, analysts should report both pointwise and uniform confidence bands, focus on cumulative or average effects rather than period-by-period tests, and use sup- $t$  bands from conformal inference when available.

## Practical Guidance

Several principles guide inference for synthetic control in practice.

**Visual Placebo-Check Comparison.** Always conduct in-space placebo checks and plot the treated unit's gap alongside all placebo-check gaps. Visual comparison is powerful and intuitive.

**Use Normalised Statistics.** Report RMSPE ratios rather than raw gaps. Ratios account for differential pre-treatment fit and make comparisons meaningful across units.

**Formal Inference via Conformal Methods.** Use conformal inference for formal confidence intervals. Conformal methods provide valid inference under symmetry conditions on residuals without requiring exchangeability of treatment assignment across units.

**Stability Assessment.** Conduct in-time placebo checks to assess stability of the synthetic control across different fitting windows.

**Report Intervals and Identified Sets.** Confidence intervals convey both statistical significance and the magnitude of uncertainty. When point identification is uncertain (for example, when DiD and SC disagree), report the identified set  $[\underline{\tau}_t, \bar{\tau}_t]$  with confidence sets. When possible, define a small number of primary estimands (for example, an average post-treatment effect or a cumulative effect) ex ante and treat period-by-period tests as exploratory. This keeps the multiple-testing burden manageable.

**Exercise Caution When Excluding Placebo Checks.** If exclusion is necessary, use RMSPE ratios or conformal inference to avoid invalidating the reference distribution.

Inference for synthetic control combines design-based and residual-based methods to quantify uncertainty. By combining permutation-based inference, conformal inference, and bounds analysis, practitioners can draw credible conclusions without relying solely on large-sample approximations. The next section develops diagnostic procedures for assessing synthetic control quality.

## 6.5 Diagnostics and Goodness of Fit

Credible synthetic control analysis requires rigorous diagnostics that assess the quality of the pre-treatment fit, the sensitivity of conclusions to specification choices, and the plausibility of identification assumptions. This section connects diagnostics to the identification theory from Sections 6.1–6.3 and provides practical guidance.

### Connection to Identification Theory

Diagnostics in synthetic control are not merely descriptive. They help you assess whether the identification assumptions are plausible in a given application. Recall the bias decomposition from Section 6.1 (equation 6.7), where the bias equals  $\mathbf{f}'_t \Delta_\lambda$  and  $\Delta_\lambda = \boldsymbol{\lambda}_1 - \sum_j w_j^* \boldsymbol{\lambda}_j$  is the factor loading mismatch.

The fundamental diagnostic question is: *How large is  $\Delta_\lambda$ ?* Since factor loadings are unobserved, we cannot compute  $\Delta_\lambda$  directly. Pre-treatment fit metrics serve as proxies.

**RMSPE as Proxy for Factor Loading Mismatch.** Under the factor model (Assumption 16), pre-treatment outcomes satisfy

$$Y_{1t} - \sum_j w_j^* Y_{jt} = \mathbf{f}'_t \Delta_\lambda + \text{noise}_t,$$

where  $\text{noise}_t = \varepsilon_{1t} - \sum_j w_j^* \varepsilon_{jt}$ . The pre-treatment RMSPE aggregates this discrepancy:

$$\text{RMSPE}_{\text{pre}}^2 = \frac{1}{T_0} \sum_{t=1}^{T_0} (\mathbf{f}'_t \Delta_\lambda + \text{noise}_t)^2.$$

A small  $\text{RMSPE}_{\text{pre}}$  is therefore necessary for small bias: it indicates that the weighted donors track the treated unit closely in pre-treatment periods. But it is not sufficient.  $\text{RMSPE}_{\text{pre}}$  is an empirical proxy for the average magnitude of  $\mathbf{f}'_t \Delta_\lambda$  on the pre-treatment path, not a direct measure of  $\|\Delta_\lambda\|$ . If  $\mathbf{f}_t$  changes after treatment, or if noise partly offsets mismatch in the pre-treatment period, a small  $\text{RMSPE}_{\text{pre}}$  can still coexist with substantial post-treatment bias. In terms of the bias bound  $|\text{Bias}(\hat{\tau}_{1t})| \leq L \cdot \delta$  from Section 6.3,  $\text{RMSPE}_{\text{pre}}$  is an empirical proxy for  $\delta$ , while  $L$  reflects how volatile the common factors  $\mathbf{f}_t$  are over the post-treatment window.

### Pre-Treatment Fit Metrics

The root mean squared prediction error in the pre-treatment period is defined in Section 6.2, equation (6.14). To interpret  $\text{RMSPE}_{\text{pre}}$ , it is useful to scale it relative to outcome variability:

$$\text{Relative RMSPE} = \frac{\text{RMSPE}_{\text{pre}}}{\text{SD}(Y_{1,\text{pre}})}, \quad (6.26)$$

where  $\text{SD}(Y_{1,\text{pre}})$  is the standard deviation of the treated unit's pre-treatment outcomes over  $t \leq T_0$ .

Relative RMSPE is easiest to interpret when you benchmark it against decision-relevant effect sizes. If the typical pre-treatment discrepancy implied by the fit is large enough to change a managerial decision, treat the design as fragile. These benchmarks are heuristics, not formal cutoffs, and they should be interpreted in context.

## Convex Hull Diagnostic

Identification requires the convex hull condition (Assumption 17):  $\lambda_1 \in \text{conv}\{\lambda_j\}$ . Since factor loadings are unobserved, we diagnose this indirectly.

When  $\text{RMSPE}_{\text{pre}}$  remains large despite optimisation, when weights pile up on a single donor, or when the treated unit's predictors are extreme relative to donors, these are signals that the convex hull condition may fail. The PCA diagnostic—plotting treated and donor units in the first two principal components of the predictor matrix—provides a visual check: if the treated unit lies outside the donor cloud in this space, extrapolation is likely. PCA on predictors diagnoses predictor-space extremeness, not factor-loading extremeness. It is a proxy for convex-hull concerns, not a direct test of the identification condition. Because PCA reduces the predictor space to a low-dimensional projection, lying inside the donor cloud in the first two components does not guarantee that the convex hull condition holds in the full predictor space. Treat this diagnostic as suggestive rather than definitive. These diagnostics complement the convex hull discussion in Sections 6.1 and 6.2.

When the convex hull condition fails, several options are available. One can expand the donor pool to include more extreme units. Alternatively, augmented synthetic control (Section 6.8) allows extrapolation with bias correction. Finally, analysts may need to acknowledge that SC may not be appropriate for the application.

## Bounds Width as Diagnostic

Section 6.3 introduced the identified set  $\mathcal{I}_t = [\underline{\tau}_t, \bar{\tau}_t]$  under Synthetic Parallel Trends. The width of the bounds provides a diagnostic for identification strength:

$$\text{Width}_t = \bar{\tau}_t - \underline{\tau}_t. \quad (6.27)$$

Narrow bounds indicate that pre-treatment data strongly constrain the counterfactual. Different weighting schemes (DiD, SC, SDID) should agree. Wide bounds indicate that many weighting schemes are consistent with the data, suggesting weak identification where conclusions depend on which weights one privileges. Note

that bounds width depends on the tolerance  $\epsilon$  and norm choice in Definition 6.1, so analysts should report the chosen tolerance and assess sensitivity to it.

If bounds width is large relative to the magnitude of plausible effects, identification uncertainty dominates estimation uncertainty. In such cases, report bounds rather than point estimates, and discuss which weighting scheme is most credible. A bound that rules out zero may still be informative even when the width is large, so bounds width should be interpreted alongside the position of the bounds, not as a standalone pass-fail rule.

## Visual Diagnostics

Visual diagnostics provide intuitive evidence on fit quality and treatment effects.

**Trajectory Plot.** Plot the treated unit’s outcomes and the synthetic control’s outcomes over the full sample period. The vertical line marks the intervention. If trajectories align pre-treatment and diverge post-treatment, this provides visual evidence of both good fit and a treatment effect.

**Gap Plot.** Plot  $\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^{\text{syn}}$  over time. The pre-treatment gap should fluctuate around zero. Post-treatment, the gap should be large and persistent if treatment had an effect.

**In-Space Placebo-Check Gap Plot.** Overlay the treated unit’s gap and all placebo-check gaps (from in-space placebo checks, Section 6.4). If the treated unit’s post-treatment gap is an outlier relative to placebo-check gaps, this provides visual evidence of effect.

## Weight Diagnostics

Report the synthetic control weights  $w_j^*$  explicitly. Compute the effective number of donors:

$$N_{\text{eff}} = \frac{1}{\sum_j (w_j^*)^2}. \quad (6.28)$$

This is the reciprocal of the Herfindahl index of weights and corresponds to the weight concentration term in the variance decomposition (equation 6.24). A small  $N_{\text{eff}}$  corresponds to a large Herfindahl index and thus inflates the idiosyncratic-variance component  $\sigma^2(1 + \sum_j (w_j^*)^2)$  in equation (6.24). When  $N_{\text{eff}}$  is close to  $N - 1$ , weights are diffuse and the synthetic control averages broadly. When  $N_{\text{eff}}$  is close to 1, weights are concentrated and the synthetic control is essentially a single-donor comparison. Sparse weights (small  $N_{\text{eff}}$ ) may be appropriate if a few donors closely match the treated unit—for example, in a natural matched-pair setting—but they increase sensitivity to individual donors.

**Predictor Balance.** Compare the treated unit's predictors to the synthetic control's predictors (weighted average of donor predictors):

$$\text{Balance}_k = X_{1k} - \sum_j w_j^* X_{jk} \quad (6.29)$$

for each predictor  $k$ , where  $X_{1k}$  is the treated unit's value on predictor  $k$  and  $X_{jk}$  are the donor values. Good balance (small discrepancy) indicates the synthetic control matches the treated unit on observed characteristics.

## Sensitivity Analyses

**Leave-One-Donor-Out.** For each donor  $j \in \mathcal{J}$ , re-estimate the synthetic control excluding  $j$  and compute the treatment effect estimate on a chosen summary estimand, for example the average post-treatment effect:

$$\hat{\tau}_1^{\text{avg}} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{\tau}_{1t}, \quad \hat{\tau}_1^{\text{avg}(-j)} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{\tau}_{1t}^{(-j)}.$$

The choice of summary estimand (for example, average post-treatment effect, cumulative effect, or an effect at a specific horizon) should be aligned with the substantive decision problem. Define the influence of donor  $j$  as

$$\text{Influence}_j = \left| \hat{\tau}_1^{\text{avg}} - \hat{\tau}_1^{\text{avg}(-j)} \right|. \quad (6.30)$$

If any donor has large influence, conclusions are sensitive to that donor's inclusion. From an identification perspective, if a single donor drives results, the no-interference and comparability assumptions effectively collapse to that donor, which is a stronger claim than "many donors average out". Report the range of  $\hat{\tau}_1^{\text{avg}(-j)}$  across all leave-one-out specifications, identify which donors have largest influence, and assess whether high-influence donors are credible comparisons.

**Sensitivity to Predictor Sets.** Re-estimate the synthetic control using alternative predictor sets: all pre-treatment periods versus selected periods, including versus excluding covariates, and levels versus logs versus detrended outcomes. Report results across specifications. Stability across specifications supports robustness.

**Sensitivity to Donor Pool.** Vary the donor pool by excluding geographic neighbours (potential spillovers), restricting to same industry or sector, and expanding to include borderline-comparable units. Report how estimates change. Sensitivity to donor pool is a form of model dependence. Large swings across donor pools or predictor sets are a symptom of the model-selection sensitivity highlighted in Section 6.1. They should prompt a more cautious interpretation of any single SC specification.

## In-Time Placebo Checks and Stability

In-time placebo checks, as defined in Section 6.4, assess whether the synthetic control provides a stable counterfactual over extended pre-treatment periods. Choose pseudo-intervention dates  $T_0^* < T_0$ . Fit synthetic control using periods  $1, \dots, T_0^*$ , compute pseudo-gaps for periods  $T_0^* + 1, \dots, T_0$ , and assess whether pseudo-gaps are near zero.

Near-zero pseudo-gaps support the stability assumption: the synthetic control continues to track the treated unit outside the fitting window. Large pseudo-gaps suggest overfitting or an unstable factor structure.

## Overfitting Diagnostics

Cross-validation within the pre-treatment period helps diagnose overfitting. Split the pre-treatment period into training and validation sets. Use periods  $1, \dots, T_{\text{train}}$  to estimate weights, then compute RMSPE on validation periods  $T_{\text{train}} + 1, \dots, T_0$ . If validation RMSPE is much larger than training RMSPE, the weights are overfitted to idiosyncratic noise.

Overfitting risk increases when  $N - 1 > T_0$  (more donors than pre-treatment periods), when the predictor set is large relative to  $T_0$ , and when the nested optimisation over  $V$  is unrestricted. Penalised synthetic control (Section 6.2) addresses overfitting by regularising the weights. Cross-validation can flag overfitting to pre-period noise, but it cannot validate post-treatment counterfactuals. It remains an in-sample diagnostic from the perspective of causal inference. Use cross-validation to flag overfitting, not to pick a single “best” specification to report. Overfitting diagnostics should feed into a broader sensitivity analysis rather than end the analysis.

## Diagnostic Summary Table

**Table 6.1** Synthetic Control Diagnostic Checklist (Illustrative Heuristics)

Diagnostic	What It Assesses	Warning Signal
$\text{RMSPE}_{\text{pre}}$	Pre-treatment fit	Large relative to $\text{SD}(Y_{1,\text{pre}})$
Relative RMSPE	Fit relative to variability	Substantial fraction of $\text{SD}$ (large relative to decision-relevant effects)
Bounds width	Identification strength	Width comparable to plausible effect magnitudes
$N_{\text{eff}}$	Weight concentration	Very small ( $\approx 1\text{-}2$ donors)
Predictor balance	Covariate matching	Large discrepancies
Leave-one-out range	Donor sensitivity	Wide range
In-time placebo-check gaps	Stability	Large pseudo-gaps
CV validation RMSPE	Overfitting	$\gg$ training RMSPE

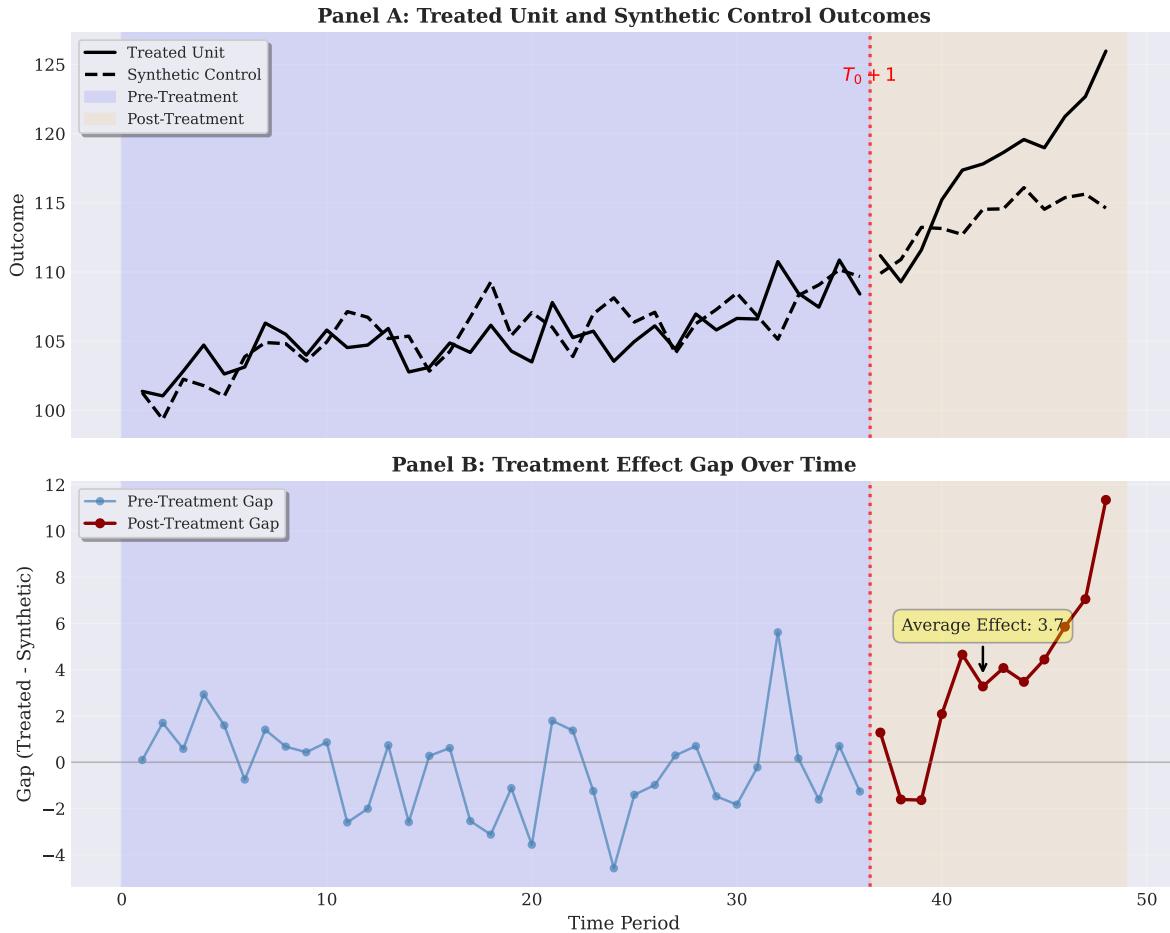
These thresholds are illustrative heuristics rather than universal cutoffs. They are intended to flag situations that merit closer scrutiny, not to provide pass–fail rules.

## Practical Workflow

The diagnostic workflow proceeds in stages. First, assess pre-treatment fit by computing  $\text{RMSPE}_{\text{pre}}$  and relative RMSPE. If fit is poor, consider expanding the donor pool or using augmented SC. Second, check the convex hull by plotting the treated unit versus donors in predictor space and checking for boundary weights. Third, if using Synthetic Parallel Trends, compute bounds and assess width.

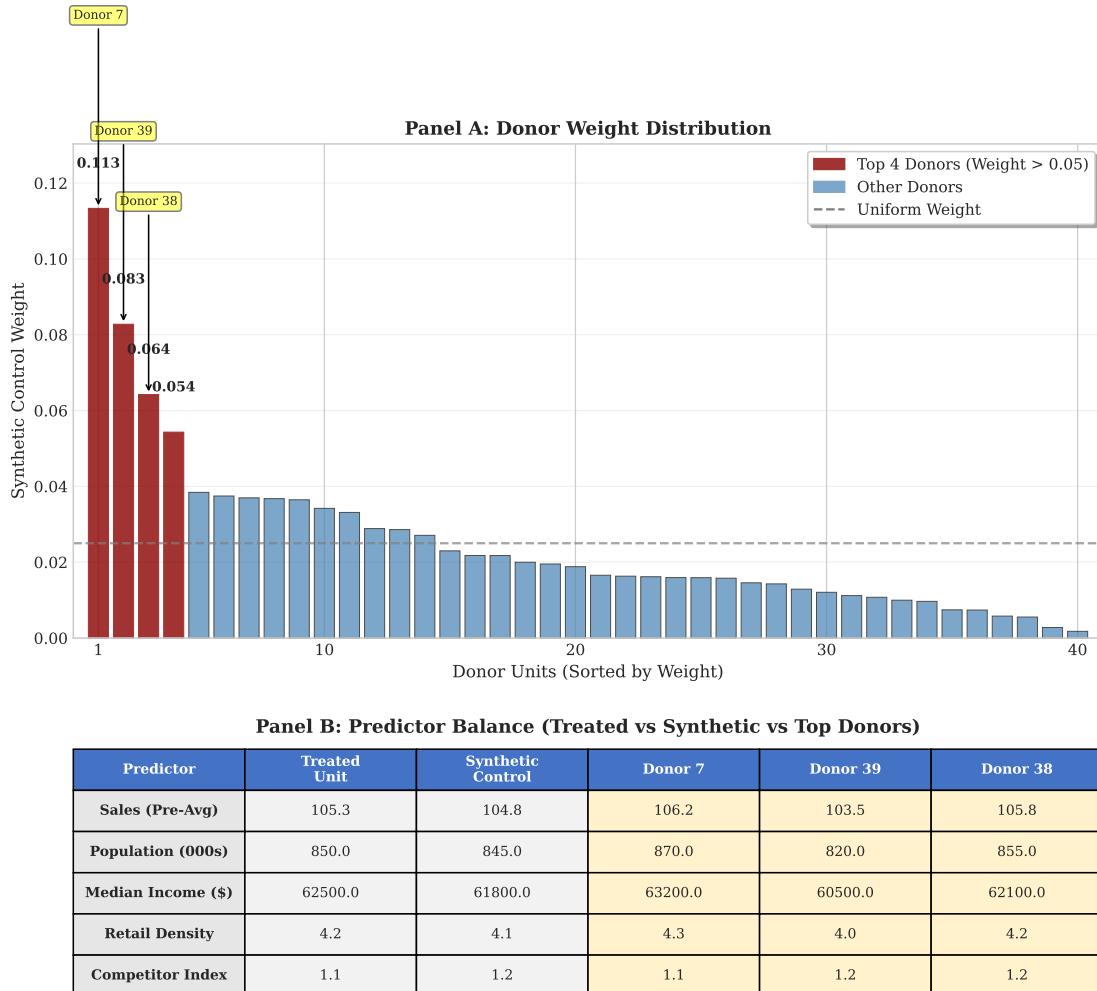
Fourth, visualise the results: plot trajectories, gaps, and placebo-check gaps. Fifth, examine weights by reporting  $N_{\text{eff}}$  and identifying high-weight donors. Sixth, conduct sensitivity analyses including leave-one-out, varying predictors, and varying the donor pool. Seventh, assess stability using in-time placebo checks with pseudo-intervention dates. Eighth, check for overfitting via cross-validation if  $N - 1 > T_0$  or the predictor set is large.

Diagnostics build a cumulative case for credibility. Each diagnostic maps back to specific identification assumptions in Section 6.3. Use the checklist to make explicit which assumptions are most plausible and which remain vulnerable. No single diagnostic is definitive. Report all diagnostics transparently and discuss implications for the validity of conclusions.



**Fig. 6.1** Pre-Treatment Fit and Post-Treatment Gap Plot (Illustrative)

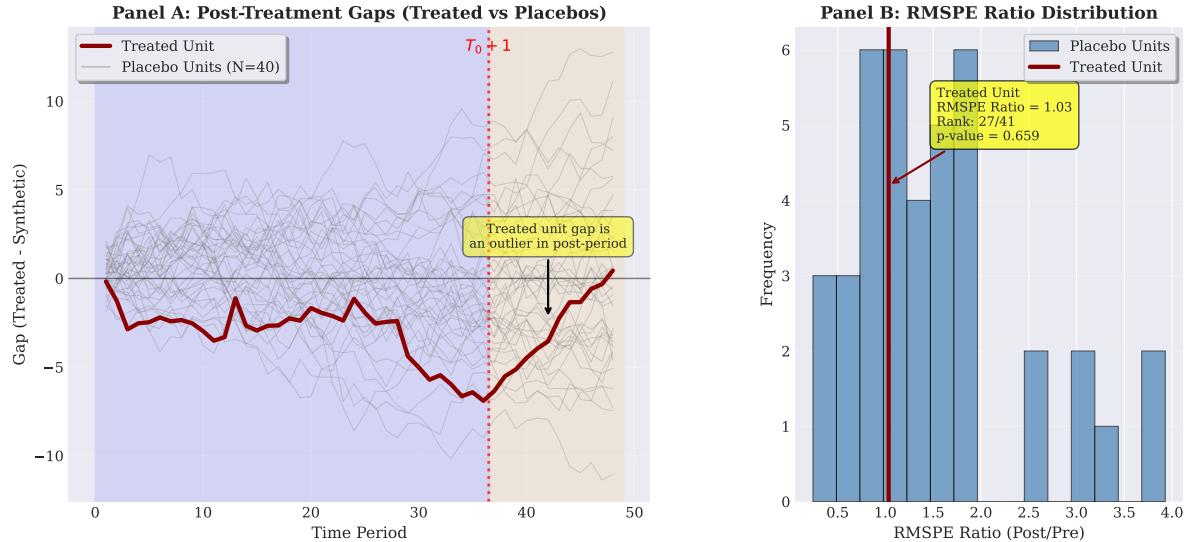
Note: Panel A shows the treated unit (solid line) and synthetic control (dashed line) outcomes over the full sample period. The vertical line marks the intervention time. Pre-treatment trajectories are closely aligned, indicating good fit. Post-treatment divergence shows the treatment effect. Panel B displays the gap ( $\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^{\text{syn}}$ ) over time. The pre-treatment gap fluctuates near zero, while the post-treatment gap is large and persistent. This figure corresponds to the trajectory and gap plot diagnostics described above.



Note: Synthetic control constructed from 40 donor units. Top 4 donors account for 31.5% of total weight. Predictor balance shows close match between treated unit and synthetic control.

**Fig. 6.2** Donor Weight Distribution and Predictor Balance (Illustrative)

Note: Panel A shows the distribution of synthetic control weights across donor units, sorted by weight magnitude. The effective number of donors  $N_{\text{eff}}$  indicates moderate concentration (illustrative example). Panel B presents predictor balance, comparing the treated unit, synthetic control (weighted average), and top individual donors. Close alignment demonstrates successful replication of the treated unit's pre-treatment characteristics.



Note: RMSPE ratio compares post-treatment to pre-treatment fit. Treated unit has ratio of 1.03, ranking 27 out of 41 units. Large ratio indicates post-treatment divergence.

**Fig. 6.3** Placebo-Check Gap Distribution and RMSPE Ratio (Illustrative)

Note: Panel A displays post-treatment gaps for the treated unit (bold red line) and all placebo-check units (thin grey lines). The treated unit's gap is an outlier in the post-treatment period. Panel B shows the distribution of RMSPE ratios. The treated unit's ratio (red line) is in the extreme right tail, yielding a small p-value (illustrative example) under the permutation distribution.

## 6.6 Practical Issues in Marketing Panels

Implementing synthetic control methods in marketing panels requires navigating practical challenges such as donor pool curation, covariate selection, missing data, anticipation, carryover, and spillovers. These choices map directly to the identification assumptions in Section 6.3. For example, donor pool curation relates to the convex hull condition (Assumption 17), spillovers relate to the no interference assumption (Assumption 15), anticipation relates to the no anticipation assumption (Assumption 14), and covariate selection relates to the factor model structure (Assumption 16).

This section provides practical guidance on each issue, showing how violations affect the bias decomposition from Section 6.1 (equation 6.7). The focus is on implementation details for marketing applications. For the underlying theory, cross-references point to the relevant earlier sections.

### Designing the Donor Pool and Convex Hull

Donor pool curation determines whether the convex hull condition (Assumption 17) holds. The donor pool must include units that are comparable to the treated unit on dimensions relevant to outcomes and must exclude units that were treated or contaminated by spillovers.

**Comparability and Factor Loadings.** Under the factor model (Assumption 16), comparability means that donor factor loadings  $\{\lambda_j\}$  span a space that includes the treated unit's loadings  $\lambda_1$ . In marketing terms, comparable units include DMAs with similar population, demographics, media markets, and competitive environments, stores with similar format, size, location characteristics, and customer base, and markets with similar baseline sales trajectories. Excluding incomparable donors improves the convex hull condition and reduces  $\|\Delta_\lambda\|$ . However, overly aggressive exclusion can itself violate the convex hull condition by removing donors that help span  $\lambda_1$ . Donor curation should therefore balance comparability against the need to maintain a rich, high-coverage donor pool that provides adequate factor-loading coverage.

**Contamination and Bias.** Contamination by spillovers extends the bias decomposition. Let  $\tau_{jt}^{\text{spill}} = Y_{jt}(0, h_j^{\text{treated}}) - Y_{jt}(0, h_j^{\text{untreated}})$  denote the spillover effect on donor  $j$  at time  $t$ , where  $h_j^{\text{treated}}$  represents donor  $j$ 's exposure state when unit 1 is treated, and  $h_j^{\text{untreated}}$  represents the exposure state when unit 1 is untreated. This notation follows the general exposure-mapping framework for spillovers developed in Chapter 11. Taking expectations over idiosyncratic errors and conditioning on the factors, loadings, and exposure pattern, we obtain

$$\mathbb{E}[\hat{\tau}_{1t} - \tau_{1t}] = \mathbf{f}'_t \Delta_\lambda + \sum_{j \in \mathcal{J}} w_j^* \tau_{jt}^{\text{spill}}. \quad (6.31)$$

If spillovers are positive (treatment increases donor outcomes), the contamination term  $\sum_j w_j^* \tau_{jt}^{\text{spill}}$  may be positive, so the observed gap  $\hat{\tau}_{1t}$  may understate or overstate the true direct effect  $\tau_{1t}$ . The direction of the bias depends on the specific context.

**Buffer Zones.** Excluding donors geographically or economically close to the treated unit can mitigate contamination, but it also shrinks the donor pool. Start with institutional knowledge about how spillovers plausibly travel, then vary the buffer size (for example, 50 km, 100 km, 200 km) and report how estimates and fit change.

**Market Re-Definitions.** When the natural donor pool is contaminated, consider redefining units. Options include product categories within markets (assuming cross-category spillovers are limited), customer segments within markets, or time-shifted comparisons (same market in different periods). Such re-definitions require careful justification and sensitivity analyses.

Practical constraints may restrict donors: same region or operating division (operational comparability), same competitive set (market structure comparability), or same retail format (channel comparability). These restrictions improve comparability but reduce the donor pool. The analyst must balance comparability, pool size, and fit.

## Covariate Selection and Factor-Model Fit

Covariate selection determines which variables enter the predictor set  $\mathbf{X}_1$  and affects how well the synthetic control matches the treated unit's factor loadings.

The guiding principle is to include covariates that proxy for factor loadings  $\lambda_i$ . Under the factor model, outcomes are driven by factor loadings. Covariates can proxy for factor loadings and help the synthetic control match the treated unit's underlying structure.

Appropriate inclusions are pre-treatment outcomes (all or selected pre-treatment periods, which are the primary proxy for factor loadings), pre-treatment covariates (variables correlated with outcomes and plausibly related to factor loadings, such as market size, baseline characteristics, and demographics), and summary statistics (pre-treatment mean, trend, and volatility, which capture different aspects of the pre-treatment trajectory).

Variables to exclude include post-treatment covariates (including post-treatment variables introduces leakage and invalidates causal interpretation), covariates affected by treatment (even if measured pre-treatment, exclude variables that may have been affected by anticipation), and irrelevant covariates (variables uncorrelated with outcomes add noise without improving fit). Including such variables effectively conditions on post-treatment information or on other units' responses, violating the no-anticipation and no-interference components of SUTVA and biasing the estimated effect.

As noted in Section 6.3, covariate selection is a form of model selection. The analyst should pre-specify the covariate set based on theory and institutional knowledge, report sensitivity to alternative covariate sets, and avoid data-mining over covariate combinations to achieve good fit.

## Time-Varying Covariates

Time-varying covariates (covariates that change over time) require care because they can be affected by treatment or by anticipation of treatment. Safe inclusions are pre-treatment values of time-varying covariates that are determined by slow-moving forces (for example, census demographics in year  $T_0 - 1$ ) and are unlikely to respond to treatment announcements. Unsafe inclusions are post-treatment covariate values (conditioning on post-treatment data invalidates causal interpretation) and any covariates, even if measured pre-treatment, that may reflect anticipatory behaviour (for example, stocking decisions or forward-buy behaviour in response to announced promotions).

Consider an example where customer demographics shift after treatment due to the campaign. Including post-treatment demographics would condition on a post-treatment variable and can bias the effect estimate.

## Missing Data

Missing data in panel outcomes or covariates must be handled carefully.

**Factor-Based Imputation.** Under the factor model (Assumption 16), missing values can be imputed using the estimated factor structure:

$$\hat{Y}_{it}^{\text{imputed}} = \hat{\alpha}_i + \hat{\lambda}_t + \hat{\lambda}'_i \hat{\mathbf{f}}_t, \quad (6.32)$$

where factors and loadings are estimated from observed data. This approach is appropriate only when the factor model provides a good description of the observed panel (see Chapter 8 for diagnostics). If factor structure fits poorly, factor-based imputation can propagate misspecification into the SC construction. This approach aligns with the identification logic of synthetic control. See Chapter 8 for estimation details.

Simple imputation methods (mean imputation, forward-filling, or interpolation) do not preserve the dynamic factor structure and may bias the synthetic control. Use only if factor-based imputation is infeasible.

If missing data are extensive (for example, when missingness is substantial, such as a large fraction of pre-treatment periods), synthetic control may not be feasible. Consider difference-in-differences (Chapter 4) or factor models (Chapter 8) as alternatives.

Factor-based imputation should be applied to donor outcomes and, if necessary, to pre-treatment outcomes. Analysts should not impute the treated unit's post-treatment outcomes and then use those imputations in place of observed  $Y_{1t}$  when constructing the synthetic control. This would undermine the design-aligned interpretation.

## Anticipation

Anticipation violates the no anticipation assumption (Assumption 14) and biases the synthetic control if pre-treatment outcomes reflect anticipatory behaviour. Section 6.5 provides diagnostics for detection.

**Detecting Anticipation.** In-time placebo checks reveal whether fit deteriorates near the intervention date. Visual inspection shows whether pre-treatment gaps widen as treatment approaches. Institutional knowledge indicates whether treatment was announced in advance.

**Addressing Anticipation.** Three strategies address anticipation. First, use the announcement date. If treatment was announced at  $T_A < T_0$ , construct the synthetic control using periods  $1, \dots, T_A - 1$  and interpret  $\hat{\tau}_{1t}$  for  $t \geq T_A$  as the combined effect of anticipation and implementation. Formally, this redefinition changes the estimand from the pure implementation effect  $\tau_{1t}$  to a combined effect that includes both anticipatory and post-implementation responses. You should state this clearly when reporting results. Second, truncate the pre-treatment period by excluding periods near the intervention where anticipation may contaminate outcomes. Third, redefine the estimand by acknowledging that the estimated effect includes anticipatory responses and interpreting accordingly.

## Carryover and Dynamic Effects

Carryover (persistent or lagged effects) is naturally accommodated in synthetic control. The treatment effect trajectory  $\{\hat{\tau}_{1t}\}_{t>T_0}$  traces the post-treatment response path for the treated unit under the realised treatment path, typically a single switch-on event. It does not by itself identify counterfactual effects of alternative treatment histories without stronger assumptions about treatment effect dynamics.

The cumulative effect over the post-treatment window is:

$$\hat{\tau}_1^{\text{cum}} = \sum_{t=T_0+1}^T \hat{\tau}_{1t}. \quad (6.33)$$

The average treatment effect is:

$$\hat{\tau}_1^{\text{avg}} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{\tau}_{1t}. \quad (6.34)$$

These coincide with the cumulative and average effects used in the event-study metrics of Chapter 5, specialised here to a single treated unit. Chapter 10 covers formal modelling of dynamic effects, including ramp-up, decay, and equilibrium responses.

## Spillovers and Interference

Spillovers violate the no interference assumption (Assumption 15) and introduce the contamination bias (equation 6.31).

**Buffer Zones.** Exclude donors within distance  $d$  of the treated unit:

$$\mathcal{J}_{\text{buffered}} = \{j \in \mathcal{J} : \text{dist}(j, 1) > d\}, \quad (6.35)$$

where  $\text{dist}(j, 1)$  measures the distance between donor  $j$  and the treated unit (for example, geodesic distance between DMA centroids, travel time, or a competitive distance metric). This reduces the contamination term  $\sum_{j \in \mathcal{J}} w_j^* \tau_{jt}^{\text{spill}}$  in equation (6.31) at the cost of fewer donors and potentially worse pre-treatment fit. The trade-off is that larger  $d$  reduces contamination bias but shrinks the donor pool and may worsen fit.

**Cluster Designs.** If treatment is applied to entire clusters and donors are drawn from separate clusters, within-cluster spillovers are eliminated. This approach requires that clusters are independent.

**Explicit Exposure Mappings.** Where the exposure structure is known, spillovers can be modelled explicitly by defining potential outcomes as a function of both own treatment and exposure to others. For donors that never receive own treatment, we can write a simple linear exposure model as

$$Y_{jt}(0, h_{jt}) = Y_{jt}(0, 0) + \gamma \cdot h_{jt}, \quad (6.36)$$

where  $h_{jt} = h_j(D_{-j,t})$  is the exposure mapping for donor  $j$  in period  $t$ , and  $\gamma$  is a parameter governing the strength of spillovers. See Chapter 11 for estimation.

Placebo checks from Section 6.4 help detect spillovers. If donor units show systematic post-treatment gaps, this suggests spillover contamination.

## Practical Workflow

This section's guidance links the abstract assumptions in Section 6.3 to concrete design choices in marketing panels. In practice, you curate the donor pool and covariates, handle missing data, define intervention timing and buffers, construct the synthetic control (Section 6.2), subject it to diagnostics (Section 6.5), and then conduct inference (Section 6.4). Document each design choice and show, through sensitivity analyses, that your main conclusions are not artefacts of a single specification. Archive the specification decisions, sensitivity analyses, and diagnostic results alongside the main results to ensure transparency and replicability.

## Design Choices and Bias-Variance Trade-Offs

Table 6.2 summarises key design choices and their implications for the bias decomposition and variance.

**Table 6.2** Design Choices: Implications for Bias and Variance

Design Choice	Impact on Bias ( $f_t' \Delta_\lambda$ )	Impact on Variance
Donor pool size	Larger pool → better convex hull coverage → smaller $\ \Delta_\lambda\ $	Effect depends on how weights distribute across donors. More donors can lower variance if they allow more diffuse weights (higher $N_{\text{eff}}$ ), but concentrated weights on a small subset can leave variance largely unchanged
Pre-treatment window	Longer window → better factor loading estimation → smaller $\ \Delta_\lambda\ $	Longer window → smaller variance (more data)
Covariate set	Covariates that proxy for $\lambda_i$ can improve fit. Irrelevant covariates add noise	More covariates → higher variance unless regularised
Convexity constraint	Prevents extrapolation. It may increase $\ \Delta_\lambda\ $ if hull condition fails	Restricting weights to be non-negative and sum to one typically stabilises variance by ruling out extreme negative weights, though it may prevent variance-reducing extrapolation in settings where the treated unit lies outside the donor hull
Regularisation	Reduces overfitting and can improve out-of-sample $\ \Delta_\lambda\ $ when the unregularised fit is noisy, but may increase bias if the true optimal weights are far from the regularised target	Shrinks weights toward uniform → lower variance
Buffer zone	Excluding contaminated donors reduces spillover bias	Smaller pool → higher variance

## 6.7 Data Fusion for Cold-Start Problems

A fundamental limitation of synthetic control is the requirement for a long pre-treatment period to learn donor weights. In “cold-start” scenarios—a crisis response, product launch in a new market, or entry into a new category—pre-intervention data for the target unit may be unavailable or sparse. This section introduces Causal Data Fusion [Yang et al., 2024], which addresses cold-start by leveraging an auxiliary reference domain.

### The Cold-Start Problem

The synthetic control estimator requires pre-treatment outcomes in the target domain,  $\{Y_{1t}^{\text{target}}\}_{t \leq T_0^{\text{target}}}$ , to estimate weights  $\mathbf{w}^*$ . When  $T_0^{\text{target}}$  is small or zero, identification fails because weight optimisation has insufficient data to distinguish factor-driven variation from noise. Cold-start scenarios arise frequently in marketing: a retailer may launch a loyalty programme in a new market with no historical data, a brand may respond to a competitor’s sudden market entry, a new product category may be introduced with no pre-launch sales, or a crisis may require immediate intervention with no time for baseline data collection.

The key question is whether data fusion can substitute for missing target-domain pre-treatment data and, if so, under what conditions. The answer depends on whether structural relationships between units are stable across domains.

### Causal Data Fusion Framework

The key insight of data fusion is that while the target domain  $Y$  may lack pre-treatment data, a related reference domain  $F$  may have rich historical information. If the relationship between the treated unit and donors is similar across domains, weights learned from the reference domain can be transferred to the target domain.

We introduce notation for the two-domain setting. Let  $Y_{it}^{\text{target}}$  denote the outcome in the target domain (for example, sales of new product) and  $F_{it}^{\text{ref}}$  denote the outcome in the reference domain (for example, sales of established product, foot traffic). We treat the reference domain as unaffected by the target intervention. No unit in the reference domain receives the target-domain treatment, and reference-domain outcomes are not affected by spillovers from the target-domain intervention. The reference domain may still have its own shocks or interventions that do not interact with the target-domain treatment. This assumption justifies writing  $F_{it}^{\text{ref}}$  without potential-outcomes notation. Let  $\mathcal{T}^{\text{ref}} = \{1, \dots, T^{\text{ref}}\}$  denote time periods observed in the reference domain and let  $\mathcal{T}^{\text{target}} = \{1, \dots, T\}$  denote observed time periods in the target domain. In the cold-start limit where no target-domain pre-treatment data exist, we set  $T_0^{\text{target}} = 0$  so that all observed target-domain periods are post-treatment.

## The Equi-ConFOUNDing Assumption

Identification relies on equi-conFOUNDing: the relationship between the treated unit and donors in the reference domain mirrors the relationship in the target domain.

**Assumption 18 (Equi-ConFOUNDing)** The target and reference domains share the same factor loading structure:

$$\begin{aligned} Y_{it}^{\text{target}}(0) &= \alpha_i^Y + \lambda_t^Y + \sum_{r=1}^R \lambda_{ir} f_{tr}^Y + \varepsilon_{it}^Y, \\ F_{it}^{\text{ref}} &= \alpha_i^F + \lambda_t^F + \sum_{r=1}^R \lambda_{ir} f_{tr}^F + \varepsilon_{it}^F, \end{aligned}$$

where  $\alpha_i^Y$  and  $\alpha_i^F$  are domain-specific unit fixed effects,  $\lambda_t^Y$  and  $\lambda_t^F$  are domain-specific time fixed effects,  $f_{tr}^Y$  and  $f_{tr}^F$  are domain-specific factor paths, and the factor loadings  $\lambda_{ir}$  are shared across domains. Write  $\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{iR})'$  for the loading vector. This is the two-domain analogue of the factor-structure Assumption 16: both target and reference domains share the same unit-specific factor loadings  $\lambda_{ir}$ , but may have different factor paths and time effects.

Assumption 18 requires that the treated and donor units share the same factor loadings across domains, while allowing the factor paths  $f_{tr}^Y$  and  $f_{tr}^F$  to differ. In some applications a stronger, approximately linear relationship between the treated–donor gaps in the two domains is plausible. We refer to this as linear equi-conFOUNDing and formalise it below.

**Definition 6.3 (Linear Equi-ConFOUNDing)** Let  $\mathcal{T}_{\text{pre}}^{\text{target}} \subseteq \mathcal{T}^{\text{target}}$  denote any target-domain pre-treatment periods for which  $Y_{it}^{\text{target}}$  is observed. The linear equi-conFOUNDing condition holds over  $\mathcal{T}_{\text{pre}}^{\text{target}}$  if, for each  $t \in \mathcal{T}_{\text{pre}}^{\text{target}}$ ,

$$\mathbb{E}[Y_{1t}^{\text{target}}(0)] - \sum_{j \in \mathcal{J}} w_j \mathbb{E}[Y_{jt}^{\text{target}}(0)] = \kappa \left( \mathbb{E}[F_{1t}^{\text{ref}}] - \sum_{j \in \mathcal{J}} w_j \mathbb{E}[F_{jt}^{\text{ref}}] \right), \quad (6.37)$$

for some time-invariant scaling factor  $\kappa$ . Under this condition, any weights  $\mathbf{w}$  that balance the reference domain in expectation (right-hand side equal to zero across  $\mathcal{T}_{\text{pre}}^{\text{target}}$ ) also balance the target domain in expectation. This condition is not testable without overlap periods where both  $Y$  and  $F$  are observed pre-treatment. If such overlap exists, analysts can estimate  $\kappa$  via a diagnostic regression and check stability across periods.

Linear equi-conFOUNDing is substantially stronger than Assumption 18: it requires not only shared factor loadings but also a stable proportional relationship between factor paths across domains (effectively a single latent factor or perfectly collinear factors). This additional structure is necessary for weights that achieve balance in the reference domain to guarantee balance in the target domain, and will be plausible only in tightly related domains.

### Identification under Equi-Conounding

We assume that the reference-domain outcomes  $F_{it}^{\text{ref}}$  are never directly treated and are not affected by spillovers from the target-domain intervention, so that they can be used to learn weights without contamination. This extends the no-anticipation and no-interference logic from standard SC to the cross-domain setting.

**Proposition 6.2 (Identification via Data Fusion)** *Under Assumption 18 (equi-conounding) and the standard SC assumptions (no anticipation, no interference), if weights  $\mathbf{w}^*$  satisfy:*

- (i) Convexity:  $w_j^* \geq 0, \sum_{j \in \mathcal{J}} w_j^* = 1$
- (ii) Factor-loading match (shared across domains):  $\sum_{j \in \mathcal{J}} w_j^* \boldsymbol{\lambda}_j = \boldsymbol{\lambda}_1$

then the data fusion estimator  $\hat{\tau}_{1t}^{\text{fusion}} = Y_{1t}^{\text{target}} - \sum_{j \in \mathcal{J}} w_j^* Y_{jt}^{\text{target}}$  identifies the treatment effect:

$$\mathbb{E}[\hat{\tau}_{1t}^{\text{fusion}}] = \tau_{1t}.$$

Condition (ii) is a theoretical requirement expressed in terms of unobserved factor loadings. In practice, it is approximated by matching reference-domain outcomes  $F_{it}^{\text{ref}}$  over  $\mathcal{T}^{\text{ref}}$  under the same rank and factor-identification conditions as in Proposition 6.1.

*Sketch of argument.* Under equi-conounding, weights that balance factor loadings in the reference domain also balance them in the target domain. The proof follows Theorem 6.1, with the reference domain providing the data to estimate weights.

### Data Fusion Algorithm

The data fusion algorithm proceeds in three steps. First, reference matching constructs synthetic control weights using the reference domain:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{t \in \mathcal{T}^{\text{ref}}} \left( F_{1t}^{\text{ref}} - \sum_{j \in \mathcal{J}} w_j F_{jt}^{\text{ref}} \right)^2 \quad \text{s.t. } w_j \geq 0, \sum_{j \in \mathcal{J}} w_j = 1.$$

This formulation uses unweighted squared-error loss. As in Section 6.2, the loss can be extended to include  $V$ -weighted predictors, covariate augmentation, or regularisation (ridge/entropy penalties) to stabilise weights.

Second, weight transfer applies the estimated weights to target domain outcomes:

$$\hat{Y}_{1t}^{\text{target}}(0) = \sum_{j \in \mathcal{J}} \hat{w}_j Y_{jt}^{\text{target}}.$$

Third, estimation computes the treatment effect:

$$\hat{\tau}_{1t}^{\text{fusion}} = Y_{1t}^{\text{target}} - \hat{Y}_{1t}^{\text{target}}(0).$$

## Bias Analysis

If equi-confounding fails (factor loadings differ across domains), the data fusion estimator is biased.

**Bias Decomposition.** Let  $\boldsymbol{\lambda}_i^Y$  and  $\boldsymbol{\lambda}_i^F$  denote the factor loadings in the target and reference domains, respectively. If equi-confounding fails:

$$\text{Bias}(\hat{\tau}_{1t}^{\text{fusion}}) = (\mathbf{f}_t^Y)' \underbrace{\left( \boldsymbol{\lambda}_1^Y - \sum_j w_j^* \boldsymbol{\lambda}_j^Y \right)}_{\text{target domain mismatch}}. \quad (6.38)$$

This mirrors the standard SC bias expression  $\mathbf{f}_t' \boldsymbol{\Delta}_\lambda$  in Section 6.1, but now the mismatch is measured in the target domain even though weights are learned in the reference domain.

Bias tends to be large when domain-specific shocks affect the target domain but not the reference domain (for example, pricing or promotion changes affecting  $Y$  but not  $F$ ), when competitive conditions differ across categories, when measurement units are misaligned (for example, sales versus traffic), or when changes in assortment or supply break the cross-domain relationship. It can also be large when the reference domain is measured with error or aggregated differently, or when structural changes occur between the reference and target observation periods.

## Diagnostics for Equi-Confoundung

Because equi-confounding involves unobserved factor loadings, it is not directly testable. The diagnostics from Section 6.5 still apply, but they must be adapted to the two-domain setting. When any target-domain pre-treatment data are available, even sparsely, they provide a crucial check on whether weights learned in the reference domain behave sensibly in the target domain.

**Partial Pre-Treatment Data.** If any target domain pre-treatment data are available (even sparse), assess whether reference-domain weights fit the target domain:

$$\text{RMSPE}_{\text{pre}}^{\text{target}} = \sqrt{\frac{1}{|\mathcal{T}_{\text{pre}}^{\text{target}}|} \sum_{t \in \mathcal{T}_{\text{pre}}^{\text{target}}} \left( Y_{1t}^{\text{target}} - \sum_j \hat{w}_j Y_{jt}^{\text{target}} \right)^2}. \quad (6.39)$$

Small  $\text{RMSPE}_{\text{pre}}^{\text{target}}$  supports equi-confounding. When any target-domain pre-treatment data exist, this check is the primary diagnostic for equi-confounding: large  $\text{RMSPE}_{\text{pre}}^{\text{target}}$  indicates that reference-derived weights do not transfer well and that equi-confounding is implausible.

**Covariate Balance.** If pre-treatment covariates are available for both domains, check whether reference-domain weights achieve balance in the target domain.

**Reference Domain Fit Quality.** Good fit in the reference domain (small  $\text{RMSPE}_{\text{pre}}^{\text{ref}}$ ) is necessary but not sufficient. It indicates that the convex hull condition is satisfied in the reference domain. On its own, however, it says nothing about whether equi-confounding holds. Good fit in  $F$  is compatible with severe mismatch in  $Y$ .

**Sensitivity Analysis.** Vary the reference domain (for example, different product categories, different outcome measures) and assess stability of conclusions. Robust results across reference domains support equi-confounding. Sensitivity is now over reference-domain definitions, not only donor pools.

## Inference for Data Fusion

Conformal inference (Section 6.4) can be adapted to data fusion. The key difference is that calibration uses the reference domain rather than target domain pre-treatment data.

The conformal data fusion procedure proceeds as follows. First, compute reference-domain residuals:  $e_{1t}^{\text{ref}} = F_{1t}^{\text{ref}} - \sum_j \hat{w}_j F_{jt}^{\text{ref}}$  for  $t \in \mathcal{T}^{\text{ref}}$ . Second, under equi-confounding, these residuals calibrate the target domain uncertainty. Specifically, conformal calibration requires that reference-domain residuals, after appropriate scaling, have comparable distribution and dependence structure to target-domain residuals. Without this comparability, coverage statements become meaningless. This is an additional assumption beyond equi-confounding. Third, construct confidence intervals by scaling reference-domain residuals to the target domain scale.

Confidence intervals should account for uncertainty in the equi-confounding assumption. Conservative intervals add a sensitivity parameter for domain mismatch. If equi-confounding or residual comparability fails, these intervals can be severely misleading. Report them alongside sensitivity analyses that relax the cross-domain stability assumptions.

## Marketing Application: New Product Launch

Consider a consumer goods company that launches a new product category (organic snacks) in a flagship DMA. No historical sales data exist for this category. However, the company has years of sales data for its established products (conventional snacks) across all DMAs.

The data fusion application proceeds as follows. The reference domain consists of sales of conventional snacks across all DMAs over 24 pre-launch months. The target domain consists of sales of organic snacks in the post-launch period. Reference matching constructs synthetic control for the flagship DMA using conventional snack sales. Weight transfer applies these weights to organic snack sales in donor DMAs. Estimation then yields the launch effect on organic snack sales.

The equi-confounding rationale is that the assumption holds if DMAs that are similar in conventional snack sales are also similar in organic snack sales—plausible if both categories are driven by common factors (population, income, health consciousness, retail density).

For diagnostics, if any organic snack pilot data exist (even a few weeks), assess whether conventional-snack weights fit organic-snack trajectories.

## When to Use Data Fusion

Data fusion is appropriate when target domain pre-treatment data are unavailable or sparse, when a related reference domain with rich historical data exists, when equi-confounding is plausible based on institutional knowledge, and when some target domain data are available for diagnostic validation.

Data fusion is inappropriate when reference and target domains have fundamentally different structures, when no institutional basis for equi-confounding exists, or when target domain data are sufficient for standard SC.

Data fusion relies on an untestable assumption (equi-confounding). Use diagnostics, sensitivity analysis, and institutional knowledge to assess plausibility. Report uncertainty honestly, acknowledging that identification depends on cross-domain stability. Data fusion is a last-resort strategy for genuine cold-start settings: when standard SC is feasible with a reasonable pre-treatment window in the target domain, it is generally preferable to rely on within-domain identification.

## 6.8 Extensions and Variants

Synthetic control methods can be extended to settings with multiple treated units, staggered adoption, regularised estimation, and continuous treatments, though many extensions require materially stronger assumptions than the baseline case. This section provides an overview of key extensions, connecting each to the bias framework from Sections 6.1–6.3 and providing forward references to detailed treatments in later chapters.

### Multiple Treated Units

When multiple units are treated at a common time  $T_0$ , estimate unit-specific treatment effects and aggregate. The same assumptions as in Section 6.3 apply unit-by-unit.

**Unit-Specific Estimation.** For each treated unit  $i \in \mathcal{I}_{\text{treated}}$ , construct a synthetic control using the donor pool  $\mathcal{J}$  (excluding all treated units) and compute:

$$\hat{\tau}_{it} = Y_{it} - \sum_{j \in \mathcal{J}} w_{ij}^* Y_{jt}, \quad t > T_0. \quad (6.40)$$

If treated units interfere with each other (for example, neighbouring DMAs, rival stores), unit-by-unit SC no longer identifies the direct effect without spillover contamination. In such cases, treat the treated set as a cluster and redefine the estimand (Chapter 11).

**Aggregation.** The average treatment effect on the treated (ATT) at time  $t$  is:

$$\widehat{\text{ATT}}_t = \sum_{i \in \mathcal{I}_{\text{treated}}} w_i \hat{\tau}_{it}, \quad (6.41)$$

where  $w_i$  are aggregation weights summing to one. Note that these are aggregation weights for combining unit-specific estimates, distinct from the donor weights  $w_{ij}^*$  used within each unit's synthetic control construction. Here  $\widehat{\text{ATT}}_t$  is an average over the currently treated units  $\mathcal{I}_{\text{treated}}$  at time  $t$ , with aggregation weights  $w_i$  that should be chosen to match the target treated population of interest. Common choices include equal weights ( $w_i = 1/|\mathcal{I}_{\text{treated}}|$ ), precision weights ( $w_i \propto 1/\text{Var}(\hat{\tau}_{it})$ ), and pre-treatment outcome weights ( $w_i \propto \bar{Y}_{i,\text{pre}}$ ). Different choices of  $w_i$  correspond to different target estimands (for example, an equally weighted ATT across treated units versus an ATT weighted by baseline outcome levels). Analysts should choose  $w_i$  to reflect the business question rather than solely for statistical convenience.

Aggregation does not change the identification conditions. Each unit-specific estimate  $\hat{\tau}_{it}$  requires the factor model, convex hull, and stability assumptions (Assumptions 14–17) for that unit. The aggregated ATT inherits the bias:

$$\text{Bias}(\widehat{\text{ATT}}_t) = \sum_{i \in \mathcal{I}_{\text{treated}}} w_i (\mathbf{f}_t)' \boldsymbol{\Delta}_{\lambda,i}, \quad (6.42)$$

where the shared factor path  $\mathbf{f}_t$  comes from Assumption 16. The aggregated ATT inherits unit-level biases through the weighted average of unit-specific mismatches  $\Delta_{\lambda,i}$ .

For inference, conduct in-space placebo checks for each treated unit and aggregate placebo-check gaps with the same weights. The distribution of aggregated placebo-check gaps provides inference for the aggregated ATT.

Aggregation assumes identification holds for each unit. If any unit violates the convex hull condition, aggregated estimates inherit that bias.

## Staggered Adoption

With staggered adoption (units treated at different times), standard SC faces challenges. Donor pools shrink as more units become treated. Already-treated units contaminate comparisons for later-treated units. Aggregation across adoption cohorts requires careful weighting. The same assumptions as in Section 6.3 apply cohort-by-cohort.

**Stacking Approach.** For each adoption cohort with adoption time  $G_i = g$  (units treated at time  $g$ ), construct synthetic controls using never-treated units and not-yet-treated units observed strictly before their own adoption dates as donors. Specifically, for cohort  $g$ , not-yet-treated units  $j$  with  $G_j > g$  contribute only their pre-adoption periods ( $t < G_j$ ) to the donor pool. Weights are re-estimated cohort-by-cohort using only these eligible pre-periods. Estimate cohort-time effects:

$$\hat{\tau}(g, t) = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \hat{\tau}_{it}, \quad t > g. \quad (6.43)$$

For each cohort  $g$ , the factor-structure, convex-hull, no-anticipation, and no-interference assumptions from Section 6.3 must hold when we treat that cohort as “treated” and the eligible donors as “controls”. Violations in any cohort can bias aggregated event-time effects.

**Event-Time Aggregation.** Aggregate to event time  $k = t - G_i$ :

$$\hat{\theta}_k = \sum_g w_g \hat{\tau}(g, g + k), \quad (6.44)$$

where  $w_g$  are cohort weights and  $\hat{\theta}_k$  denotes the event-time effect at relative time  $k$  after adoption.

Hybrid methods such as Synthetic Difference-in-Differences [Arkhangelsky et al., 2021] and Augmented Synthetic Control [Ben-Michael et al., 2021] combine SC weights with DiD adjustments to address staggered adoption more robustly. See Chapter 7. Event study aggregation and interpretation are developed in Chapters 4 and 5.

Cohort-specific identification becomes harder as donor pools shrink with each wave of adoption.

## Ridge Synthetic Control

Ridge regularisation [Doudchenko and Imbens, 2016] addresses overfitting when the pre-treatment period is short or the donor pool is large.

**Regularised Objective.** Add a quadratic penalty to the weight optimisation:

$$\hat{\mathbf{w}}^{\text{Ridge}} = \arg \min_{\mathbf{w}} \|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}\|_V^2 + \eta \|\mathbf{w} - \bar{\mathbf{w}}\|^2 \quad \text{s.t. } w_j \geq 0, \sum_j w_j = 1, \quad (6.45)$$

where  $\eta > 0$  is the regularisation parameter and  $\bar{\mathbf{w}} = (1/(N-1), \dots, 1/(N-1))'$  is the uniform convex weight vector over the donor pool. The penalty shrinks  $\mathbf{w}$  toward uniform weights. Shrinking toward  $\bar{\mathbf{w}}$  implicitly uses the equally weighted DiD estimator (simple average of donors) as the regularisation target.

**Bias–Variance Trade-Off.** In the usual bias–variance decomposition, regularisation introduces bias but reduces variance. Schematically:

$$\text{MSE}(\hat{\tau}^{\text{Ridge}}) = \underbrace{(\mathbf{f}'_t \Delta_{\lambda}^{\text{Ridge}})^2}_{\text{Bias}^2(\eta)} + \underbrace{\text{Var}(\hat{\tau}^{\text{Ridge}})}_{\text{Var}(\eta)}. \quad (6.46)$$

As  $\eta \rightarrow 0$ , Ridge SC approaches unregularised SC (low bias, higher variance). As  $\eta \rightarrow \infty$ , weights approach the uniform vector  $\bar{\mathbf{w}}$  (higher bias, lower variance). Cross-validation on pre-treatment periods can be used as a heuristic to choose  $\eta$ . Cross-validation can reduce overfitting to pre-period noise but cannot validate post-treatment counterfactuals. It operates entirely on pre-treatment data and trades off in-sample fit against weight stability.

Ridge SC is particularly useful when  $N - 1 > T_0$  (more donors than pre-treatment periods), when many donors are similar (so unregularised SC may assign large weights to a few donors by chance), or when pre-treatment fit is poor (regularisation can stabilise weights).

Regularisation reduces variance but introduces bias. The optimal trade-off depends on unknown post-treatment counterfactuals.

## Augmented Synthetic Control (ASCM)

Augmented Synthetic Control [Ben-Michael et al., 2021] combines SC with regression adjustment to correct for imperfect pre-treatment fit.

**Bias Correction.** After constructing SC weights  $\mathbf{w}^*$ , ASCM estimates a regression of pre-treatment outcomes on covariates and uses it to adjust the counterfactual:

$$\hat{\tau}_{1t}^{\text{ASCM}} = Y_{1t} - \sum_j w_j^* Y_{jt} - \hat{m} \left( \mathbf{X}_1 - \sum_j w_j^* \mathbf{X}_j \right), \quad (6.47)$$

where  $\hat{m}(\cdot)$  is a regression function—estimated, for example, by linear regression or flexible machine learning methods—that maps predictor imbalance into outcome imbalance. In practice,  $\hat{m}$  is estimated using pre-treatment data (often on donors and possibly on the treated unit), with cross-fitting or sample splitting recommended in high-dimensional settings to avoid overfitting the regression adjustment to the same noise used to determine  $\mathbf{w}^*$ . We write  $\Delta_X = \mathbf{X}_1 - \sum_j w_j^* \mathbf{X}_j$  for the predictor imbalance vector.

Under the factor model, the ASCM correction targets the residual bias from imperfect balance. Schematically:

$$\text{Bias}(\hat{\tau}^{\text{ASCM}}) \approx \mathbf{f}'_t \Delta_\lambda - \hat{m}(\Delta_X), \quad (6.48)$$

so that if either the SC weights achieve near-exact balance ( $\Delta_X \approx 0$ ) or the regression function  $\hat{m}$  accurately captures the relationship between predictor imbalance and outcome imbalance, residual bias is reduced.

**Bias Reduction Property.** ASCM can reduce bias when either SC weights achieve good balance, or the regression function  $\hat{m}$  is well specified. Formal results in the ASCM literature show double-robustness properties under specific modelling assumptions (for example, correctly specified linear factor structures and linear regression adjustments); here we emphasise the intuitive idea that ASCM combines weighting and outcome regression to hedge against misspecification. See Chapter 7 for detailed treatment.

The double-robustness property requires correctly specifying either the SC weighting model or the regression adjustment. Misspecification of both can still lead to substantial bias.

## Factor Models and Matrix Methods

The connection to factor models (Chapter 8) clarifies the role of low-rank structure in SC identification.

Under the factor model (equation 6.4), SC identification requires  $\lambda_1 = \sum_j w_j^* \lambda_j$ . Factor models estimate  $\lambda_i$  and  $\mathbf{f}_t$  directly, relaxing the convexity constraint.

SC can be viewed as a constrained factor model estimator with two constraints: convexity (weights are non-negative and sum to one) and sparsity (loadings for the treated unit are a weighted combination of donor loadings).

Factor models may outperform SC when the treated unit is outside the convex hull of donors, when the donor pool is small relative to the factor dimension, or when the noise-to-signal ratio is low (factor structure is strong). In bias terms, when the treated unit lies outside the donors' convex hull or the factor dimension  $R$  is large relative to  $N - 1$ , any convex combination of donors will induce a sizeable  $\Delta_\lambda$ . Unconstrained factor models can sometimes achieve a much smaller mismatch and therefore lower bias, at the cost of relying more heavily on parametric low-rank structure.

Unconstrained factor models rely more heavily on low-rank structure and may extrapolate aggressively when the treated unit lies outside the donor hull.

## High-Dimensional Settings

High-dimensional predictors or donors require regularisation or selection methods.

**Lasso Synthetic Control.** Apply L1 penalties [Doudchenko and Imbens, 2016] to induce sparsity in convex weights:

$$\hat{\mathbf{w}}^{\text{Lasso}} = \arg \min_{\mathbf{w}} \|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}\|_2^2 + \lambda \|\mathbf{w} - \bar{\mathbf{w}}\|_1 \quad \text{s.t. } w_j \geq 0, \sum_j w_j = 1. \quad (6.49)$$

Lasso selects a small number of donors by shrinking many weights back toward the target  $\bar{\mathbf{w}}$  while allowing a subset to deviate, rather than directly penalising  $\|\mathbf{w}\|_1$  under the simplex constraint, which would be constant. As with Ridge SC, cross-validation on pre-treatment periods can be used as a heuristic to select the tuning parameter  $\lambda$ . Cross-validation can reduce overfitting to pre-period noise but cannot validate post-treatment counterfactuals.

**Double Machine Learning.** Double Machine Learning (DML, Chapter 12) treats SC-style weighting and outcome models as nuisance components, estimated with flexible machine learning, and then constructs debiased estimators for target causal parameters. This extends the design-based ideas in SC to settings with very high-dimensional covariates and complex prediction models, but represents an extension beyond vanilla SC rather than a simple wrapper. Identification in DML setups typically relies on different orthogonality conditions and nuisance-parameter structure than pure SC. Treat these methods as a separate class that borrows SC-style weighting intuition, not as minor tweaks to SC itself.

See Chapter 13 for comprehensive treatment of high-dimensional methods in panel data.

## Continuous and Multivalued Treatments

Recall from the global notation that  $Y_{it}(d)$  denotes the potential outcome for unit  $i$  at time  $t$  under dose  $d$ , and  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  is the average dose-response function. Dose-response SC aims to construct estimators of  $Y_{it}(d)$ , and hence of  $\mu(d)$ , by reweighting donors at each dose level.

**Dose-Response Synthetic Control.** For treated unit  $i$  receiving dose  $d_i$ , construct counterfactual outcomes at alternative dose levels. Let  $D_j$  denote an aggregated treatment intensity for unit  $j$ , constructed from the panel treatment path  $\{D_{jt}\}_t$  over a specified window  $\mathcal{T}_D$  as  $D_j = |\mathcal{T}_D|^{-1} \sum_{t \in \mathcal{T}_D} D_{jt}$ . For continuous  $d$ , exact matching is typically infeasible, so weights are estimated within narrow dose bands or using kernel weighting:

$$\hat{Y}_{it}(d) = \sum_{j: |D_j - d| < h} w_{ij}^*(d) Y_{jt}, \quad (6.50)$$

where  $h$  is a bandwidth parameter and weights  $w_{ij}^*(d)$  are estimated within the dose band around  $d$ , using the same SC logic as for binary treatment applied to units at that dose level. Alternatively, kernel weighting can replace the sharp band cutoff.

Identification then requires adequate support at each dose (enough donors at all relevant dose levels and in each dose band), a factor model structure at each dose level, and a convex hull condition within each dose stratum, as well as the usual no-anticipation and no-interference assumptions. Chapter 14 develops these ideas using generalised propensity scores and dose-response curves. Dose-response SC can be viewed as a design-based weighting component within that broader framework.

Identification requires adequate donor support at all dose levels and a convex hull condition within each dose stratum. These conditions become harder to satisfy as the number of distinct dose levels increases.

## Summary

Extensions and variants broaden SC applicability, but many require materially stronger assumptions than the baseline design.

**Table 6.3** SC Extensions: Key Features

Extension	Addresses	Detailed Treatment
Multiple units	Aggregation, inference	This section
Staggered adoption	Time-varying treatment	Chapter 7
Ridge SC	Overfitting, short $T_0$	This section
ASCM	Imperfect fit	Chapter 7
Factor models	Convex hull failure	Chapter 8
High-dimensional	Many predictors/donors	Chapter 13
Dose-response	Continuous treatment	Chapter 14

The core logic remains: construct a transparent counterfactual by reweighting control units to match the treated unit's pre-treatment trajectory, assess credibility through diagnostics, and quantify uncertainty through permutation-based or conformal inference.

## 6.9 Marketing Applications

Synthetic control methods are particularly well-suited to marketing applications where a single unit or a small number of units receive treatment and transparency and executive communication are priorities. This section develops synthetic control designs in four common marketing settings: flagship city campaign, exclusive retail partnership, regional regulation, and platform entry, and then discusses an empirical application linking offline advertising to online chatter.

### Flagship City Campaign

Flagship city campaigns provide a canonical setting for synthetic control. A consumer packaged goods brand launches a major television advertising campaign in a designated market area (DMA), investing heavily in television spots over a six-month period. The goal is to estimate the causal effect on sales. The treated DMA is the flagship market. The initial donor pool consists of 40 other DMAs that did not receive the campaign. Pre-treatment data span three years (36 months), which gives a long window to establish the baseline trajectory and seasonal pattern. Post-treatment data span 12 months, which allows us to study both immediate and medium-run effects.

Design starts with donor curation. DMAs with incomplete data, idiosyncratic local shocks, or overlapping media interventions during the pre-treatment window are removed from the donor pool. Neighbouring DMAs that share substantial media spillover are flagged for later sensitivity analysis. The analyst then selects predictors: lagged monthly sales in the pre-treatment period, along with covariates such as population, median income, retail distribution, and competitive intensity. This mirrors the general guidance from Section 6.2. The synthetic control should match the treated unit's path over time using lagged outcomes, while the structural covariates help anchor long-run levels.

The synthetic control is constructed using these pre-treatment outcomes and covariates. The optimisation typically produces weights that assign substantial mass to three or four donor DMAs with similar demographics and pre-treatment sales trajectories. In the benchmark specification the relative RMSPE is around 5%, well below typical levels we regard as concerning (Section 6.5). The gap plot shows that the synthetic control tracks the treated DMA closely across all 36 pre-treatment months rather than only in a short sub-window. Stable alignment over the full pre-treatment window is consistent with the factor-structure and convex-hull assumptions in Section 6.3; if fit deteriorated markedly in early periods, this would suggest either factor instability or that the treated DMA lies near the edge of the donor hull. If the fit were weak, or if the synthetic control matched only the last year of data but not the earlier part of the pre-period, the analyst would revisit donor curation and predictor choice, or conclude that the design does not support a credible synthetic control.

In-space placebo checks are then conducted for all donor DMAs using the placebo-check procedures from Section 6.4. For each DMA the analyst re-assigns treatment at the campaign launch date and re-estimates a synthetic control using the same predictors and donor pool. In a successful design the treated DMA's

post-treatment RMSPE ratio is noticeably larger than those of the placebo-check units. In our example it is the largest among all units, yielding a rank-based p-value of  $p = 1/41 \approx 0.024$ . This p-value is conditional on the chosen donor set and pre-treatment window; it does not by itself validate the identification assumptions if donor curation or timing choices are questionable. The gap plot overlaying placebo-check gaps shows that the treated DMA's gap is an outlier, standing above the cloud of placebo-check gaps. If instead the treated DMA's RMSPE ratio lay in the middle of the distribution, the analyst would be cautious about interpreting the observed post-treatment gap as evidence of an effect.

This design rests on an identifying assumption that, conditional on matching the pre-treatment sales path and covariates, there are no DMA-specific shocks that line up exactly with the campaign launch for the treated DMA but not for the donors. Potential violations include concurrent national brand campaigns that differentially affect the flagship market, chain-wide operational changes such as distribution or pricing adjustments that coincide with the campaign launch, and macro shocks such as local economic conditions or competitor actions that affect the treated DMA differently from donors. Spillovers are another important potential violation if the campaign generates awareness in neighbouring DMAs through media overlap or customer mobility. To probe this, donor DMAs geographically adjacent to the treated DMA are excluded in a sensitivity analysis. Stability under adjacency exclusion is consistent with limited spillovers, but does not rule out broader media overlap. Additional falsification checks examine whether other campaign-like events or distribution changes occur in the treated or donor DMAs around the launch date.

Once design and diagnostics are satisfactory, the analyst translates the cumulative effect, defined as the sum of monthly gaps over the 12 post-treatment months, into incremental revenue and compares it to campaign cost to estimate return on investment. The final report combines gap plots, weight tables, placebo-check distributions, and a narrative interpretation that explains any ramp-up dynamics and long-run persistence. This structure illustrates how synthetic control can support an expansion decision to other DMAs while making the identifying assumptions and diagnostic evidence explicit.

## Exclusive Retail Partnership

Exclusive retail partnerships illustrate synthetic control in settings with strategic, non-replicable interventions. Consider a retailer that signs an exclusive partnership with a popular brand, agreeing to carry the brand's full product line in exchange for favourable terms. The partnership is piloted in a single flagship store. The goal is to estimate the effect on store-level revenue and foot traffic. The treated store is the flagship. The donor pool consists of 50 stores in the same chain with similar format, size, and location characteristics. Pre-treatment data span two years (24 months). Post-treatment data span one year (12 months).

The synthetic control is constructed using pre-treatment monthly revenue, foot traffic, and covariates such as store size, local demographics, and proximity to competitors, following the workflow from Section 6.2. Because the pilot store is chosen strategically, one concern is that management selects a store that is already on an improving trajectory for unobserved reasons. The analyst therefore examines pre-treatment trends carefully. In a credible design the synthetic control matches both the level and the slope of revenue and

traffic for the flagship store, and pre-treatment RMSPE is small. If the treated store already exhibits a strong upward drift that cannot be reproduced by any convex combination of donors, the synthetic control design is unlikely to recover the causal effect of the partnership. In terms of the bias expression from Section 6.1, this corresponds to a large factor-loading mismatch: no choice of convex weights can make the mismatch small, so post-treatment bias is likely large even if pre-treatment fit looks good over a short window.

More subtly, pre-treatment trend matching cannot rule out selection on anticipated future performance: if management chose the flagship store because they expected it to outperform, the SC estimate will be upward-biased even with perfect pre-treatment fit. In that case, synthetic control attributes to the partnership a component of performance that would have materialised even without the intervention, unless those expectations are fully captured by observed covariates and pre-treatment dynamics. Perfect pre-treatment fit cannot rule out selection on unobserved future shocks or managerial expectations about performance trajectories that differ from historical trends. Synthetic control cannot in general resolve selection on unobserved expectations about future performance; at best it can show that the observed ramp-up is inconsistent with what similar stores experienced, conditional on observed covariates and pre-treatment paths.

In the benchmark specification the post-treatment gap is positive and growing over time, indicating that the partnership increases revenue with a ramp-up period as customers discover the new product line. In-space placebo checks again show that the treated store's gap is in the top 5% of placebo-check gaps, which, together with the pre-treatment fit, supports the presence of an effect. The analysis includes a sensitivity check excluding stores that may have been indirectly affected by the partnership, for example stores that experience customer substitution from the flagship. If the estimated effect is robust to these exclusions, the analyst gains confidence that the effect is not driven by within-chain reallocation rather than genuine incremental demand.

The retailer uses the synthetic control analysis to assess whether to expand the partnership to additional stores. The gap plot and the cumulative revenue gain are presented to executives, who value the transparency. The blueprint from this example underscores that, in strategic pilot settings, synthetic control is most convincing when strong pre-treatment alignment is combined with robustness to alternative donor pools that address substitution and other selection concerns.

## Regional Regulation

Regional regulations provide a setting where synthetic control is the natural method because treatment is applied to a single geographic or administrative unit by external policy rather than firm choice. Suppose a city enacts a regulation restricting retail promotions, for example banning loss-leader pricing or limiting discount frequency. The goal is to estimate the effect on retail prices and sales. The treated city is the regulated market. The donor pool consists of 30 unregulated cities in the same country. Pre-treatment data span three years (36 months). Post-treatment data span two years (24 months).

The synthetic control is constructed using pre-treatment monthly prices and sales, along with city-level covariates such as population, income, and retail concentration. The pre-treatment fit is very tight, and the

gap plot for prices and sales shows no systematic divergence before the regulation comes into force. After the regulation, the synthetic control counterfactual tracks what would have happened in the absence of the policy, and the post-treatment gap shows that prices increase and sales decline, consistent with economic theory, but still vulnerable to coincident shocks. In-space placebo checks provide strong evidence that the observed pattern is unusual relative to the donor cities.

This design relies on the assumption that, once we control for pre-treatment trends and covariates, there are no other city-specific shocks coinciding with the regulation that move prices and sales in the treated city but not in the donors. These checks operationalise the identifying assumption from Section 6.3 that, conditional on matching pre-treatment paths and covariates, no other city-specific shocks coincide with the regulation in a way donors cannot reproduce. Sensitivity analyses therefore check whether the effect could be driven by concurrent macroeconomic or sectoral shocks. The analyst examines whether donor cities experienced similar changes in national unemployment, inflation, and sector-wide shocks, and verifies that these are either common across treated and donor cities (and thus absorbed by the synthetic control) or negligible in magnitude. The absence of parallel breaks in the donor cities supports a causal interpretation of the regulation's effect.

The analysis is used by policymakers to assess the welfare implications of the regulation and by industry stakeholders to argue for or against expanding the regulation to other markets. The transparency of the synthetic control method, which makes clear exactly which cities contribute to the counterfactual and how closely they match the treated city, helps the analysis speak to both audiences.

## Platform Entry with Spillovers

This example is deliberately constructed as a cautionary case: it illustrates the challenges of synthetic control when interference is present and shows when the method reaches its limits. A food delivery platform enters a major metropolitan market, partnering with restaurants and launching a marketing campaign to attract users. The goal is to estimate the direct effect on restaurant revenues in the treated city and the spillover effect on nearby restaurants and markets. Note that “platform entry” is defined as city-level market access, while restaurant participation is unit-level. This multi-level exposure structure requires an exposure mapping to precisely define the estimand (Chapter 11). The treated market is the city where the platform entered. Potential donors are 25 cities without platform entry. However, spillovers are likely. Restaurants within the treated city that do not join the platform may experience demand shifts, and restaurants in adjacent cities may be affected by cross-border customer mobility.

In a first pass the analyst applies the standard synthetic control recipe. The synthetic control is constructed using pre-treatment monthly restaurant revenues and city-level covariates, and donor cities within a buffer distance of the treated city are excluded to reduce contamination. The pre-treatment fit is acceptable and the post-treatment gap suggests a positive effect on restaurant revenues in the treated city. However, in-space placebo checks reveal that some donor cities exhibit unusual post-treatment gaps that align with the timing of entry, even though they were nominally untreated. This pattern indicates a violation of the no-

interference assumption (Assumption 15). Once Assumption 15 fails at this scale, the bias term  $\sum_j w_j^* \tau_{jt}^{\text{spillover}}$  in equation (6.31) becomes first-order, and vanilla SC no longer targets a well-defined direct effect for the treated city.

Rather than treating this as a nuisance to be fixed by ever more aggressive donor pruning, the analyst reframes the problem using the explicit spillover models developed in Chapter 11. The refined analysis redefines the donor pool to include only cities far from the treated city, models exposure as a function of distance and platform penetration, and estimates direct and spillover effects jointly. In this framework the synthetic control for the treated city provides one ingredient for the counterfactual path, but identification of spillovers relies on the richer exposure mapping and additional structure from the interference setting.

In a representative application the analysis concludes that platform entry raises revenues for participating restaurants in the entry city, generates positive spillovers for some non-participating restaurants through category expansion, and may reduce revenues in adjacent markets through customer substitution. The example is deliberately constructed to show that vanilla synthetic control, even with buffering and placebo checks, can struggle in the presence of substantial interference, and that credible answers require methods that treat spillovers as a first-order design feature rather than a minor complication.

These four applications illustrate the versatility and transparency of synthetic control methods in marketing and provide blueprints for designing and diagnosing studies in practice. The method can provide credible counterfactuals when diagnostics support the design, along with intuitive visualisations and permutation-based inference procedures that communicate results clearly to technical and non-technical stakeholders. Tailor donor curation, predictor selection, and the diagnostic workflow to the substantive context, and document assumptions and sensitivity analyses transparently.

## Offline Advertising and Online Chatter

Tirunillai and Tellis apply synthetic control to measure how offline television advertising affects online user-generated content [Tirunillai and Tellis, 2017]. The unit of analysis is a brand (in a given product category), and the treatment time  $T_0$  is the campaign launch week. The setting is a major consumer brand launching a television advertising campaign. The outcome is online chatter on social media and review platforms, measured along multiple dimensions including popularity (volume of mentions), negativity (sentiment), and visibility (reach). The challenge is that campaign timing is endogenous to brand strategy. Brands launch campaigns when they anticipate favourable conditions or need to counteract negative trends.

The authors construct synthetic controls for treated brands by matching on pre-campaign chatter metrics over 24 weeks. Donor brands are selected from the same product category but did not launch campaigns during the study window. Weights are chosen to minimise pre-treatment root mean squared prediction error across all chatter dimensions simultaneously. This is implemented by stacking the pre-period vectors of each chatter metric into the predictor set and choosing a diagonal  $V$  to balance their relative importance. The method reveals that television advertising increases online chatter popularity by roughly 15 per cent and reduces

negativity by around 8 per cent, with effects persisting for approximately three weeks before decaying to baseline (exact magnitudes from Tirunillai and Tellis [2017]).

This application illustrates how synthetic control can mitigate concerns about endogenous campaign timing when we observe a long, well-fitted pre-treatment path and have a rich donor pool from the same product category, but it does not eliminate the possibility that unobserved brand-specific shocks coincide with launch. The design is most credible when there are no emerging pre-trends in chatter immediately before the campaign and no brand-specific shocks coinciding with launch that donor brands cannot reproduce. The multi-dimensional outcome structure requires careful aggregation. When testing effects on multiple chatter dimensions (popularity, negativity, visibility), inference should account for multiplicity, for example by using joint tests or family-wise error control as discussed in Chapter 16. The authors report separate effects for each chatter dimension and conduct in-space placebo checks by applying the method to each donor brand as if it were treated. The treated brand's post-treatment RMSPE ratio ranks in the top decile of the placebo-check distribution (see Section 6.4 for the RMSPE ratio and ranking logic), supporting the inference that the observed effects are unusual relative to the donor pool. The analysis links offline advertising investment to online engagement metrics, demonstrating cross-channel effects that inform integrated marketing strategies.

## 6.10 Workflow Checklist

This section provides a compact end-to-end protocol for conducting synthetic control analyses in marketing panels. The workflow integrates design, construction, diagnostics, inference, and reporting. Following these nine steps systematically helps practitioners avoid common pitfalls and produce analyses that withstand scrutiny.

**Step 1: Define the Treated Unit and Intervention Time.** Identify the treated unit (or units) and the calendar period in which treatment begins. If treatment is announced in advance, decide whether the intervention time is the announcement date or the implementation date. Base this decision on whether anticipation is plausible. Document the treatment and its timing clearly. Provide institutional context on why the unit received treatment and how it was implemented.

**Step 2: Curate the Donor Pool.** Assemble the donor pool by excluding units that received treatment themselves, units that were contaminated by spillovers, and units that are fundamentally incomparable to the treated unit. Document donor selection criteria (for example, geographic region, market size, operational characteristics). Provide summary statistics comparing donors to the treated unit. If spillovers are plausible, define a buffer zone and exclude donors within the buffer. Report the final donor pool size and the characteristics of included donors.

**Step 3: Select Predictors and Pre-Treatment Window.** Define the pre-treatment window (all periods before the intervention time). Select predictors including pre-treatment outcomes (monthly or quarterly outcomes for the full pre-treatment window) and key covariates (pre-treatment or time-invariant characteristics). Avoid including post-treatment covariates or covariates affected by treatment. Document the predictor set and the rationale for including each predictor. If the pre-treatment period is short, consider using longer lags of outcomes or differenced outcomes to capture trends.

**Step 4: Fit Synthetic Control and Assess Pre-Treatment Fit.** Solve the weight optimisation problem to obtain synthetic control weights  $w_j^*$  for  $j \in \mathcal{J}$ . Compute the pre-treatment RMSPE and assess whether the fit is tight relative to the scale of the outcome and the variability across units. Plot the treated unit's pre-treatment outcomes and the synthetic control's pre-treatment outcomes on the same axes to visualise the fit. Check predictor balance by comparing the predictor values for the treated unit to the weighted average of predictor values for the donors. If the fit is poor or predictor balance is weak, revisit the donor pool or the predictor set. Iterate until satisfactory, but recognise that iterating on donors and predictors improves pre-treatment fit while also opening the door to overfitting on noise and hidden specification search. Where possible, archive an analysis plan (donor pool, predictors, and tuning choices) before you look at post-treatment outcomes; treat any later changes as exploratory analyses that require clear labelling. Section 6.5 discusses how to use RMSPE, relative RMSPE, and sensitivity analyses to distinguish genuine structure from overfitting.

Report the synthetic control weights explicitly. Identify which donors receive substantial weight and assess whether the weights are concentrated on a few units or distributed broadly. Sparse weights (few donors with large weights) can indicate that the treated unit is well-matched by a small set of donors, but they also increase variance and make results sensitive to those donors. Diffuse weights reduce variance by averaging more broadly, but may signal that no single donor is a close match and that identification relies on weaker interpolation. What matters is not only sparsity but also whether the high-weight donors are substantively plausible comparators.

**Step 5: Produce Gap Plots and Estimate Post-Treatment Effects.** Extend the plot of outcomes into the post-treatment period, showing the treated unit's observed outcomes and the synthetic control's counterfactual outcomes. Compute the gap (treated minus synthetic) for each post-treatment period and plot the gap over time. Compute the cumulative gap (sum of gaps over the post-treatment window) and the average gap (mean over the post-treatment window) as summary measures of the treatment effect. Interpret the gap plot carefully. Does the gap open immediately after treatment, or is there a ramp-up period? Does the gap persist, or does it decay?

**Step 6: Run In-Space and In-Time Placebo Checks.** Conduct in-space placebo checks (Section 6.4) by applying the synthetic control method to each donor unit as if it received treatment at the intervention time. Compute post-treatment gaps and RMSPE ratios for all placebo-check units. Plot the treated unit's gap alongside placebo-check gaps to assess whether the treated unit's gap is an outlier. Compute the rank-based p-value and report it.

Conduct in-time placebo checks by applying the synthetic control method using a pseudo-intervention date in the middle of the pre-treatment period. Assess whether the pseudo post-treatment gap is near zero. Report placebo-check results transparently, including plots and rank statistics, recognising that in-time placebo checks are diagnostics of stability rather than formal hypothesis tests (see Section 6.4). Note that in-time placebo checks do not diagnose spillovers or anticipation directly; they mainly flag overfitting or instability in the factor structure.

**Step 7: Select Inference Procedure and Report Uncertainty.** Choose an inference procedure based on the sample size and the design. Permutation-based, rank-based p-values (Section 6.4) are natural when there is a reasonably large donor pool. With 20 donor units, the smallest achievable p-value is about 1/21, so more donors improve the resolution of the permutation distribution but p-values remain discrete and can be coarse. More generally, with  $|\mathcal{J}|$  donors, the smallest achievable p-value is  $1/(|\mathcal{J}|+1)$ . In observational settings, interpret rank p-values as measures of relative extremeness in the donor pool, not as probabilities derived from known random assignment. When the donor pool is small or when you care about effects over many post-treatment periods, conformal inference (Section 6.4) provides a principled way to construct confidence intervals and uniform bands that account for serial correlation and weight uncertainty; prefer conformal methods when confidence intervals rather than hypothesis tests are the primary goal. When different weighting schemes such as DiD, SC, and SDID yield divergent estimates, consider reporting Synthetic Parallel Trends bounds and their confidence sets (Section 6.3) alongside point estimates.

In all cases, report RMSPE ratios, ranks, and p-values, and interpret them alongside the substantive magnitude and persistence of the gap rather than as binary pass–fail criteria. If multiple outcomes or multiple post-treatment periods are tested, discuss multiplicity and consider adjustments or joint tests (for example, sup- $t$  bands or family-wise error control across post-treatment periods rather than interpreting each pointwise interval in isolation).

**Step 8: Conduct Sensitivity Analyses.** Vary the donor pool (Section 6.5) by excluding influential donors one at a time (leave-one-donor-out) and re-estimating the synthetic control. Report the range of post-treatment gaps across leave-one-donor-out specifications. Vary the predictor set by including or excluding specific covariates or pre-treatment periods, and re-estimate the synthetic control. If spillovers or anticipation are concerns, vary the buffer zone or the intervention date and assess robustness. Construct a specification curve showing the distribution of estimated effects across defensible modelling choices. If estimates cluster tightly, conclusions are robust. If estimates vary widely, document the sensitivity and discuss which specifications are most credible. The goal of the specification curve is to map model dependence across a pre-specified set of defensible modelling choices, not to search for a single “best” specification. Section 6.5 provides concrete diagnostic thresholds and examples for interpreting such specification curves. Readers should see how conclusions respond to defensible changes in donor pools, predictors, and timing choices.

**Step 9: Document Assumptions and Report Results.** Prepare a comprehensive report that includes the research question, the treated unit and intervention time, the donor pool curation process, the predictor set, the pre-treatment fit metrics, the synthetic control weights, the gap plot, the placebo-check results, the inference procedure, the sensitivity analyses, and the substantive interpretation. Articulate the identification assumptions clearly. When articulating identification assumptions, refer back to the formal Assumptions 14–17 so that readers can see exactly which potential violations each diagnostic is intended to address. State the no anticipation, no interference, and stable factor structure (Assumption 16) assumptions explicitly. Provide evidence that they are plausible. Discuss threats to validity (poor fit, contamination, structural breaks) and the robustness of conclusions. If the analysis was pre-registered, document any deviations from the pre-analysis plan and provide justification.

Translate the estimated gap into business metrics relevant for marketing decisions. Ensure that the chosen summary metrics (for example, cumulative effects, average post-treatment effects, or long-run multipliers) match the decision problem, and compute them using the event-time frameworks from Chapters 5 and 4 where appropriate. Express effects in terms of incremental revenue, market share changes, or return on investment. Discuss the ramp-up dynamics, persistence, and long-run implications. Relate findings to decision-relevant questions such as whether to expand the intervention, adjust its design, or discontinue it.

Provide replication materials including cleaned data (or simulated data if proprietary), analysis scripts, and documentation of software versions. This enables readers to verify results, assess the identification assumptions in light of the data, and, if needed, conduct alternative analyses.

By following this workflow, practitioners can conduct synthetic control analyses that are transparent, rigorous, and aligned with modern best practices. The workflow integrates design-aligned reasoning, careful donor curation, diagnostics, permutation-based inference, and sensitivity analysis. When identification

assumptions are plausible and sensitivity analyses are reported transparently, this approach can support credible conclusions.

## Chapter 7

# Generalised and Augmented Synthetic Control

This chapter develops hybrid methods that blend design-based weighting with outcome modelling. They target the same unit-specific effects and treated-unit averages as synthetic control, including ATT, cohort-time effects  $\tau(g, t)$ , and event-time effects  $\theta_k$ . Relative to Chapter 6, the key change is how we construct counterfactuals by combining donor weights with a structured outcome model motivated by time fixed effects  $\lambda_t$  and latent factors  $f_{tr}$  with unit-specific loadings  $\lambda_{ir}$ . You will learn how to formalise augmented synthetic control (ASCM), regularised synthetic control (including ridge and elastic-net penalties), and synthetic difference-in-differences (SDID). We state the identification assumptions these hybrids rely on, then explain how tuning, donor curation, diagnostics, and inference support a disciplined workflow for multiple treated units and staggered adoption. We close with marketing settings where pre-treatment histories are short, treated units are few, and data are messy.

## 7.1 Motivation and Setup

Credibility and efficiency often pull in opposite directions. Design-based methods earn trust by avoiding strong functional-form assumptions, but they pay a price. When the treated unit lies outside the donor pool's convex hull, as discussed in Section 6.1, weighting alone cannot close the gap. The fit deteriorates, bias creeps in, and pure synthetic control, for all its transparency, sometimes fails. In the factor-model perspective from Chapter 6, this failure corresponds to a mismatch in the treated unit's latent factor loadings relative to any convex combination of donor loadings, so that the remaining imputation error can be first-order.

This chapter confronts that failure and offers a resolution. Hybrid methods blend weighting with outcome modelling to gain flexibility while staying anchored to the design logic of synthetic control. They can improve pre-treatment fit and reduce variance, but they also introduce modelling and stability assumptions. The question is whether the compromise works, and when it does not.

Consider what happens when a retailer pilots a loyalty programme in five flagship stores. The pilot runs for two years. The retailer has perhaps twenty control stores, but none resembles the flagships. These stores anchor high-traffic urban centres, draw different customer segments, and generate revenue patterns that no convex combination of suburban or regional stores can replicate. Standard synthetic control produces weights, but the pre-treatment fit is poor. The gap between the synthetic and actual trajectories signals trouble. Any post-treatment estimate carries that residual bias forward.

Hybrid methods attack this problem from three directions. Augmented synthetic control pairs the weighting estimator with a regression adjustment that corrects for the residual imbalance. If the weights produce a synthetic control that undershoots the treated unit's pre-treatment revenue by five per cent, the regression model estimates that gap and uses it to re-centre the synthetic control's counterfactual trajectory before computing the treatment effect. Regularised synthetic control takes a different path: it shrinks the weights toward simplicity, trading a small increase in bias for a larger reduction in variance. Synthetic difference-in-differences introduces time weights alongside unit weights, aligning both the cross-sectional and temporal dimensions before computing contrasts.

Each approach relaxes a constraint that pure synthetic control imposes. They respond to two distinct failure modes. The first is interpolation failure: when the treated unit sits outside the donor convex hull, no choice of non-negative weights can match its untreated path. The second is overfitting: when the donor pool is large or the pre-period is noisy, the optimisation can chase idiosyncratic fluctuations and produce a synthetic path that does not generalise. You gain flexibility, but flexibility has costs. Adding a regression adjustment introduces model dependence. Regularisation requires tuning choices that practitioners must justify. Time weights rely on a reweighted parallel-trends-style comparison that may or may not be credible. Hybrid methods are not strictly superior to their predecessors. They occupy a different point on the bias-variance frontier, and the right choice depends on the data structure you face.

The factor model foundation developed in Chapter 6 remains central. As defined there, untreated potential outcomes decompose into unit-specific loadings and time-varying factors. When the treated unit's loadings lie outside the convex hull of donor loadings, pure synthetic control cannot recover the counterfactual without bias. Hybrid methods modify the constraints that govern this imputation. Augmented synthetic control uses a regression adjustment to repair residual mismatch after weighting. Regularised variants pe-

nalise extreme weight vectors and can stabilise estimates when plain synthetic control overfits. Synthetic difference-in-differences reweights both units and pre-treatment periods, changing which comparisons carry identifying weight.

Return to the loyalty programme. The five flagship stores differ from controls not just in observable characteristics such as square footage, foot traffic, and product mix, but in unobservable ways that load onto latent factors. Urban flagship stores may respond more strongly to macroeconomic shocks, and their customer base may exhibit different seasonal patterns. Pure synthetic control tries to match on pre-treatment outcomes, hoping the match implicitly captures these factor loadings. When it fails, the treated unit sits outside the convex hull, and no combination of donors can replicate its trajectory.

Augmented synthetic control (ASCM) offers a partial fix. You fit the best synthetic control you can, then estimate a regression of outcomes on covariates or lagged outcomes within the control group. You apply that regression to adjust the synthetic control's predictions for the treated unit. Under conditions we make precise later (for example, correctly specified factor structure or correctly specified regression adjustment within the control group), this estimator has a doubly robust flavour: it can remain close to unbiased if either the weighting scheme or the regression adjustment is well specified. But “doubly robust” does not mean “doubly correct”. In small samples, and marketing panels are almost always small in the relevant dimension, misspecification on either side can still generate finite-sample bias. The insurance policy has limits.

Regularisation addresses a different pathology. When the donor pool is large, synthetic control can overfit. The optimisation finds weights that match pre-treatment outcomes precisely, but those weights may be chasing noise. A donor that happens to correlate with the treated unit's idiosyncratic pre-treatment fluctuations receives weight it does not deserve. Out of sample, the synthetic control performs poorly. In variance terms, highly concentrated weight vectors (small effective number of donors) inflate the idiosyncratic component  $\sigma^2(1 + \sum_j(w_j^*)^2)$  from Section 6.4, making overfitted synthetic controls particularly unstable. Ridge-type penalties shrink weights toward uniformity or toward zero. Elastic-net variants combine sparsity with shrinkage. The bias-variance trade-off shifts: you accept some pre-treatment imbalance to avoid fitting noise.

Synthetic difference-in-differences (SDID) makes a structural innovation. Identification in SDID relies on a reweighted common-trend condition: after applying the estimated unit and time weights, untreated potential outcomes for treated and control units follow parallel trends in the reweighted space. SDID does not make parallel trends true by construction. It shifts the identifying burden onto whether the reweighted comparison is credible in your application. Standard synthetic control constructs a weighted average of donor units. SDID also constructs a weighted average of pre-treatment time periods. The estimator computes a double contrast: treated minus synthetic control, post-period minus a weighted pre-period. This structure nests both difference-in-differences, with equal unit weights and equal time weights, and synthetic control, with optimised unit weights and equal time weights, as special cases. By optimising both sets of weights, SDID can handle settings where neither pure method would succeed, provided its reweighted common-trend assumptions are credible.

Now consider a brand launching campaigns in twenty markets over three years. Markets adopt treatment at different times. Some start early, others late, some never. This is staggered adoption, a setting introduced in Chapter 4, and it complicates everything. You cannot simply pool post-treatment observations, because early adopters contribute to the donor pool for late adopters. The safe way to organise a staggered analysis

is to define cohort–time effects  $\tau(g, t)$  relative to not-yet-treated or never-treated donors, and then aggregate them into event-time summaries  $\theta_k$  with transparent, non-negative weights. Hybrid methods aim to improve the quality of each  $\tau(g, t)$  estimate by constructing better counterfactuals within each group–time cell. They do not, by themselves, eliminate the aggregation pitfalls discussed in Chapter 4.

The design-first philosophy remains central, but it needs reinterpretation in hybrid settings. Pure synthetic control embodies the idea that identification lives in the assignment process and the comparability of donors, not in a fitted regression. Hybrid methods introduce modelling, which creates dependence on additional stability and specification assumptions. The resolution is not to abandon design thinking but to use it as discipline. You choose models that respect the structure of the problem, validate them through diagnostics, and report sensitivity to specification. The goal is transparency about what you are assuming, not purity about what you are not.

What follows develops these ideas systematically. We begin with augmented synthetic control in Section 7.2, which addresses bias from poor pre-treatment fit by combining synthetic control weighting with outcome regression. We then examine regularised and balancing approaches in Section 7.3, which stabilise weights and improve robustness to overfitting. Synthetic difference-in-differences receives extended treatment in Sections 7.4–7.10, showing how unit and time weights reshape the comparison that underpins a parallel-trends-style argument. Finally, we briefly survey matrix completion and other machine learning hybrids that relax linearity assumptions further.

The core message is pragmatic. Hybrid methods are not a replacement for design. They are an extension, a way to make design-disciplined inference work in settings where the data would otherwise defeat it. Use them when pure methods fail. Understand what you gain and what you give up. Report your choices transparently. That is how credible evidence gets built.

## 7.2 Augmented Synthetic Control (ASCM)

### The Problem That Motivates Augmentation

Return to the five flagship stores piloting a loyalty programme. Standard synthetic control searches for weights that make a convex combination of control stores match the flagships' pre-treatment revenue trajectory. In this case the optimisation stalls short of that goal. The flagships anchor high-traffic urban centres with customer profiles that no suburban or regional store replicates. Even the best synthetic control undershoots the flagships' baseline revenue by around 8%. That gap carries forward into the post-treatment period, so the synthetic control estimator mixes the treatment effect with residual bias from imperfect pre-period fit.

Augmented synthetic control attacks this problem directly. It pairs the weighting estimator with a regression adjustment that corrects the gap in expected outcomes. If the synthetic control undershoots by 8% in periods where the regression predicts the treated store should be higher, the augmentation shifts the counterfactual up to reflect that systematic difference. The benefit is that we can improve fit when the treated unit sits near, but not squarely inside, the donor convex hull. The cost is model dependence: once we add a regression component, the estimator inherits whatever misspecification that model carries.

### The Estimator

Let unit 1 be treated and let  $\mathcal{J}$  index donor stores. Denote by  $\hat{\mathbf{w}}$  the weights from the standard synthetic control optimisation, with components  $\hat{w}_j$ , and let  $\hat{m}_{it}$  be predictions from an auxiliary outcome model for  $Y_{it}(0)$ , estimated on pre-treatment data for donors and (where available) the treated unit. In practice  $\hat{m}_{it}$  often comes from a ridge regression of outcomes on a covariate vector  $X_{it}$ . In many applications  $X_{it}$  is time-invariant and the  $t$  subscript simply keeps the notation consistent with the panel setup.

For a post-treatment period  $t > T_0$ , the augmented counterfactual for the treated unit takes the form

$$\hat{Y}_{1t}^{\text{ASCM}}(0) = \sum_{j \in \mathcal{J}} \hat{w}_j Y_{jt} + \left( \hat{m}_{1t} - \sum_{j \in \mathcal{J}} \hat{w}_j \hat{m}_{jt} \right).$$

The first term is the standard synthetic control prediction. The second term corrects for any residual imbalance that the auxiliary model attributes to systematic differences between the treated unit and its synthetic control. When, according to the regression model, the treated unit and the synthetic control have the same expected outcome in period  $t$ , this correction vanishes and the estimator collapses back to standard synthetic control.

A useful diagnostic is to report the size of the augmentation term,  $\hat{m}_{1t} - \sum_{j \in \mathcal{J}} \hat{w}_j \hat{m}_{jt}$ , relative to the weighted-donor term,  $\sum_{j \in \mathcal{J}} \hat{w}_j Y_{jt}$ . If the augmentation dominates, the estimate is driven primarily by extrapolation through the outcome model rather than by donor interpolation.

The period- $t$  treatment effect estimator is

$$\hat{\tau}_{1t}^{\text{ASCM}} = Y_{1t} - \hat{Y}_{1t}^{\text{ASCM}}(0) = (Y_{1t} - \hat{m}_{1t}) - \sum_{j \in \mathcal{J}} \hat{w}_j (Y_{jt} - \hat{m}_{jt}).$$

You can read this as reweighting residuals from the auxiliary outcome model. We strip out the component of revenue that the regression explains and then apply synthetic control to the remaining, unexplained part. This representation makes clear that any misspecification in the regression model flows directly into the residuals we reweight.

## Why Augmentation Helps — and When It Hurts

Augmentation is attractive because it gives us another route to good counterfactuals. If the synthetic control weights alone already capture the relationship between treated and donor units, the augmentation term is small and ASCM behaves much like standard synthetic control. If the weights are imperfect but the regression model captures how covariates predict outcomes across stores and time, the augmentation can correct much of the bias from imperfect pre-treatment fit.

This “two chances to get it roughly right” story is sometimes described as a form of double robustness [Ben-Michael et al., 2021]. In our setting, the analogy is that if either (i) the SC weights approximate the untreated counterfactual well or (ii) the regression adjustment  $\hat{m}_{it}$  is close to  $\mathbb{E}[Y_{it}(0) | X_{it}]$ , then ASCM can substantially reduce bias relative to pure SC. In the strict econometric sense, however, ASCM is not generically doubly robust in the way classical missing-data estimators are. In finite marketing panels both components are estimated, both are noisy, and both can be misspecified. The practical message is more modest: augmentation can reduce bias when one of the components tracks the untreated potential outcomes well, but it can increase variance and even bias when both components are off in different directions.

A deeper issue is extrapolation. Standard synthetic control constrains the treated unit’s counterfactual to live inside the convex hull of donor outcomes. When the treated store is an outlier relative to donors, this constraint forces you to admit that no credible synthetic control exists. ASCM keeps the convex-hull restriction for the weighted donor outcomes, but then adds a regression correction that can push the augmented counterfactual outside that hull. For the flagship stores, the regression might infer that urban locations systematically load more heavily on an “urban-consumer” factor than any individual donor store. The augmentation then shifts the counterfactual store up to reflect that higher loading. If this model of how covariates map into outcomes is right, extrapolation buys you a better counterfactual. If it is wrong, the same mechanism projects you into the wrong part of outcome space.

## Identification Assumptions

ASCM targets the same estimand as standard synthetic control: the treatment effect on the treated unit in each post-treatment period,  $\tau_{1t} = Y_{1t}(1) - Y_{1t}(0)$ . The estimator replaces the unobserved  $Y_{1t}(0)$  with the

augmented counterfactual  $\hat{Y}_{1t}^{\text{ASCM}}(0)$ . The assumptions extend those for synthetic control by adding structure on the regression component.

First, the no-anticipation and no-interference conditions introduced in Chapter 6 continue to apply. Pre-treatment outcomes for the treated unit must equal their untreated potential outcomes,  $Y_{1t} = Y_{1t}(0)$  for  $t \leq T_0$ , and donor outcomes must not be affected by the treated unit's loyalty programme, so  $Y_{jt} = Y_{jt}(0)$  for all donors and all periods. In the loyalty-programme example this rules out substantial competitive responses that meaningfully shift donor revenue paths.

Second, we require that the outcome model and the weighting step together approximate the treated unit's untreated path. A convenient way to express this is to define an outcome model

$$m_{it} := \mathbb{E}[Y_{it}(0) | X_{it}]$$

and residualised untreated outcomes  $u_{it} := Y_{it}(0) - m_{it}$ . The ASCM logic is that the weights should approximately balance these residualised outcomes in the pre-treatment period,

$$u_{1t} \approx \sum_{j \in \mathcal{J}} w_j^* u_{jt}, \quad t \leq T_0,$$

while the fitted model  $\hat{m}_{it}$  provides a stable approximation to  $m_{it}$  for imputing untreated outcomes in the post-treatment period.

Finally, ASCM needs stability of the regression relationship across the treatment boundary. The outcome model that underlies  $\hat{m}_{it}$ , estimated from pre-treatment donor data (and, where relevant, the treated unit), must continue to describe how covariates relate to untreated outcomes in the post-treatment period. Structural breaks that change this mapping—such as a major platform algorithm shift or a sharp macroeconomic shock—will cause the regression component to extrapolate incorrectly. Because we never observe  $Y_{1t}(0)$  after treatment, this stability condition cannot be verified directly. You must justify it using institutional knowledge and auxiliary evidence about how  $X_{it}$  and outcomes co-move around the intervention.

Some parts of this structure can be partially assessed. You can check whether the augmentation improves fit within the pre-treatment period by splitting it into training and validation segments. Estimate the weights and regression on the training segment, predict the validation segment, and compare prediction errors with and without augmentation. If the augmented model consistently worsens pre-period fit, it is unlikely to repair problems after treatment. What you cannot observe is whether the same regression relationship holds once the loyalty programme goes live. That extrapolation across the treatment boundary is, as in many causal designs, the part you must defend with economic argument rather than direct data.

## Implementation

Implementing ASCM requires three linked choices: the predictors for synthetic control, the covariates for augmentation, and the form of the regression model. The predictors for synthetic control typically include a vector of pre-treatment outcomes and a small set of time-invariant store characteristics, mirroring the

baseline synthetic control set-up from Chapter 6. The regression covariates may overlap this set or extend it with transformations, trends, or interactions that capture how store attributes relate to revenue dynamics.

Ridge regression is a natural default because it handles collinearity in marketing covariates and shrinks coefficient estimates when we only have a short pre-treatment history. When the covariate set is rich relative to  $T_0$ , use sample splitting or cross-fitting for the regression step so that the same noise in pre-treatment outcomes does not drive both weight choice and regression adjustment. In the flagship example, the weighting step might use monthly revenue over twenty-four pre-treatment months along with store characteristics such as square footage, average foot traffic, and a product-mix index. The regression step then uses these same covariates to predict each store's revenue. You choose the ridge penalty  $\lambda$  by cross-validation within the pre-treatment period, trading off in-sample fit against stability of predictions.

Short pre-treatment periods create a tension. With few time points, the regression has limited data and will overfit unless regularised. Ridge and elastic net penalties help by shrinking coefficients towards zero and, in the elastic net case, encouraging sparsity in covariate selection. At the same time, a large donor pool can make the synthetic control weights diffuse, spreading small positive weights across many donors whose characteristics only loosely resemble the flagship stores. In that setting the regression correction does much of the work, effectively pulling the counterfactual towards what the model thinks an “urban flagship” should look like given its covariates. When the treated unit's characteristics sit far from the donor distribution, that correction becomes a strong extrapolation and should be interpreted with care.

A sensible validation strategy mirrors what you did for basic synthetic control. Split the pre-treatment period into a training block and a holdout block. Estimate synthetic control weights and the regression on the training block, use them to predict the holdout block, and compare root mean squared prediction error with and without augmentation. This checks predictive stability within the untreated regime. It does not validate that the same relationship will hold once treatment starts. If ASCM does not improve prediction in the holdout, there is little reason to trust it in the post-treatment period. You can also run placebo checks with pseudo intervention dates inside the pre-period, exactly as in Chapter 6, and examine whether the resulting pseudo treatment effects concentrate near zero. In the ASCM setting, placebo checks stress-test the weighting and regression components: large pseudo-effects indicate that at least one of them is capturing noise or unstable structure rather than the untreated trajectory.

## Comparing ASCM to Alternatives

ASCM is most attractive when the treated unit lies near the boundary of the donor convex hull and simple synthetic control cannot achieve good pre-treatment fit. In that setting the regression adjustment can use observed differences in store attributes and pre-trends to repair part of the mismatch. The trade-off is that ASCM is more model dependent: any misspecification in the regression component directly feeds into the adjusted counterfactual.

Good empirical practice is to treat ASCM as one estimator in a small ensemble rather than as a replacement for simpler designs. You can run standard synthetic control, event-study difference-in-differences, and ASCM

on the same loyalty-programme experiment, all targeting the same post-treatment ATT for the flagship stores. If the three estimators tell a consistent story, you gain confidence that conclusions are not driven by the modelling choices specific to any one method. If they diverge, that divergence is itself information. For example, large differences between ASCM and standard SC with similar pre-period fit point to sensitivity to the regression specification, whereas differences between SC and event-study DiD with similar covariates highlight tension between convex-hull and parallel-trends assumptions. In that case the right response is not to pick the most “favourable” estimate but to diagnose why the methods disagree and to reflect that uncertainty in how you present the results.

The next section turns to a different way of addressing poor fit. Rather than adding a regression correction on top of synthetic control, regularised synthetic control modifies the weight construction itself, shrinking towards simpler weighting schemes or enforcing explicit balance constraints.

### 7.3 Regularised and Balancing Variants of SC

Standard synthetic control can overfit. When the donor pool is large and the pre-treatment period short, the optimisation can find weights that chase noise rather than signal. A donor that happens to share an idiosyncratic seasonal blip with the treated unit may receive high weight for the wrong reason. The synthetic control then fits pre-treatment outcomes tightly but predicts post-treatment outcomes poorly.

Return to the flagship stores. The retailer has twenty potential control stores. Standard synthetic control assigns weight 0.45 to one suburban store that happens to experience a similar seasonal spike in month fourteen, weight 0.35 to a regional store with an unrelated promotional calendar, and spreads the remaining 0.20 across three others. The pre-treatment root mean squared prediction error looks impressively low. Yet the weights reflect coincidence, not structural similarity. When the loyalty programme launches, the synthetic control diverges from the flagships' trajectory because the weighted donors share noise, not fundamentals.

Regularised synthetic control attacks this problem by penalising weight configurations that are overly concentrated. Instead of letting the optimisation pursue pre-treatment fit at any cost, regularisation shrinks weights towards simpler, more diffuse patterns. This typically sacrifices a small amount of in-sample fit in exchange for weights that are more stable and less sensitive to sampling variation.

#### Ridge Synthetic Control

Ridge synthetic control adds an L2 penalty to the standard optimisation problem [Doudchenko and Imbens, 2016]. Let  $\mathbf{X}_1$  collect pre-treatment outcomes and covariates for the treated unit and let  $\mathbf{X}_0$  stack the corresponding donor values, so that  $\mathbf{X}_1 \in \mathbb{R}^k$  and  $\mathbf{X}_0 \in \mathbb{R}^{k \times N_0}$ , where  $N_0$  is the number of donors. Let  $\mathbf{V}$  weight the predictors as in Chapter 6. Ridge SC solves

$$\min_{\mathbf{w}} \|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}\|_{\mathbf{V}}^2 + \eta \|\mathbf{w}\|^2 \quad \text{subject to} \quad w_j \geq 0, \sum_j w_j = 1,$$

where  $\|a\|_{\mathbf{V}}^2 := a' \mathbf{V} a$ ,  $\eta > 0$  controls the strength of the penalty, and  $\|\mathbf{w}\|^2 = \sum_j w_j^2$ . Under the simplex constraints, minimising  $\|\mathbf{w}\|^2$  discourages putting all the mass on one or two donors and instead favours more diffuse weights. In fact, if  $\bar{\mathbf{w}}$  denotes the uniform vector with components  $1/N_0$ , then  $\|\mathbf{w} - \bar{\mathbf{w}}\|^2 = \|\mathbf{w}\|^2 - 1/N_0$ , so penalising  $\|\mathbf{w}\|^2$  is equivalent (up to an additive constant) to shrinking towards uniform weights. In terms of the variance decomposition in equation (6.24), shrinking  $\|\mathbf{w}\|^2$  reduces the Herfindahl index  $\sum_j (w_j^*)^2$  and thereby the idiosyncratic-variance component  $\sigma^2(1 + \sum_j (w_j^*)^2)$ . Uniform weights  $w_j = 1/N_0$  minimise the penalty, while highly concentrated weights incur a large penalty.

The parameter  $\eta$  governs the trade-off. If you set  $\eta = 0$  you recover standard synthetic control. As  $\eta$  grows, the objective gives more weight to simplicity and less to pre-treatment fit, and the solution moves towards uniform weights. The goal is to choose an intermediate  $\eta$  that retains enough structure from the pre-treatment data while damping out idiosyncratic noise. A practical way to do this is to use the same cross-validation logic introduced for ASCM. Split the pre-treatment period into training and validation blocks, estimate ridge

SC over a grid of  $\eta$  values on the training block, and select the  $\eta$  that minimises prediction error on the validation block.

Cross-validation can reduce overfitting to pre-period noise, but it cannot validate that the resulting weights will produce unbiased post-treatment counterfactuals.

Applied to the flagship stores, ridge SC with a cross-validated  $\eta$  will typically spread weights across more donors. Pre-treatment RMSPE might increase from 0.03 to 0.05, which is a modest deterioration in fit. In return, the synthetic control reflects a broader set of stores with similar customer demographics and format characteristics, rather than a narrow pair that merely share a transient seasonal spike.

Regularisation is not free. If the true counterfactual is genuinely well-approximated by a sparse combination of donors that closely resemble the flagships, ridge SC dilutes those weights and introduces bias. You trade accuracy in those best-case designs for robustness in the more common case where the data are noisy and the pre-period is short. From a design perspective, standard SC and ridge SC both construct counterfactuals as weighted averages of observed donors, using the same predictor set. Ridge SC simply adds a stronger structural assumption on the weights: that you prefer diffuse combinations unless the data provide a compelling reason to concentrate.

## Balancing Synthetic Control

Ridge SC regularises via the objective function. Balancing SC takes a complementary route and builds explicit covariate balance into the constraints. The optimisation problem becomes

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to} \quad w_j \geq 0, \quad \sum_j w_j = 1, \quad \|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}\|_B \leq \delta,$$

where  $\|\cdot\|_B$  measures imbalance in a chosen set of balance statistics and  $\delta$  is a tolerance parameter chosen ex ante as a credibility constraint. For example, you might take  $\|a\|_B := \sqrt{a' \mathbf{B} a}$  for a positive semidefinite weighting matrix  $\mathbf{B}$  built from standardised mean differences over a selected covariate subset. The objective again prefers diffuse weights, while the balance constraint insists that the synthetic control match the treated unit on key predictors within tolerance  $\delta$ .

This formulation inverts the emphasis of standard SC. Standard SC minimises a measure of covariate imbalance subject to convexity constraints on the weights. Balancing SC minimises weight complexity while treating covariate balance as non-negotiable. The inversion matters when you have strong prior views about which characteristics must be matched for the design to be credible.

Consider a DMA-level advertising study where the treated DMA (San Francisco) has population 4.7 million, median household income 112,000 dollars, and baseline monthly sales of 2.3 million dollars. Suppose the analyst believes that any credible counterfactual must match these characteristics within about 10 per cent. Standard SC might achieve an impressive match on pre-treatment sales by assigning large weight to a smaller DMA whose sales path happens to track San Francisco but whose demographics look nothing like it. Balancing SC encodes the demographic requirements directly: it looks for weights that bring standardised

differences in population, income, and baseline sales below a chosen threshold, even if that forces some deterioration in the fit of the pre-treatment sales path.

The tolerance parameter  $\delta$  tunes this tension. Very small  $\delta$  values demand near-exact balance and may make the problem infeasible if the treated unit lies outside the donor convex hull. Very large  $\delta$  values relax the constraint so much that the solution drifts back towards uniform weights. In practice you choose  $\delta$  by inspecting standardised mean differences for critical covariates and aiming for thresholds (for example below 0.1 in absolute value) as a heuristic, calibrated to how predictive those covariates are for outcomes.

Balancing SC and ridge SC serve different purposes. Balancing SC is useful when particular covariates, such as demographics or baseline volume, are non-negotiable for credibility and you want that requirement enforced mechanically. Ridge SC is better suited when you have a large donor pool relative to the pre-period and are primarily worried about overfitting idiosyncratic noise, without strong prior views about specific covariates beyond the predictor set already used in SC.

## Elastic Net and Other Variants

Elastic net synthetic control combines L1 and L2 penalties in the objective. In regression settings the L1 term induces sparsity by pushing many coefficients exactly to zero. In synthetic control designs with convexity constraints, many donors already receive zero weight even without an L1 penalty, so in typical marketing panels the additional sparsity from an explicit L1 term is modest. Under the non-negativity and sum-to-one constraints,  $\|\mathbf{w}\|_1 = 1$  for all feasible  $\mathbf{w}$ , so any L1 penalty must operate on deviations from a target (for example,  $\|\mathbf{w} - \bar{\mathbf{w}}\|_1$ ) or be embedded in a formulation without the simplex constraint. In practice, this limits the incremental impact of L1 penalties beyond what convexity already imposes. Elastic net SC can be useful when the donor pool is extremely large and you want aggressive regularisation, but ridge SC will suffice in most applications covered in this book.

Other weighting schemes push in similar directions. Entropy-balancing weights, for example, minimise a divergence measure from a reference distribution subject to moment constraints on covariates [Hainmueller, 2012], and can be adapted to panel settings by balancing on pre-treatment summaries (for example, means and trends) and then applying the resulting weights to the panel. Kernel-based approaches weight donors by similarity in a richer feature space, while Bayesian formulations place priors directly on weights or counterfactual trajectories. Each of these methods imposes a particular structure on how weights should behave. They differ in how they trade fit against stability and interpretability, but they share the core principle that some regularisation of weights is often necessary to obtain reliable out-of-sample performance.

## Extrapolation and Interpolation

Regularisation also changes where the synthetic control sits inside the donor space. Unregularised SC interpolates inside the convex hull of donors and often places substantial weight on boundary donors that most

closely resemble the treated unit (that is, donors whose predictor vectors lie near the edge of the convex hull of  $\mathbf{X}_0$  in predictor space), as discussed in Chapter 6. Ridge SC shrinks those weights towards the centre of the donor distribution and reduces reliance on any single boundary donor. This is helpful when the treated unit is well represented in the middle of the donor cloud, because the synthetic control becomes less sensitive to outlier donors. It can hurt when the treated unit genuinely lies near the boundary. Shrinking weights towards the centre then pulls the synthetic control away from the treated unit's true untreated trajectory.

Balancing SC brings its own potential failure mode. When covariate constraints are tight and the donor pool does a poor job of spanning the treated unit's characteristics, the optimisation may be forced onto donors that match on demographics but differ sharply on outcome dynamics. The resulting synthetic control satisfies the covariate balance criteria but predicts pre-treatment outcomes poorly.

As in the earlier SC sections, diagnostics start with pre-treatment fit and covariate balance. If regularisation causes a dramatic increase in RMSPE or clearly worsens the alignment of pre-treatment paths, the method is fighting the data and should be treated with suspicion. If it modestly increases RMSPE while stabilising weights and preserving acceptable balance on key covariates, the trade-off is more likely to be favourable.

## Choosing Among Variants

The choice among standard, ridge, and balancing synthetic control depends on both the data structure and the credibility constraints of the application. In practice, choices among variants should be driven by pre-treatment RMSPE, effective number of donors  $N_{\text{eff}}$ , covariate balance, and sensitivity of estimated effects to small changes in the donor pool and predictor set, as discussed in Section 6.5.

Standard SC is well suited to settings where the donor pool is modest, the pre-treatment period is long enough to pin down stable weights, and the treated unit lies comfortably inside the donor convex hull. In those designs, unregularised weights often achieve tight pre-treatment fit without obvious instability.

Ridge SC is more appropriate when the donor pool is large relative to the pre-treatment period and you see signs of overfitting, such as highly concentrated weights on donors that share only idiosyncratic patterns with the treated unit. The penalty spreads weight across more donors and typically improves out-of-sample performance at the cost of a small increase in pre-treatment RMSPE.

Balancing SC is most attractive when certain covariates are critical to the story you will tell. In the DMA advertising study, for instance, a marketing executive is unlikely to accept a counterfactual that differs sharply from San Francisco on population or income, regardless of how well it matches sales trends. In that case you should encode demographic balance as a hard design requirement via balancing SC.

In practice you do not need to commit to a single variant. A robust workflow applies standard SC and one or two regularised variants to the same campaign and compares both pre-treatment diagnostics and estimated treatment paths. When the estimates and diagnostics line up across methods, you gain confidence that conclusions are not driven by a particular regularisation choice. When they diverge, that disagreement is itself an important finding. It signals that some combination of convex-hull coverage, covariate balance,

and regularisation structure is failing, and that your substantive conclusions should reflect this uncertainty explicitly rather than resting on a single preferred specification.

## 7.4 Synthetic Difference-in-Differences (SDID)

### The Problem That Motivates Time Weights

Let us return to the brand launching campaigns in twenty markets over three years. Markets adopt treatment at different times: some in Q1 2022, others in Q3 2022, still others in Q2 2023. Standard synthetic control matches each treated market to a weighted combination of controls based on pre-treatment outcomes. The difficulty is that pre-treatment windows differ across cohorts. Early adopters have long pre-treatment histories. Late adopters have short ones. Their pre-treatment periods also overlap different seasonal patterns, macro conditions and competitive environments.

Standard synthetic control treats all pre-treatment periods equally. If a treated market experiences an unusual spike in month eight because of a local event, that spike receives as much weight in the matching objective as any other month. The optimisation then seeks donors that also spiked in month eight, perhaps for entirely unrelated reasons. The resulting match reflects coincidence rather than structural similarity.

Synthetic difference-in-differences, introduced by Arkhangelsky et al. [2021], addresses this by weighting time periods as well as units. Periods with idiosyncratic shocks receive lower weight. The matching objective focuses on periods when treated and control markets behave more comparably. The result is a synthetic control that reflects stable patterns rather than particular quirks of the calendar.

### The Estimator

Let  $Y_{it}$  denote the outcome for unit  $i$  in period  $t$ . Let  $\mathcal{I}$  be the set of treated units and  $\mathcal{J}$  the set of control units. SDID constructs two sets of weights from the pre-treatment data: unit weights  $w_j$  for control units  $j \in \mathcal{J}$  and time weights  $v_t$  for pre-treatment periods  $t \in \mathcal{T}_{\text{pre}}$ . Both sets satisfy convexity constraints, with non-negative weights that sum to one.

One convenient formulation chooses weights by minimising a regularised, weighted squared-error loss over the pre-treatment panel,

$$\min_{\mathbf{w}, \mathbf{v}} \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}_{\text{pre}}} \left( Y_{it} - \sum_{j \in \mathcal{J}} w_j Y_{jt} \right)^2 v_t + \eta_w \|\mathbf{w}\|^2 + \eta_v \|\mathbf{v}\|^2,$$

subject to  $w_j \geq 0$  and  $\sum_{j \in \mathcal{J}} w_j = 1$ , and similarly  $v_t \geq 0$  and  $\sum_{t \in \mathcal{T}_{\text{pre}}} v_t = 1$ . Here  $\|\mathbf{w}\|^2 := \sum_j w_j^2$  and  $\|\mathbf{v}\|^2 := \sum_t v_t^2$ . The regularisation parameters  $\eta_w$  and  $\eta_v$  control how strongly we penalise concentrated unit and time weights. Under the simplex constraints, minimising these norms pushes the solution towards more diffuse weights, down-weighting reliance on any single donor or any single pre-treatment period.

Given the estimated unit weights  $\hat{\mathbf{w}}$  and time weights  $\hat{\mathbf{v}}$ , SDID estimates an average treatment effect on the treated by comparing doubly-differenced means. The SDID estimator targets a single scalar ATT summary over a user-chosen post-treatment window.

To make the averaging explicit, let  $\mathcal{T}_{\text{post}}$  denote the post-treatment periods included in the summary, and define

$$\bar{Y}_{\text{treated},\text{post}} := \frac{1}{|\mathcal{I}| |\mathcal{T}_{\text{post}}|} \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}_{\text{post}}} Y_{it}, \quad \bar{Y}_{j,\text{post}} := \frac{1}{|\mathcal{T}_{\text{post}}|} \sum_{t \in \mathcal{T}_{\text{post}}} Y_{jt}, \quad \bar{Y}_{\text{treated},t} := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Y_{it}.$$

One convenient representation of the estimator is

$$\hat{\text{ATT}}^{\text{SDID}} = \left( \bar{Y}_{\text{treated},\text{post}} - \sum_{j \in \mathcal{J}} \hat{w}_j \bar{Y}_{j,\text{post}} \right) - \left( \sum_{t \in \mathcal{T}_{\text{pre}}} \hat{v}_t \bar{Y}_{\text{treated},t} - \sum_{j \in \mathcal{J}} \hat{w}_j \sum_{t \in \mathcal{T}_{\text{pre}}} \hat{v}_t Y_{jt} \right).$$

The first term compares treated units with their synthetic control in the post-treatment period. The second term subtracts the baseline difference between treated and synthetic control in a weighted version of the pre-treatment period. The double differencing removes unit-specific levels and time-specific shocks, provided the weights succeed in balancing those components.

This structure is helpful to picture. The unit weights pull the synthetic control towards control markets that resemble the treated ones. The time weights down-weight periods dominated by idiosyncratic shocks. Together they construct a counterfactual path that aims to track the treated markets' common trend more faithfully than either standard SC or unweighted DiD.

## What SDID Does Differently

The central innovation is the time weights. Standard synthetic control assigns equal importance to every pre-treatment period and asks which donors match the treated unit's entire trajectory. When that trajectory includes strong idiosyncratic shocks, the resulting match may be driven by local coincidences rather than by enduring relationships. Difference-in-differences, by contrast, assigns equal weight to all units and periods and relies on unit and time fixed effects to absorb level differences. When units differ in their exposure to time-varying shocks, the unweighted parallel trends assumption can fail.

SDID relaxes both rigidities. The unit weights allow the estimator to focus on control units that look like the treated markets on observables and pre-treatment dynamics. The time weights allow the estimator to focus on periods when treated and control units are most comparable. This double flexibility is attractive in marketing applications where treated and control markets face different seasonal or macro environments, or where the pre-treatment period includes one-off shocks.

Apply this to the campaign launch. Suppose treated markets include both Sun Belt cities with strong summer sales and Midwestern cities with strong holiday sales. A single unregularised synthetic control will often struggle to match both patterns, because no fixed set of donor weights fits the entire year for both cohorts. SDID instead estimates time weights that down-weight the months when Sun Belt and Midwest diverge most sharply and puts more emphasis on spring and autumn, when their patterns converge. The resulting synthetic control captures the common trend without being dominated by region-specific seasonality.

## Identification

SDID targets the same type of estimand as DiD, namely an average treatment effect on the treated over the chosen post-treatment window, but under a weighted version of the parallel trends assumption. Let  $Y_{it}(0)$  denote the untreated potential outcome. Weighted parallel trends requires that, for each post-treatment period  $t$ , there exist population weights  $w_j^*$  on control units and  $v_s^*$  on pre-treatment periods such that the doubly differenced untreated outcomes satisfy

$$\mathbb{E}\left[Y_{it}(0) - \sum_{s \in \mathcal{T}_{\text{pre}}} v_s^* Y_{is}(0) \mid i \in \mathcal{I}\right] = \mathbb{E}\left[\sum_{j \in \mathcal{J}} w_j^* \left(Y_{jt}(0) - \sum_{s \in \mathcal{T}_{\text{pre}}} v_s^* Y_{js}(0)\right)\right]$$

Expectations are taken over the sampling distribution (or a superpopulation model) for outcomes, with  $i \in \mathcal{I}$  indicating the treated group. In words, after reweighting pre-treatment periods and control units, treated units and their synthetic control would display the same change from the weighted pre-treatment mean in the absence of treatment.

This condition weakens unweighted parallel trends along one dimension and strengthens it along another. It is weaker in that we require parallelism only after reweighting, which can accommodate differential exposure to shocks that are balanced by the unit and time weights. It is stronger in that it depends on the existence of suitable weights and on the estimator's ability to approximate them from the pre-treatment data. As in the generalised parallel trends discussion in Chapter 4, you trade an unconditional assumption for a weighted one and pay for that flexibility by estimating the weights. By contrast, unweighted parallel trends would require, for a fixed reference pre-treatment period  $t_0$ ,

$$\mathbb{E}[Y_{it}(0) - Y_{it_0}(0) \mid i \in \mathcal{I}] = \mathbb{E}[Y_{jt}(0) - Y_{jt_0}(0) \mid j \in \mathcal{J}]$$

for all post-treatment  $t$ . SDID replaces this with a reweighted version that holds only after applying  $w_j^*$  and  $v_s^*$ .

Two further ingredients are implicit and should be remembered from earlier chapters. First, the usual no-anticipation and no-interference conditions apply. Write  $G_i$  for the adoption time of unit  $i$ . Units must not adjust behaviour in anticipation of the campaign, so outcomes must satisfy  $Y_{it} = Y_{it}(0)$  for all  $t < G_i$ . Control markets must remain unaffected by others' campaigns. Second, the mapping from unit and time characteristics into untreated outcomes must be sufficiently stable across the pre- and post-treatment periods that weights estimated from pre-treatment data remain informative after treatment.

Whether this trade-off is favourable depends on the data. When the donor pool is large and the pre-treatment period contains enough informative variation, the optimisation can find weights that balance pre-treatment paths in a way that plausibly generalises. When the donor pool is thin or the pre-treatment window is short, the weight estimates become noisy and the weighted parallel trends condition becomes more of a modelling claim than a restriction you can pressure-test.

## Implementation

Implementing SDID in the campaign example requires the same design choices as DiD and SC, plus regularisation parameters. You must specify the donor pool, the pre-treatment window, the post-treatment evaluation window, and how to tune  $\eta_w$  and  $\eta_v$ . For a staggered launch over three years, the donor pool includes markets that never receive the campaign and those that have not yet received it at a given event time, mirroring the staggered DiD set-up in Chapter 4. The pre-treatment window should be long enough to identify stable weights; in marketing panels this typically means at least a year of monthly data. The post-treatment window covers the horizon over which you care about effects, for example the first year after launch.

The regularisation parameters control the bias–variance trade-off for the weights. Larger values shrink unit and time weights towards something close to uniform, improving stability but reducing the estimator’s ability to exploit heterogeneity across markets and time. A practical choice again uses cross-validation on the pre-treatment panel. Split the pre-treatment period into training and validation segments, estimate SDID weights over a grid of  $\eta_w, \eta_v$  values on the training segment, and select the pair that minimises prediction error on the validation segment. This can reduce overfitting to pre-period noise, but it cannot validate post-treatment counterfactuals. After tuning, verify that the chosen weights deliver acceptable pre-treatment RMSPE, covariate balance, and placebo behaviour, using the diagnostics in Section 6.5.

In a typical campaign application, this procedure yields moderate regularisation. The resulting unit weights concentrate on a handful of donor markets with similar demographics and baseline sales, while the time weights down-weight months dominated by unrelated promotions or macro shocks. The SDID estimate then typically sits between the DiD and SC estimates in this campaign example, with tighter standard errors than DiD when the weights succeed in removing noise and a more plausible level than SC when plain SC suffers from poor pre-treatment fit. This is not a general guarantee, but rather a diagnostic pattern in many marketing panels. The point is not that SDID will always land between the two, but that its performance is diagnostic of how much the double weighting buys you in a particular design.

## Connection to Factor Models

SDID fits naturally into the factor model perspective developed in Chapter 6. If untreated potential outcomes admit a low-rank representation with unit-specific loadings on a small number of time-varying factors, then unit weights can be interpreted as approximating the treated units’ factor loadings by convex combinations of donor loadings, while time weights emphasise pre-treatment periods that are most informative about the factors that matter for the post-treatment comparison. Together, the weights aim to reconstruct the relevant part of the low-rank structure that governs untreated outcomes for the treated units.

This perspective clarifies when SDID is likely to work well. When the outcome matrix is close to low rank and a moderate number of factors explain most of the variation, the doubly weighted averages can track the treated units’ latent factors and produce credible counterfactuals. When outcomes are driven by many

independent shocks with weak common structure, no choice of unit and time weights can summarise the data effectively, and SDID may fail to fit pre-treatment paths. As with SC and ASCM, pre-treatment fit and balance diagnostics provide the main empirical check on whether the factor-structure story is plausible.

## Costs and Limitations

SDID buys flexibility at a cost. The estimator is more complex than basic SC or DiD because it requires two sets of weights, two regularisation parameters and an additional set of diagnostics. Explaining a time-weighted, unit-weighted double difference to a non-technical stakeholder is harder than explaining a simple before-after comparison or a single-unit SC gap plot.

The dependence on estimated weights also creates fragility. When the pre-treatment period is short, the optimisation problem for the weights is underpowered and the resulting  $\hat{w}$  and  $\hat{v}$  are imprecise. In that case, the weighted parallel trends condition becomes largely untestable in practice. When the donor pool is small, the unit weights have limited room to move and tend towards simple averages, pushing SDID back towards a DiD-like comparison with only modest gains from the synthetic component.

Staggered adoption adds further layers. In designs with multiple cohorts, you are effectively estimating separate sets of weights by cohort and then aggregating cohort-specific effects into an overall effect. As discussed in Chapter 4, the aggregation scheme matters: event-time averages, population-weighted averages and variance-weighted averages can all yield different summaries. SDID does not remove these aggregation issues. It provides an alternative set of cohort-level estimates that you can plug into the same aggregation framework.

## When to Use SDID

SDID is most useful when you have a moderate number of treated units, a donor pool rich enough to support meaningful unit reweighting, and a pre-treatment period long enough to support meaningful time reweighting. It shines in settings where treated and control markets face different seasonal patterns or macro conditions, and where pre-treatment periods contain idiosyncratic shocks that you would like the design to down-weight rather than fit.

It is not a default. When the donor pool is very small or the pre-treatment period is short, the weight estimates will be noisy and the extra complexity may not buy you much beyond carefully specified DiD or augmented SC. In those cases, the simpler methods discussed earlier in the chapter often deliver more stable answers and are easier to explain.

In practice, you can treat SDID as another estimator in the same ensemble that already includes SC, ASCM and DiD. Applying all of them to the same campaign, inspecting pre-treatment diagnostics and comparing estimated treatment paths gives you a richer picture of how sensitive your conclusions are to the way you weight units and time. When SDID and the simpler methods agree, you gain confidence that your

substantive conclusions are not driven by the extra modelling structure. When they diverge, that divergence is itself informative and should be reflected openly in how you present and interpret the results.

## 7.5 Triply Robust Panel (TROP) Estimators

The synthetic control, augmented synthetic control and synthetic difference-in-differences methods in the preceding sections each extend basic parallel trends in a different direction. Standard synthetic control relies on a convex combination of donors that matches the treated unit's pre-treatment path. Augmented synthetic control adds an explicit outcome model to correct residual imbalance. SDID reweights both units and time periods to focus on more comparable comparisons. Each design is robust within its own framework, but each can fail badly when its primary identification mechanism breaks down. SDID, for example, remains biased if weighted parallel trends does not hold, no matter how well we tune unit and time weights.

Triply Robust Panel (TROP) estimators push this logic one step further by combining all three ingredients: unit weights, time weights and a flexible outcome model based on interactive fixed effects. Athey et al. [2025b] propose a TROP estimator that learns unit weights to balance treated and control groups, learns time weights to down-weight less informative periods, and fits a low-rank factor structure to capture heterogeneous responses to common shocks. The core theoretical claim is a conditional bias bound: under a factor-structured model for untreated outcomes, the leading bias term is bounded by a constant times the product of three errors, one for unit imbalance, one for time imbalance, and one for misspecification in the regression adjustment. This is the sense in which the estimator is "triply robust".

To connect the paper's formalism to the monograph's causal language, it is useful to state the target explicitly. A natural estimand in the paper is a treated-cell average over the analysis window,

$$\tau^{\text{TROP}} := \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1].$$

This is an ATT-style object, but the expectation is taken over treated unit-period cells rather than over treated units.

The formal analysis in Athey et al. [2025b] is also developed under strong restrictions that are often questionable in marketing settings. In particular, it assumes no interference (no spillovers into donors) and no dynamic effects, so that potential outcomes can be written as  $Y_{it}(d)$  rather than depending on the treatment path  $d_i^t$ . We use the paper's results as a guide to how bias behaves inside this restricted model class, not as a mechanical guarantee in settings with carryover or interference.

TROP does not escape the need for structure. It replaces the additive parallel trends assumption that underlies DiD and SDID with a factor-model assumption for untreated potential outcomes. We state this assumption here for reference and refer you back to Chapter 6 for a full discussion.

**Assumption 19 (Factor Model for Untreated Potential Outcomes)** For all units  $i$  and periods  $t$ , untreated outcomes satisfy

$$Y_{it}(0) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon_{it} \mid \{\lambda_{ir}, f_{tr}\}_{r=1}^R] = 0,$$

where  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects,  $f_{tr}$  are latent factors and  $\lambda_{ir}$  are unit-specific factor loadings, and  $\varepsilon_{it}$  is idiosyncratic noise. The low-rank matrix with elements  $L_{it} = \sum_{r=1}^R \lambda_{ir} f_{tr}$  captures interactive fixed effects.

This factor structure is more flexible than additive parallel trends but more restrictive than a fully non-parametric model. It allows units to load differently on common shocks – something we often see in marketing panels – but still assumes that a small number of latent factors explain most of the systematic co-movement in outcomes.

## What Triple Robustness Means

It is easy to misuse the language of robustness in panel settings. Earlier in the chapter we saw two different notions of “double robustness.” Augmented synthetic control offers insurance across two different modelling components: if either the weighting model or the regression model is correctly specified, the estimator remains consistent. SDID’s “double robustness” is weaker. Within an additive two-way fixed-effects model, its bias is proportional to the product of unit imbalance and time imbalance. If either imbalance term is small, the bias is small, but both are defined relative to the same underlying structural model.

TROP’s promise is to extend this product-of-errors logic to a factor-structured setting. The key result in Athey et al. [2025b] is a conditional bias bound stated for fixed (non-data-dependent) weights, interpreted as probability limits of estimated weights.

In applications, the weights and tuning parameters are chosen from the data. The theorem is therefore best read as describing how the leading bias behaves when the estimated weights concentrate around stable limits, not as a finite-sample identity.

**Theorem 7.1 (Triple robustness bias bound; Athey et al. [2025b])** *Under Assumption 19, suppose unit weights  $\mathbf{w}$  and pre-treatment time weights  $\mathbf{v}$  (both non-negative and summing to one) are fixed. Let  $\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{iR})'$  denote the factor-loading vector for unit  $i$  and let  $\mathbf{f}_t = (f_{t1}, \dots, f_{tR})'$  denote the factor vector at time  $t$ . Define unit and time imbalance terms*

$$\Delta^u = \sum_{j \in \mathcal{J}} w_j \boldsymbol{\lambda}_j - \boldsymbol{\lambda}_{i^*}, \quad \Delta^t = \sum_{s \in \mathcal{T}_{pre}} v_s \mathbf{f}_s - \mathbf{f}_{t^*},$$

for a target treated unit  $i^*$  and target post-treatment period  $t^*$ . If the regression adjustment for the low-rank component has bias controlled by an operator  $B$  with operator norm  $\|B\|_{op}$ , then the conditional bias of the resulting counterfactual contrast is bounded by

$$|\mathbb{E}[\hat{\tau} - \tau | \mathbf{L}]| \leq \|\Delta^u\|_2 \|\Delta^t\|_2 \|B\|_{op}.$$

In the paper’s formal analysis,  $B$  captures systematic shrinkage in the estimated low-rank component (for example, bias induced by nuclear-norm regularisation or rank truncation), and  $\|\cdot\|_{op}$  is the operator (spectral) norm.

The theorem clarifies what “triply robust” does and does not mean. It does not say that arbitrary modelling errors cancel out. It says that, under a shared factor structure, the leading bias term is small when at least one of the three components is small: unit imbalance, time imbalance, or regression-adjustment misspecification.

This is the sense in which TROP is “triply robust.” It does not mean that you can be cavalier about all three ingredients. It means that you now have three levers which, if any one of them is correctly tuned relative to the underlying factor structure, can salvage identification. In practice, all three will be estimated with error and all three may be misspecified to some degree. The value of TROP is that small mistakes in several places multiply rather than add, so that moderate imbalance and moderate model error can still produce modest bias.

## Why Factor Structure Matters in Marketing Panels

The factor model is not an abstract technicality. It is designed to capture a pattern we routinely see in marketing data: outcomes co-move because of common shocks, but different units respond with different intensity.

Consider again a national recession in a campaign evaluation. Under additive parallel trends, every market experiences the same time effect  $\lambda_t$ , so recession quarters shift all units by the same amount. That is implausible. Affluent DMAs with high discretionary spending might see a sharp fall in premium-category sales. Industrial regions hit by plant closures see a different pattern again. Budget markets where consumers trade down to cheaper brands may see smaller or even opposite shifts in category volume. These heterogeneous responses are exactly what cause additive DiD and SDID designs to struggle.

In a factor framework, the recession is represented by a common factor  $f_t$  that spikes during downturn quarters. Affluent, industrial and budget DMAs carry different loadings  $\lambda_i$  on that factor. A treated DMA’s untreated path during the recession therefore depends on both the factor and its own loading. TROP’s low-rank component aims to learn these heterogeneous loadings from the donor panel, while its unit and time weights try to align the treated DMA with a weighted combination of donors that shares a similar factor profile. When this works, the method can deliver credible counterfactuals in settings where additive two-way fixed effects are clearly inadequate.

The same story applies to seasonality, category trends and platform shocks in digital advertising. Common shocks exist, but units react differently. TROP’s model structure is built to capture exactly that pattern.

## Estimator Sketch and Relation to Existing Methods

At a high level, TROP works by fitting a two-way fixed-effects plus factor model to the donor panel, while reweighting both units and time periods to focus on observations that are most informative for a given treated unit and treatment period. In the most general formulation in Athey et al. [2025b], the weights can depend on the target treated cell  $(i, t)$ . For exposition, we suppress this dependence and describe the three components

in a simpler global-weight form. Let  $D_{it}$  denote treatment,  $Y_{it}$  the outcome and assume we observe a panel with many untreated  $(i, t)$  pairs. The estimator proceeds in three steps.

First, it defines unit weights  $w_j$  that give more weight to donors whose pre-treatment paths lie close to the treated unit's path. Conceptually, these play the same rôle as the unit weights in SC and SDID. Distance can be measured by root mean squared differences in pre-treatment outcomes, much as in synthetic control. A tuning parameter controls how sharply the weights decay with distance; when the parameter is zero the weights become uniform and the method behaves like a global factor model, whereas large values make the estimator focus on close neighbours.

Second, it defines time weights  $v_s$  that give more weight to pre-treatment periods close to the treatment date. This reflects the idea that recent history is often more informative about current behaviour than distant history. A separate tuning parameter controls how quickly the weights decay as you move away from the treatment date. When this parameter is zero all pre-treatment periods are weighted equally. When it is large the estimator behaves more like a local time-differencing design.

Third, given these weights, it fits a low-rank factor model to the untreated panel by solving a nuclear-norm-regularised least-squares problem over unit and time fixed effects and the interactive component. The nuclear norm penalty controls the effective rank of the factor structure and is tuned by cross-validation. The fitted model yields predictions of untreated outcomes for any unit and time, including treated units in post-treatment periods. The TROP treatment effect estimate for unit  $i$  in period  $t$  is the difference between the observed outcome and this predicted counterfactual.

This framework connects to several familiar designs. If you set all weights equal and choose a large rank, you obtain something close to matrix completion on the donor panel. If you turn off the factor component and keep only unit and time fixed effects with data-driven weights, you obtain an estimator similar in spirit to SDID. If you remove time weighting and focus on unit weighting and a simple outcome model, you move back towards ASCM. TROP is best thought of as a unifying language and a flexible class of estimators rather than a single closed-form formula.

## Design Choices and Tuning

In applied work you face three main design choices with TROP: how sharply to concentrate unit weights, how local to make time weights, and how complex to make the factor structure. Athey et al. [2025b] propose choosing these tuning parameters by cross-validation on untreated cells, with an objective designed so that pseudo treatment effects on untreated cells are close to zero. This tuning logic is appealing because it directly targets the imputation problem that drives TROP.

Good predictive performance on held-out untreated outcomes is necessary for credible identification but not sufficient. Cross-validation cannot, on its own, guarantee that the factor structure and weighting scheme satisfy the causal assumptions needed for post-treatment counterfactuals.

In practice, you would define a grid of candidate tuning parameters for unit weights, time weights and the nuclear norm penalty. For each combination you fit the model on a training subset of untreated observations

and measure prediction error on a validation subset. You then select the parameters that minimise this error. If the data exhibit strong interactive fixed effects, cross-validation will favour a non-trivial factor structure. If additive two-way fixed effects suffice, it will push the nuclear norm penalty high and effectively shut down the factor component. If all history is informative, it will choose slowly decaying time weights. If only recent periods matter, it will pick fast decay.

This tuning mechanism is conceptually appealing because it lets the data decide which components matter most, rather than forcing the analyst to commit *ex ante* to SC, SDID or matrix completion. It is also computationally intensive. Each tuning combination requires solving a large convex optimisation problem. For small-to-medium marketing panels (say a few dozen units over a few dozen periods) this is feasible with modern optimisation toolkits. For very large panels, staged or random-search strategies are needed to keep computation reasonable.

## When to Use TROP

TROP is an ambitious method. It makes sense in applications where you have good reason to suspect interactive fixed effects, where the sample is large enough to estimate a low-rank structure and where simpler methods struggle to achieve credible pre-treatment fit.

Marketing panels with many units and moderate time depth are a natural candidate. Think of 40 DMAs followed for 20 quarters, with rich common shocks and clear heterogeneity in how markets respond. In such settings, standard SC may fail because the treated unit sits outside the convex hull of donors, ASCM may struggle because a simple regression cannot capture the full pattern of co-movement and SDID may fail because additive time effects are too crude. If, in that context, a carefully tuned TROP model achieves substantially better pre-treatment fit than SC, ASCM or SDID while using plausible weights and a modest factor rank, it deserves serious consideration.

By contrast, when the donor pool is small or the time dimension is short, the extra complexity of TROP is unlikely to pay off. Estimating a factor structure with only eight pre-treatment months or with only five control markets is more art than science. In those cases, you are better off with the simpler designs developed earlier in the chapter, which are easier to estimate and explain.

At the time of writing, TROP is a frontier method. The underlying research is at preprint stage and stable, well-documented software implementations are not yet widely available. Implementing it today requires custom optimisation code, typically built on general-purpose tools such as cvxpy in Python or convex optimisation libraries in R. That makes it more suitable for methodological work or for teams with strong in-house econometrics capability than for routine marketing analytics.

## Positioning TROP in the Hybrid Methods Hierarchy

From a strategic point of view, the main value of TROP for this book is conceptual. It shows how weighting, differencing and factor modelling can be combined in a single framework and clarifies what “triple robustness” means in panels. It also gives you a way to think about why SC, ASCM, SDID and matrix completion succeed or fail in particular designs. For example, SC failures typically reflect convex-hull and unit-imbalance problems, ASCM failures reflect regression misspecification on top of those, SDID failures reflect violations of (weighted) parallel trends, and matrix-completion failures reflect poor low-rank approximation.

In empirical marketing work, the right approach is to treat TROP as one more estimator in a small ensemble. You start with methods that are transparent and well understood – SC, ASCM, regularised SC and SDID – and only move to TROP when those methods either fail to fit the pre-treatment data or produce estimates that are clearly at odds with economic common sense. When you do estimate TROP, you should present it alongside these simpler benchmarks, explain what its tuning parameters are doing and show how its pre-treatment diagnostics compare. Agreement across methods strengthens confidence in your conclusions. Disagreement is a signal to dig deeper, not a licence to cherry-pick the method that tells the most appealing story.

In that sense, TROP completes the conceptual hierarchy rather than replacing the tools you already have. It is a powerful idea – and potentially a powerful estimator – but its practical role in marketing analytics will depend on how the method and its software ecosystem evolve beyond the time this book is written.

## 7.6 Identification and Assumptions

Hybrid methods build on the same basic causal ingredients as synthetic control and difference-in-differences. They still ask what the treated units' outcomes would have been in the absence of treatment, still rely on comparisons to donor units, and still use pre-treatment data to learn relationships they then extrapolate into the post-treatment period. What changes is how these relationships are modelled and how much structure is imposed. This section states the identification assumptions that hybrids inherit from earlier chapters, highlights what each hybrid adds on top, and connects these requirements to the factor-model perspective developed in Chapter 6.

For clarity we focus on a single treated unit with index 1 and donor units indexed by  $j \in \mathcal{J}$ , as in the basic synthetic control setup in Chapter 6. Extension to multiple treated units proceeds by averaging unit-specific effects, as in the DiD and SC chapters.

Throughout, let  $T_0$  denote the last pre-treatment period. Treatment begins after  $T_0$ .

The primary estimand is the post-treatment path effect  $\tau_{1t} = Y_{1t}(1) - Y_{1t}(0)$  for  $t > T_0$ . Later sections summarise this path by averaging over post-treatment windows or aggregating across treated units.

The core identification logic for hybrid estimators mirrors that of synthetic control. We require that, in expectation, the untreated potential outcome for the treated unit can be represented by a combination of donor outcomes and a model-based adjustment that we can estimate from pre-treatment data and plausibly extrapolate to post-treatment periods. Schematically, we can write untreated outcomes for the treated unit as a sum of a weighted donor component and a model-based adjustment. For example, for augmented synthetic control we work with a representation of the form

$$Y_{1t}(0) \approx \sum_{j \in \mathcal{J}} w_j^* Y_{jt}(0) + X'_{1t} \beta^* + \varepsilon_{1t},$$

where  $w_j^*$  are population weights and  $\beta^*$  are outcome-model parameters. For SDID, we work with a representation that combines unit weighting with additive time adjustments,

$$Y_{1t}(0) \approx \sum_{j \in \mathcal{J}} w_j^* Y_{jt}(0) + \alpha_1 + \lambda_t + \varepsilon_{1t},$$

with  $\alpha_1$  a unit effect and  $\lambda_t$  a time effect. In SDID the time weights  $v_t$  defined in Section 7.4 determine which pre-treatment periods matter most for learning the comparison, rather than redefining time effects. These expressions are approximations that summarise the rôle of weights and adjustments. They are best read as imputation decompositions used to motivate estimation, not as causal structural models for  $Y_{it}(0)$ . In the pre-treatment period we estimate the relevant weights and adjustment parameters from observed data. In the post-treatment period we hold those relationships fixed and use them to construct counterfactuals for  $Y_{1t}(0)$ .

## Generic Assumptions

All hybrid methods in this chapter inherit a common set of identification conditions from the general framework in Chapter 2, the synthetic control chapter and the DiD chapter. We restate them here in compact form to fix ideas and to emphasise that hybrids do not relax these fundamentals.

**Assumption 20 (No Anticipation in Pre-Treatment Periods)** For the treated unit, pre-treatment outcomes coincide with untreated potential outcomes:

$$Y_{1t} = Y_{1t}(0) \quad \text{for all } t \leq T_0.$$

No anticipation, introduced in Chapter 2, rules out behavioural responses before the recorded treatment date. If there is evidence that outcomes start to move in response to the campaign before  $T_0$ , then those periods do not provide clean information about the untreated trajectory. In practice, you respond by redefining the intervention to include the anticipation period or by restricting the pre-treatment window used to estimate weights and adjustments to earlier periods.

**Assumption 21 (No Interference or Explicit Exposure Modelling)** Either the treatment applied to the treated unit does not affect donor outcomes,

$$Y_{jt} = Y_{jt}(0) \quad \text{for all } j \in \mathcal{J}, t = 1, \dots, T,$$

or spillovers are represented through an exposure mapping  $h_j(D_{-j,t})$  with spillover-aware potential outcomes  $Y_{jt}(d, h)$ , and the analysis targets effects holding the exposure process fixed.

This is the same no-spillovers component of SUTVA that underpins synthetic control and DiD. If the treated unit's campaign materially changes competitors' behaviour or market conditions in donor units, then donor outcomes no longer represent valid counterfactuals. Hybrid designs cannot solve this problem mechanically. The remedy remains design-based: curate the donor pool, impose geographic or competitive buffers, or build an explicit exposure model.

**Assumption 22 (Pre/Post Stability of the Imputation Relationship)** The relationships used to impute  $Y_{1t}(0)$  from donor outcomes, covariates, and any adjustment model remain stable across the treatment boundary. In particular, interpret the weights and adjustment parameters used by a given hybrid method as converging to population limits  $w^*$  and  $\psi^*$  learned from pre-treatment data. We require that the same imputation rule that fits the pre-treatment untreated outcomes continues to apply to  $Y_{1t}(0)$  after  $T_0$ .

Stability says that the way we map donor outcomes, predictors, and weights into counterfactual outcomes does not jump at the treatment date. For ASCM, this means the outcome regression extrapolates sensibly beyond the pre-treatment window. For SDID, it means that the reweighted two-way fixed-effects structure that fits pre-treatment data remains a good approximation afterwards. For factor-based hybrids such as TROP, it means that the factor structure learnt from pre-treatment data continues to describe untreated outcomes in the post-treatment period. We cannot test this directly for  $Y_{1t}(0)$ , but placebo checks and cross-validation within the pre-treatment period provide indirect evidence on whether the mechanism generalises.

**Assumption 23 (Overlap and Feasibility)** There exist unit weights  $w_j^*$ , augmentation parameters  $\psi^*$  and, where relevant, time weights  $v_t^*$  such that the hybrid representation achieves good pre-treatment fit for the treated unit, with stable and interpretable weights.

Overlap here means that the treated unit is not so extreme that no combination of donors and adjustment terms can approximate its untreated path. In the pure SC case this reduces to the convex-hull condition on factor loadings; in hybrids it adds the requirement that the chosen adjustment model and time weights can repair any residual mismatch without creating implausible extrapolation. In practice you diagnose this by inspecting pre-treatment RMSPE relative to the scale of  $Y$ , by plotting pre-treatment paths for the treated unit and its synthetic counterpart and by examining how concentrated and specification-sensitive the weights are. There is no universal numerical threshold that guarantees identification. The rule of thumb is that pre-treatment discrepancies should be small relative to both the natural volatility of outcomes and the size of the effects you care about, and that small perturbations in the design should not cause weights or fits to swing wildly.

## Method-Specific Assumptions

On top of these generic conditions, each hybrid introduces its own additional structure. These method-specific assumptions are where identification gains or breaks.

For ASCM, the key extra ingredient is the outcome regression. As discussed in Section 7.2, within a suitable factor-model framework the estimator can remain close to unbiased if either the SC weights achieve good balance or the regression model accurately captures the remaining imbalance, even if the other component is imperfect. In finite marketing panels both components will typically be imperfect, so the practical interpretation is more modest: ASCM buys you another route to good counterfactuals, not a guarantee that one of the two is correct.

For a formal treatment of the augmentation logic, see [Ben-Michael et al., 2021].

For ridge and balancing SC, the additional structure comes from regularisation [Doudchenko and Imbens, 2016]. Ridge SC shrinks weights towards more diffuse configurations; balancing SC enforces explicit covariate-balance constraints. Identification then hinges on choosing the penalty or tolerance so that you do not regularise away the very heterogeneity you need to capture. The stability assumption in Assumption 22 must hold for the particular regularised weights chosen by cross-validation or other tuning rules, not just for some hypothetical unregularised solution.

For SDID, the central new requirement is weighted parallel trends. After applying population unit weights and time weights  $w_j^*$  and  $v_s^*$ , the doubly differenced untreated outcomes for treated units must evolve in parallel to those of the weighted donors, as set out formally in Section 7.4. This assumption is weaker than unweighted parallel trends, because it allows you to reweight the comparison group and pre-treatment periods to balance differential exposure to shocks. It is stronger than simply observing good pre-treatment fit, because it asserts that the reweighted relationship continues to hold for  $Y_{1t}(0)$  after treatment. See Section 7.4 for the formal weighted parallel-trends condition.

For TROP, the method-specific assumption is the factor structure discussed in Section 7.5. Untreated potential outcomes must admit a low-rank decomposition into unit and time effects plus a small number of latent factors. Triple robustness – in the sense that bias can be expressed as a product of unit imbalance, time imbalance and factor-model error – operates inside this factor framework. If outcomes in fact do not exhibit a strong low-rank structure, the factor-model error component of the product-of-errors bound is large, so even small unit and time imbalance can generate substantial bias.

### Connection to Factor and Imputation Models

The factor-model perspective from Chapter 6 provides a unifying lens for these identification assumptions. Suppose untreated potential outcomes follow a factor structure

$$Y_{it}(0) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

with a small number of common factors  $f_{tr}$  and unit-specific loadings  $\lambda_{ir}$ . Synthetic control implicitly constrains the treated unit's loadings to lie in the convex hull of donor loadings:  $\lambda_{1r} = \sum_j w_j \lambda_{jr}$  for each factor  $r$ , with non-negative weights summing to one. This convexity delivers interpretability but can fail when the treated unit sits near or outside the hull.

Interactive fixed-effects and matrix-completion methods estimate loadings freely without convexity constraints. They can match pre-treatment paths very closely, but rely on regularisation and factor-rank restrictions to avoid overfitting and to extrapolate credibly. They shift the burden of identification from geometric coverage (convex hull) to structural assumptions about low rank and stability of factors.

Hybrid estimators sit between these poles. ASCM keeps the convex-hull logic but adds a regression adjustment that can soak up residual factor-structure differences. Ridge and balancing SC adjust the geometry of the hull by regularising weights or enforcing covariate balance. SDID adds an additive structure that reweights units and periods while remaining within an extended fixed-effects framework. TROP combines unit weights, time weights and a factor model, letting the data decide which component carries most of the explanatory power. Athey et al. [2021] and Arkhangelsky and Imbens [2024] formalise this view by treating these procedures as versions of imputation with different constraints on how the counterfactual surface may vary across units and time.

### Practical Guidance

From an identification standpoint, hybrid methods are most useful when neither pure synthetic control nor pure DiD gives a comfortable answer on its own. If the treated unit lies well inside the donor convex hull and unregularised synthetic control achieves excellent pre-treatment fit, the extra structure of ASCM, SDID or TROP is unlikely to change the substantive conclusions and may add unnecessary complexity. If unconditional

parallel trends between treated and control groups is plausible and panel structure is simple, classical DiD or event-study designs may suffice.

Hybrids earn their keep in the intermediate regions. When the treated unit is near but not squarely inside the donor hull, or when treated and control groups violate unweighted parallel trends but can be brought into alignment by reweighting, hybrids allow you to trade additional modelling structure for better pre-treatment balance and more credible counterfactuals. The price is stronger assumptions about how that extra structure behaves out of sample.

The right way to use these methods is cumulative. Start with the simplest design whose identification assumptions you can defend. Use hybrid estimators to see whether conclusions are sensitive to relaxing convexity or unweighted parallel trends, and to check whether richer factor structures improve pre-treatment fit in a way that aligns with your marketing context. When hybrids and simpler methods agree, you gain confidence that the underlying causal story is robust. When they disagree, the identification assumptions in this section tell you exactly which features of the data-generating process you need to scrutinise to explain why.

## 7.7 Multiple Treated Units and Staggered Adoption

Hybrid methods extend naturally to settings with multiple treated units and staggered adoption. The main challenge is no longer constructing a counterfactual for a single treated store or market, but doing so coherently across many treated units that adopt at different times. The gains from hybrids in this setting are the same as before — better pre-treatment fit through augmentation, robustness to trend violations through time weighting and more stable weights through regularisation — but we now have to combine unit-level estimates into group-time and event-time summaries.

Throughout this section, let  $\mathcal{I}$  denote the set of treated units, grouped into cohorts  $g$  by their adoption date, and let  $\mathcal{J}$  denote the set of potential donors. For a given calendar period  $t$ , let  $\mathcal{J}_{-t}$  be the set of units that have not yet been treated by period  $t$  and can therefore serve as valid donors. Let  $\bar{Y}_{g,t}$  denote the average outcome for cohort  $g$  in period  $t$ , and similarly for other bar-notation quantities.

### Common Intervention Times: Unit-Level vs Pooled Estimation

When several treated units adopt at a common time, there are two natural ways to use hybrids. The first is unit-level estimation. You estimate a separate synthetic control, augmented synthetic control or SDID for each treated unit using the same donor pool and predictor set. Let  $\hat{\tau}_i$  denote a chosen summary effect for treated unit  $i$  (for example, the average post-treatment effect over a specified horizon). This yields unit-specific effects that you can aggregate into an overall average using policy-relevant weights,

$$\hat{\tau}^{\text{pooled}} = \sum_{i \in \mathcal{I}} w_i \hat{\tau}_i,$$

with  $w_i$  reflecting, for example, equal importance across treated units, market size or revenue contribution.

The unit-level route has the advantage of transparency. In a store-format trial with five flagships, you can show five separate hybrid counterfactuals and see directly that the format lifts sales in dense urban centres but has little effect in smaller suburban locations. That heterogeneity matters for decisions about scaling. The cost is computational — especially for more complex hybrids — and practical: some treated units will inevitably be harder to match than others, and unit-level estimates for those cases will be fragile.

The second route is pooled estimation. Here you estimate a single set of donor weights that minimises an aggregate pre-treatment loss across all treated units, for example

$$\min_{\mathbf{w}} \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}_{\text{pre}}} \left( Y_{it} - \sum_{j \in \mathcal{J}} w_j Y_{jt} \right)^2 + \eta \|\mathbf{w}\|^2,$$

subject to the usual convexity constraints,  $w_j \geq 0$  and  $\sum_{j \in \mathcal{J}} w_j = 1$ . The resulting  $\mathbf{w}$  defines a common synthetic control path, which you compare to the average treated path. Pooled estimation is simpler to implement and produces a single set of weights to interpret, but it can fit some treated units well and others

poorly. If treatment effects or untreated dynamics vary sharply across treated units, a single pooled synthetic control will obscure that variation.

In practice, many applications start with unit-level hybrids to diagnose heterogeneity and then report both unit-level estimates and a pooled summary based on policy-motivated weights.

## Aggregation and Policy-Relevant Weighting

Once you have unit-level or cohort-level hybrid estimates, you must decide how to weight them in forming aggregate effects. The right weights depend on the question.

If you care about the average effect across the treated units in your sample — for example, for an internal cost–benefit calculation — equal weights or sample-size weights are natural. If you care about extrapolating to a larger target population — such as a national store network — aggregation weights should reflect how representative each treated unit or cohort is of that population.

Hybrid methods do not change this logic. What they change is the quality of each unit- or cohort-level effect. The aggregation step is the same as in Chapter 4: choose weights that answer your substantive question, check that they are non-negative and interpretable, and report them.

## Staggered Adoption and Event-Time Effects

When units adopt at different times, we index cohorts by their adoption date  $g$ . For each cohort  $g$  and calendar period  $t \geq g$ , we can estimate a cohort–time treatment effect  $\tau(g, t)$  by applying a hybrid estimator to the comparison between cohort  $g$  and donors in  $\mathcal{J}_{-t}$  [Ben-Michael et al., 2022]. Here  $\mathcal{J}_{-t}$  denotes the set of units that are not yet treated at time  $t$  and therefore remain valid donors for cohort  $g$  at that date. This choice of donor set matters for the estimand. It targets the effect for cohort  $g$  at time  $t$  relative to not-yet-treated donors (and never-treated donors, when available).

For SDID, for instance, we estimate unit weights  $\hat{w}_j^g$  and time weights  $\hat{v}_s^g$  on the pre-treatment panel for cohort  $g$  and donors [Arkhangelsky et al., 2021], and then form a doubly differenced comparison between the cohort’s average outcome and its synthetic control in period  $t$ , exactly as in Section 7.4 but now applied to cohort means rather than a single unit.

We can convert these cohort–time effects into event-time profiles by defining event time  $k = t - g$  and averaging over the cohorts that contribute data at each  $k$ . If  $\mathcal{G}_k$  is the set of cohorts with observations at event time  $k$ , we can form

$$\hat{\theta}_k = \sum_{g \in \mathcal{G}_k} w_g \hat{\tau}(g, g + k),$$

with weights  $w_g \geq 0$  that sum to one within each  $\mathcal{G}_k$ . This is the same event-time aggregation logic developed in Chapter 5. The only difference is that hybrid methods improve the construction of  $\hat{\tau}(g, t)$  inside each cohort–time cell by using unit and time weights and, for ASCM, regression adjustments.

A practical detail that matters in staggered designs is donor-pool evolution. Early-adopting cohorts enjoy large donor pools but relatively short pre-treatment windows. Late-adopting cohorts have longer pre-periods but smaller sets of not-yet-treated controls. Hybrid estimators help in both directions — ASCM and ridge SC stabilise estimation with short pre-periods, while SDID’s time weights focus comparisons on more comparable segments of the pre-period — but they cannot create donors where none exist. You should always inspect which cohorts contribute to each event-time estimate and how the donor pool changes over event time.

## Connection to Group–Time Effects

The modern difference-in-differences literature, reviewed in Chapter 4, organises staggered designs around cohort–time effects  $\tau(g, t)$  [Callaway and Sant’Anna, 2021, Sun and Abraham, 2021]. Many papers describe these as group–time ATT objects. Hybrid methods plug directly into this framework. For each  $g$  and  $t$ , you construct a hybrid counterfactual for cohort  $g$  using donors in  $\mathcal{J}_{-t}$ ; the gap between observed and synthetic outcomes in period  $t$  is an estimate of  $\tau(g, t)$ . Formally, if  $\hat{Y}_{g,t}^{\text{hyb}}(0)$  is the hybrid counterfactual for cohort  $g$  at time  $t$ , then  $\hat{\tau}(g, t) = \bar{Y}_{g,t} - \hat{Y}_{g,t}^{\text{hyb}}(0)$  plugs directly into the aggregation formulas in Chapter 4. Aggregating these cohort–time effects over cohorts and time yields the same overall, calendar-time or event-time summaries as in Chapter 4; hybrid methods do not change the aggregation algebra.

This also clarifies the relationship to the negative-weighting problem. As that chapter shows, certain naive aggregation schemes over cohorts and time can assign negative weights to some  $\tau(g, t)$  terms when treatment effects are heterogeneous [Sun and Abraham, 2021]. Estimating  $\tau(g, t)$  with hybrids does not, by itself, eliminate negative weighting. What prevents negative weighting is structuring the analysis cohort-by-cohort and using aggregation schemes that keep weights non-negative and interpretable. The contribution of hybrids is to improve the quality of each  $\tau(g, t)$  estimate by constructing better counterfactuals from not-yet-treated donors.

## Practical Challenges in Staggered Panels

Two practical issues recur in staggered applications. The first is variation in pre-treatment length across cohorts. Early-adopting cohorts may have only a handful of pre-periods, while late adopters have many. With very short pre-periods, synthetic control weights are under-identified and can overfit. In those cases, SDID with time regularisation or ridge SC can be more stable than unregularised SC, because they shrink weights and, in SDID’s case, focus attention on the most informative parts of the pre-period. At the extreme, when a cohort has only a handful of pre-treatment observations, you should be cautious about relying on any weighting-based method and consider reporting that cohort separately with appropriate caveats.

The second issue is shrinking donor pools. As more cohorts adopt treatment, fewer not-yet-treated units remain available as donors. By the last adoption wave, only never-treated units are left. When the donor pool drops to only a handful of units, hybrid estimators are forced to lean heavily on those remaining donors, and

weight stability deteriorates. Monitoring donor-pool size and composition across cohorts, and being explicit about which cohorts have robust donor support, is essential.

## Choosing Between Hybrid Methods in Staggered Designs

The basic trade-offs between SC, ASCM and SDID carry over to staggered adoption. SDID is attractive when you believe that, after reweighting units and periods, treated and not-yet-treated cohorts satisfy a weighted parallel-trends condition, and when pre-treatment dynamics differ across cohorts. Its time weights can down-weight months or quarters where cohort trends diverge, making it easier to isolate a common component.

ASCM is more appealing when certain cohorts are clear outliers relative to donors and simple reweighting cannot achieve good fit, or when you have strong prior beliefs about which observables drive outcome differences across cohorts. In those cases, the regression adjustment gives you an additional lever to repair residual imbalances that SC and SDID leave behind.

Standard SC remains competitive when cohorts lie comfortably within the donor convex hull and achieve excellent pre-treatment fit. In those rare but pleasant cases, the extra structure and tuning parameters of SDID or ASCM may not buy you much.

In practice, the best way forward is empirical. For a subset of cohorts, estimate SC, ASCM and SDID using the same donor sets  $\mathcal{J}_{-t}$  and compare pre-period fit, event-time pre-trends and post-period estimates. If all three methods tell a consistent story, you can confidently report the simplest design. If they diverge, that divergence is itself diagnostic. In terms of identification, severe discrepancies between SC and SDID point to tension between convex-hull coverage and (weighted) parallel trends, whereas discrepancies between SC/SDID and ASCM highlight sensitivity to the outcome-regression component. It tells you that some combination of convex-hull coverage, weighted parallel trends and outcome modelling is failing and that any conclusions about long-run dynamic effects in that part of the event-time profile should be presented with appropriate caution.

## Reporting and Transparency

Staggered hybrid designs place a premium on clear reporting. At minimum, you should report how each cohort contributes to aggregate and event-time estimates. In practice, this means tabulating the cohort weights  $w_g$  used in each event-time estimate  $\hat{\theta}_k$ , listing  $\mathcal{G}_k$  for key values of  $k$ , and indicating where only early-adopting cohorts contribute. You should show pre-treatment fit diagnostics for each cohort — plots of treated vs hybrid counterfactual paths and summary measures of fit — and flag cohorts with visibly poor match.

Finally, you should treat alternative aggregation schemes and estimator choices as sensitivity analyses, not afterthoughts. Reweight cohorts in ways that reflect different policy questions, rerun the analysis with SC, ASCM and SDID where feasible, and show how the dynamic treatment effect profile changes. When

estimates prove stable across these perturbations, you have a much stronger case that your findings about campaign ramp-up, peak and decay are not artefacts of a particular hybrid specification.

## 7.8 Tuning, Implementation, and Donor Curation

Implementing hybrid methods in marketing panels means making a series of design choices: which predictors to include, how to tune penalties, and how to curate the donor pool. These choices interact with the estimator and affect both pre-treatment fit and the credibility of the counterfactual. The central tension is that pushing too hard on prediction accuracy in the pre-period can harm causal identification if the model then extrapolates poorly. This section offers practical guidance on navigating that trade-off in a way that respects the identification logic developed earlier in the chapter.

### Predictor Selection

Predictor selection plays different roles for different hybrid methods.

For synthetic control and ridge SC, the predictor set defines the space in which the treated unit must be approximated. Including more pre-treatment outcome periods and key covariates expands the dimensionality of the matching problem and makes the convex hull richer. In marketing panels, a small number of latent demand and seasonality factors often drive most of the variation, so a modest number of carefully chosen pre-periods can go a long way. In factor-model terms (Chapter 6), you need the pre-treatment window to be long enough that  $T_0$  (the last pre-treatment period) is comfortably larger than the number of factors  $R$  for pre-periods to distinguish signal from noise; in typical marketing panels  $R$  is small, so 12–24 well-chosen months can be enough even when the full panel is longer. This rule of thumb is safest when seasonality is stable, the pre-period contains no major breaks, and the donor pool is rich enough to support matching. If the panel is long and outcomes are highly autocorrelated, using every second or third period, or summarising history through moving averages, can reduce dimensionality without losing the structure that matters, provided these choices still capture the relevant seasonal and business-cycle patterns. Holding out a block of pre-treatment periods for validation, as in Section 7.3, helps check that the chosen predictor set generalises within the pre-period.

For ASCM, the augmentation model uses covariates to correct residual differences between the treated unit and its synthetic control. Here predictor selection is closer to standard outcome regression. You should include variables that plausibly predict outcomes and differ between treated and control units — such as store size, location demographics, competitive intensity or channel mix — and be cautious about mechanical variable selection that loads the model with weak predictors. Pre-specifying this covariate set based on institutional knowledge is essential, with a small number of pre-planned sensitivity checks that add or remove candidate variables.

Do not include predictors that are themselves affected by treatment (or by anticipation), since that can build post-treatment information into the counterfactual.

For SDID, predictors enter implicitly through outcomes rather than through a separate covariate matrix. The algorithm constructs unit and time weights to balance pre-treatment paths directly, so the main design choice is the length and placement of the pre-treatment window. You still need enough pre-treatment periods

to identify stable weight patterns — typically at least several periods that span the relevant seasonal or business cycle variation — but you do not have to specify a separate predictor set.

For TROP and related factor-based hybrids, predictors matter through the distance metrics used for unit weighting and through the factor model itself. The considerations mirror those for synthetic control and interactive fixed-effects models in Chapter 6: you need enough pre-treatment periods to identify a low-rank structure and enough variation across units to distinguish factor loadings. In short marketing panels with only a handful of pre-periods, the factor component will necessarily be simple, and you should treat any rich factor structure as exploratory rather than definitive.

## Penalty Parameter Tuning

Regularisation penalties control the balance between fit and stability. Tuning them well is critical for hybrids that rely on weighting and augmentation.

For ridge SC, the ridge penalty on the weights governs how concentrated or diffuse the donor weights become. Small penalties allow the optimisation to place large weight on one or two donors and chase tight pre-treatment fit. Larger penalties spread weight across more donors and sacrifice some fit in exchange for stability. A practical way to choose the penalty is to split the pre-treatment period into a training block and a validation block, estimate ridge SC over a reasonable grid of penalty values on the training block and pick the value that predicts the validation block best. Cross-validation here tunes out-of-sample prediction within the pre-treatment period. Good predictive performance is necessary for credible identification, but it does not on its own guarantee that the same weights produce unbiased counterfactuals after treatment. The exact grid is not sacred; what matters is that you explore a range from very little shrinkage to quite strong shrinkage and that the selected value is not at an extreme of the grid.

For ASCM, the outcome regression often includes its own regularisation, such as ridge or elastic net on the covariate coefficients. The same training-validation logic applies. Given the typically short pre-periods in marketing panels, it is safer to err on the side of modest complexity in the augmentation model. The double-robustness logic discussed in Section 7.2 provides some insurance across the weighting and regression components, but in finite samples both components are estimated and often both are somewhat wrong. Regularisation should therefore be viewed as a way to tame variance and overfitting in the regression step, not as a licence to greatly increase covariate or functional-form complexity.

For SDID, the main design choice is the implementation of the optimisation problem that delivers unit and time weights. Standard implementations incorporate stabilisation internally, so you do not typically search over penalty values by hand. You do, however, need to check that the optimisation has converged and that the resulting weights are interpretable — not all mass on one donor, not all mass on one pre-treatment period, and broadly in line with your understanding of which markets and periods are comparable.

Even when you do not tune a single penalty explicitly, SDID still regularises implicitly through its objective and constraints. Treat the resulting unit and time weights as tuned objects, not as fixed design inputs.

For TROP and other flexible factor-based hybrids, penalty tuning is more demanding: you must decide how sharply to concentrate unit weights, how local to make time weights and how complex to make the factor structure. One approach, discussed in Section 7.5, is to use staged cross-validation [Athey et al., 2025b], tuning time weights first, then unit weights, then the factor penalty, each time using held-out untreated observations to assess predictive performance. This staged scheme is a pragmatic heuristic, not a theoretically guaranteed route to causal validity. Given the method’s current research status and lack of off-the-shelf software, we view these tuning rules as guidance for methodological work rather than prescriptions for routine marketing analysis.

## Connection Between Tuning and Diagnostics

Tuning choices show up directly in the diagnostics you report, and diagnostics should in turn inform how you interpret tuning.

Pre-treatment RMSPE will vary with the strength of regularisation. Tight fit with very small penalties can signal that the weights are overfitting idiosyncratic donor fluctuations. Loose fit with very large penalties can signal that the estimator is effectively averaging over donors and ignoring meaningful structure. The key is not to maximise or minimise RMSPE mechanically, but to look at how RMSPE behaves across a range of penalties and how that behaviour lines up with weight patterns and institutional knowledge. If only extreme penalties deliver good validation performance or plausible weights, that is a warning sign.

Weight dispersion is another informative diagnostic. You can summarise how many donors meaningfully contribute by using, for example, the inverse of the sum of squared weights, which behaves like an “effective number of donors.” Because  $N_{\text{eff}}^{-1} = \sum_j \hat{w}_j^2$  appears inside the idiosyncratic-variance term  $\sigma^2(1 + \sum_j (w_j^*)^2)$  from equation (6.24), very small  $N_{\text{eff}}$  mechanically inflates variance even when pre-period fit is tight. Very low values indicate near-singleton donor reliance and high sensitivity to that donor’s idiosyncrasies. Very high values indicate near-uniform weighting. In most marketing panels you want an interior solution where a modest number of donors carry most of the weight and those donors make sense given the business context.

Sensitivity to tuning choices completes the picture. If your estimated treatment effect changes little across a broad range of penalty values that all deliver acceptable pre-treatment fit, your conclusions are unlikely to hinge on fine details of the regularisation. If estimates swing sharply as you move the penalty, especially in regions of the grid where fit and weights look similar, then any point estimate should be presented with caution and accompanied by a discussion of that instability. Complement these summaries with simple weight plots (sorted bar charts for donor weights, line plots for SDID time weights) so that outliers are obvious. Flat numerical diagnostics can hide pathologies that are visually clear.

## Donor Pool Curation

Donor curation is at least as important as tuning. As in Chapter 6, donors must be comparable to treated units, free of treatment and unaffected by spillovers.

Comparability comes first. In a retail setting, it is rarely sensible to use outlets from very different formats, regions or customer segments as donors. Restrict the pool to units that plausibly face similar demand, competition and operational constraints. This often means staying within the same banner or region, or at least within a relatively homogeneous subset of markets.

Absence of treatment and contamination is next. Donor units must not themselves receive the intervention during the window you use for estimation and evaluation. More subtly, they should not be exposed to strong spillovers through competition, supply chains or shared customers. Buffered designs, where you drop donors within a certain geographic or competitive radius of treated units, are a simple way to reduce contamination, but they also shrink the donor pool. Hybrid estimators can sometimes tolerate a smaller, cleaner donor pool by using augmentation or regularisation to repair fit, but they cannot recover information that is simply not there. Overly aggressive donor restrictions can violate the convex-hull or overlap conditions in Section 7.6 by removing donors that help span the treated unit in predictor space. Curation should therefore balance business comparability against the need for geometric coverage.

If spillovers are plausible, you either curate donors so the exposure mapping  $h_i(D_{-i,t})$  is effectively constant (often zero), or you commit to an explicit exposure model.

These design choices interact with the method. Synthetic control needs donors that span the treated unit in the space of predictors. Excluding too many donors can push the treated unit outside the convex hull and force the estimator into extrapolation. ASCM can absorb small extrapolations through its regression component, but still relies on donors that share the main drivers of outcomes. SDID requires that, after reweighting, treated and donor groups can plausibly satisfy weighted parallel trends. Factor-based hybrids such as TROP assume that donors and treated units share a common factor structure. In each case, donor curation should be justified both by business logic and by how the curated pool supports the specific hybrid's assumptions.

## Prediction vs Identification

It is tempting to treat pre-treatment prediction as the primary goal and to judge hybrids solely by how well they fit the pre-period. That is not what we want for causal work. The objective is to construct a counterfactual that approximates the treated unit's untreated path in the post-treatment period, not to win a forecasting competition on pre-treatment data.

Overfitting the pre-period by piling on predictors, using highly flexible models or driving penalties towards zero can produce impressive pre-treatment fit. But if the underlying factor structure or covariate relationships change at treatment (for example, a structural break in  $f_t$  or a new channel shock), that same flexibility will happily track noise and then extrapolate it forward. Underfitting by imposing strong penalties or sparse

models can produce looser pre-treatment fit but may extrapolate more stably when the pre-period is short or only partially representative of the post-period environment.

The practical lesson is to combine automatic tools, such as cross-validation within the pre-period, with economic judgement about how the intervention might change the data-generating process. Placebo checks, which treat pre-treatment periods as pseudo post-treatment periods, are especially valuable: they show whether an estimator that fits early pre-period data well can also predict later pre-period outcomes it has not seen.

## A Practical Workflow

A disciplined workflow for hybrids brings these elements together.

Start by pre-specifying a predictor set based on the business context, focusing on pre-treatment outcome histories and covariates that are plausibly related to both treatment assignment and outcomes. Decide in advance on a small number of alternative specifications that add or drop predictors to test robustness, rather than searching post hoc for combinations that deliver striking results.

Next, curate the donor pool. Exclude units that receive treatment during the estimation window, units that are clearly incomparable on business grounds and units that are likely to be heavily contaminated by spillovers. Document these choices and, where possible, illustrate how they affect pre-treatment fit and weight patterns.

Then tune regularisation using pre-treatment data. For ridge-based hybrids such as ridge SC and many ASCM implementations, use a training-validation split and a broad grid of penalties that span from very light to quite strong shrinkage. For SDID, focus on checking convergence and inspecting the resulting unit and time weights rather than searching over penalties. For more complex hybrids like TROP, recognise that current tuning guidance is still research-oriented and only pursue that route if you have both the data and the technical capacity to justify it.

Finally, estimate treatment effects and present diagnostics. Report pre-treatment RMSPE, weight dispersion and sensitivity to tuning choices, not just point estimates. If a hybrid method achieves materially better pre-treatment fit than standard SC while producing plausible weights and stable estimates across tuning choices, the extra complexity is likely earning its keep. Better pre-treatment fit is supportive, but it is not sufficient if spillovers, anticipation, or breaks in the untreated outcome process remain plausible. If the gains in fit are marginal and estimates are sensitive to tuning, the simpler design is usually preferable for transparency and ease of explanation.

In all cases, the role of tuning and implementation is to support the identification arguments made earlier in the chapter, not to replace them. Careful predictor selection, thoughtful donor curation and transparent regularisation choices help ensure that hybrid methods deliver counterfactuals that are both statistically well behaved and substantively credible in real marketing applications.

## 7.9 Diagnostics and Goodness of Fit

Hybrid methods are only as credible as the diagnostics that accompany them. In marketing panels, where data are noisy and designs are often tight, you cannot afford to treat a hybrid estimator as a black box. You need to check how well it fits the pre-treatment path, how well it balances covariates, how weights are distributed across donors and time, how sensitive estimates are to individual donors or periods, and how stable the design is under placebo checks. This section outlines a diagnostic workflow tailored to hybrids and connects each diagnostic to specific remedial actions, with Chapter 17 providing a more general framework. These are diagnostics and falsification checks. Passing them does not prove identification, but failing them is strong evidence against it.

### Pre-Treatment Fit

Pre-period fit is the first line of defence. For any hybrid method, define the pre-treatment RMSPE as

$$\text{RMSPE}_{\text{pre}} = \sqrt{\frac{1}{T_0} \sum_{t=1}^{T_0} (Y_{1t} - \hat{Y}_{1t}^{\text{syn}})^2},$$

where  $\hat{Y}_{1t}^{\text{syn}}$  is the method-specific hybrid counterfactual in period  $t$ , and  $T_0$  is the last pre-treatment period. For SC, this is the weighted donor average; for ASCM, it is the augmented synthetic control; for SDID, it is the reweighted, intercept-shifted comparison.

RMSPE is not a goal in itself but a way to compare specifications and methods on a common scale. Expressing RMSPE relative to the pre-treatment standard deviation of  $Y_{1t}$  helps with interpretation: an RMSPE that is small relative to the natural variability in outcomes suggests that the synthetic path tracks the treated unit closely; an RMSPE of the same order as the outcome SD indicates that the counterfactual is effectively noise. Comparing RMSPE across SC, ASCM, SDID and DiD shows whether the extra structure in hybrids is buying you materially better pre-treatment fit. In the factor-model notation from Chapter 6,  $\text{RMSPE}_{\text{pre}}$  is best viewed as an empirical proxy for pre-period imputation error, not a direct bound on post-period bias. Small  $\text{RMSPE}_{\text{pre}}$  is therefore necessary but not sufficient for small post-treatment bias. If RMSPE is of the same order as the post-period effect you hope to detect, the design cannot cleanly separate treatment from misspecification.

The key is to combine RMSPE with other diagnostics. A hybrid that shaves a few percentage points off RMSPE by pushing weights onto one odd donor is less credible than a slightly looser-fitting design with more stable weights and better covariate balance.

## Covariate Balance

Outcome paths tell only part of the story. You also care about whether the hybrid synthetic control matches the treated unit on covariates that predict outcomes and treatment. Standardised mean differences (SMDs) are a useful summary:

$$\text{SMD}_x = \frac{X_1 - \sum_j \hat{w}_j X_j}{s_X},$$

where  $X_1$  is the treated unit's covariate value,  $\sum_j \hat{w}_j X_j$  is the weighted donor mean, and  $s_X$  is a reference standard deviation (for example, the pooled standard deviation of  $X$  across treated and donor units). Here  $X$  represents either baseline (time-invariant) covariates or pre-period averages of time-varying covariates. Values close to zero indicate good balance; large absolute values indicate imbalance.

Thresholds from the matching literature — such as SMD below 0.1 in absolute value as “good balance” and above about 0.2 as concerning — are useful guides rather than hard rules. In marketing contexts you should focus especially on covariates that are central to the business story: store size, income, channel mix, competitive intensity, and so on. If one of these remains substantially imbalanced even under a hybrid design, you need to decide whether that imbalance plausibly biases the estimated effect.

ASCM gives you a natural lever to correct residual imbalances by including problematic covariates in the augmentation model. Ridge and balancing SC let you tilt the weighting scheme towards better balance at the cost of some pre-treatment fit. If, after reasonable adjustments, a key covariate remains badly imbalanced, you should report that fact and discuss its implications for identification.

## Weight Dispersion

Hybrid estimators work by reweighting donors and, in SDID, time periods. Diagnostics should therefore include a view of how concentrated or diffuse those weights are. For unit weights, an informative summary is the “effective number of donors”

$$N_{\text{eff}} = \frac{1}{\sum_j \hat{w}_j^2},$$

which is the inverse of the Herfindahl index of the weights. A value of one means all weight sits on a single donor; a value equal to the number of donors  $N_0$  means weights are uniform. For SDID, an analogous effective number of pre-treatment periods

$$T_{\text{eff}} = \frac{1}{\sum_{t=1}^{T_0} \hat{v}_t^2}$$

shows how many periods meaningfully contribute.

Extremes in either direction are informative. Very low  $N_{\text{eff}}$  indicates that the hybrid is leaning heavily on one or two donors and may be vulnerable to their idiosyncrasies. Very high  $N_{\text{eff}}$  indicates that the estimator is close to uniformly averaging over donors and may be ignoring structure that could help sharpen the counterfactual. In many marketing panels, an interior solution where a modest number of sensible donors

carry most of the weight is what you hope to see. If diagnostics show otherwise, revisit the penalty strength, donor pool and predictor set.

## Leverage and Influence

Even when overall weight dispersion looks reasonable, individual donors or periods can have outsized influence. Simple leave-one-out checks are powerful here. Re-estimate the hybrid estimator excluding each of the largest-weight donors in turn and record how much the estimated treatment effect changes. Do the same for pre-treatment periods in SDID by excluding one period at a time when estimating time weights.

If dropping a single donor or period barely moves the estimate, the design is robust along that dimension. If excluding one donor or one pre-period causes large swings, dig deeper. Is that donor truly comparable, or does it have a unique pattern that the algorithm is over-using? Does the influential period contain an unusual local promotion, a data glitch or a macro shock that should be handled explicitly? The point is not to mechanically delete influential observations, but to understand why they matter. Large influence for a donor or period that is marginal to the business question (for example, a very small donor market or a holiday period outside the main evaluation window) may be less concerning than large influence for a core donor or a period central to the decision; diagnostics should be interpreted in that light.

## Placebo Checks

Placebo checks probe the stability assumptions that hybrids rely on. The idea is to pretend that treatment occurred earlier than it did and see whether the estimator generates spurious “effects” inside the pre-treatment window.

Concretely, choose a pseudo-intervention date  $T^* < T_0$ . Treat periods up to  $T^*$  as pre-treatment, estimate the hybrid estimator on that shorter pre-period, and then compute pseudo treatment effects for periods between  $T^* + 1$  and  $T_0$ . If the model extrapolates well, these pseudo effects should be small and centred near zero. If you see large, systematic pseudo effects, the specification is not stable even within the pre-period, and there is little reason to trust it out of sample. Large, systematic pseudo effects indicate violations of the stability condition in Assumption 22: the pre-treatment mechanism linking weights, covariates and untreated outcomes is not even stable within the pre-period.

These checks are particularly revealing for complex hybrids such as ASCM, SDID and TROP, where overfitting the early pre-period is easy. When placebo checks fail, the appropriate reaction is not to hunt for a different pseudo date that “passes”, but to rethink the specification: shorten the pre-period used for estimation, simplify the augmentation model, strengthen regularisation, or reconsider whether the design can support a hybrid at all. Pre-specify a small set of pseudo-dates (for example, one or two meaningful cut points based on seasonality or business cycles) rather than scanning the entire pre-period.

## Residual Diagnostics

Residuals provide a complementary view. For a given hybrid estimator, define residuals as

$$e_{1t} = Y_{1t} - \hat{Y}_{1t}^{\text{syn}}$$

for all periods  $t$ , and plot them over time with the intervention date marked. In a well-specified design, pre-treatment residuals should fluctuate around zero without obvious trend or seasonality. Systematic drifts, seasonal cycles or clusters of large outliers in the pre-period signal misspecification.

Seasonal residual structure suggests the donor pool is mismatched on seasonality or the pre-period window is misaligned. Structural breaks suggest redefining the pre-period or excluding donors with breaks.

Post-treatment residuals will reflect both treatment effects and any remaining misspecification. A clean treatment story looks like a discernible level shift or pattern emerging after treatment against a background of pre-period noise that has been well controlled. If pre-period residuals already exhibit strong non-random structure, it becomes hard to argue that post-period deviations are due to treatment rather than to model failure.

## Method-Specific Considerations

Different hybrids bring their own diagnostic nuances.

For ASCM, the size of the augmentation correction relative to the pure synthetic-control component is informative. If the augmentation correction routinely accounts for a large share of the level of  $\hat{Y}_{1t}^{\text{syn}}$  across pre- and post-treatment periods, the estimator is heavily reliant on the outcome model. Inspecting coefficient paths for the augmentation regression, and how sensitive the correction is to tuning choices and alternative specifications of the augmentation model, helps you judge whether that reliance is defensible.

For SDID, the pattern of time weights  $\hat{v}_t$  is revealing. Concentration on the very last pre-treatment period effectively turns the design into a first-difference comparison. Near-uniform time weights put equal emphasis across the whole pre-period. Combined with diffuse unit weights, the estimator moves closer to a conventional DiD-style comparison, but it is not identical to DiD in levels; both extremes should be justified by the seasonal and macro environment. When you see extreme patterns, you should ask whether they make sense given seasonality, promotions and macro conditions, and whether alternative specifications of the pre-period window would yield more interpretable weights.

For TROP and other factor-based hybrids, the tuning parameters and inferred factor rank play a diagnostic role. When cross-validation pushes the factor penalty so high that the factor component is effectively shut down, the data are telling you that a simpler hybrid (closer to SDID) suffices. When unit-weight decay is very strong, the estimator is effectively using nearest neighbours. Reporting these tuning outcomes alongside fit metrics gives readers a sense of how complex the underlying structure really needs to be. Given the current research stage of TROP, treat these tuning patterns as exploratory diagnostics rather than as definitive evidence that a rich factor structure is required.

## Putting Diagnostics Together

In practice, you should view diagnostics as a package rather than as a sequence of hurdles. A credible hybrid design is one where pre-period RMSPE is small relative to outcome variability and clearly better than simpler alternatives; covariates central to the business story are well balanced; unit and time weights are neither excessively concentrated nor implausibly uniform; estimates are not unduly driven by a single donor or period; placebo checks show no large spurious effects; and residuals in the pre-period look like noise rather than trend.

When several of these diagnostics point in the same positive direction, you can be reasonably confident that the hybrid method is constructing a credible counterfactual. When they send mixed signals — for example, excellent RMSPE but poor balance on key covariates, or good balance but unstable placebo performance — you should say so and treat any causal claims with appropriate caution. The goal is not to certify a method as “valid” once a checklist is ticked, but to build a cumulative case that your design and estimator are fit for purpose in the marketing context at hand. Chapter 17 provides a complementary, design-level checklist; you should interpret the hybrid-specific diagnostics here within that broader framework.

## 7.10 Inference

Hybrid methods do not change the basic purpose of inference. We still want to know how much uncertainty surrounds our estimated treatment effects and whether those effects are unusually large relative to what the design would produce by chance. Uncertainty bands can target either the counterfactual path  $\hat{Y}_{1t}(0)$  itself or the treatment effect  $\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}(0)$ ; the interpretation and width differ. What hybrids do change is the structure of that uncertainty: weights are estimated, time weights  $\hat{v}_t$  may be estimated, augmentation models contribute additional noise, and, for TROP, factor structure is tuned. Inference must respect this extra layer of estimation while remaining honest about the limits imposed by small marketing panels.

In this section we sketch inference strategies that are well suited to the hybrid estimators in this chapter, and point to Chapter 16 for the full development of panel inference tools.

### Method-Specific Variance Structures

Each hybrid estimator has its own variance structure, reflecting which objects are treated as fixed and which as estimated.

For SDID, both unit weights and time weights  $\hat{v}_t$  are estimated from the pre-treatment panel. Arkhangelsky et al. [2021] derive an analytic variance formula that treats these weights as functions of the data and accounts for their sampling variability when the control group is large. In small- $N$  marketing panels, this large-control-group asymptotics can be fragile; in those cases it is safer to treat the analytic formula and bootstrap estimates as complementary checks rather than to rely on one alone. In practice, many applications rely either on that formula as implemented in software or on bootstrap methods that resample units or cohorts, as described below.

For ASCM, uncertainty comes from two sources: the synthetic control weights and the augmentation model. The more weight the design puts on the regression adjustment, the more variance (and potential model dependence) flows through that channel; the more weight it puts on the synthetic control component, the more variance comes from which donors carry weight and how their outcomes fluctuate. Closed-form variance expressions exist under specific modelling assumptions, but they depend on nuisance quantities that are hard to estimate reliably in small samples. In particular, variance contributions from the augmentation model depend on how well its residual structure is captured; when the regression is high-dimensional or heavily regularised, asymptotic formulas can be especially misleading. For this reason, resampling-based approaches are the natural default.

For ridge SC and related regularised SC variants, the ridge penalty stabilises weights and typically reduces variance relative to unregularised SC, at the cost of some bias. Because the weights are complicated functions of the data and the penalty, analytic variance expressions quickly become unwieldy. Again, cluster-robust or bootstrap-based inference is the practical choice.

For TROP and other factor-based hybrids, the variance structure is more complex still because unit weights, time weights  $\hat{v}_t$  and the factor component are all tuned from the data. Here, inference is best treated

as a research topic rather than a settled routine. Proposals in the literature include cross-validation-style variance estimates based on how effects vary across folds. At the time of writing, these ideas are promising but not yet standard, and we recommend using TROP primarily as a robustness check alongside simpler methods whose inference is better understood.

## Placebo and Permutation Tools

Placebo and permutation tools play a central role in hybrid inference because they make minimal modelling assumptions and directly probe the stability of the design.

In-space placebos apply the hybrid estimator to each donor unit in turn, pretending that donor  $j$  was treated at the same time as the actual treated unit. Comparing the resulting placebo gaps to the treated unit's gap gives a sense of how unusual the observed effect is relative to what the method produces when applied to untreated units. Ranking statistics such as the ratio of post- to pre-treatment RMSPE, introduced in Chapter 6, can be turned into randomisation-style tail probabilities under symmetry assumptions. In observational marketing panels, it is safer to treat these as descriptive diagnostics — “our treated unit is more extreme than all but one placebo” — rather than as literal frequentist p-values. In other words, they tell you how unusual your treated unit looks when processed through the same estimator and design, conditional on the realised panel, not how likely such an effect would be under a random-assignment mechanism that did not in fact operate.

In-time permutations treat intervention timing as if it were unknown within the pre-treatment window. By shifting the pseudo-intervention date forward and backward and recomputing pseudo effects, you obtain a distribution of “effects” that arise purely from fitting and extrapolating within the pre-period. If the actual post-treatment effect sits well outside the range of these pseudo effects, that strengthens the case that you are seeing treatment rather than model artefacts. If it falls squarely inside that range, your design is not discriminating strongly between real and pseudo treatment. Large pseudo-effects under in-time permutations are direct evidence against the pre/post stability condition in Assumption 22: the mechanism that fits early pre-period data does not extrapolate even to later pre-periods.

Both procedures stress-test the hybrid's weighting and augmentation components without additional parametric assumptions. They are especially valuable when conventional asymptotics are unreliable because the number of units or pre-periods is small.

## Bootstrap Approaches

Bootstrap methods, introduced in Chapter 16, provide a flexible route to standard errors and confidence intervals when analytic formulas are unavailable or untrustworthy.

In panel hybrids, the natural resampling unit is the unit or cohort, not individual observations. A unit-level (or cohort-level) block bootstrap proceeds by resampling units with replacement within the treated

and donor groups separately, re-estimating the hybrid estimator on each bootstrap sample and recording the resulting treatment effect. The dispersion of these bootstrap replicates provides a standard error, and percentile intervals provide confidence bands. When there is only a single treated unit, resampling treated units is not meaningful; single-treated-unit settings typically rely on placebo distributions, time permutations, or residual-based procedures rather than a naive “treated-unit bootstrap”. If you do bootstrap in such settings, it is usually via resampling donors (and/or blocks of time under additional assumptions), with interpretation as a sensitivity device. This preserves the within-unit time-series structure while acknowledging that treated and donor units are the primary sources of sampling variability. When the number of treated units is extremely small (for example, one or two flagships), bootstrap variability in pooled effects will be dominated by between-unit heterogeneity rather than sampling noise; in such cases, simple unit-level bootstraps can underestimate uncertainty and should be interpreted conservatively.

Wild bootstrap variants adapt the same idea to settings with few units and heteroscedastic residuals. Wild bootstrap variants either hold weights fixed and perturb residuals, or re-estimate the full procedure in each bootstrap draw; both aim to approximate sampling variation under a null. This is particularly useful when you pool effects across cohorts or event times and want to account for heterogeneity in noise levels across markets and over time.

In both cases, you must respect the design: resampling should not break the treatment–control structure (for example, treated units should remain treated) or the staggered-adoption pattern.

## Conformal and Prediction-Interval Views

Conformal inference, discussed in Chapter 16 [Chernozhukov et al., 2021, Cattaneo et al., 2021], offers an alternative perspective by constructing prediction intervals for the treated unit’s untreated outcomes based on the distribution of pre-treatment residuals. For hybrids, the idea is simple. You compute residuals between observed and synthetic outcomes in the pre-period, treat their empirical distribution as a reference for what “noise” looks like under no treatment, and then form intervals around the hybrid counterfactual in the post-period by adding and subtracting suitable quantiles of those residuals.

If the observed post-treatment outcomes consistently lie outside these bands in one direction, that is evidence of a treatment effect. If they oscillate inside the bands, the data are consistent with noise around the estimated counterfactual. The strength of this approach is that it makes minimal distributional assumptions beyond a form of exchangeability: conditional on no treatment, the distribution of residuals in the post-period is assumed to match the distribution of pre-treatment residuals. In panels with serial correlation, you need a blocked or exchangeable-by-block residual assumption, not i.i.d. residuals. Its weakness is that those assumptions are not guaranteed in short, seasonal marketing panels. We recommend using conformal-style bands as a complement to, not a replacement for, placebo and bootstrap-based checks.

## Randomisation Inference and Multiplicity

Randomisation inference has a clean interpretation in genuine experiments, where treatment is assigned by design. In that case, permuting treatment labels among units and recomputing the hybrid estimator for each permutation reproduces the exact randomisation distribution under a sharp null of no effect. Comparing the observed effect to this distribution yields p-values with a straightforward causal meaning. In genuine randomised geo-experiments, where treatment assignment across markets is known and under the experimenter's control, this interpretation is exact. In observational panels, any such test must be viewed as a hypothetical sensitivity exercise conditional on strong assumptions about the assignment mechanism.

In observational hybrids, treatment is not randomly assigned, so randomisation inference can only be interpreted as a sensitivity tool under strong assumptions about selection. A small permutation tail probability (interpreted as a diagnostic measure of relative extremeness) tells you that, given the observed pattern of outcomes and your hybrid estimator, it would be rare to see an effect as large as the one you estimate if treatment labels were randomly shuffled. It does not, by itself, prove that treatment caused the effect. Use such tests as one piece of a larger inferential puzzle, not as a standalone verdict.

Multiplicity is unavoidable when you examine many post-treatment periods, cohorts or outcomes. Testing each post-treatment event-time effect  $\theta_k$  at 5% inflates the chance of at least one false positive. Chapter 16 discusses formal tools such as Bonferroni corrections, false discovery rate control and stepdown procedures. In the hybrid context, it is often more informative to combine these with graphical event-study plots and joint tests (for example, testing that all post-treatment event-time effects  $\theta_k$  are zero). For hybrid estimators that feed into event-time profiles, the uniform-band and joint-test ideas from the event-study chapter (Chapter 5) apply directly. In marketing applications with small samples, you will frequently be underpowered for strict family-wise error control; the focus should be on patterns that are consistent across periods, methods and specifications, rather than on individual stars in a table.

## Small-Sample Considerations

Most marketing panels are small enough that textbook asymptotics are, at best, rough guidance. With few treated units, modest donor pools and short pre-treatment histories, any inferential procedure must be interpreted with care.

Chapter 16 discusses the distinction between sampling-based and design-based uncertainty in more detail [Abadie et al., 2020], which is especially relevant when inference is layered on top of observational hybrid designs.

With a small donor pool, rank-based placebo tests can only take on a coarse set of tail probabilities, and bootstrap distributions may be jagged. With short pre-periods, placebo checks in time have limited power and conformal intervals are based on few residuals. With few treated units, pooled estimates will have wide intervals and considerable sampling noise. In these settings, the most honest stance is to de-emphasise binary significance thresholds and focus on the size, sign and robustness of estimated effects across methods

and specifications. In practice, this means showing effect paths with confidence or prediction bands, placebo distributions, and bootstrap intervals side by side, rather than relying on asterisks in tables.

Whatever inferential tools you deploy, good practice is to be explicit about sample sizes (numbers of treated units, donors, pre- and post-periods), about which inference method you used and why, and about how multiple testing was handled if you report many coefficients. This transparency lets readers calibrate their own confidence in your conclusions and see where the evidence is strong, where it is suggestive and where it is simply too thin to support sharp claims.

## 7.11 Marketing Applications

Hybrid methods are particularly well suited to marketing problems where treated units are few, pre-treatment periods are short, adoption is staggered and decision-makers care about both transparency and robustness. This section sketches four stylised applications drawn from common marketing settings. The examples are illustrative rather than empirical case studies. They show how hybrid designs are put together in practice and what credible findings typically look like when the methods are used carefully.

### Application 1: Multiple-Market Loyalty Programme Rollout

Suppose a retailer pilots a loyalty programme in ten flagship stores, chosen for their size and affluent urban locations, with forty comparable stores never receiving the programme during the study window. Quarterly revenue is observed for three years before the pilot and two years after. The goal is to estimate the average effect on quarterly revenue for the treated stores.

Flagship stores are larger and more urban than most donors, so a hybrid that can correct systematic covariate imbalances is natural. ASCM fits this structure well [Ben-Michael et al., 2021]. For each flagship, the analyst constructs a synthetic control using pre-treatment revenue trajectories and store characteristics such as floor space, local income and urbanisation. A regularised outcome regression then adjusts for any remaining level differences linked to these covariates.

Diagnostics show that ASCM tracks pre-treatment revenue more closely than standard SC: pre-period RMSPE is noticeably smaller relative to revenue volatility, covariate imbalances on store size and urbanisation are substantially reduced, and weights are spread across a moderate number of sensible donor stores rather than concentrating on one or two idiosyncratic outlets. Leave-one-donor-out checks indicate that no single donor dominates the estimates.

When the pilot goes live, the hybrid counterfactual reveals an uplift in revenue that builds over the first few quarters as customers enrol and then levels off. The magnitude is economically meaningful but not explosive — think of a mid-single-digit percentage increase on a quarterly basis, with intervals that suggest a positive effect while still reflecting meaningful sampling noise. In-space placebo analyses, treating donor stores as if they had launched loyalty, yield much smaller gaps on average, suggesting that the observed pattern is not easily reproduced by chance. The retailer uses these results, together with cost information, to form ROI projections for a wider rollout.

### Application 2: Staggered Advertising Campaign Launch

Consider a brand that rolls out an advertising campaign in thirty regional markets over six quarters. Ten markets adopt in Q1, ten in Q3 and ten in Q5. Twenty markets never receive the campaign and serve as

donors. The brand wants to understand how effects evolve over event time and whether early and late adopters respond differently.

The staggered structure and potential differences in pre-treatment trends across cohorts make SDID a natural choice [Arkhangelsky et al., 2021]. For each adoption cohort, the analyst applies SDID using not-yet-treated and never-treated markets as donors, estimating unit weights that align markets with similar pre-trends and time weights  $\hat{v}_t$  that emphasise pre-treatment periods where treated and donor markets move together. The cohort-time effects  $\tau(g, t)$  are aggregated into event-time effects  $\hat{\theta}_k$  using the framework from Chapter 5.

Diagnostics show that, for each cohort, the SDID synthetic paths match pre-treatment sales reasonably well, with time weights  $\hat{v}_t$  concentrating on the few quarters just before adoption — a pattern that fits the idea that recent history is most informative. When seasonality is strong, you should check that the time weights are not simply picking one season and ignoring the rest. Event-time estimates tell a coherent story: effects are small on impact, build over a few quarters as awareness accumulates and then settle at a roughly constant lift. Pre-treatment leads in the event-time plot hover close to zero, supporting weighted parallel trends, and cohort-specific profiles look broadly similar rather than wildly heterogeneous.

From a marketing perspective, the key insights are about timing and persistence: the campaign appears to ramp up over a few quarters and then sustain a modest but non-trivial sales lift, with little evidence that early or late adopters behave differently once you condition on pre-trends.

### Application 3: Overlapping Loyalty and Promotional Programmes

Now imagine a retailer that introduces two interventions over the same horizon: a loyalty programme launching in Q1 and a new promotional strategy starting in Q4. Ten markets receive both, ten receive only loyalty, ten only promotions and twenty receive neither. Management wants to know not just whether each programme works, but also whether they interact.

The overlapping treatment structure creates identification challenges. For markets that receive both programmes, the relevant comparison is markets that receive neither, not those receiving only one programme. A hybrid extension of ASCM can help by constructing synthetic controls separately for each treatment sequence group, using never-treated markets as donors and including programme indicators in the augmentation model.

In practice, the analyst would first check that each treatment-sequence group can be matched reasonably well in the pre-period using donors that never receive any programme. Augmentation then helps correct remaining level differences linked to store and market characteristics. This approach effectively treats “treatment sequence group” as the exposure, requiring that sequence assignment is not driven by unobserved shocks to untreated outcomes (beyond the no-anticipation and no-carryover conditions already assumed). Under the same identification assumptions that would justify a richer DiD or event-study model — no unobserved shocks tied systematically to the timing of loyalty or promotions — ASCM can help disentangle the average effect of each programme and their overlap.

A stylised outcome might be that loyalty produces a fairly stable, medium-sized sales lift that accumulates over several quarters, while promotions produce shorter-lived spikes around the launch period. Markets with both programmes might show combined effects that are smaller than the sum of individual effects, suggesting some substitution or saturation. The value of the hybrid here is in sharpening the counterfactuals for each group and making these patterns visible, not in eliminating the need for strong assumptions about how overlapping treatments are assigned.

### **Application 4: Flagship City Launch with Potential Spillovers**

Finally, consider a platform that launches a new service in a flagship city, with concern that neighbouring cities may be indirectly affected through word-of-mouth, competition or changes in supply. Five neighbouring cities are plausibly exposed to spillovers. Thirty more distant cities form a candidate donor pool. The platform wants to estimate both the direct effect in the flagship and any spillovers among neighbours.

For the direct effect, ASCM applied to the flagship city with only distant donors in the pool is a natural starting point. The key design step is donor curation: neighbours are excluded from the donor set because they may be contaminated by spillovers, while distant cities are retained if there is no compelling reason to think they are affected by the launch. As in earlier examples, diagnostics focus on pre-treatment fit, balance on relevant city characteristics and weight stability.

Estimating spillovers requires adopting the exposure-mapping framework from Chapter 11. One simple mapping treats neighbours as “exposed” (using the exposure-mapping notation  $h_i(D_{-i,t})$  from Chapter 11) once the flagship launches. For each neighbour, the analyst constructs a synthetic control from distant cities and includes exposure indicators in an augmentation model that captures how similar the neighbour is to the flagship on demographics and baseline adoption. Under the assumption that distant donors are unaffected and that exposure captures the main channel linking the flagship to neighbours, differences between neighbours and their synthetic controls after launch can be interpreted as spillover effects.

In a plausible outcome, the flagship city shows a relatively large, sustained jump in adoption following launch, while neighbours exhibit smaller, delayed increases that are stronger for nearer cities than for those further away. The platform can then form a regional impact measure that combines the direct effect in the flagship with spillover effects in neighbours, weighted by population. That informs decisions about where to launch next: cities with dense clusters of nearby markets may generate more total value than isolated cities with similar own-market potential.

### **Summary**

These stylised applications show how hybrid methods can be woven into realistic marketing analyses. The common elements are straightforward. You choose a hybrid that matches the structure of the problem — ASCM when covariate imbalance is central, SDID when staggered timing and differential trends matter, more

complex factor-based hybrids only when the data and team can support them. You curate donors carefully to respect comparability and spillovers. You lean heavily on diagnostics — pre-treatment fit, covariate balance, weight patterns, placebos and sensitivity checks — to assess whether the resulting counterfactuals are credible.

Quantitatively, credible marketing findings from these designs often tend to be modest in size (single- to low-double-digit percentage lifts), with uncertainty intervals that acknowledge real sampling noise and with patterns that are consistent across neighbouring periods and alternative estimators. Used in this way, hybrids provide a disciplined way to extract causal insight from complex marketing panels without promising more precision than the data can support.

## 7.12 Workflow Checklist

This section distils the chapter into a compact protocol for using hybrid methods in marketing panels. The workflow is deliberately high level. It tells you what to decide and in what order, and points back to earlier sections where you can find detail on design, tuning, diagnostics, inference and reporting. The aim is not to turn hybrid methods into a mechanical checklist, but to provide a disciplined way to organise an analysis.

### Step 1: Define the Estimand and Cohorts

Start by fixing the substantive question and the estimand. For a single treated unit, this is usually an average treatment effect on that unit over a chosen post-treatment window. With several treated units that adopt at the same time, you must decide whether you care about unit-specific effects, an overall ATT, or both. With staggered adoption, define cohorts by adoption time and decide whether you will focus on cohort-specific effects, event-time profiles or calendar-time effects. In staggered designs, the building blocks are cohort-time effects  $\tau(g, t)$ , which can be aggregated into event-time effects  $\theta_k$  as described in Section 7.7.

Make this explicit before you look at estimates. The aggregation schemes in Section 7.7 show how unit-level or cohort-time effects map into different summaries. Pre-specifying the target helps prevent you from drifting toward whichever summary looks most flattering ex post.

### Step 2: Curate the Donor Pool

Next, assemble a donor pool that can plausibly stand in for the treated units in the absence of treatment. Exclude units that receive treatment during the estimation window, units that are clearly incomparable on business grounds, and units that are likely to be heavily contaminated by spillovers, drawing on the guidance in Section 7.8 and Chapter 11.

Document these choices. Summarise how donors differ from treated units on key characteristics such as region, size, format or demographics. If donors are systematically smaller, poorer or otherwise different, that will influence your choice of method: designs such as ASCM that can correct covariate imbalances become more attractive.

In staggered designs, remember that the donor pool for a given cohort and period consists of not-yet-treated and never-treated units. Track how this pool shrinks as later cohorts adopt. Late adopters may have far fewer valid donors than early ones.

### Step 3: Choose Candidate Hybrid Methods

Choose one or more hybrid estimators that match the structure of your problem. Section 7.3 lays out the main trade-offs.

ASCM is a natural candidate when treated units differ systematically from donors on observables that predict outcomes and when pure SC cannot achieve good pre-treatment fit. Ridge SC is helpful when the donor pool is large relative to the pre-period and you worry about unstable weights. SDID fits staggered adoption and settings where treated and control units have different pre-trends but can be aligned after reweighting. More complex factor-based hybrids such as TROP make sense only when interactive fixed effects are plausible, the panel is reasonably large and you have the technical capacity to implement and scrutinise them.

If more than one method seems appropriate — which is common — plan from the outset to estimate several and compare their diagnostics and estimates.

### Step 4: Select Predictors and Tuning Rules

Specify the predictors you will use for weighting and, for ASCM, for augmentation, drawing on Section 7.8. Pre-treatment outcomes over a reasonably long window are usually central; a small number of well-chosen covariates that capture scale, demographics and competition often add value. Avoid building predictor sets by trial and error on the post-treatment data.

Decide how you will tune regularisation parameters using only pre-treatment information. For ridge-type hybrids, this typically means splitting the pre-period into a training block and a validation block and searching over a broad grid of penalties. For SDID, most implementations solve the underlying optimisation problem once given the design; tuning is implicit rather than explicit. For TROP, staged cross-validation is one option, but given its research status it should be treated as experimental rather than routine.

Write down these rules before inspecting post-treatment outcomes to reduce the temptation to chase specifications that produce the most appealing effects.

### Step 5: Assess Pre-Treatment Fit and Balance

Estimate your chosen hybrids and compute pre-treatment RMSPE for each, as defined in Section 7.9. Express RMSPE relative to the standard deviation of pre-period outcomes and compare across methods. Designs that cannot track the treated unit’s pre-treatment path at all are not credible; designs that match the path much better than simpler alternatives, without implausible weights, are promising, conditional on donor validity (no spillovers) and stable pre/post relationships.

At the same time, examine covariate balance for variables you believe matter for both treatment and outcomes. Standardised mean differences provide a convenient summary; values close to zero indicate good

balance, and large absolute values flag imbalances that may threaten identification. You do not need to enforce a rigid cutoff, but imbalances on central covariates should be rare in a design you trust.

Finally, look at weight dispersion. Effective numbers of donors or periods that are either extremely low (all mass on one donor or one period) or extremely high (near-uniform weights) deserve scrutiny. Visualising weight distributions over donors and periods alongside pre-treatment fit helps you see whether a design is relying too heavily on a few units or times.

If pre-treatment fit is poor, key covariates are badly imbalanced, or weights look extreme, regard this as feedback, not as failure. Revisit the donor pool, predictors or tuning and iterate until you either achieve acceptable diagnostics or conclude that the design cannot support a reliable hybrid.

## Step 6: Run Placebos and Sensitivity Checks

Once pre-period diagnostics look reasonable, stress-test the design. Apply in-space placebos by treating donors as if they were exposed and comparing their placebo gaps to the treated unit's gap. Apply in-time placebos by shifting the intervention date into the pre-period and checking whether the method spuriously produces "effects" where none should exist.

Conduct leave-one-out analyses by re-estimating after removing influential donors or pre-periods. Large swings in estimates when a single donor or period is omitted indicate fragility and call for explanation. Residual plots and weight-distribution plots help you see whether such donors and periods are truly special or artefacts of the optimisation.

The goal of these checks is not to pass a checklist, but to understand how your design behaves under small perturbations.

## Step 7: Compute and Interpret Inference

Choose inference tools that respect your sample size and design, drawing on Chapter 16 and Section 7.10. In small marketing panels, unit-level or cohort-level block bootstraps, wild bootstraps and placebo-based distributions are often more reliable than asymptotic formulas. When there is only one treated unit, placebo-based and time-permutation tools often carry more interpretive weight than a naive unit bootstrap.

Be explicit about what each inferential device assumes. Rank-based placebo statistics rely on symmetry or exchangeability arguments; bootstrap procedures treat your sample as a stand-in for a wider population; conformal-style prediction intervals lean on residual exchangeability. In observational work, randomisation-style p-values are best viewed as sensitivity tools, not as definitive causal tests.

Report interval estimates and uncertainty summaries alongside point estimates, and interpret them in the context of effect sizes that matter for decisions.

### Step 8: Aggregate and Explore Heterogeneity

If your design involves multiple treated units or cohorts, aggregate unit-level or group–time estimates into the estimand you specified in Step 1. Use weights that reflect the policy question — equal weights for “average treated unit” questions, population or revenue weights for questions about aggregate impact — and report these weights so readers can see which units drive the results.

For staggered adoption, present event-time profiles with confidence bands, mark the omitted event time (the reference period that defines the baseline for  $\theta_k$ ) and indicate which cohorts contribute at each horizon. For common-adoption designs, show unit-specific estimates alongside the aggregate to reveal heterogeneity. In both cases, sensitivity analyses that vary aggregation weights or restrict attention to better-matched units help assess how conclusions depend on these choices.

### Step 9: Document Assumptions and Make the Analysis Auditable

Finally, write down the identification assumptions you are relying on: no anticipation, no interference (or an explicit exposure mapping if spillovers are modelled, using  $h_i(D_{-i,t})$  as in Chapter 11), stability of the predictor–outcome relationship across pre- and post-periods, and sufficient overlap between treated units and donors. Connect these to the economic context of your application. Explain why you believe, for example, that donors were not affected by the campaign or that there were no structural breaks at the time of treatment.

Document any deviations from your pre-specified plan and justify them. Where possible, provide replication materials — cleaned data or suitably anonymised proxies, code, and details of software versions — so that others can reproduce your results or apply alternative specifications. Transparency about design choices, diagnostics and inference strengthens the credibility of your conclusions more than any single “significant” estimate.

Used together with the workflow above, the diagnostics and examples in this chapter provide a practical blueprint for deploying hybrid methods in marketing panels in a way that is both rigorous and transparent.

**Table 7.1** Hybrid Methods at a Glance (see [Abadie et al., 2010, Ben-Michael et al., 2021, Arkhangelsky et al., 2021, Athey et al., 2025b] for methodological details)

Method	Key Assumptions	Tuning Focus	Typical Use-Cases
Standard SC	Convex-hull approximation; no anticipation; stability of the imputation mapping learned in the pre-period	Predictor set; choice of pre-period	Single treated unit; long pre-period; treated unit well represented in donor pool
ASCM	Same as SC plus outcome-model stability; bias is small when weighting imbalance and augmentation error are both small under the maintained model	Choice of covariates; regularisation strength for augmentation and, where used, weights	Treated units near but not clearly inside donor hull; important covariate imbalances; short pre-periods
Ridge SC	Approximate convex-hull match with explicit shrinkage towards diffuse weights	Ridge penalty on weights; pre-period split for validation	Large donor pools; concern about unstable or overly concentrated weights
SDID	Weighted parallel trends after reweighting units and periods; overlap between treated and not-yet-treated units	Implementation details for weight estimation; choice of pre-period window	Staggered adoption; differential pre-trends across units; rich donor pool of not-yet-treated markets
TROP	Low-rank factor structure for untreated outcomes; shared factor space between treated and donor units	Penalties on unit weights, time weights and factor rank (typically tuned by cross-validation)	Panels with strong interactive fixed effects; moderate sample sizes; teams comfortable with factor models

**Part IV**

**Factor Models and Matrix Methods**



## Chapter 8

# Interactive Fixed Effects and Matrix Completion

This chapter develops factor models and matrix completion methods for counterfactual imputation and treatment-effect estimation in marketing panels. We treat factor models as low-rank statistical approximations to panel outcomes, capturing shared shocks and heterogeneous exposure across units, and we show how this structure motivates interactive fixed effects (IFE) estimators.

We formalise IFE models, including identification up to rotation (that is, identification of the factor space and fitted values, not the factors themselves) and selection of the number of factors. We then show how to apply matrix completion with nuclear-norm regularisation to impute counterfactual outcomes for treated or otherwise missing cells. We develop diagnostics, tuning rules, and inference procedures that work in marketing panels with serial dependence and a limited number of clusters. By the end of the chapter, you should be able to (i) specify an IFE model for a marketing panel, (ii) select the factor rank, (iii) implement matrix-completion estimators, and (iv) diagnose when a low-rank approximation is not credible.

Factor-based designs sit alongside parallel-trends and synthetic control approaches introduced earlier in the book. As throughout the book, credible use requires a clear estimand, identification assumptions about the assignment mechanism and stability of  $\hat{Y}_{it}(0)$ , an estimator for imputing counterfactual outcomes, and diagnostics and inference that assess sensitivity to these choices. We map connections to synthetic control and SDID (Chapters 6 and 7), clarifying when factor designs are preferable to parallel-trends designs and when you should instead rely on difference-in-differences (Chapter 4). Later sections refer back to event-study diagnostics and inference procedures developed in Chapters 5, 17, and 16.

## 8.1 Motivation and Setup

Factor models provide a framework for understanding how panel outcomes co-move over time and across units. Rather than imposing the parallel-trends restriction from Chapter 4 or relying solely on the synthetic control reweighting from Chapter 6, we posit that untreated outcomes are driven by a small number of common time-varying shocks, or factors. These factors affect different units with heterogeneous intensities, captured by unit-specific loadings. This view explains patterns of co-movement across markets, stores, products, or customers and motivates estimators that impute counterfactual outcomes by predicting  $Y_{it}(0)$  for treated unit-period cells using the estimated factor structure.

The factor model view is particularly natural in marketing panels. Regional sales co-move because they respond to national macroeconomic conditions, seasonal patterns, and industry-wide trends, but the sensitivity to these shocks varies across regions due to differences in demographics, competitive intensity, or retail infrastructure. Store-level revenue displays common patterns driven by holidays, weather, and category-wide demand shifts, yet stores differ in their exposure based on format, size, and location. By decomposing outcomes into common shocks and heterogeneous loadings, factor models provide a parsimonious representation of high-dimensional panel data that captures these regularities. For single-unit synthetic control designs we refer to Chapter 6, and for hybrid estimators and staggered adoption we refer to Chapter 7.

### Trade-offs with Alternative Approaches

The key advantage of factor models over parallel trends and synthetic control lies in how they model common shocks. Under parallel trends (Chapter 4), time shocks enter as an additive fixed effect  $\lambda_t$  common to all units. This is restrictive when units differ in their sensitivity to aggregate conditions, for example when a recession depresses revenue in urban stores more than in rural stores. Synthetic control (Chapter 6) requires that the treated unit lies within the convex hull of the controls, which can fail when the treated unit is larger, more urban, or otherwise structurally different from all available controls.

Factor models relax both restrictions. They allow time shocks to affect units with heterogeneous intensities through interactive fixed effects of the form  $\lambda'_i f_t$  rather than purely additive time effects  $\lambda_t$ . They approximate a treated unit's outcomes using its estimated loading vector, without convexity constraints linking that vector to the loadings of control units. The treated unit need not lie within the convex hull of the controls; instead, its loading vector  $\lambda_i$  is estimated from pre-treatment data, while post-treatment factors  $f_t$  are learned from untreated controls. This logic requires stability of the untreated factor structure across time.

The cost of this flexibility is the assumption that untreated outcomes are well approximated by a low-rank factor structure. We assume that the  $N \times T$  outcome matrix for untreated potential outcomes can be approximated by the product of an  $N \times R$  loading matrix  $\Lambda$  and an  $R \times T$  factor matrix  $F'$ , where the rank  $R$  is small relative to both  $N$  and  $T$ . Formally, most of the variation in outcomes is attributed to a small number of common factors. This low-rank assumption is an empirical approximation rather than an economic principle. We can assess its plausibility using diagnostic tools such as scree plots of the eigenvalues of the

outcome matrix, eigenvalue ratio tests that compare adjacent eigenvalues [Ahn and Horenstein, 2013], and information criteria that penalise model complexity [Bai and Ng, 2002, Bai, 2009]. Section 8.2 discusses rank selection and these diagnostics in detail.

When a small number of macroeconomic shocks, seasonal patterns, and category trends drive most of the variation in outcomes, the low-rank structure is a plausible approximation and factor models can deliver accurate counterfactuals and efficient estimates. When outcomes are instead driven by many idiosyncratic unit-time shocks with little common structure, factor models risk overfitting the pre-treatment data and extrapolating poorly to the post-treatment period. In practice, failure of the low-rank approximation shows up as slow decay of the singular values of the outcome matrix, highly erratic estimated factors with no interpretable macro or seasonal pattern, or large residual serial correlation after removing the factor structure. When we see these patterns in marketing panels, the simpler parallel-trends structure from Chapter 4 is often a safer choice.

## Connection to the Potential Outcomes Framework

Connecting to the potential outcomes framework introduced in Chapter 2 and grounded in the foundational work of Rubin [1974], Imbens and Rubin [2015], we use factor models to approximate the evolution of untreated potential outcomes  $Y_{it}(0)$ , with treatment effects entering additively on top of  $Y_{it}(0)$ . (In staggered-adoption chapters we used  $Y_{it}(g)$  and  $Y_{it}(\infty)$ . Here we use the two-potential-outcome notation 0/1 because treatment patterns can be arbitrary.) The most general specification in this chapter is the interactive fixed effects (IFE) model

$$Y_{it}(0) = \alpha_i + \lambda_t + \lambda'_i f_t + \varepsilon_{it},^1$$

where  $\alpha_i$  and  $\lambda_t$  are the unit and time fixed effects from the canonical panel specification in Chapter 4. The term  $\lambda_i \in \mathbb{R}^R$  is the loading vector for unit  $i$ ,  $f_t \in \mathbb{R}^R$  is the factor vector for period  $t$ , and  $\varepsilon_{it}$  is an idiosyncratic error with mean zero. Writing  $\lambda_i = (\lambda_{i1}, \dots, \lambda_{iR})'$  and  $f_t = (f_{t1}, \dots, f_{tR})'$ , the interactive term  $\lambda'_i f_t$  coincides with the factor representation in the notation guide,  $\sum_{r=1}^R \lambda_{ir} f_{tr}$ . The interactive term  $\lambda'_i f_t$  captures unit-specific responses to common time-varying shocks.

Several familiar models arise as special cases. If we set  $\lambda_i = 0$  for all  $i$ , there are no interactive effects and the model reduces to the two-way fixed effects structure underlying difference-in-differences. To see how DiD fits within the factor framework, note that the additive time effect  $\lambda_t$  can be written as an interactive term with a single factor,  $\lambda'_i f_t$ , by setting  $R = 1$ ,  $\lambda_i = 1$  for all  $i$ , and  $f_t = \lambda_t$ . In this sense, DiD corresponds to a rank-one factor model in which all units share identical loadings on the common time factor. When  $\alpha_i = 0$  and  $\lambda_t = 0$ , we obtain a pure factor model with no additive fixed effects. For treatment effect estimation, we impute the counterfactual untreated outcome for a treated unit  $i$  in a post-treatment period  $t$  as  $\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\lambda}_t + \hat{\lambda}'_i \hat{f}_t$ , and define the treatment effect as the difference between the observed outcome  $Y_{it}$

---

<sup>1</sup> In this chapter we denote the untreated potential outcome as  $Y_{it}(0)$ . Unlike the staggered adoption setting in other chapters where  $Y_{it}(\infty)$  is used, factor models allow for arbitrary treatment patterns, so the two-potential-outcome notation  $Y_{it}(0)$  and  $Y_{it}(1)$  is standard in this literature [Xu, 2017].

and this imputed counterfactual. Throughout this chapter we treat the factor structure that drives untreated potential outcomes,  $\alpha_i$ ,  $\lambda_t$ ,  $\lambda_i$ , and  $f_t$ , as unaffected by the treatment. Treatment changes outcomes only through an additive treatment effect on top of  $Y_{it}(0)$ . This exclusion restriction is the factor-model analogue of the parallel-trends assumption in difference-in-differences. Later sections state it formally and show how to diagnose violations.

## Connection to Matrix Completion

Factor models are closely related to matrix completion methods. Viewed as a matrix, the panel of untreated outcomes  $\{Y_{it}(0)\}$  has missing entries at the treated cells, the unit-time combinations where treatment was applied. Under the IFE model we can write the signal part of this matrix as  $\alpha_i + \lambda_t + \lambda'_i f_t$ . After partialling out or absorbing the additive fixed effects  $\alpha_i$  and  $\lambda_t$ , the remaining signal matrix induced by  $\lambda'_i f_t$  has the form  $AF'$  and is therefore low-rank.

Nuclear-norm regularisation can recover this low-rank structure under suitable conditions on the rank, noise level, and pattern of observed entries. In the matrix completion literature, these conditions typically require that the rank  $R$  is small relative to  $\min(N, T)$ , that the factors are not too "spiky" (so information about them is spread across units and periods), and that the observed entries cover the matrix in a reasonably diffuse way rather than concentrating in a few rows or columns; see, for example, Candès and Recht [2009], Candès [2010], Mazumder et al. [2010], Athey et al. [2021]. Matrix completion methods recover the underlying low-rank matrix from the observed (control) entries and then use it to impute the missing (treated) entries. This connects causal panel methods to a broader literature on low-rank matrix estimation, including nuclear-norm regularisation, singular value decomposition, and soft-thresholding.

This perspective also clarifies when imputation is feasible. The observed entries must be informative enough about the low-rank structure that the missing treated cells can be inferred. If treated units or treated periods are too different from the available controls, or if large blocks of the matrix are unobserved, the imputation becomes unreliable. For example, if a national campaign treats all large urban stores at once, leaving no large urban controls in the post-treatment period, the low-rank structure cannot be learned for that segment and matrix completion will necessarily extrapolate based on smaller or rural stores. Beyond approximate low rank, credible imputation requires that the pattern of missing entries induced by treatment does not destroy our ability to recover the low-rank structure from untreated cells. Intuitively, the treated block must be predictable from the observed control block using the common factors and loadings. Later in the chapter we formalise this requirement through restrictions on how large the treated block can be and how similar treated units and periods must be to their controls. Section 8.3 develops the matrix completion perspective in detail.

## Unification with Synthetic Control and SDID

The framework developed by Arkhangelsky and Imbens [2024] situates factor models within the modern panel-data literature and clarifies their connection to synthetic control, difference-in-differences, and hybrid methods discussed in Chapter 7.

The relationship is most transparent when we compare how each method constructs the counterfactual for a treated unit. In a simple difference-in-differences design, we can write the counterfactual as  $\hat{Y}_{it}(0) = \bar{Y}_{i,\text{pre}} + (\bar{Y}_{\text{control},t} - \bar{Y}_{\text{control,pre}})$ , which relies on the parallel-trends assumption. Here  $\bar{Y}_{i,\text{pre}}$  denotes the average pre-treatment outcome for treated unit  $i$ , and  $\bar{Y}_{\text{control},t}$  and  $\bar{Y}_{\text{control,pre}}$  are defined as in Chapter 4. Synthetic control constructs the counterfactual as  $\hat{Y}_{it}(0) = \sum_j w_j Y_{jt}$ , where the weights  $w_j$  are chosen to match pre-treatment outcomes and satisfy  $w_j \geq 0$  and  $\sum_j w_j = 1$ . This is equivalent to assuming that the treated unit's loading vector is a convex combination of the control loadings,  $\lambda_1 = \sum_j w_j \lambda_j$ , so that its potential outcomes lie in the convex hull of the controls (conditional on pre-treatment fit). Factor models estimate the counterfactual as  $\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\lambda}_t + \hat{\lambda}'_i \hat{f}_t$ , which relies on the assumption that untreated outcomes are well approximated by a low-rank factor structure that remains stable from the pre-treatment to the post-treatment period.

This unification highlights when each method is appropriate. When the treated unit is similar to controls and lies within their convex hull, synthetic control is natural. When units share similar loadings and respond similarly to common shocks, the parallel-trends structure behind difference-in-differences is adequate. When the treated unit is an outlier with different loadings, or when we expect rich heterogeneity in responses to aggregate conditions, factor models become necessary. Hybrid methods in Chapter 7 combine these ideas through augmentation or regularisation.

## Chapter Outline

This chapter develops factor models and matrix completion methods in detail, with a focus on practical implementation in marketing panels. We present the interactive fixed effects model first, including identification, rank selection, and estimation. We then develop the matrix completion perspective, emphasising nuclear-norm regularisation and the role of missingness patterns. Formal identification assumptions are articulated using Assumption environments, and we map explicit connections to synthetic control and SDID. We conclude with guidance on tuning, inference, and diagnostic workflows that help you assess whether the low-rank assumption is plausible in your application.

## 8.2 Interactive Fixed Effects (IFE)

The interactive fixed effects (IFE) model decomposes untreated potential outcomes into unit-specific loadings and time-specific factors. In this section we formalise the model, discuss identification up to rotation, present methods for selecting the number of factors, and describe estimation procedures that recover loadings and factors from untreated observations in marketing panels.

### Model Specification

The general IFE model for untreated potential outcomes, introduced in Section 8.1, combines additive and interactive components:

$$Y_{it}(0) = \alpha_i + \lambda_t + \lambda'_i f_t + \varepsilon_{it},$$

where  $\alpha_i$  and  $\lambda_t$  are the unit and time fixed effects from the canonical panel specification in Chapter 4. The term  $\lambda_i \in \mathbb{R}^R$  is the loading vector for unit  $i$ ,  $f_t \in \mathbb{R}^R$  is the factor vector for period  $t$ , and  $\varepsilon_{it}$  is an idiosyncratic error with mean zero and finite variance. Writing  $\lambda_i = (\lambda_{i1}, \dots, \lambda_{iR})'$  and  $f_t = (f_{t1}, \dots, f_{tR})'$ , the interactive term  $\lambda'_i f_t$  coincides with the factor representation  $\sum_{r=1}^R \lambda_{ir} f_{tr}$  used in the notation guide.

In practice, we first remove the additive fixed effects by demeaning. Define the demeaned outcome

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot t} + \bar{Y}_{\cdot\cdot},$$

where  $\bar{Y}_{i\cdot}$  is the unit mean,  $\bar{Y}_{\cdot t}$  is the time mean, and  $\bar{Y}_{\cdot\cdot}$  is the grand mean, computed over the untreated estimation sample  $\Omega$  defined below. If the IFE model holds, then the demeaned untreated potential outcomes satisfy

$$\tilde{Y}_{it}(0) = \tilde{\lambda}'_i \tilde{f}_t + \tilde{\varepsilon}_{it},$$

where  $\tilde{\lambda}_i$  and  $\tilde{f}_t$  are the loadings and factors after demeaning. For notational simplicity, in the remainder we drop tildes and interpret  $Y_{it}(0)$ ,  $\lambda_i$ ,  $f_t$ , and  $\varepsilon_{it}$  as demeaned objects unless stated otherwise.

For most of this section we therefore work with the demeaned factor model

$$Y_{it}(0) = \lambda'_i f_t + \varepsilon_{it},$$

and (to reduce notation) we denote the demeaned outcome by  $Y_{it}(0)$  in displayed equations. We recover level outcomes by adding back estimated unit and time effects when constructing counterfactuals. This connects directly to the approximate factor model literature surveyed in Bai [2009], Bai and Ng [2002].

## Interpretation of Loadings and Factors

The factor vector  $f_t = (f_{t1}, \dots, f_{tR})'$  represents common time-varying shocks that affect all units. The  $r$ -th factor  $f_{tr}$  captures a distinct source of variation such as a macroeconomic shock, a seasonal pattern, or a category-specific trend. The loading vector  $\lambda_i = (\lambda_{i1}, \dots, \lambda_{iR})'$  represents how sensitive unit  $i$  is to each factor. If unit  $i$  has a large positive loading  $\lambda_{ir}$  on factor  $r$ , then shocks in factor  $r$  have a large positive effect on that unit's outcomes.

In applied work you will often start by plotting the estimated factors over time to identify patterns such as seasonality, trends, and structural breaks. You can then plot loadings against unit characteristics (for example, store size, customer demographics, or region) to see which units are most sensitive to which factors. Figure 8.3 illustrates a typical factor structure with three factors: a macro factor that affects all units similarly, a seasonal factor with heterogeneous loadings, and a regional factor that affects only a subset of units.

## Identification Up to Rotation

Factors and loadings are identified only up to rotation and scale. If  $(\lambda_i, f_t)$  satisfy the model, then so do  $(H\lambda_i, H^{-1}f_t)$  for any invertible  $R \times R$  matrix  $H$ :

$$\lambda_i' f_t = (H\lambda_i)'(H^{-1}f_t).$$

This rotational invariance means that the estimated factors and loadings are not uniquely determined without normalisation.

Following Bai and Ng [2002] and Bai [2009], a standard normalisation imposes two conditions. First, the factors are orthonormal in sample,  $T^{-1} \sum_t f_t f_t' = I_R$ . Second, the loading covariance is diagonal with distinct diagonal elements,  $N^{-1} \sum_i \lambda_i \lambda_i'$  is diagonal. These constraints pin down scale and rotation up to sign flips. Under this normalisation the first factor explains the largest share of variance, the second explains the next largest share conditional on the first, and so on.

For causal inference the implication is that we should treat factors and loadings as summary representations of common structure rather than structural causal objects. The product  $\lambda_i' f_t$  is identified (up to the model), but the individual components are not separately identified. This is not a problem for treatment effect estimation because only the product enters the counterfactual prediction.

## Selecting the Number of Factors

Choosing the rank  $R$  is critical. If we include too few factors, we underfit the data, leave systematic co-movement unexplained, and risk biased counterfactuals. If we include too many factors, we overfit idiosyncratic noise and produce unstable estimates. Several approaches guide the choice.

Information criteria, such as the Bai–Ng criteria IC1, IC2, and IC3, penalise model complexity and select  $R$  to balance fit and parsimony [Bai and Ng, 2002]. IC2 is commonly used and takes the form

$$\text{IC2}(R) = \log(\hat{\sigma}_R^2) + R \cdot \frac{(N+T)}{NT} \log(\min(N, T)),$$

where  $\hat{\sigma}_R^2$  is the mean squared residual with  $R$  factors.

Eigenvalue-based diagnostics provide a complementary view. Plot the eigenvalues  $\mu_1 \geq \mu_2 \geq \dots$  of a sample covariance matrix such as  $\mathbf{Y}'\mathbf{Y}/T$  or  $\mathbf{Y}\mathbf{Y}'/N$  and look for an “elbow” where eigenvalues drop sharply. Eigenvalue ratio rules formalise this idea by selecting  $R$  where the ratio  $\mu_R/\mu_{R+1}$  is largest, indicating a sharp drop in explained variance between the  $R$ -th and  $(R+1)$ -th eigenvalue.

You can also use cross-validation. Split the pre-treatment period into training and validation sets. For each candidate  $R$ , estimate factors on the training set and compute prediction error on the validation set. Choose  $R$  to minimise validation error. In all cases, the loss function in cross-validation should target prediction of untreated outcomes, not overall fit, because the goal is to minimise counterfactual error on cells that are missing by design. A practical implementation is to hold out the last 20% of pre-treatment periods as the validation set and treat earlier periods as the training set.

Finally, economic reasoning should discipline purely statistical selection. Count the plausible sources of common variation: one or two factors for seasonality, one factor for macroeconomic shocks, one factor per major category trend, and possibly one factor per large region or region group. This gives a prior range for  $R$ . In many marketing panels it is sensible to explore values between  $R = 3$  and  $R = 10$  and to report results for several choices (for example,  $R = 3, 5, 7$ ) as a robustness check rather than committing to a single preferred rank. This range is problem-dependent; in granular panels with many distinct product categories or regions, higher  $R$  may be needed to capture the richer structure.

## Least Squares Estimation

We estimate the factor structure by minimising the sum of squared residuals on observed untreated cells. Let  $\Omega$  denote the set of observed unit-period pairs used for estimation. The least squares estimator solves

$$\min_{\Lambda, \mathbf{F}} \sum_{(i,t) \in \Omega} (Y_{it} - \lambda'_i f_t)^2 \quad \text{subject to the normalisation constraints.}$$

Here  $\mathbf{Y}$  is the  $N \times T$  matrix of demeaned outcomes,  $\Lambda$  stacks the loadings, and  $\mathbf{F}$  stacks the factors.

For balanced panels, where all pre-treatment cells used in estimation are observed, we obtain the solution via principal components analysis (PCA). Stack outcomes in  $\mathbf{Y}$ . The estimated factor matrix  $\hat{\mathbf{F}}$  consists of the first  $R$  eigenvectors of  $\mathbf{Y}'\mathbf{Y}$ , scaled by  $\sqrt{T}$ . The estimated loadings are  $\hat{\Lambda} = \mathbf{Y}\hat{\mathbf{F}}/T$ . Equivalently, we can apply singular value decomposition (SVD),  $\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}'$ , and take  $\hat{\mathbf{F}} = \sqrt{T}\mathbf{V}_R$  and  $\hat{\Lambda} = \mathbf{U}_R\mathbf{S}_R/\sqrt{T}$ , where the subscript  $R$  denotes the first  $R$  components. This scaling ensures consistency with the normalisation  $T^{-1}\mathbf{F}'\mathbf{F} = I_R$  (equivalently,  $T^{-1} \sum_t f_t f_t' = I_R$ ).

The computational cost of a full SVD is  $O(\min(N, T) NT)$ . For typical marketing panels, such as a few hundred to a few thousand units observed over a few hundred time periods, this is usually feasible on standard hardware. For very large panels, with both  $N$  and  $T$  in the tens of thousands or more, randomised SVD or iterative methods provide faster approximations.

When the panel is unbalanced because some cells are missing (for example, under staggered treatment adoption), PCA on the full matrix is not directly applicable. In that case we use an alternating least squares algorithm, which alternates between updating loadings given factors and updating factors given loadings. Box 8.2 describes a practical implementation. Each step weakly decreases the objective, so the sequence of objective values converges, and in practice the algorithm converges to a useful stationary point. Multiple starting values help avoid poor local solutions; we choose the solution with the smallest objective. Because the objective is not jointly convex in  $\Lambda$  and  $\mathbf{F}$ , the algorithm can converge to different local minima depending on starting values. In practice, you should initialise from several random or PCA-based starting points and check that the implied counterfactuals and aggregate treatment effects are stable across runs.

#### Box 8.1: Iterative Least Squares Algorithm for IFE

**1. Initialise:** Set initial factors  $\hat{\mathbf{F}}^{(0)}$  using PCA on a balanced subset of control units or via random initialisation.

**2. Update Loadings:** For each unit  $i$ , regress its observed outcomes on the current factors:

$$\hat{\lambda}_i^{(k+1)} = \left( \sum_{t \in \Omega_i} \hat{f}_t^{(k)} \hat{f}_t^{(k)'} \right)^{-1} \sum_{t \in \Omega_i} \hat{f}_t^{(k)} Y_{it},$$

where  $\Omega_i$  is the set of observed untreated periods for unit  $i$ , provided the matrix is invertible (in practice you may add a small ridge term  $\epsilon I_R$  if the pre-treatment window is short or ill-conditioned).

**3. Update Factors:** For each period  $t$ , regress observed outcomes on the current loadings:

$$\hat{f}_t^{(k+1)} = \left( \sum_{i \in \Omega_t} \hat{\lambda}_i^{(k+1)} \hat{\lambda}_i^{(k+1)'} \right)^{-1} \sum_{i \in \Omega_t} \hat{\lambda}_i^{(k+1)} Y_{it},$$

where  $\Omega_t$  is the set of observed untreated units in period  $t$ , provided the matrix is invertible.

**4. Normalise and Check:** Apply the Bai–Ng normalisation to the updated matrices. Check convergence of the objective (sum of squared residuals). Repeat steps 2–4 until the relative change in the objective falls below a tolerance such as  $10^{-6}$ . In practice, 10 to 50 iterations are usually sufficient.

## Consistency and Asymptotics

In the classical approximate factor model with all  $NT$  cells observed, Bai and Ng [2002], Bai [2009] show that under weak cross-sectional and serial dependence in  $\varepsilon_{it}$ , and with strong factors whose associated eigenvalues grow with  $N$  and  $T$ , the principal components estimator consistently recovers the factor space. In our causal panel setting, only the untreated cells in  $\Omega$  are observed. Consistency then also requires that the untreated cells are dense enough across units and periods to identify both the loading and factor spaces. Intuitively, we need sufficiently many untreated periods for each treated unit, and sufficiently many untreated units in

each period, so that the space spanned by the observed cells approximates the full factor structure. Later sections translate this requirement into design diagnostics. Most formal asymptotic results for factor models, such as Bai and Ng [2002], Bai [2009], assume a fully observed panel. Extending these results to patterns of missingness induced by treatment requires additional assumptions on the density and design of untreated cells; we view our large- $N$ , large- $T$  arguments here as heuristic guidance rather than a complete theory, especially because the missingness pattern is induced by treatment rather than by generic sampling. As  $N, T \rightarrow \infty$ , the space spanned by the estimated factors converges to the true factor space:

$$\|\mathbf{P}_{\hat{\mathbf{F}}} - \mathbf{P}_{\mathbf{F}^0}\| \xrightarrow{P} 0,$$

where  $\mathbf{P}_{\mathbf{F}}$  is the projection matrix onto the column space of  $\mathbf{F}$  and  $\mathbf{F}^0$  denotes the true factor matrix.

The strong-factors assumption requires that each factor explains a non-negligible share of variance. Weak factors, whose eigenvalues do not grow with  $N$  or  $T$ , are harder to estimate consistently and may require regularisation. In marketing panels with clear seasonal and macroeconomic structure, we typically expect at least some strong factors.

## Treatment Effect Estimation

For causal inference we must restrict estimation to untreated cells to avoid contaminating factor estimates with treatment effects. Identification of treatment effects from the IFE model requires that the factor structure for untreated potential outcomes remains stable across the pre-treatment and post-treatment periods. Treatment can change outcomes only through an additive treatment effect on top of  $Y_{it}(0)$ ; it cannot alter the unit effects  $\alpha_i$ , time effects  $\lambda_t$ , or factor loadings and shocks  $\lambda_i$  and  $f_t$ . If treatment changes the factor structure itself—for example, by inducing a new common shock that affects both treated and control units—then the IFE-based counterfactuals will generally be biased.

In a single treatment timing design where treatment begins in period  $T_0 + 1$  for a subset of units, we estimate factors and loadings using all periods for never-treated control units and pre-treatment periods  $t \leq T_0$  for treated units. The estimated loadings  $\hat{\lambda}_i$  for treated units come entirely from their pre-treatment data. The estimated factors  $\hat{f}_t$  for post-treatment periods come from control units' outcomes in those periods.

With staggered adoption, each unit  $i$  has its own treatment start period  $T_{0i}$ . The observed untreated cells form the set

$$\Omega = \{(i, t) : t \leq T_{0i} \text{ or unit } i \text{ is never treated}\}.$$

We apply the iterative least squares algorithm to this irregular observation pattern. The algorithm naturally handles unit-specific treatment timing by including only untreated cells in each unit's and period's regression.

For a treated unit  $i$  in a post-treatment period  $t > T_{0i}$ , we construct the counterfactual untreated outcome as

$$\hat{Y}_{it}^{\text{IFE}}(0) = \hat{\alpha}_i + \hat{\lambda}_t + \hat{\lambda}'_i \hat{f}_t,$$

where the fixed effects are recovered from the demeaning step when they are part of the specification. The treatment effect estimate is then

$$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}^{\text{IFE}}(0).$$

## Practical Considerations

Several practical issues arise in implementation.

Scaling matters when units operate on very different outcome scales. Standardising outcomes (for example, dividing by the within-unit standard deviation) prevents high-variance units from dominating factor estimation, but it also changes the implicit weighting of units in the objective. Report the scaling choice.

Seasonality is often prominent in marketing panels. You can include seasonal dummies in the demeaning step or estimate an explicit seasonal factor. Cross-validation over specifications with and without explicit seasonal adjustment helps determine whether these adjustments improve out-of-sample fit.

Software implementation is widely available. The `gsynth` package in R implements IFE estimation with cross-validation for rank selection, and the `fect` package provides IFE and matrix completion methods for causal panel analysis. Both handle staggered adoption and missing data patterns similar to those discussed here.

More advanced specifications—including nuclear-norm regularisation, grouped factors, and tighter connections to synthetic control—are developed in Section 8.3 and Chapter 9.

### 8.3 Matrix Completion Perspective

The matrix completion perspective views the outcome matrix as partially observed, with missing or treated cells, and seeks to impute the unobserved entries by exploiting low-rank structure. In this section we present nuclear-norm minimisation as a convex surrogate for rank minimisation, discuss missingness patterns and their implications for recovery, and address the bias–variance trade-off induced by regularisation. For advanced matrix and tensor methods that handle multi-way panels, non-stationarity, or outliers, we refer to Chapter 9.

#### Formalisation and Rank Minimisation

Matrix completion formalises the imputation problem as follows. The  $N \times T$  outcome matrix  $\mathbf{Y}$  is partially observed. We observe entries  $Y_{it}$  for  $(i, t) \in \Omega$ , where  $\Omega$  is the set of observed cells, typically untreated unit-period pairs after two-way demeaning as in Section 8.2. We seek to impute entries  $Y_{it}$  for  $(i, t) \notin \Omega$ , which correspond to treated or otherwise missing cells.

If  $\mathbf{Y}$  has low rank, as under the IFE model with rank  $R$  where the noiseless component  $\Lambda F'$  has rank  $R$ , then we can pose recovery as the optimisation problem

$$\min_{\mathbf{M}} \text{rank}(\mathbf{M}) \quad \text{subject to} \quad M_{it} = Y_{it} \text{ for all } (i, t) \in \Omega.$$

This problem searches for the lowest-rank matrix  $\mathbf{M}$  that matches the observed entries exactly. Under appropriate conditions on the rank, the singular vectors, and the observation pattern, the solution can accurately approximate unobserved entries.

Figure 8.1 illustrates the matrix completion setup. The observed cells, corresponding to control units and pre-treatment periods, are shaded, while the treated cells to be imputed are blank. When the low-rank structure and observation pattern are favourable, the information in the shaded cells is sufficient to recover the missing entries in the blank region.

#### Nuclear-Norm Minimisation

Direct rank minimisation is computationally intractable (NP-hard), which motivates a convex relaxation. The nuclear norm of a matrix  $\mathbf{M}$ , defined as the sum of its singular values,

$$\|\mathbf{M}\|_* = \sum_{k=1}^{\min(N,T)} \sigma_k(\mathbf{M}),$$

serves as a convex surrogate for rank [Candès and Recht, 2009, Candes, 2010]. A regularised matrix completion estimator solves

$$\hat{\mathbf{M}} = \arg \min_{\mathbf{M}} \left( \frac{1}{2} \sum_{(i,t) \in \Omega} (Y_{it} - M_{it})^2 + \lambda \|\mathbf{M}\|_* \right),$$

where  $\lambda > 0$  is a regularisation parameter that controls the complexity, or effective rank, of the solution. (Here  $\lambda$  is a tuning parameter. Its scale depends on how  $Y$  is standardised.) Large  $\lambda$  encourages low-rank solutions by shrinking singular values more aggressively. Small  $\lambda$  allows more complex structure and prioritises fitting the observed entries.

### Soft-Impute Algorithm

We solve the nuclear-norm optimisation using an iterative singular value thresholding (SVT) algorithm often referred to as Soft-Impute. Define the projection operator  $P_\Omega(\mathbf{Y})$ , which keeps the observed entries of  $\mathbf{Y}$  and sets unobserved entries to zero, and  $P_{\Omega^\perp}(\mathbf{M})$ , which keeps unobserved entries of  $\mathbf{M}$  and sets observed entries to zero. Box 8.3 describes the procedure.

#### Box 8.2: Soft-Impute Algorithm

**1. Initialise:** Set  $\mathbf{M}^{(0)} = \mathbf{0}$  or warm-start from a solution at a nearby value of  $\lambda$ .

**2. Fill Missing Entries:** Construct

$$\mathbf{Z}^{(k)} = P_\Omega(\mathbf{Y}) + P_{\Omega^\perp}(\mathbf{M}^{(k)}),$$

replacing observed entries with data and missing entries with the current imputations.

**3. Apply SVT:** Compute the singular value decomposition  $\mathbf{Z}^{(k)} = \mathbf{U}\mathbf{S}\mathbf{V}'$  and soft-threshold the singular values:

$$\mathbf{M}^{(k+1)} = \mathbf{U} \operatorname{diag}((\sigma_j - \lambda)_+) \mathbf{V}',$$

where  $(x)_+ = \max(x, 0)$  denotes the positive-part operator.

**4. Check Convergence:** Iterate steps 2–3 until

$$\|\mathbf{M}^{(k+1)} - \mathbf{M}^{(k)}\|_F / \max\{\|\mathbf{M}^{(k)}\|_F, 10^{-12}\} < 10^{-5}.$$

In practice, convergence usually occurs within a few dozen iterations.

The algorithm is efficient when the solution is low rank because you can compute a truncated SVD, since large singular values below  $\lambda$  are zeroed out anyway. Warm-starting from a solution at a nearby  $\lambda$  speeds up computation when you search over a grid of regularisation parameters.

## Regularisation and Bias–Variance Trade-Off

The regularisation parameter  $\lambda$  governs a familiar bias–variance trade-off. Small values of  $\lambda$  prioritise fitting the observed entries and yield low bias but high variance. The imputed matrix can overfit idiosyncratic noise. Large values of  $\lambda$  enforce stronger shrinkage toward a low-rank structure. This reduces variance but introduces bias by oversmoothing genuine patterns, especially local or unit-specific shocks. From a causal perspective, large  $\lambda$  can oversmooth trajectories and attenuate estimated effects in many settings (though the direction and magnitude depend on the design and where the treated cells lie), while too little regularisation (small  $\lambda$ ) inflates variance and produces highly unstable cell-level estimates of  $\hat{\tau}_{it}$ .

You can choose  $\lambda$  using cross-validation on the observed cells. Partition  $\Omega$  into a training set and a validation set, for example an 80–20 split. For each candidate  $\lambda$  on a grid such as  $\lambda \in \{0.1, 0.5, 1, 2, 5, 10\} \times \sigma_1$ , where  $\sigma_1$  is the leading singular value of the demeaned matrix with missing entries filled with zeros via  $P_\Omega(\mathbf{Y})$ , estimate  $\hat{\mathbf{M}}(\lambda)$  on the training cells and compute prediction error on the validation cells. Select the  $\lambda$  that minimises validation error.

## Connection to IFE

Matrix completion with nuclear-norm regularisation is closely related to IFE estimation in Section 8.2. Both approaches exploit low-rank structure to impute counterfactuals. The differences are mainly in how rank is controlled and how singular values are treated.

IFE requires you to specify the rank  $R$  explicitly. Estimation then keeps only the first  $R$  singular values and discards the rest, which corresponds to hard thresholding. Matrix completion instead uses the penalty  $\lambda$  to determine the effective rank. The SVT step shrinks all singular values by  $\lambda$  and sets those below  $\lambda$  to zero, which is soft thresholding. Both methods benefit from two-way demeaning to remove additive fixed effects before estimation so that the low-rank component captures interactive variation rather than simple level differences.

IFE is natural when the rank is well determined and you care about interpreting the loadings, for example when you want to relate factor exposure to observable unit characteristics. Matrix completion is more attractive when the rank is uncertain or when the missingness pattern is complex and you prefer a penalised formulation that trades fit against complexity smoothly.

## Recovery Conditions

The quality of matrix completion depends critically on the missingness pattern. Candès and Recht [2009] show that if a rank- $R$  matrix has sufficiently incoherent singular vectors and the observed cells are sampled uniformly at random, then exact recovery is possible with high probability provided the number of observed cells satisfies

$$|\Omega| \geq C R(N + T) \log^2(N + T),$$

for a constant  $C$ . This condition implies that, on average, you need on the order of  $R$  observations per row and per column, up to logarithmic factors.

Under random missingness, where each cell is observed independently with probability  $p$ , recovery is feasible when  $p \gtrsim C R \log(NT) / \min(N, T)$ . For illustration, consider a rank-5 matrix with  $N = T = 100$ . Plugging these values into the bound suggests that you need to observe on the order of half of all entries, depending on the constant  $C$ , to have good chances of recovery. This back-of-the-envelope calculation is illustrative rather than a strict design rule.

In causal inference, missingness is not random but driven by treatment assignment. Treated cells are missing by design. Matrix recovery conditions describe when imputation is feasible. Causal interpretation still requires the identification assumptions in Section 8.4. Recovery then depends on whether the observed cells span both the row space (loadings) and the column space (factors). You need control units whose loadings span the same subspace as the treated units, and pre-treatment periods whose factors resemble those in the post-treatment period, so that the low-rank structure estimated from observed cells can plausibly extend to the treated region. In practice, this means that designs which treat entire blocks of the matrix without overlap—for example, all large urban stores after a certain date, with no comparable controls remaining—violate the recovery conditions. The treated block then lies outside the span of the observed control block, and matrix completion can only extrapolate rather than interpolate. By contrast, designs with staggered adoption and rich pre-treatment overlap across all relevant store types are much closer to the idealised random-observation setting studied by Candès and Recht [2009].

## Implications for Causal Inference

For causal panel analysis, the quality of counterfactual imputation depends on two main conditions. First, the treated units' loadings should lie in, or at least close to, the span of the control units' loadings. This is analogous to the support and convex hull conditions in synthetic control (Chapter 6), although matrix completion does not impose convexity of weights. Second, the post-treatment factors should be comparable to the pre-treatment factors. Large structural breaks in the factor structure at the treatment date undermine the premise that pre-treatment data inform post-treatment imputation.

Matrix completion is more flexible than synthetic control because it does not require the treated unit's loading to be a convex combination of control loadings, but in exchange it relies more heavily on the low-rank assumption and on the richness of the observed pattern.

## Treatment Effect Estimation

Once you have imputed the missing entries and obtained  $\hat{\mathbf{M}}$ , you can compute treatment effects at the cell level. For a treated cell  $(i, t) \notin \Omega$ , define

$$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0),$$

where  $\hat{Y}_{it}(0)$  denotes the estimated untreated potential outcome. If the estimation was carried out on demeaned data, we write

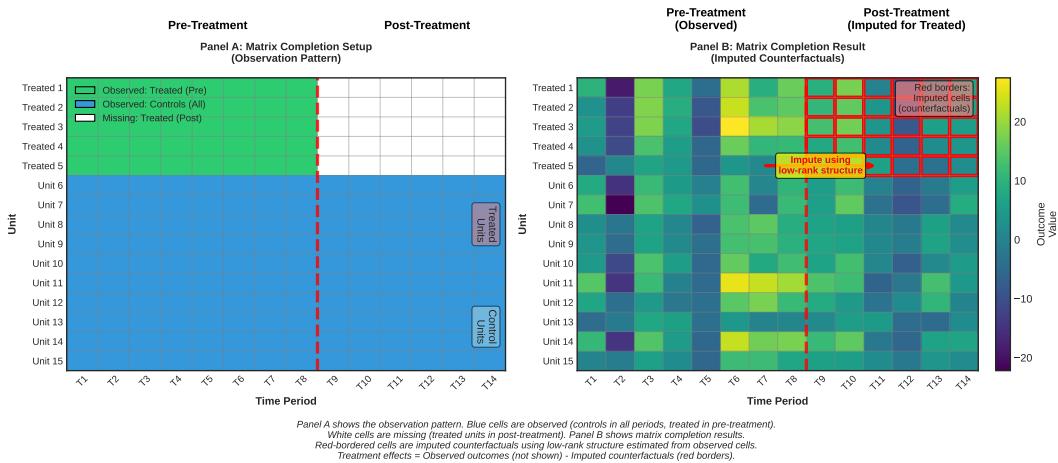
$$\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\lambda}_t + \hat{M}_{it},$$

where  $\hat{M}_{it}$  is the imputed demeaned low-rank component and  $\hat{\alpha}_i$  and  $\hat{\lambda}_t$  are the recovered unit and time fixed effects from the demeaning step.

You obtain an average treatment effect on the treated (ATT) by averaging  $\hat{\tau}_{it}$  over the treated cells. You can construct event-time and cohort-specific effects by aggregating over subsets of treated cells defined by time since treatment and treatment cohort, as in Chapter 7.

## Practical Implementation

Practical implementation involves several choices. You start by constructing the outcome matrix with units as rows and periods as columns, flagging treated cells, and applying two-way demeaning to remove additive unit and time effects. You then apply the Soft-Impute or related SVT algorithm on the demeaned matrix, using only untreated cells as observed entries. You choose  $\lambda$  via cross-validation as described above, and once the algorithm converges you add back the row and column means to recover the imputed untreated outcomes in levels. You typically search over  $\lambda/\sigma_1$  in a modest interval (for example 0.05–0.5) and then adapt based on validation error.



**Fig. 8.1** Matrix Completion Setup for Causal Panel Analysis

*Note:* The  $N \times T$  outcome matrix is partially observed. Shaded cells (control units across all periods; treated units in pre-treatment periods) are observed. Blank cells (treated units in post-treatment periods) are imputed using the low-rank structure estimated from observed cells. The nuclear-norm penalty encourages low-rank solutions that generalise from observed to missing cells.

## 8.4 Identification, Assumptions, and Design Implications

Causal identification in factor models and matrix completion requires assumptions about (i) the structure and stability of untreated potential outcomes and (ii) how treatment assignment determines which cells are observed as untreated. In this section we articulate these assumptions using formal Assumption environments, clarify their empirical implications, and discuss when factor designs are preferable to alternative approaches.

### Formal Identification Assumptions

We state the identification assumptions using the potential outcomes framework introduced in Chapter 2.

**Assumption 24 (Low-Rank Structure)** After removing additive unit and time effects (for example, via two-way demeaning on untreated cells), the remaining signal in the matrix of untreated potential outcomes admits a low-rank representation:

$$\mathbb{E}[\tilde{\mathbf{Y}}(0) | \mathbf{F}, \Lambda] = \Lambda \mathbf{F}',$$

where  $\text{rank}(\Lambda \mathbf{F}') = R \ll \min(N, T)$ . Here  $\mathbf{Y}(0)$  is the  $N \times T$  matrix of untreated outcomes with units as rows and periods as columns,  $\tilde{\mathbf{Y}}(0)$  is the corresponding two-way-demeaned matrix constructed over untreated cells,  $\Lambda$  is the  $N \times R$  matrix of loadings, and  $\mathbf{F}$  is the  $T \times R$  matrix of factors. The idiosyncratic errors  $\mathbf{E} = \tilde{\mathbf{Y}}(0) - \Lambda \mathbf{F}'$  satisfy  $\mathbb{E}[E_{it} | \Lambda, \mathbf{F}] = 0$  and weak dependence conditions.

**Interpretation.** The low-rank assumption asserts that the  $N \times T$  matrix of untreated outcomes, after removing idiosyncratic errors, has effective rank  $R$ . Outcomes are driven by  $R$  common factors rather than  $N \times T$  unrelated shocks. In marketing panels, outcomes often display strong common patterns such as seasonality, macroeconomic shocks, and category trends, which makes a low-rank structure plausible. The condition  $R \ll \min(N, T)$  ensures that the factor representation is genuinely low-dimensional relative to the panel. Elementwise, this representation corresponds to  $Y_{it}(0) = \lambda'_i f_t + E_{it}$ , where the  $i$ -th row of  $\Lambda$  is  $\lambda'_i$  and the  $t$ -th row of  $\mathbf{F}$  is  $f'_t$ . In practice, diagnostics in Section 8.2 typically point to a small number of factors. Values of  $R$  between 3 and 10 are a sensible starting range in many marketing panels, subject to panel frequency, seasonality, and diagnostics in Section 8.2.

**Empirical implication.** Scree plots of the eigenvalues of the outcome matrix, together with information criteria such as IC2 and eigenvalue ratio diagnostics, help assess whether a small number of factors explains most of the variance and whether the residual spectrum decays quickly. A sharp elbow in the scree plot and a small value of IC2 at moderate  $R$  both support the view that a handful of strong factors, rather than many weak ones, drive most of the common variation.

**Assumption 25 (Strong Factors)** The factors and loadings are pervasive. As  $N, T \rightarrow \infty$ :

$$\frac{1}{T} \mathbf{F}' \mathbf{F} \xrightarrow{p} \Sigma_F > 0, \quad \frac{1}{N} \Lambda' \Lambda \xrightarrow{p} \Sigma_\Lambda > 0,$$

where  $\Sigma_F$  and  $\Sigma_A$  are positive definite matrices.

**Interpretation.** Pervasiveness means that each factor affects many units with non-trivial intensity and that each unit is influenced by several factors. Factor-driven variation grows with sample size, while idiosyncratic shocks tend to average out. Strong factors generate a clear separation between a few large signal eigenvalues and a bulk of smaller noise eigenvalues. Weak factors, which affect only a few units or have small intensity, are harder to estimate consistently and often require regularisation, as discussed in Section 8.3. In applications where the eigenvalue spectrum decays slowly and no clear gap emerges, it is safer to treat the factor structure as weak and to favour simpler designs such as parallel trends or SC over heavily parameterised factor models.

**Empirical implication.** In finite samples, strong factors manifest as a handful of large eigenvalues that stand apart from the rest of the spectrum, together with loading patterns that are not concentrated on a tiny subset of units. Examining the eigenvalue spectrum and the distribution of estimated loadings helps assess whether factors are pervasive.

**Assumption 26 (Pre-Treatment Validity)** For any treated unit  $i$ :

1. **Invertibility:** The pre-treatment factor matrix  $\mathbf{F}_{\text{pre}}$  has full rank  $R$ , so that the treated unit's loading  $\lambda_i$  is identified from its pre-treatment outcomes.
2. **No anticipation:** Outcomes in the pre-treatment period are not affected by future treatment. For all  $t \leq T_{0i}$ ,  $Y_{it} = Y_{it}(0)$ , as defined in Chapter 2.

**Interpretation.** Invertibility ensures that the pre-treatment data contain enough variation in the factors to identify each treated unit's loading vector. This requires not just that  $T_0 \geq R$  but that the pre-treatment factors exhibit enough variation so that the smallest eigenvalue of  $T_0^{-1}\mathbf{F}'_{\text{pre}}\mathbf{F}_{\text{pre}}$  is bounded away from zero. Practically,  $T_0$  should be comfortably larger than  $R$ .  $T_0 \approx R$  is typically unstable. The no-anticipation condition rules out pre-treatment behaviour that is distorted by knowledge of upcoming treatment. Together, these conditions ensure that the estimated loadings for treated units reflect the true untreated factor structure.

**Empirical implication.** You can check that the pre-treatment period is long relative to the chosen rank and that factor estimates and loadings are stable when you vary the pre-treatment sample. A typical diagnostic is to estimate the model using an early subset of pre-treatment periods and then re-estimate including later pre-treatment periods. If estimated loadings for treated units change markedly when you drop the last few pre-treatment periods, or when you re-estimate using only an early subwindow, then either invertibility fails or anticipation effects are present.

**Assumption 27 (Design Support for Treated Units)** The treated units' loadings lie in (or close to) the span of the control units' loadings. Formally, for each treated unit  $i$  there exists a vector  $a_i$  such that

$$\lambda_i \approx \sum_{j \in \mathcal{C}} a_{ij} \lambda_j,$$

where  $\mathcal{C}$  indexes control units and the approximation error is small relative to the scale of factor variation.

**Interpretation.** This assumption ensures that the information contained in control units' outcomes is rich enough to reconstruct the factor exposure of treated units. It is the factor-analogue of support and convex-hull conditions in synthetic control designs: if treated units have loading patterns that are completely outside the span of controls, then neither IFE nor matrix completion can recover their counterfactuals reliably. Unlike synthetic control, we do not require the weights  $a_{ij}$  to be non-negative or to sum to one; only the span of the control loadings must contain the treated loadings. A practical way to interpret “close” is that the implied pre-treatment prediction error is small relative to the scale of pre-treatment variation.

**Empirical implication.** Good pre-treatment fit for treated units, small reconstruction error when you leave units out of the estimation sample, and robustness of estimated effects to dropping individual controls all support Assumption 27. Because loadings are latent, we can never verify Assumption 27 directly. The best we can do is to treat good pre-treatment fit and robust leave-one-out performance as indirect evidence that treated units are not extreme outliers in loading space. Large and systematic pre-treatment misfit for some treated units indicates that their loadings may not be well represented by the control pool.

**Assumption 28 (SUTVA or Explicit Interference Modelling)** Either:

1. **SUTVA:** The treatment applied to treated units does not affect the untreated potential outcomes of control units, as defined in Chapter 2; or
2. **Explicit modelling:** Interference is modelled through exposure mappings as in Chapter 11, so that  $Y_{it}$  depends on other units' treatment assignments through known exposure functions.

**Interpretation.** Under SUTVA, control units provide clean information about untreated outcomes because their potential outcomes do not change when other units are treated. If treated units spill over to controls through competitive effects, customer migration, or supply chain linkages, the factor estimates become contaminated and bias the counterfactuals. If substantial spillovers are present but ignored, the outcomes of nominally untreated controls embed part of the treatment effect. Factor estimates then absorb some of the treatment response, so that the imputed counterfactuals for treated units are biased toward the observed treated outcomes and estimated effects are attenuated. Factor models inherit the same exposure issues as synthetic control and difference-in-differences; low-rank structure does not by itself cure interference.

Low-rank structure improves prediction but does not immunise the design to spillovers.

**Empirical implication.** Comparing “near” controls that are plausibly exposed to spillovers with “far” controls that are unlikely to be affected can reveal interference. If near controls exhibit different post-treatment dynamics than far controls, especially when treatment intensity is high nearby, you should treat SUTVA as suspect and consider the explicit spillover designs in Chapter 11.

**Assumption 29 (Factor Stability)** The factor structure for untreated potential outcomes is stable across time. There exist common loadings and factors such that

$$Y_{it}(0) = \lambda'_i f_t + \varepsilon_{it} \quad \text{for all } i, t,$$

with the same rank  $R$  and the same factor space before and after treatment and across treated and control units.

**Interpretation.** Factor stability rules out structural breaks in the factor system for untreated potential outcomes that would affect treated units differently from controls. If a macro shock, regulatory change, or competitive disruption alters the way some units load on the factors, or introduces new factors that matter only post-treatment, then extrapolating pre-treatment factor structure into the post-treatment period is invalid and factor-imputed counterfactuals are biased.

**Empirical implication.** Splitting the pre-treatment period into early and late subperiods and estimating factors separately provides a basic stability check. The factor-stability checks in Section 8.2—such as re-estimating factors on early and late pre-treatment windows and examining control RMSPE over time—are designed precisely to probe Assumption 29. If the factor trajectories and fitted values for controls look similar across subperiods, stability is more plausible. Examining post-treatment residuals for controls, and tracking how diagnostics change when you vary the rank or estimation window, provides additional evidence.

## Staggered Adoption

When treatment timing varies by unit (each unit  $i$  has its own treatment start  $T_{0i}$ ), the assumptions adapt as follows.

The low-rank structure and strong-factor conditions remain unchanged: the untreated potential-outcome matrix obeys the same factor representation regardless of treatment timing. Pre-treatment validity becomes unit-specific: each unit's loading is identified from its own pre-treatment data, and units with longer pre-treatment histories provide sharper estimates. The no-anticipation condition applies relative to each unit's own treatment date, not a common  $T_0$ .

SUTVA is more delicate in staggered designs. Earlier-treated units may affect later-treated or never-treated units through spillovers, contaminating the control pool for later cohorts. Design choices should reflect this risk, for example by restricting controls to units and periods where interference is unlikely (see Section 8.2 and Chapter 11). Factor stability now requires that the factors estimated from the pooled untreated cells

$$\Omega = \{(i, t) : t \leq T_{0i} \text{ or unit } i \text{ is never treated}\}$$

continue to govern untreated potential outcomes in the missing region. This is a stronger requirement than in simple single-adoption designs, because the untreated region can be sparse and irregular. In particular, late cohorts contribute few pre-treatment periods, so their loadings are weakly identified unless the common factors are very strong and well estimated from earlier cohorts and controls. Designs in which large cohorts are treated shortly after entering the panel therefore pose serious challenges for factor-based identification. A simple diagnostic is to compute pre-treatment reconstruction error separately by cohort. Cohorts with

high pre-treatment error are those for which the staggered design provides weak support for factor-based counterfactuals.

## When Factor Models Dominate

Factor models are often preferable to parallel trends (DiD) and synthetic control (SC) under specific conditions that reflect the assumptions above.

**Strong common shocks with heterogeneous exposure.** When outcomes co-move tightly due to a few macro, seasonal, or category factors, but units differ in their sensitivity to these shocks, the parallel-trends assumption that all units respond identically to time shocks is implausible. Factor models accommodate heterogeneous exposure by estimating loadings freely, and they provide accurate counterfactuals so long as the design support and stability conditions hold. When Assumptions 24, 25, and 29 are well supported by diagnostics, factor models exploit this structure more efficiently than methods that impose homogeneous trends.

**Treated units outside the convex hull.** Synthetic control relies on the treated unit's loading being a convex combination of control loadings. When treated units are structural outliers—larger, more urban, or otherwise different from all controls—good SC matches do not exist. Factor models estimate loadings without convexity constraints, so treated units can lie outside the convex hull of controls as long as Assumption 27 holds.

**Few controls but long pre-treatment.** SC requires a donor pool of never-treated units. Factor models can leverage long pre-treatment periods for treated units to estimate loadings, which reduces reliance on large control pools. In panels with many treated units, few controls, and long pre-treatment histories, IFE and matrix completion often deliver more precise counterfactuals than SC. This advantage disappears if the factor structure shifts at the treatment date or if the post-treatment shocks affecting treated units differ systematically from those affecting the few controls, violating Assumption 29.

**When alternatives dominate.** Parallel trends or SC dominate when low-rank structure is weak and outcomes are driven by many idiosyncratic shocks. In such settings, factor models risk overfitting noise, while simpler models that rely on homogeneous trends or convex combinations of controls can be more robust. Empirically, such settings show up as slowly decaying eigenvalue spectra, poor pre-treatment fit for treated units even at high ranks, and factor trajectories with no interpretable macro or seasonal pattern. In these cases, the low-rank assumption behind IFE and matrix completion is simply not credible. The choice between factor-based methods and alternatives should be guided by diagnostics: pre-treatment reconstruction error, factor stability, the behaviour of eigenvalues, and the sensitivity of estimates to the chosen rank.

**Table 8.1** Assumptions, Empirical Implications, and Diagnostics

Assumption	Empirical Implication	Diagnostic
Low-Rank Structure	A few leading eigenvalues capture most variance	Scree plot; IC2; eigenvalue ratio rules
Strong Factors	Clear separation between signal and noise eigenvalues	Eigenvalue spectrum; loading patterns not concentrated on a few units
Pre-Treatment Validity	Loadings stable across pre-treatment subperiods	Early vs late pre-treatment estimation; placebo pre-trends
Design Support	Good pre-treatment fit for treated units	Out-of-sample pre-treatment reconstruction; unit- and cohort-level leave-one-out diagnostics
SUTVA / Interference	Control dynamics are unrelated to proximity to treated units once you condition on observed characteristics	Comparisons of near vs far controls; spillover placebos
Factor Stability	Control post-treatment residuals small and structure stable	Control RMSPE; factor trajectory plots; sensitivity to rank and window

## 8.5 Connections to SC and SDID

Factor models, synthetic control (SC) and synthetic difference-in-differences (SDID) are closely related approaches for constructing counterfactual outcomes  $\hat{Y}_{it}(0)$  in panels. This section maps those connections. It shows how SC and SDID can be interpreted as constrained or approximate factor models, and how factor-based imputation methods link back to the hybrid estimators in Chapter 7. Algorithmic details for SC and SDID themselves remain in Chapter 6 and Chapter 7. Here the goal is conceptual. Understanding these connections helps you choose between methods in applications and interpret differences in their estimates as consequences of their underlying modelling and design assumptions rather than as arbitrary discrepancies.

### Synthetic Control as Constrained Factor Model

Recall the basic synthetic control set-up from Chapter 6. We choose weights  $w_j$  on donor units such that the weighted donor outcomes match the treated unit's pre-treatment trajectory,  $\sum_{j \in \mathcal{J}} w_j Y_{jt} \approx Y_{1t}$  for  $t \leq T_0$ . For simplicity we focus on a single treated unit; the notation extends straightforwardly to treated groups. We then use  $\hat{Y}_{1t}(0) = \sum_{j \in \mathcal{J}} w_j Y_{jt}$  as the counterfactual after treatment.

In this discussion we assume donor units are untreated, so  $Y_{jt} = Y_{jt}(0)$  for  $j \in \mathcal{J}$ .

Now place this inside a factor model. Suppose untreated outcomes follow

$$Y_{it}(0) = \lambda_i' f_t + \varepsilon_{it},$$

with unit-specific loadings  $\lambda_i$  and common factors  $f_t$ . The SC prediction for the treated unit can be written as

$$\sum_{j \in \mathcal{J}} w_j Y_{jt}(0) = \sum_{j \in \mathcal{J}} w_j (\lambda_j' f_t + \varepsilon_{jt}) \approx \left( \sum_{j \in \mathcal{J}} w_j \lambda_j \right)' f_t.$$

To approximate the treated unit's untreated outcome  $Y_{1t}(0) = \lambda_1' f_t + \varepsilon_{1t}$  across pre-treatment periods, we therefore need

$$\lambda_1 \approx \sum_{j \in \mathcal{J}} w_j \lambda_j, \quad w_j \geq 0, \quad \sum_j w_j = 1.$$

Geometrically, SC behaves as if it were constraining the treated unit's loading to lie inside the convex hull of donor loadings. This is the synthetic-control analogue of Assumption 27: the treated unit's loading must lie in the convex hull of donor loadings, not just in their linear span. It interpolates among donors rather than extrapolating beyond them. When  $\lambda_1$  lies far outside this convex hull, no choice of  $w_j$  can recover  $Y_{1t}(0)$  even under a correctly specified factor model, so SC is fundamentally misspecified in loading space. An unconstrained factor model, by contrast, would estimate  $\lambda_1$  freely in  $\mathbb{R}^R$ , allowing extrapolation.

Two main contrasts follow. First, convexity: SC restricts the treated unit to a convex combination of donors; factor models do not. Second, estimation: SC never estimates factors explicitly, but works directly with observed outcomes. Factor models estimate both  $A$  and  $\mathbf{F}$  and then form counterfactuals from these estimates.

SC's convexity constraint acts as a built-in regulariser; factor models often need explicit regularisation (for example, nuclear norms) to achieve similar stability.

## SDID as a Weighted Factor Perspective

Synthetic difference-in-differences (SDID, Section 7.4) extends SC by introducing time weights alongside unit weights. Let  $w_j$  denote unit weights for donors and  $v_t$  denote weights on pre-treatment periods. A convenient way to write the SDID estimand is

$$\hat{\tau}^{\text{SDID}} = \left( \bar{Y}_{\text{treated},\text{post}} - \sum_j w_j \bar{Y}_{j,\text{post}} \right) - \sum_{t \leq T_0} v_t \left( \bar{Y}_{\text{treated},t} - \sum_j w_j \bar{Y}_{jt} \right),$$

where  $\bar{Y}_{\text{treated},\text{post}}$  is the average outcome for treated units over post-treatment periods,  $\bar{Y}_{j,\text{post}}$  is the corresponding average for donor  $j$ , and  $\bar{Y}_{\text{treated},t}$  is the cross-sectional mean among treated units at time  $t$ .

Here "post" denotes the set of post-treatment periods, and  $\bar{Y}_{j,\text{post}}$  averages  $Y_{jt}$  over those periods.

From a factor perspective, SDID is easiest to visualise in a simple model with unit and time intercepts and one dominant factor,

$$Y_{it}(0) = \alpha_i + \lambda_t + \lambda_i f_t + \varepsilon_{it}.$$

Ordinary DiD removes the common time effect  $\lambda_t$  but leaves heterogeneous loadings  $\lambda_i f_t$  in place. SDID retains the DiD differencing but chooses unit and time weights to make these heterogeneous components align as well as possible across treated and donor units. The time weights  $v_t$  tilt attention towards pre-treatment periods where the factor structure relevant for post-treatment comparisons is most visible.

In that sense, SDID can be interpreted as emphasising a low-dimensional, factor-like component of the data without explicitly estimating factors. This is an interpretive device rather than a literal factor estimator: SDID does not recover  $f_t$  or  $\lambda_i$  directly, but its weighting scheme tends to emphasise directions in the data that look like low-dimensional common shocks. Causal interpretation still relies on the standard design assumptions (no anticipation, no interference, and a suitable comparability condition), with weighting acting to make those assumptions more plausible in the reweighted sample.

## Matrix Completion in the Same Spectrum

Matrix completion (Section 8.3) offers another vantage point. We can view the outcome panel as a matrix with missing entries for treated cells and seek to impute those entries under a low-rank structure.

Interactive fixed-effects (IFE) models correspond to imposing a hard rank constraint: we approximate the outcome matrix by a rank- $R$  matrix and choose  $R$  based on diagnostics. Nuclear-norm-penalised matrix completion replaces this hard rank constraint with a soft one, shrinking singular values towards zero and letting the data determine an effective rank through the penalty parameter.

SC fits into this picture as a more constrained imputation scheme: instead of approximating the entire matrix, it constructs the treated unit's counterfactual as a convex combination of donor rows, effectively restricting attention to a particular part of the low-rank structure spanned by donors. Viewed heuristically, SC avoids extrapolation beyond the donor convex hull, while IFE and nuclear-norm completion allow extrapolation but require explicit rank or penalty choices. This extrapolation is justified only under the low-rank structure, factor stability, and design-support conditions from Section 8.4. Without them, rank constraints and nuclear norms can formalise the imputation problem but do not by themselves guarantee causal identification. In designs where treated units lie near the boundary of the donor hull and low-rank structure is weak, this extrapolation risk can dominate any variance gains from the more flexible models.

## Hybrid Methods as Intermediate Approaches

The hybrid methods in Chapter 7 deliberately sit between pure weighting and pure factor-based imputation.

Augmented synthetic control (ASCM) starts from SC's weights and adds an outcome regression that corrects residual differences not captured by the convex-hull approximation. From the factor-model standpoint, the augmentation term can be viewed as approximating the part of the treated unit's loading that lies just outside the donor hull, under the assumption that this residual is predictable from observed covariates.

Ridge SC modifies SC's convexity constraint by shrinking weights towards more diffuse configurations. This smoothing across donors can be seen as a way of approximating a lower-rank structure in the donor space without explicitly estimating factors. It trades a small amount of bias for a larger gain in stability, particularly when the donor pool is large and the pre-period is short.

SDID, as already discussed, augments SC with time weights in a way that mimics aspects of a factor adjustment while retaining the interpretability of unit weights anchored in observed donors.

These hybrids therefore form intermediate points along a spectrum from pure design-based weighting (SC) to fully model-based imputation (IFE, nuclear-norm completion). Where you should stand on that spectrum depends on how plausible the convex-hull restriction is, how strong the factor structure appears to be, and how much you value transparency over flexibility.

## Comparison Table

### When to Prefer Each Approach

The choice among these approaches ultimately comes back to data structure and identification.

SC is most attractive when the treated unit looks like a typical member of the donor pool, when the pre-treatment period is long enough to diagnose fit, and when you value a transparent design that never extrapolates beyond observed donors. In practice, SC is most compelling when pre-treatment RMSPE for the

**Table 8.2** Comparison of SC, SDID, IFE, and Matrix Completion

Method	Convexity	Rank Structure	Time Weights	Key Tuning
SC (Ch 6)	Convex weights on donors	Implicit: convex hull of donor loadings	No	Predictor set; choice of pre-period
SDID (Ch 7)	Convex weights on donors	Emphasises low-dimensional trends	Yes ( $v_t$ )	Implementation details; pre-period window
IFE (Sec 8.2)	None	Fixed rank $R$	No	Rank $R$ ; regularisation if used
Matrix Completion (Sec 8.3)	None	Soft rank via nuclear norm	No	Penalty parameter on nuclear norm

treated unit is small relative to typical control-control gaps, SC weights are reasonably diffuse rather than concentrating on a single donor, and placebo tests in Chapter 6 show no systematic pre-treatment gaps.

SDID is appealing when you suspect that unweighted parallel trends fails but that, after appropriate reweighting of units and time periods, treated and donor groups can be made comparable. It lets you adjust for differential trends without committing to a specific factor rank.

IFE and nuclear-norm matrix completion are natural when exploratory work suggests a strong, low-rank factor structure that explains much of the variation in untreated outcomes, and when you are willing to pay the price of greater model dependence in exchange for a flexible imputation of counterfactuals. For IFE and matrix completion, you should see a clear eigenvalue gap, good pre-treatment reconstruction for treated units at moderate ranks, and stable factors across pre-treatment subsamples.

Hybrids such as ASCM and ridge SC are useful when you are near the boundary between these regimes: the treated unit is close to, but not comfortably inside, the donor hull, or the donor pool is large enough that unconstrained SC weights look unstable. In those cases, adding augmentation or regularisation can correct modest departures from ideal SC conditions without giving up entirely on the design-based intuition that makes SC attractive in the first place. Empirically, this shows up as good but imperfect pre-treatment fit for SC, with noticeable but not catastrophic gaps and highly concentrated SC weights. In that regime, ASCM and ridge SC often reduce bias and variance relative to both pure SC and unconstrained factor-based methods.

Once counterfactuals are constructed—whether by SC, SDID, IFE or matrix completion—treatment effects are aggregated using the group-time and event-time machinery from Chapters 4 and 5. You compute gaps  $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$  at the unit-time level, then average within cohorts and across time according to the estimand (for example, ATT,  $\tau(g, t)$ , or  $\theta_k$ ). The factor perspective helps you understand where those counterfactuals come from. The DiD and event-study chapters tell you how best to summarise them.

## 8.6 Tuning and Implementation

Implementing factor models and matrix-completion methods in marketing panels means making a series of design choices: how long a pre-treatment window to use, which controls to include, how many factors to retain or how strong a penalty to apply, and what pre-processing to perform. These decisions affect bias, variance and interpretability. This section offers practical guidance on those choices and points back to earlier sections where the underlying theory is developed.

### Pre-Treatment Period Length

The length of the pre-treatment period is a first-order determinant of how well factor structures can be estimated. Longer pre-periods give you more information about common shocks and unit-specific responses, making estimated loadings and factors more stable.

A useful starting point is to have several pre-treatment observations per factor you hope to identify. As a rough guide, if you expect only a handful of dominant factors (say, rank  $R$  between 3 and 10), you often want at least  $5R-10R$  pre-treatment periods as a starting point for stable estimation. From an identification standpoint (Assumption 26), the pre-treatment period must be long enough that the pre-treatment factor matrix has full rank and well-separated eigenvalues. In practice this means not just  $T_0 \geq R$ , but  $T_0$  considerably larger than  $R$ , so that treated units' loadings are identified with reasonable precision. With weekly data, two to three years of history often provide enough variation to identify seasonal and business-cycle components; with monthly data, several years of pre-period are usually needed to see enough holiday and macro cycles.

When the pre-period is very short relative to the candidate rank (for example,  $T_0 \leq 3R$ ), unconstrained factor models become weakly identified and numerically unstable. In such designs it is usually safer to impose a much smaller rank, to rely on stronger penalties, or to fall back on two-way fixed effects or SC rather than forcing a high-rank factor structure onto weak data. These rules of thumb assume that the underlying common shocks are reasonably strong and that the donor pool is large enough to support Assumption 27. When either condition is doubtful, you should treat even long pre-periods as only partial evidence for a reliable factor structure.

### Donor Diversity and Curation

Factor-based methods draw strength from diversity in the control sample. If donors span a wide range of characteristics and outcome patterns, it becomes easier to express treated units' factor loadings as combinations of donor loadings. If donors are all very similar to each other and very different from the treated units, extrapolation risk rises.

The same donor-curation principles from Chapter 6 and Chapter 7 apply here. First, exclude treated cells from estimation; for units that are treated later, retain their pre-treatment periods. Second, exclude units

that are plausibly affected by spillovers from treated units, for example by defining geographic or competitive buffer zones around treated markets (see Chapter 11). Third, exclude units that are clearly incomparable in business terms, such as mixing big-box flagships with tiny convenience stores in the same donor pool without a strong substantive rationale.

Simple plots of donor characteristics against baseline outcomes or pre-trends help assess whether treated units lie inside the “cloud” of donors or far outside it. The closer treated units sit to donors in these plots, the more comfortable you can be relying on factor-based imputation or interactive fixed effects. These plots are suggestive diagnostics, not tests of the latent loading support condition. This visual check is an informal way to assess Assumption 27: if treated units sit far outside the donor cloud in terms of baseline levels or pre-trends, it is unlikely that their factor loadings lie in the span of donor loadings, and factor-based imputation will extrapolate aggressively.

## Seasonality Controls

Seasonality is pervasive in marketing data. You can handle it explicitly, implicitly, or both.

Explicit controls introduce seasonal indicators or smooth functions—for example, month or quarter dummies, day-of-week dummies, or sine and cosine terms at seasonal frequencies—into the outcome model. This removes additive seasonal patterns that are common across units and reduces the burden placed on the factor structure.

Implicit controls let the factor model pick up seasonal patterns through time-varying factors and heterogeneous loadings. This is attractive when the strength of seasonality varies across units—for instance, when some stores experience pronounced holiday spikes and others do not.

A pragmatic approach is to combine the two: include seasonal dummies defined mechanically (calendar-based) to remove obvious calendar effects, then estimate factors on residuals that still contain heterogeneous seasonal responses. Section 8.2 discusses how this improves interpretability and reduces the number of factors needed.

## Demeaning and Detrending

Pre-processing choices shape both how factor models behave and how you interpret them. As discussed in Section 8.2, factor estimation is usually carried out on demeaned data, with fixed effects recovered afterwards.

Unit demeaning subtracts each unit’s mean over time, so factors capture co-movement around unit-specific levels rather than levels themselves. Two-way demeaning additionally removes time means, shifting the focus to deviations from both unit and time averages. These transformations generally make it easier to interpret factors as common shocks and loadings as heterogeneous sensitivities.

Detrending goes further by removing deterministic unit-specific trends, either via explicit regressions on time or by first differencing. Detrending is appropriate when smooth trends are clearly driven by forces

unrelated to the intervention and you do not want the factor model to spend degrees of freedom fitting them. It is less appropriate when trends themselves may be part of the treatment effect or of the confounding structure. As a default, two-way demeaning is a good starting point; detrending or differencing should be motivated by clear evidence of deterministic trends that you are willing to treat as nuisance. Over-aggressive detrending can also mask violations of Assumption 29 by removing genuine shifts in the factor structure. For example, differencing can remove a level shift that is itself evidence of a structural break coinciding with treatment. As a result, diagnostics may overstate how well a low-rank model fits untreated outcomes.

## Rank Selection for IFE

Choosing how many factors to retain in an interactive fixed-effects model is a balance between fit and parsimony. Too few factors leave systematic variation in the residuals; too many pick up noise.

Information criteria tailored to panels, such as those proposed by Bai and Ng [2002], Bai [2009], provide one principled route (see Section 8.2). They evaluate models with different ranks by combining a measure of in-sample residual variance with a penalty that increases with the number of factors and the panel dimensions. Minimising such a criterion over a reasonable range of candidate ranks often yields a sensible choice.

Visual tools are also valuable. Scree plots of eigenvalues of the demeaned outcome covariance matrix reveal where marginal eigenvalues start to flatten out; elbows in these plots suggest natural cut-offs. Eigenvalue-ratio ideas formalise this intuition by looking for large drops between successive eigenvalues.

Finally, cross-validation within the pre-treatment period can guard against overfitting. Estimate IFE models with different ranks on an early part of the pre-period and compare their ability to predict held-out later pre-period outcomes. Ranks that fit the training data slightly less well but predict the validation block more accurately are usually preferable. Validation blocks should consist of untreated cells only and should respect time ordering: for example, estimate factors on early pre-treatment periods and validate on later pre-treatment periods, rather than randomly shuffling time. This mimics the actual forecasting problem of predicting post-treatment untreated outcomes.

## Penalty Selection for Matrix Completion

For nuclear-norm-penalised matrix completion (see Section 8.3 and Candès and Recht [2009], Candes [2010], Athey et al. [2021]), the penalty parameter controls how aggressively singular values are shrunk and therefore how low-rank the imputed matrix becomes. In practice, you rarely know the right penalty a priori.

A simple approach is to define a grid of penalty values relative to the leading singular value of the demeaned outcome matrix—for example, a handful of values ranging from “light” to “strong” shrinkage—and to select among them using cross-validation on observed entries. Partition the observed cells into a training set and a validation set, fit the model on the training cells for each candidate penalty and choose the penalty that

predicts validation cells best. Plots of validation error against the penalty help you see whether there is a clear optimum or a region where performance is flat.

After choosing a penalty, inspecting the singular values of the fitted matrix gives a sense of the effective rank: the number of singular values that are meaningfully above zero. If a small increase in the penalty dramatically reduces the effective rank or suddenly changes residual patterns or treatment-effect estimates, the design is highly sensitive to tuning and factor-based identification should be treated with caution, and reported as tuning sensitivity in the main results. This links the penalised estimator back to the more concrete notion of “how many factors matter here.”

## Guarding Against Overfitting

Whatever rank or penalty you choose, you should check that the fitted factor structure is not simply chasing noise. Residual diagnostics and stability checks are the main tools.

Plot residuals over time and across units. If systematic patterns remain—for example, clear autocorrelation, seasonality or structural breaks—the model is underfitting. If residuals look roughly like noise around zero, the factor structure is doing most of its job. Re-estimating the model on different subsets of the data, such as early versus late pre-periods or different subsets of donors, reveals whether estimated factors and loadings are stable. Large shifts across such splits suggest that the factor structure is weak relative to noise or that important changes in the environment are not being modelled. In addition to residual diagnostics, you should track how aggregate treatment-effect estimates (for example, ATT or event-time effects) vary across reasonable ranks, penalties, and donor subsets. When small tuning changes lead to large swings in ATT or in the shape of event-study paths, model dependence is high and the resulting causal claims are fragile, even if in-sample fit looks good.

Rolling-window estimates offer another perspective. Estimating factors on overlapping windows of pre-treatment periods and tracking how the leading factors evolve can help detect instability and structural breaks. If the first factor’s direction flips or its correlation across windows collapses, you should question whether a global low-rank model is appropriate.

## Software Considerations

Several software packages implement interactive fixed-effects and matrix-completion methods with built-in tuning support.

In R, packages such as `gsynth`, `fect` and `MCPanell` provide interfaces for IFE and related causal-panel estimators, often including routines for rank selection and cross-validation. General-purpose matrix-completion tools such as `softImpute` can be used to implement nuclear-norm penalties when you want more control over the imputation step.

In Python and other environments, numerical libraries for singular-value decomposition and nuclear-norm optimisation play a similar role. Generic matrix-completion libraries can be integrated into custom code that handles the causal design—treatment timing, donor pools, and aggregation—following the principles in this book.

Given the pace of software development, specific package recommendations will date. The key is to understand which capabilities you need—interactive fixed effects with rank selection, nuclear-norm-penalised completion, cross-validation infrastructure—and to verify that any package you use implements them in a way that respects your design.

## Summary

Tuning factor models and matrix-completion estimators is as much about judgement as about formulas. Rules of thumb on pre-period length, rank and penalties provide starting points, but they must be tempered by diagnostics, stability checks and domain knowledge. When in doubt, prioritise designs that are simple, robust across reasonable tuning choices and easy to explain to non-technical stakeholders. The goal is not to maximise in-sample fit, but to capture stable common movement that supports credible counterfactual imputation.

## 8.7 Inference

Inference for factor models and matrix completion quantifies uncertainty in counterfactuals and in treatment effect estimates such as the aggregate ATT or event-time effects  $\theta_k$ , conditional on the identifying assumptions and the chosen imputation model. We adapt the bootstrap, asymptotic approximations, placebo-in-time tests, and conformal-style prediction intervals to this setting. Throughout this section,  $\hat{Y}_{it}(0)$  denotes the imputed untreated potential outcome from the chosen factor or matrix-completion estimator, and  $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$  the corresponding cell-level effect estimate. We emphasise small-sample properties and serial dependence, and refer to Chapter 16 for a systematic treatment of these inference tools.

### Block Bootstrap

Block bootstrap methods, introduced in time-series settings by Künsch [1989] and Politis and Romano [1994], approximate the sampling distribution of estimators under serial dependence by resampling contiguous time blocks. In our context, we resample blocks of time periods, re-estimate the factor model on each bootstrap sample, and recompute treatment effect estimates such as the ATT.

**Algorithm.** Implement block bootstrap for factor-model-based ATT or event-time effects as follows.

1. **Choose block length.** Set the block length  $\ell$  to capture the main serial dependence in outcomes. A common rule is  $\ell = \lfloor 1.75 \cdot T^{1/3} \rfloor$  as a heuristic starting point. For marketing panels with quarterly data, blocks of 4 to 8 quarters (one to two years) are typical. For weekly data, blocks of roughly 26 to 52 weeks are common. You should base the block length on the dependence structure of the residuals from the factor model, not on raw outcomes. Plot autocorrelation functions (ACFs) of residuals and choose  $\ell$  large enough to cover the main decay of the ACF, then verify that conclusions about ATT and  $\theta_k$  are stable over a plausible range of  $\ell$ .
2. **Form blocks.** Partition the time periods into  $\lceil T/\ell \rceil$  non-overlapping blocks:  $\{1, \dots, \ell\}, \{\ell+1, \dots, 2\ell\}, \dots$
3. **Resample blocks.** Draw  $\lceil T/\ell \rceil$  blocks with replacement. Concatenate them to form a bootstrap sample of  $T^* \geq T$  periods, and truncate or trim to exactly  $T$  if needed.
4. **Re-estimate the factor model.** Estimate the factor model on the bootstrap sample, obtaining loadings  $\hat{A}^*$  and factors  $\hat{F}^*$ . In causal panels, treatment timing and the set of treated cells are part of the design and must remain fixed across bootstrap samples. When resampling time blocks, we keep the treatment assignment as a function of calendar time fixed (treated cells remain treated at their original dates), and we re-estimate the factor model using the untreated cells defined by that original assignment. For staggered adoption, blocks that fall entirely in a unit's treated period contribute only to recomputing  $\hat{\tau}^{*(b)}$ , not to factor estimation.
5. **Recompute treatment effects.** For each bootstrap sample, recompute the target estimand, such as the ATT or a vector of event-time effects. Denote this bootstrap estimate by  $\hat{\tau}^{*(b)}$  for replication  $b$ .
6. **Repeat.** Perform  $B$  bootstrap replications, with  $B$  typically between 500 and 1000.

7. **Construct intervals.** For a scalar target, such as ATT or a single event-time coefficient, form the 95% confidence interval as the 2.5th to 97.5th percentiles of the bootstrap distribution  $\{\hat{\tau}^{*(b)}\}_{b=1}^B$ .

Here  $\hat{Y}_{it}(0)$  denotes the factor-model estimate of the untreated potential outcome  $Y_{it}(0)$ , and cell-level realised effects are  $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ . In practice you aggregate these cell-level effects within each bootstrap replication to construct  $\hat{\tau}^{*(b)}$  for the ATT or for event-time averages.

**Resampling units vs time.** The procedure above resamples time blocks, preserving cross-sectional correlation within each period. An alternative is to resample units (for example, stores) with replacement and keep all their time series intact, which preserves temporal dependence within units but breaks contemporaneous correlation across units. Choose between these schemes based on the dependence structure in the application. Resample time if serial correlation dominates after conditioning on factors. Resample units if contemporaneous correlation across stores is more important. When outcomes are clustered by geography or brand, resample at the cluster level rather than at the individual-store level so that within-cluster correlation is preserved. This mirrors the cluster-robust logic in Chapter 16.

## Wild Bootstrap

Wild bootstrap methods, starting with Wu [1986] and Mammen [1992], handle heteroskedastic and clustered errors by resampling residuals with randomly drawn signs or weights. In the factor model setting, we generate pseudo-outcomes by perturbing the estimated idiosyncratic component. In each replication we typically re-estimate the factor model to reflect factor-estimation uncertainty.

**Algorithm.** Implement wild bootstrap as follows.

1. **Estimate residuals.** Fit the factor model to obtain loadings  $\hat{\lambda}_i$  and factors  $\hat{f}_t$ , and compute residuals  $\hat{\varepsilon}_{it} = Y_{it} - \hat{\lambda}'_i \hat{f}_t$  for all untreated cells and pre-treatment periods.
2. **Generate wild weights.** Draw weights  $s_{it}$  independently. Common choices are Rademacher weights,  $s_{it} \in \{-1, +1\}$  with equal probability, and Mammen weights, where  $s_{it} = (1 - \sqrt{5})/2$  with probability  $(1 + \sqrt{5})/(2\sqrt{5})$  and  $s_{it} = (1 + \sqrt{5})/2$  otherwise. Rademacher weights are simple and often adequate. Mammen weights can improve small-sample performance under heteroskedasticity.
3. **Construct bootstrap outcomes.** Form

$$Y_{it}^* = \hat{\lambda}'_i \hat{f}_t + s_{it} \hat{\varepsilon}_{it},$$

for all  $(i, t)$  in the estimation sample.

4. **Re-estimate and recompute.** Re-estimate the factor model on  $Y^*$ , impute counterfactuals  $\hat{Y}_{it}^*(0)$  for treated cells, and recompute the target estimand  $\hat{\tau}^{*(b)}$ . Re-estimating the factor model in each bootstrap replication is computationally intensive but essential if you want intervals that reflect uncertainty

about both the idiosyncratic shocks and the estimated factor structure. Treating  $\hat{\lambda}_i$  and  $\hat{f}_t$  as fixed and resampling only residuals underestimates uncertainty, especially when  $T$  is modest.

5. **Repeat and construct intervals.** Repeat this procedure  $B$  times and form percentile intervals for the target estimand from the empirical distribution of  $\{\hat{\tau}^{*(b)}\}_{b=1}^B$ .

**Cluster wild bootstrap.** When units are clustered, such as stores nested within regions or brands, you can draw weights at the cluster level so that all observations in a cluster share the same sign. For example, draw  $s_c$  for each region  $c$  and set  $s_{it} = s_c$  whenever unit  $i$  belongs to region  $c$ . This preserves within-cluster dependence and mirrors the logic of cluster-robust standard errors. See Cameron et al. [2008] and MacKinnon and Webb [2017] for discussion.

## Asymptotic Approximations

Asymptotic theory for principal components estimators in large panels with interactive fixed effects is developed by Bai [2003]. Under standard regularity conditions and strong factors (Assumption 25), the principal components estimators of factors and loadings are consistent and asymptotically normal as  $N, T \rightarrow \infty$ .

**Convergence rate.** Let  $H$  be the non-singular rotation matrix that aligns estimated and true factors and loadings, as in Section 8.2. Under strong factors,

$$\|\hat{f}_t - H f_t\| = O_p(N^{-1/2} + T^{-1/2}),$$

$$\|\hat{\lambda}_i - H^{-1'} \lambda_i\| = O_p(N^{-1/2} + T^{-1/2}).$$

In typical marketing panels with many stores but shorter time series, the  $N^{-1/2}$  term tends to dominate, so factor estimates are relatively precise even when  $T$  is modest.

**Asymptotic variance.** For a given unit  $i$ ,

$$\sqrt{T}(\hat{\lambda}_i - H^{-1'} \lambda_i) \xrightarrow{d} N(0, \mathbf{V}_\lambda),$$

where  $\mathbf{V}_\lambda$  depends on the covariance of the idiosyncratic errors. In principle you can estimate  $\mathbf{V}_\lambda$  from the sample covariance of residuals and propagate this uncertainty through the imputation of  $\hat{Y}_{it}(0)$  using the delta method to obtain standard errors for ATT and event-time effects. For the panel sizes typical in marketing applications (for example,  $N$  between 20 and 200,  $T$  between 8 and 40), the large- $N$ , large- $T$  approximations of Bai [2003] should be treated as rough diagnostics rather than as the basis for reported intervals. Chapter 16 gives implementation details.

**When asymptotics fail.** Asymptotic approximations rely on both  $N$  and  $T$  being large and on a correctly specified factor structure. This holds only when the strong-factors condition (Assumption 25) is met. In

panels with only a few dozen stores or short pre-treatment histories, or when the factor rank is uncertain, we recommend treating analytic standard errors as rough diagnostics and basing reported intervals on the bootstrap.

## Placebo-in-Time Diagnostics

Placebo-in-time diagnostics, introduced for synthetic control by Abadie et al. [2010] and extended to generalised synthetic control by Xu [2017], assess whether the imputation procedure extrapolates well by pretending that treatment happened earlier than it did. This section adapts the idea to factor-model imputation. For an overview of placebo diagnostics in synthetic control and DID settings, see Chapter 6. You should avoid pseudo-dates so close to  $T_0$  that only a handful of post-pseudo-treatment periods remain. Otherwise the RMSPE statistics are extremely noisy and difficult to compare with the true post-treatment RMSPE.

**Procedure.** Fix a set of pseudo-intervention dates  $\tau \in \{R + 1, R + 2, \dots, T_0 - h\}$ , where  $R$  is the number of factors and  $h \geq 5$  is the minimum post-pseudo-treatment window.

1. For a given pseudo-date  $\tau$ , estimate the factor model using only periods  $t \leq \tau$ .
2. Impute counterfactuals for the subsequent periods  $t = \tau + 1, \dots, T_0$  as  $\hat{Y}_{it}(0 | \tau)$  for all units that remain untreated up to  $T_0$ .
3. Compute pseudo-treatment effects  $\tilde{\tau}_{it}(\tau) = Y_{it} - \hat{Y}_{it}(0 | \tau)$  for  $t > \tau$ , and aggregate them into a pseudo-ATT for that  $\tau$ .

**Diagnostic statistic.** For each  $\tau$ , compute the root mean squared pseudo-effect (RMSPE),

$$\text{RMSPE}(\tau) = \sqrt{\frac{1}{|\{(i, t) : t > \tau, t \leq T_0\}|} \sum_{i,t:t>\tau,t\leq T_0} \tilde{\tau}_{it}(\tau)^2}.$$

Plot the distribution of  $\text{RMSPE}(\tau)$  across pseudo-dates and compare it with the post-treatment RMSPE computed using the actual treatment date. If the post-treatment RMSPE is an outlier, for example above the 95th percentile of the placebo distribution, this supports the presence of a treatment effect. If the pseudo-effects are already large, the factor model is not extrapolating reliably and post-treatment estimates may be biased. Placebo-in-time diagnostics can encourage data snooping if you repeatedly tweak the model until placebo RMSPEs look small. When possible, reserve part of the pre-treatment period for out-of-sample validation and treat placebo diagnostics on the remaining window as diagnostic, not as formal hypothesis tests.

## Conformal Intervals

Conformal-style prediction intervals use the empirical distribution of residuals from the factor model to construct prediction sets for counterfactual outcomes. The idea is to treat the residuals from untreated cells as approximate draws from the distribution of imputation errors and then construct bands around  $\hat{Y}_{it}(0)$  that contain a new observation with high probability.

**Construction.** For each treated unit-period pair  $(i, t)$ :

1. Compute residuals  $\hat{\varepsilon}_{js}$  for all untreated cells  $(j, s) \in \Omega$ , where  $\Omega$  indexes the donor pool and pre-treatment periods.
2. Compute the  $(1 - \alpha/2)$ -quantile of the absolute residuals,  $q_{1-\alpha/2} = \text{Quantile}_{1-\alpha/2}(|\hat{\varepsilon}_{js}|)$ .
3. Construct the  $(1 - \alpha)$ -level prediction interval for the untreated potential outcome as

$$\text{PI}_{it}^{1-\alpha} = [\hat{Y}_{it}(0) - q_{1-\alpha/2}, \hat{Y}_{it}(0) + q_{1-\alpha/2}].$$

4. If the observed outcome  $Y_{it}$  lies outside  $\text{PI}_{it}^{1-\alpha}$ , this suggests the treated outcome is unusual relative to the model's residual variation for that cell. In dependent panels and with many treated cells, interpret this as a diagnostic rather than as a formal level- $\alpha$  test.

Interpret these intervals primarily as calibrated prediction bands rather than as a fully multiple-testing-correct procedure.

Under exact exchangeability of residuals across  $(j, s) \in \Omega$  and the post-treatment treated cells, such intervals achieve nominal coverage. For background on conformal prediction see Shafer and Vovk [2008] and Lei and Wasserman [2014]. In panels with strong serial and cross-sectional dependence, this exchangeability assumption is only approximately true even after conditioning on the factors, so coverage can deviate from the nominal level. Approximate exchangeability is most plausible when Assumptions 29 and 27 hold and when residual autocorrelation is weak after conditioning on factors. In panels with strong remaining dependence, conformal bands become more heuristic. These intervals target approximate marginal coverage for individual cells, not joint coverage across all treated cells. In applications with many treated unit-periods, some intervals will exclude  $Y_{it}$  by chance even in the absence of treatment effects.

## Small-Sample Considerations

Small-sample and dependence issues are central in marketing panels. Many campaigns involve a few dozen stores observed for a few years, with strong autocorrelation in sales and common shocks at the brand or region level.

**Few clusters.** When outcomes are correlated within clusters  $c = 1, \dots, G$  and clusters are the independent sampling unit, the effective sample size for inference is  $G$ , not  $N \times T$ . In such settings asymptotic approxi-

mations are fragile, even when  $N$  is large. We recommend wild bootstrap with cluster-level weights, drawing  $s_c$  at the cluster level to preserve within-cluster dependence. With very few clusters (for example,  $G < 20$ ), even cluster wild bootstrap can be unreliable because the reference distribution of  $\hat{\tau}^{*(b)}$  is too discrete and percentile intervals can be distorted. In such cases, report both cluster-wild results and simpler unit-level wild bootstrap as sensitivity checks, and interpret all intervals as approximate. In these designs it is often more honest to treat factor-based inference as exploratory and to cross-check it against simpler DiD or SC estimates.

**Serial dependence.** When outcomes exhibit strong autocorrelation, choose block lengths in the block bootstrap that are long enough to capture the dependence structure in residuals. As a robustness check, vary the block length over a plausible range and verify that conclusions about the ATT and key event-time effects are stable. Chapter 16 discusses heteroskedasticity- and autocorrelation-consistent (HAC) estimators in more detail. In factor models we rely primarily on bootstrap-based adjustments.

**Reporting.** When presenting results from factor-model-based causal analyses, report at least five elements. First, state the inference method, including the type of bootstrap, the block length (if applicable), and the number of replications. Second, report 95% confidence intervals for the ATT and for the main event-time effects rather than only point estimates. Third, summarise placebo-in-time diagnostics by showing the distribution of RMSPEs across pseudo-dates and indicating where the actual treatment date falls. Fourth, describe key sensitivity checks, such as varying the block length, the assumed factor rank, or trimming influential stores or periods. Fifth, show how the ATT and main event-time effects change across a reasonable range of factor ranks or penalties. Large swings indicate high model dependence.

## Software Implementation

Several software packages implement these inference procedures for factor-based causal estimators. In R, the `gsynth` package [Xu, 2017] offers parametric bootstrap and placebo-in-time tests for generalised synthetic control, while `fект` implements block and wild bootstrap for interactive fixed effects and related designs. More general bootstrap utilities are available in `boot`, which you can combine with custom factor-estimation code. In Python, the `arch` package provides block and wild bootstrap tools for dependent data, and you can pair these with factor estimation written in `numpy` and `scipy`.

## Summary Table

**Table 8.3** Inference Methods for Factor Models

Method	Assumption	When to Use	Software
Block Bootstrap	Serial dependence preserved	Autocorrelated outcomes, moderate $N$ and $T$	<code>fект, boot</code>
Wild Bootstrap	Heteroskedasticity and clustering	Small numbers of clusters, heteroskedastic errors	<code>fект, arch</code>
Asymptotic	Large $N, T$ , strong factors	$N, T > 100$ , quick approximation (primarily as a diagnostic in typical marketing panels)	<code>gsynth</code> (standard errors)
Placebo-in-Time	Stability of factor structure	Assess extrapolation and model validity	<code>gsynth, fект</code>
Conformal	Approximate exchangeability and weak dependence of residuals	Prediction intervals for treated cells	Custom

## 8.8 Diagnostics and Robustness

Credible factor-model analysis in marketing panels rests on three pillars, conditional on the identifying assumptions in Section 8.4: the low-rank structure must fit pre-treatment outcomes well, the estimated factors must be stable enough to extrapolate into the post-treatment period, and substantive conclusions must not hinge on arbitrary specification choices. This section outlines a diagnostic workflow tailored to interactive fixed effects and matrix completion. Chapter 17 develops a general design-diagnostics protocol. Here we focus on what is specific to factor-based identification. For complementary diagnostics when using synthetic control (SC) or SDID instead of factor models, see Chapter 6 and Chapter 7.

### Pre-Treatment Reconstruction Error

Pre-treatment reconstruction error measures how well the factor model fits untreated outcomes before the campaign starts. Let  $T_0$  denote the last pre-treatment period. Compute the mean squared error of the factor-model fit in the pre-treatment window,

$$\text{MSE}_{\text{pre}} = \frac{1}{|\Omega_{\text{pre}}|} \sum_{(i,t) \in \Omega_{\text{pre}}} (Y_{it} - \hat{\lambda}'_i \hat{f}_t)^2,$$

and compare it with the variance of outcomes to obtain the proportion of variance explained,

$$R^2 = 1 - \frac{\text{MSE}_{\text{pre}}}{\text{Var}(Y_{it} : (i,t) \in \Omega_{\text{pre}})}.$$

Here  $\text{Var}(Y_{it} : (i,t) \in \Omega_{\text{pre}})$  denotes the sample variance computed over the untreated pre-treatment cells.

In typical retail panels, an  $R^2$  in the high 0.8s or above often suggests that a small number of factors accounts for most systematic variation in, for example, weekly store-level sales. Values in the 0.5–0.8 range indicate that the factor model captures some structure but leaves substantial idiosyncratic noise, which may be acceptable for noisy daily conversion data. When  $R^2$  remains below about 0.5 even after adding a few factors, the low-rank assumption is weakly supported and you should generally treat factor-based designs as exploratory and report them with heightened sensitivity analysis. These ranges are heuristic rather than hard thresholds and should be calibrated against placebo fits among controls. What matters for Assumption 24 is that a small number of factors explains a large share of the variance in untreated pre-treatment outcomes, relative to designs of similar granularity.

Context matters for interpreting these numbers. In panels with inherently volatile outcomes, such as daily sales for individual stores subject to local shocks, even a moderate  $R^2$  can be informative. In contrast, for smooth aggregated outcomes such as monthly regional brand share, you should expect the fit to be very tight if a low-rank structure is appropriate.

## Factor Stability Across Subsets

Factor-based counterfactuals rely on the assumption that the latent factors and loadings that govern pre-treatment dynamics continue to operate after treatment. A simple stability check is to re-estimate the factor model on different slices of the pre-treatment period and compare the resulting factors.

Partition the pre-treatment period into an early and a late subset, such as the first and second half of weeks before a national advertising campaign. Estimate the factor model separately on each subset, align the estimated factors by sign and rotation, and compute, for each factor  $r$ , the correlation between  $\hat{f}_{t,r}^{\text{early}}$  and  $\hat{f}_{t,r}^{\text{late}}$  over their overlapping periods. Writing  $f_t = (f_{t1}, \dots, f_{tR})'$ , let  $\hat{f}_{t,r}^{\text{early}}$  and  $\hat{f}_{t,r}^{\text{late}}$  denote estimates of component  $r$  from the early and late subsamples. We align early and late factor estimates by an orthogonal Procrustes rotation so that their factor spaces are comparable (individual factors remain labelled only up to rotation and sign). Correlations near one suggest that the factor captures a persistent pattern, such as a national demand trend or a stable promotion cycle. Correlations materially below 1 (for example, below about 0.7 as a rough heuristic) signal that the factor structure is shifting over time. In that case, extrapolating the factor model to the post-treatment period risks biasing treatment-effect estimates, because the “background” dynamics are not stable. Assumption 29 requires that the factor space for untreated outcomes is stable over time.

Reporting these correlations factor by factor, and commenting on any unstable components, makes the stability assumption concrete rather than purely verbal.

## Residual Diagnostics

Residual diagnostics assess whether the idiosyncratic errors  $\varepsilon_{it}$  behave like noise after accounting for factors, or whether systematic structure remains that could threaten identification. Compute residuals  $\hat{\varepsilon}_{it} = Y_{it} - \hat{\lambda}'_i \hat{f}_t$  and examine them across time and across stores.

Start by plotting residuals over time for a sample of treated and control stores. Persistent runs of positive or negative residuals indicate remaining autocorrelation, suggesting that the factor model underfits temporal dynamics. In a supermarket chain, for example, you may see residuals that spike every December because year-end shopping is not fully captured by the factors. In that case, adding seasonal controls or extra factors that track holiday cycles is essential because those cycles can masquerade as treatment effects if your campaign overlaps them. Strong residual autocorrelation or unmodelled seasonality means that important common shocks are still in the error term rather than in the factor structure, weakening both Assumption 24 and the reliability of extrapolated counterfactuals.

Next, inspect how residual variance varies across units and over time. If some regions exhibit much larger residual volatility than others, heteroskedasticity is present. For inference this pushes you towards heteroskedasticity-robust methods such as wild bootstrap, discussed in Section 8.7. From a design perspective, highly volatile regions may contribute more noise than information to the effect estimate.

Outliers in the residual distribution often correspond to one-off events, such as a stock-out in a flagship store or an exceptional local promotion by a competing brand. Sensitivity checks that re-estimate effects after

excluding periods or stores with extreme residuals are a simple way to assess whether conclusions hinge on such events. Note that such exclusions change the estimand, and this should be reported (for example, “We estimate ATT after excluding weeks with residuals above the 99th percentile”). Finally, plotting residuals by calendar month or quarter helps detect unmodelled seasonality; systematic high residuals every November in a fashion retailer, for example, signal the need for richer seasonal structure before you rely on the factor-based counterfactual.

Figure 8.2 illustrates typical residual diagnostic plots, showing time-series patterns, cross-sectional distributions, and autocorrelation functions.

## Placebo-in-Time Diagnostics

Placebo-in-time diagnostics, introduced in Section 8.7, treat pre-treatment dates as if they were intervention dates and recompute factor-based counterfactuals. As a diagnostic they address a single question. If we apply our imputation procedure when we know there was no campaign, do we still see “effects” of comparable magnitude to the post-treatment estimates?

Compute the root mean squared pseudo-effect (RMSPE) across a grid of pseudo-intervention dates, as defined in Section 8.7. As in Section 8.7, choose pseudo-dates that leave a reasonably long pseudo-post window and compute RMSPE only over cells that are untreated in the actual design. Then compare the RMSPE for the actual treatment date with the distribution of placebo RMSPEs. If the actual RMSPE sits well inside the placebo distribution, the factor model is generating discrepancies of similar magnitude even when no campaign occurred, which weakens the case for a causal interpretation. If the actual RMSPE is unusually large, for example above the 90th or 95th percentile of the placebo distribution, that pattern is more consistent with a genuine treatment effect layered on top of model noise. These comparisons are diagnostic, not formal tests, but they provide a visual and quantitative check on the stability and extrapolation quality of the factor model. Placebo-in-time plots are powerful diagnostics, but repeatedly tweaking ranks, donor sets, or pre-period windows until placebo RMSPE looks small will understate uncertainty. When design choices are informed by placebo behaviour, interpret any subsequent inference as conditional on those choices rather than as the result of a fixed pre-specified procedure.

## Sensitivity to Rank Selection

The number of factors  $R$  (or the regularisation parameter in matrix completion) controls how flexible the factor model is. If substantive conclusions change sharply as you adjust  $R$  within a plausible range, they are fragile by construction. We therefore recommend a simple rank-sensitivity analysis.

Estimate the factor model for several values of  $R$  that you consider reasonable given the scree plot and information criteria, for example  $R \in \{3, 5, 7, 10\}$  in a weekly sales panel. For each choice of  $R$ , recompute

the primary estimand, such as the average post-treatment effect on treated stores,  $\text{ATT}(R)$ . Plot  $\text{ATT}(R)$  and key event-time effects against  $R$ . If the curve is flat or varies only modestly, say by less than roughly 10

## Leave-One-Out Sensitivity

Leave-one-out analyses assess whether a small number of units or periods drive the estimated effect. In a marketing application this might correspond to a single very large region or a holiday period during which the campaign coincides with seasonal peaks.

In a leave-one-unit-out analysis, re-estimate the factor model repeatedly, each time excluding one unit (such as a store, region, or brand) from the sample, and recompute the ATT. Summarise how much the estimated ATT changes across these specifications. A narrow range, for instance changes within about 5–10

A similar exercise over time removes one period at a time. Large shifts when you drop specific weeks or months often indicate structural breaks (such as a competitor’s national campaign) or severe data issues. These checks do not, on their own, validate identification; they only reveal whether particular units or periods have disproportionate influence on ATT. They do, however, reveal whether a headline effect is really a story about one unusually sensitive part of the panel.

## Cross-Method Comparison

Comparing factor-based estimates with those from SC and SDID provides a cross-method robustness check. The goal is not to pick a “winner” by majority vote, but to understand which identification assumptions matter in a given design.

Estimate treatment effects using interactive fixed effects as described in this chapter, then re-estimate effects using SC (Chapter 6) and SDID (Chapter 7) on the same design. Examine whether the implied counterfactual paths are similar. If all three methods tell a similar story about the campaign’s impact—a modest uplift during the TV flight that decays afterwards—that coherence makes the causal narrative more persuasive, especially when combined with good pre-treatment fit and stable factors. When methods disagree, you gain information about where assumptions bite. For example, SC may fail to fit pre-treatment outcomes if the treated region is outside the convex hull of donor regions, while IFE may fit well but rely heavily on a low-rank structure that extrapolates across brands. Disagreement between methods should prompt further diagnostics and possibly a redesign rather than a mechanical choice of one estimator.

Agreement across methods does not prove that identification holds. If all three methods share a violated assumption, for instance spillovers that contaminate the control pool or a structural break that affects all estimators, they can all be biased in the same direction. SC, SDID, and IFE all assume that controls are unaffected by treatment (SUTVA) and that key background trends are stable. If either fails, all three methods can agree and still be jointly biased. Cross-method comparison raises confidence but does not substitute for careful design and assumption checking.

## Diagnostic Workflow and Reporting

In practice, you will not treat each diagnostic in isolation. A coherent workflow runs through pre-treatment fit, factor stability, residual structure, placebo-in-time behaviour, tuning sensitivity, and cross-method comparison, and then reports a concise summary. Chapter 17 sets out a general protocol. For factor models in marketing panels, a reasonable workflow is to report pre-treatment  $R^2$  and scree plots, document factor stability correlations, summarise key residual patterns, show the placebo RMSPE distribution with the actual treatment date marked, present rank-sensitivity and leave-one-out ranges for the ATT, and juxtapose IFE estimates with SC and SDID where feasible.

Transparent reporting of these elements does not guarantee that identification holds, but it allows readers—and your future self—to see which assumptions carry the causal claim and how strongly the data support them.

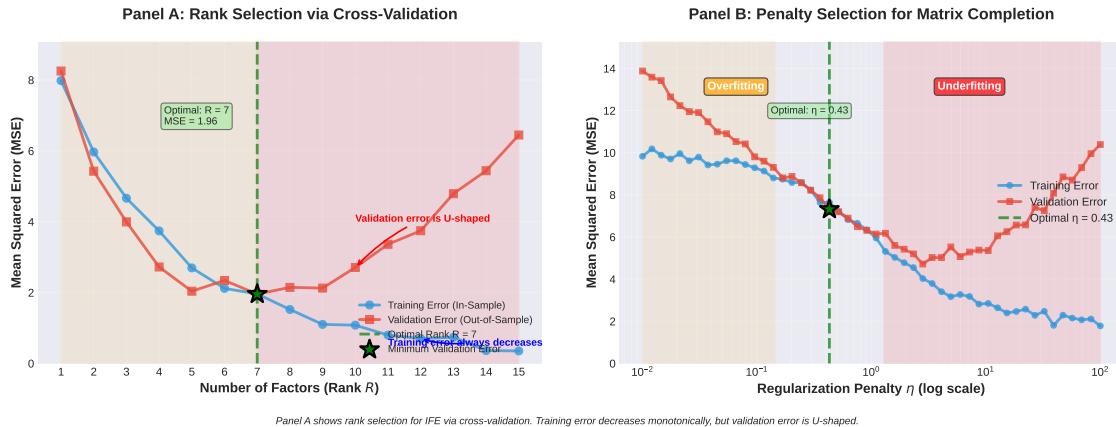
## Software Implementation

Several packages help implement these diagnostics. In R, `gsynth` [Xu, 2017] reports pre-treatment fit measures, placebo-in-time diagnostics, and factor-extraction summaries for generalised synthetic control designs that are close to the factor models in this chapter (see Section 8.5). The `fект` package provides residual diagnostics, cross-validation tools for rank selection, and leave-one-out analyses that are useful in staggered adoption settings common in retail experiments. The `Synth` package remains the workhorse for classical SC, and its standard plots for pre-treatment fit and placebo studies are natural comparators when you benchmark factor-based designs. When built-in functionality is insufficient—for example, to compute factor stability correlations or to compare IFE with SDID—you can work directly with the estimated factors and loadings and implement the plots and summaries using standard graphics libraries.

## Diagnostic Summary Table

**Table 8.4** Diagnostic Checklist for Factor Models

Diagnostic	Metric	Indicative Range	Primary Question
Pre-treatment fit	$R^2$	High (for example $\gtrsim 0.8$ for smooth data)	Is a low-rank structure plausible? (Assumption 24)
Factor stability	Correlation (early vs late)	Close to 1, markedly low values signal change	Can we extrapolate factors into post-treatment? (Assumption 29)
Residual patterns	Autocorrelation, seasonality, outliers	Weak structure after factoring	Are important dynamics or shocks left out?
Placebo-in-time	RMSPE percentile of actual date	Actual RMSPE unusually large vs placebo	Are post-treatment discrepancies larger than placebo noise?
Rank sensitivity	Range of ATT( $R$ ) across $R$	Modest variation across plausible $R$	Do conclusions hinge on rank choice?
Leave-one-out	Change in ATT by unit/period	Small changes for individual deletions	Is the effect driven by a few influential units or periods?
Cross-method	IFE vs SC vs SDID	Similar qualitative patterns	Are results robust across identification strategies?

**Fig. 8.2** Factor Model Diagnostic Plots

Note: Panel (a) shows residuals over time for selected units, revealing autocorrelation patterns. Panel (b) shows the distribution of residuals across units, checking for outliers. Panel (c) shows the scree plot of eigenvalues with the selected rank marked. Panel (d) compares factor estimates from early vs late pre-treatment periods, assessing stability.

## 8.9 Marketing Applications

Factor models and matrix completion are particularly well-suited to marketing settings where outcomes share strong common patterns—seasonality, macro shocks, category trends—and where units differ in their exposure to these patterns. This section illustrates factor-model designs in three common marketing scenarios (with illustrative magnitudes only) and shows how the methods in this chapter play out in practice. In each case we connect design choices to the identification arguments, diagnostics, and inference tools developed earlier. Throughout this section,  $\hat{Y}_{it}(0)$  denotes the imputed untreated potential outcome from the chosen factor or matrix-completion estimator, and  $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$  the corresponding cell-level effect estimate.

### Application 1: Retail Category Demand with Shared Seasonality

Consider a supermarket chain operating 80 stores across eight regions, with weekly sales for a beverage category such as carbonated soft drinks observed over four years (208 weeks). Sales display clear weekly seasonality (weekend peaks, weekday troughs), annual seasonality (higher in summer, lower in winter), and holiday spikes around Christmas, New Year, and Independence Day. The chain introduces a multi-week promotional strategy in 20 stores starting in week 105, and the goal is to estimate the effect on category sales.

**Estimation.** Estimate the factor model using weeks 1 through 104 (pre-treatment) for all 80 stores. After two-way demeaning (subtracting store and week means, Section 8.2), apply principal components to the residual matrix. The scree plot shows five eigenvalues substantially larger than the rest, suggesting rank  $R = 5$ . In each case we report a representative rank choice (for example,  $R = 5$ ), selected via the scree plots, information criteria, and cross-validation procedures in Section 8.2. The five-factor space often aligns with interpretable patterns (for example, weekday–weekend cycles, seasonality, holidays) after choosing a rotation and sign convention. Loadings for treated stores are estimated from their pre-treatment outcomes, and post-treatment factors are estimated from the post-treatment outcomes of the 60 never-treated stores.

This last step is central for identification: never-treated stores anchor the post-treatment factor structure. This design directly supports Assumptions 24, 27, and 29: the high pre-treatment  $R^2$  and stable factors suggest a strong low-rank structure, the donor pool of 60 never-treated stores provides rich support for treated stores' loadings, and the agreement of early and late pre-treatment factors supports stability. This structure helps separate common shocks from campaign effects, conditional on no interference and stability of the untreated factor structure, because the factor estimates in weeks 105–208 are driven by units whose potential outcomes remain untreated.

**Treatment effects.** For each treated store and post-treatment week, construct counterfactual outcomes  $\hat{Y}_{it}(0) = \hat{\lambda}'_i \hat{f}_t$  using the estimated loadings and post-treatment factors. Here we work with demeaned outcomes, so  $\hat{\lambda}'_i \hat{f}_t$  already represents the counterfactual in the demeaned space; when reporting effects in levels, we add back the estimated unit and week means as in Section 8.2. The estimated cell-level effects are gaps

$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ , and the primary estimand is the average effect over treated stores and post-treatment weeks, ATT. In a typical implementation, the analysis might yield an estimate on the order of single-digit percentage points (for example, 5–10%), depending on the category and campaign.

**Diagnostics.** Applying the diagnostics from Section 8.8, the pre-treatment fit is high (for example,  $R^2$  around 0.85), factor stability correlations between early and late pre-treatment periods exceed 0.9, and residuals show no remaining seasonal pattern. Estimates of ATT are stable when the rank varies over a plausible range such as  $R \in \{4, 5, 6\}$ . Inference uses block bootstrap with blocks of several weeks to accommodate serial dependence, as discussed in Section 8.7.

**Why factor models.** The factor model replaces a very high-dimensional set of weekly and seasonal dummies with a small number of latent factors that capture shared demand patterns across stores. Because loadings differ by store, the model allows store-specific seasonality and holiday sensitivity, which is critical when promotions coincide with holidays. Never-treated stores provide post-treatment information on the common factors, so factor-based counterfactuals inherit the same identification logic as a well-designed panel with untreated controls.

## Application 2: National Campaign with Heterogeneous Market Response

Now consider a consumer packaged goods brand launching a national television campaign that airs in all 50 designated market areas (DMAs) over a year. The campaign’s gross rating points (GRPs) vary across markets: large DMAs receive heavier exposure, while smaller DMAs receive lighter flights. The goal is to estimate heterogeneous effects across markets, relating them to advertising intensity.

**Estimation.** Use 24 pre-campaign months for all 50 markets to estimate the factor structure that captures shared dynamics such as seasonality, macro trends, and regional demand shifts. Extract a modest number of factors (for example,  $R = 5$ ) and interpret their loadings as measuring each market’s baseline sensitivity to these common shocks. Let  $D_{it} = \text{GRP}_{it}$  denote the monthly advertising intensity in DMA  $i$  at month  $t$ , so that  $D_{it}$  plays the role of a continuous treatment in the global notation.

To identify heterogeneous treatment effects when all units are treated, you cannot rely on post-treatment controls. Instead, treat GRPs as an observed continuous treatment in an interactive fixed effects model, where outcomes are driven jointly by factors and by campaign intensity. Conceptually, you can write the outcome as  $Y_{it} = \alpha_i + \lambda_t + \lambda'_i f_t + \beta D_{it} + \varepsilon_{it}$ , with the factors partialling out common shocks while residual variation in GRPs across markets identifies the response to advertising.

Identification here relies on a selection-on-observables or ignorability condition for the continuous treatment: after conditioning on fixed effects and the factor structure, residual variation in  $D_{it}$  is independent of  $Y_{it}(0)$ , and overlap holds. Because all units are exposed, post-period factor estimates are learned from treated outcomes, increasing reliance on correct specification.

**Treatment effects.** For each DMA, construct a counterfactual path by setting  $D_{it}$  to zero (or to the pre-campaign baseline) in the post-campaign months while keeping the factors and loadings fixed. This yields an imputed  $\hat{Y}_{it}(0)$  under “no campaign” and a corresponding time path of effects  $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ . Aggregating over the post-campaign year produces market-specific average effects. In an illustrative scenario, large markets might show larger percentage changes than smaller markets, with the pattern aligning with cumulative GRPs.

**Diagnostics.** Where feasible (for example, using partially untreated subperiods, placebo windows, or submarkets with differential exposure), compare factor-based estimates with SC or SDID-style benchmarks, as discussed in Section 8.5. Rank-sensitivity and residual diagnostics from Section 8.8 apply as before.

**Why factor models and what they assume.** When treatment intensity varies continuously and all markets are exposed, factor models are attractive because they absorb rich common structure and let you use cross-sectional variation in GRPs to identify a dose–response relationship. The trade-off is that, without never-treated DMAs, post-campaign factors are no longer anchored by untreated units. Post-campaign factor trajectories are learned entirely from treated outcomes. If the campaign induces or coincides with a new common demand shock, the factor estimator cannot separate that shock from the campaign effect, and Assumption 29 may fail in ways that push part of the true effect into the background factor component. Identification now rests on the assumption that the campaign effect operates through the observed GRPs and that, after conditioning on factors and GRPs, there is no structural break in the remaining error. If the campaign changes the underlying demand trend in ways that resemble a new common factor, the factor model may treat part of the effect as background and underestimate the true impact. This design is therefore weaker than Application 1 and depends critically on how credible those assumptions are in the specific campaign.

### Application 3: Platform Policy Change Without Controls

Finally, consider a digital platform that implements a policy change—such as a fee increase for marketplace sellers or a change in recommendation algorithms—affecting all users starting in month 13. The platform observes engagement metrics for 100 000 users. No users are exempt, so there are no never-treated controls, and even partial-treatment designs (as in the national campaign) are unavailable.

**Estimation.** To keep computation manageable, aggregate users into 100 cohorts based on sign-up date and engagement level (for example, early adopters with high activity, recent sign-ups with low activity). Aggregating users into cohorts improves computational feasibility but also averages over heterogeneous responses and possible spillovers within cohorts. Any cohort-level effect should therefore be interpreted as an average over potentially quite different user trajectories. Estimate the factor model on months 1–12 (pre-treatment) for all cohorts. A small number of factors typically capture time-of-day and day-of-week patterns, seasonal variation, platform-wide trends, and broad cohort differences.

**Factor extrapolation and assumptions.** Because there are no controls, post-treatment factors cannot be estimated from untreated outcomes. Instead, you extrapolate the factor trajectories estimated in the pre-treatment window into the post-treatment period. One possible extrapolation strategy is to project seasonal components forward by repeating the monthly cycle and to evolve trend components using a fitted trend (for example, linear or smooth growth), while assuming cohort-specific loadings remain stable. This procedure imposes a strong assumption: in the absence of the policy change, the latent factors and loadings that govern engagement would have evolved after month 12 exactly as implied by their pre-treatment dynamics. Formally, we assume that the factor structure for untreated engagement would have continued after month 12 exactly as implied by the pre-policy estimates—Assumption 29 extrapolated beyond the observed window. Any platform-wide shock that coincides with the policy date is therefore indistinguishable from the policy in this design.

**Caution and sensitivity.** This design is much closer to structured forecasting than to the panel identification strategies used earlier. Any platform-wide shock that coincides with the policy date—a competitor’s launch, macro downturn, or unrelated product change—is observationally indistinguishable from the policy in this design. Sensitivity analyses should therefore vary extrapolation rules (for example, flat trend versus continued growth, different seasonal patterns) and report the resulting range of estimated effects. Placebo-in-time diagnostics, discussed in Sections 8.7 and 8.8, play a central role. Treat an earlier month such as month 7 as a pseudo-policy date and check whether extrapolated counterfactuals track months 8–12 reasonably well.

**Treatment effects.** Within a given extrapolation scheme, counterfactual outcomes remain  $\hat{Y}_{it}(0) = \hat{\lambda}'_i \hat{f}_t$  with extrapolated factors. Treatment effects for each cohort and month are the gaps  $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ , which can be aggregated across cohorts using cohort sizes as weights. A typical pattern might be a change on the order of a few percentage points, with heterogeneity across cohorts. These estimates, however, are only as credible as the extrapolation assumptions that underpin them.

**Why factor models and why the design is weak.** When no controls exist, factor models with explicit extrapolation are one of the few tools that let you articulate a counterfactual path consistent with pre-policy dynamics. On its own, this design does not deliver a credible causal estimate in the sense of Chapter 2. Its value lies mainly in stress-testing narratives that are also supported by stronger evidence, such as randomised experiments on subsets of users or staggered rollouts in earlier cohorts (see Chapter 4).

## Summary and Comparison

Table 8.5 summarises the three applications and highlights how the panel structure and availability of controls affect design strength.

**Table 8.5** Factor Model Applications in Marketing

Note: Numbers in the "Key Finding" column are illustrative and chosen to reflect typical magnitudes rather than specific em-

	<b>Application</b>	<b>N</b>	<b>T</b>	<b>Controls</b>	<b>Rank</b>	<b>Key Finding</b>
empirical results.	Retail promotion	80 stores	208 weeks	60 never-treated stores	$R = 5$	Illustrative single-digit percentage change
	National campaign	50 DMAs	36 months	No never-treated DMAs, continuous GRPs	$R = 5$	Illustrative heterogeneity by exposure
	Platform policy	100 cohorts	24 months	None, factor support and stability are imposed, not tested	$R = 5$	Illustrative change under extrapolation

## Choosing the Right Design

The three applications illustrate a hierarchy of designs ordered by identification strength.

**Never-treated controls available (Application 1).** When some units are never treated and are comparable to treated units, post-treatment factors can be estimated from their outcomes. Under the stability assumptions in this chapter, this lets you separate common shocks from treatment effects in a way that can be probed with residual diagnostics and placebo diagnostics. Whenever the business setting allows it, this is the preferred design. If the low-rank structure were weak, a TWFE DiD or SC design using the 60 never-treated stores would be the natural fallback.

**All units treated but treatment intensity varies (Application 2).** When treatment timing is common but intensity differs across units, factor models can still contribute by absorbing shared shocks and using cross-sectional variation in intensity to identify a response curve. The cost is that, without never-treated units, you have no direct post-treatment benchmark for the common factors. Validity now hinges on modelling treatment intensity explicitly and on assuming that, after conditioning on factors and observed GRPs, there is no structural break in the remaining component. This sits between the first and third designs in terms of credibility. If GRPs were only weakly informative, you might instead rely on simpler cross-sectional designs or market-level experiments.

**No controls and factor extrapolation required (Application 3).** When all units are treated and there is no variation in intensity, post-treatment counterfactuals must come from extrapolation. In this setting, stability of factor dynamics across the policy date is untestable in the post-treatment window, and any concurrent structural change is confounded with the policy. Diagnostics and sensitivity analysis become essential, but they cannot fully substitute for the information that untreated units provide. This is the weakest design and should be used only when more credible alternatives are unavailable. Without any variation

in timing or intensity, there is no DiD or SC analogue; experimentation is the only way to strengthen identification.

Across all three scenarios, the diagnostic workflow in Section 8.8, the tuning choices in Section 8.6, and the inference procedures in Section 8.7 apply unchanged. The main difference lies in how much the data allow you to probe and challenge the core assumptions behind factor-based counterfactuals.

## 8.10 Workflow Checklist

This section provides a compact protocol for factor-model and matrix-completion analyses in marketing panels. The workflow integrates design, estimation, diagnostics, inference, and reporting so that causal conclusions about campaigns, price changes, and policy shifts rest on transparent assumptions and reproducible steps.

### Step 1: Define the Estimand and Treated Cells

Begin with the substantive question and the target estimand. Decide whether you seek an average treatment effect across all treated units and post-treatment periods, ATT, unit-specific or cohort-specific effects, or event-time effects that trace dynamics around the intervention. Define the treated cells, the unit-period pairs exposed to treatment, and the untreated cells, which include pre-treatment periods for treated units and all periods for never-treated units. In the global notation, these are the cells with  $D_{it} = 1$  (treated) and  $D_{it} = 0$  (untreated), where  $D_{it}$  denotes the binary or continuous treatment as in Chapter 2. Document treatment timing, staggered adoption patterns, and the observation window. Section 8.2 sets out the formal panel and potential-outcomes notation; we build directly on that setup here.

### Step 2: Choose IFE or Matrix Completion

Choose between interactive fixed effects (IFE) and matrix completion based on the data and the role of interpretability. Use IFE when you are willing to commit to a specific rank, either because it is suggested by information criteria or economic reasoning, and when it is helpful to interpret factors and loadings as latent demand drivers and unit sensitivities. From an identification perspective, both IFE and matrix completion rely on the same low-rank, design-support, and factor-stability assumptions in Section 8.4; the main difference is whether you enforce a hard rank constraint (IFE) or a soft rank through the nuclear norm (matrix completion). If diagnostics suggest the low-rank assumption is weak, both approaches may be inappropriate. Use matrix completion when the rank is unclear or when you need regularisation to stabilise imputation, and when prediction of missing cells matters more than interpreting the factors themselves. Sections 8.2 and 8.3 provide the underlying estimation details; here we treat them as alternative engines for producing counterfactuals.

### Step 3: Select the Rank or Penalty

Select the number of factors  $R$  (for IFE) or the regularisation parameter  $\lambda$  (for matrix completion) using the tools in Section 8.6. In practice this means combining information criteria such as IC1 and IC2, visual

inspection of the scree plot for clear eigenvalue gaps, and cross-validation that holds out part of the pre-treatment period and evaluates predictive fit. Plot prediction error as a function of  $R$  or  $\lambda$  and choose either the minimiser or an elbow where additional complexity yields little improvement. Where possible, pre-specify the rank or penalty selection rule before inspecting post-treatment outcomes and treat any post-hoc tuning as exploratory analysis rather than as part of the main confirmatory specification. Remember that cross-validation blocks must consist of untreated cells and respect time ordering.

#### Step 4: Fit on Untreated or Pre-Treatment Data

Estimate the factor model using only untreated outcomes: pre-treatment periods for treated units and all periods for never-treated control units. For IFE, apply principal components or iterated least squares to the appropriately demeaned outcome matrix to obtain loadings  $\hat{\lambda}_i$  and factors  $\hat{f}_t$ . For staggered adoption, restrict estimation to cells  $(i, t)$  with  $D_{it} = 0$ ; treated cells never contribute to factor estimation. For matrix completion, solve the nuclear-norm regularisation problem on the observed cells to obtain an estimated low-rank matrix  $\hat{M}$ . Pre-processing steps such as two-way demeaning and detrending, discussed in Section 8.6, should be carried out before factor estimation.

When all units are treated or when only treatment intensity varies, adaptations are needed. In those designs, outlined in Section 8.9, you either incorporate observed treatment intensity directly into the outcome model or extrapolate factors forward in time, at the cost of stronger assumptions. These adaptations rely on substantially stronger assumptions—notably extrapolation of factor dynamics or ignorability of continuous treatment given factors—and should be viewed as weaker designs than those with never-treated controls.

#### Step 5: Impute Counterfactuals

Use the estimated factor structure to construct counterfactual outcomes for treated cells. For each treated unit-period pair  $(i, t)$ , define the estimated untreated outcome  $\hat{Y}_{it}(0)$  as either  $\hat{Y}_{it}(0) = \hat{\lambda}'_i \hat{f}_t$  under IFE or  $\hat{Y}_{it}(0) = \hat{M}_{it}$  under matrix completion. The cell-level treatment-effect estimator is then

$$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0).$$

In designs with never-treated controls, post-treatment factors are identified from units whose potential outcomes remain untreated, so  $\hat{Y}_{it}(0)$  is anchored by observed control paths. In all-treated or extrapolation designs, the same formula is purely model-based:  $\hat{Y}_{it}(0)$  reflects extrapolation from pre-treatment dynamics and strong assumptions about factor stability rather than direct information from untreated units.

### Step 6: Aggregate and Plot Treatment Effects

Aggregate cell-level effects into the estimands defined in Step 1. The overall estimator for the average treatment effect on the treated is

$$\widehat{\text{ATT}} = \frac{1}{|\mathcal{T}|} \sum_{(i,t) \in \mathcal{T}} \widehat{\tau}_{it},$$

where  $\mathcal{T}$  indexes treated cells. Event-time estimators average effects at a given relative time  $k$  after treatment,

$$\widehat{\theta}_k = \frac{1}{|\mathcal{T}_k|} \sum_{(i,t) \in \mathcal{T}_k} \widehat{\tau}_{it},$$

with  $\mathcal{T}_k$  denoting the set of treated cells at event time  $k$ . Here  $k = t - G_i$  is event time relative to the adoption time  $G_i$  of unit  $i$ , and  $\mathcal{T}_k = \{(i,t) : D_{it} = 1, t - G_i = k\}$  as in Chapter 5.

Plotting event-time profiles helps visualise dynamics and check that estimated effects emerge after, rather than before, treatment. As in the event-study chapter (Chapter 5), including leads of treatment provides a simple pre-trend diagnostic: under factor stability and correct specification, lead coefficients should be close to zero. In factor models this is necessary but not sufficient. Lead coefficients close to zero support Assumption 26 (no anticipation) but do not by themselves verify Assumption 29 or the low-rank structure; they complement, rather than replace, the factor-specific diagnostics in Section 8.8.

### Step 7: Conduct Diagnostics

Apply the diagnostic workflow from Section 8.8 to assess whether the factor structure and specification support credible counterfactuals. In practice this means reporting pre-treatment reconstruction error and  $R^2$ , assessing factor stability by re-estimating on early versus late pre-treatment subsets, examining residuals across time and units for remaining autocorrelation, heteroskedasticity, and unexplained seasonality, varying the rank over a plausible range to gauge sensitivity, and carrying out leave-one-unit-out and leave-one-period-out analyses to identify influential observations. Where feasible, compare factor-based counterfactuals to those from SC and SDID to understand how much your conclusions depend on the factor-model assumptions.

These diagnostics are not optional embellishments. They are the main way to interrogate the low-rank and stability assumptions that underpin factor-based identification. When diagnostics point clearly against low rank, factor stability, or design support—for example, low pre-treatment  $R^2$ , unstable factors, or high rank sensitivity—the appropriate response is often to revise the design (for example, simplify the model, change the donor pool, or switch to a DiD or SC specification), not merely to report weaker diagnostics.

## Step 8: Quantify Uncertainty

Choose inference procedures that match the panel dimensions and dependence structure, using Section 8.7 as the primary reference. For sampling uncertainty in estimators such as  $\widehat{\text{ATT}}$  and event-time effects, block bootstrap and wild bootstrap are usually the workhorses. In panels with few units or short pre-treatment periods (for example,  $N < 50$  or  $T < 20$ ), analytic standard errors and asymptotic approximations are fragile. In these settings, rely primarily on block or wild bootstrap and treat all intervals as approximate, cross-checking results against simpler designs where possible. Block bootstrap resamples contiguous stretches of time to preserve serial dependence in outcomes, while wild bootstrap resamples residuals with random signs or weights to handle heteroskedastic and clustered errors.

When outcomes are correlated within clusters  $c = 1, \dots, G$  and clusters are the independent sampling unit, the effective sample size for inference is  $G$ . In those settings, use cluster-aware procedures, such as cluster wild bootstrap, and be explicit about the clustering unit in reported standard errors.

Placebo-in-time diagnostics and conformal-style intervals complement, rather than replace, these tools. Placebo-in-time diagnostics, defined in Section 8.7, treat pre-treatment dates as pseudo-interventions and assess whether the factor model produces “effects” when none should exist. They are best interpreted as diagnostics for extrapolation quality, not as direct confidence intervals. Conformal intervals provide prediction bands for individual counterfactual outcomes by using the distribution of residuals from untreated cells. They help quantify how far individual treated outcomes deviate from what the model regards as normal variation. Report 95% intervals for the main estimands and relate them to the scale and persistence of estimated effects.

## Step 9: Report Sensitivity and Cross-Method Comparisons

Use sensitivity analysis and cross-method comparison to pressure-test your conclusions. Compare factor-based estimates to those from SC (Chapter 6) and SDID (Chapter 7) where the design allows it. When all three methods tell a similar story about the sign, timing, and rough magnitude of effects, the evidence is more consistent with a robust causal interpretation. Agreement across methods is reassuring but not decisive: if they share a violated assumption—such as untreated units being unaffected by spillovers—all can be biased in the same direction. When they diverge, investigate why: for example, SC may struggle when treated units are far from the donor convex hull, while IFE may rely heavily on a low-rank structure that extrapolates across units.

Report results for multiple plausible ranks or penalties and explain which specification you regard as most credible in light of diagnostics and marketing context. Where close comparison groups exist and you wish to avoid committing to a single identification strategy, methods discussed in Section 20.5 can combine evidence across designs. In all cases, provide replication materials—data, code, and documentation—so that others can verify and extend your analysis.

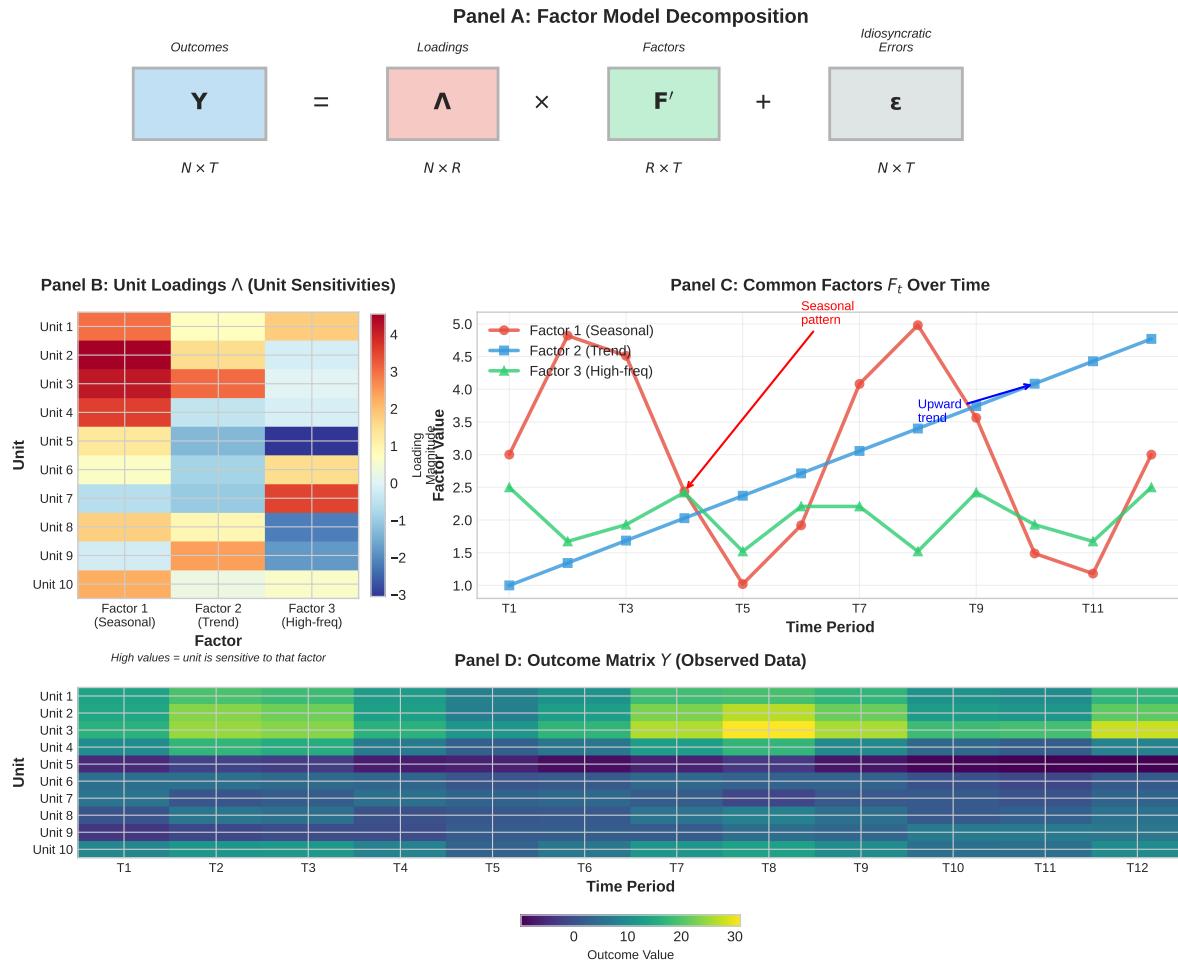
## Concluding the Analysis

Working through this workflow from design to sensitivity analysis ensures that factor-model and matrix-completion studies in marketing panels put identification first and estimation second. The earlier steps clarify what effect you are trying to learn about and which units identify it. The middle steps use the low-rank structure to impute counterfactuals and summarise effects. The final steps—diagnostics, uncertainty quantification, and cross-method comparison—confront the design with the data and make explicit how much your conclusions rely on the assumptions in this chapter.

Chapter 9 extends these ideas to dynamic factor models, tensor decompositions, and non-linear structures that can accommodate richer panel data without changing the basic logic of design, imputation, diagnostics, and inference.

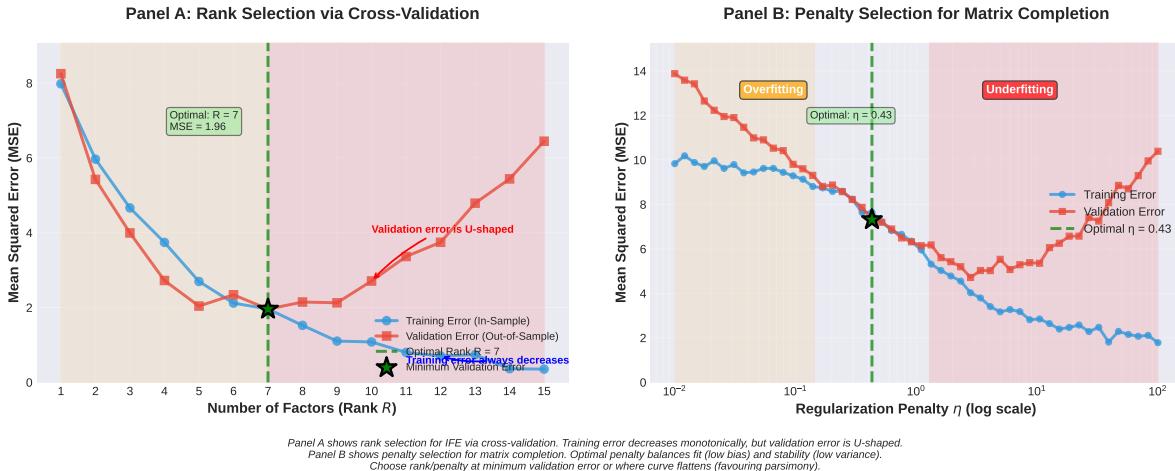
**Table 8.6** When to Prefer SC, SDID, IFE, or Matrix Completion Given Data Features

Method	Data Features	Key Assumptions (Section 8.4)	Advantages / Disadvantages
SC (Chapter 6)	Few treated units, good donor matches, long pre-period	Treated unit inside donor convex hull, no anticipation, no interference	Transparent weights and no extrapolation, but methods can fail when the treated unit is an outlier.
SDID (Chapter 7)	Staggered adoption with overlap, moderate pre-period	Parallel trends after reweighting in the weighted sample	Combines weighting and differencing and is robust to modest misspecification, but is more complex to implement and explain than SC.
IFE (this chapter)	Strong co-movement, long pre-period, diverse controls, rank well-defined	Assumptions 24–29	Flexible and accommodates heterogeneous exposure with interpretable factors, but requires rank selection and credible diagnostics for low-rank adequacy.
Matrix Completion (this chapter)	Missing cells, unclear rank, need for regularisation, possibly shorter pre-period	Assumptions 24–29, plus observation-pattern conditions	Provides regularised imputation without committing to a single rank and is useful with complex missing data, but offers less direct interpretability of factors than IFE.

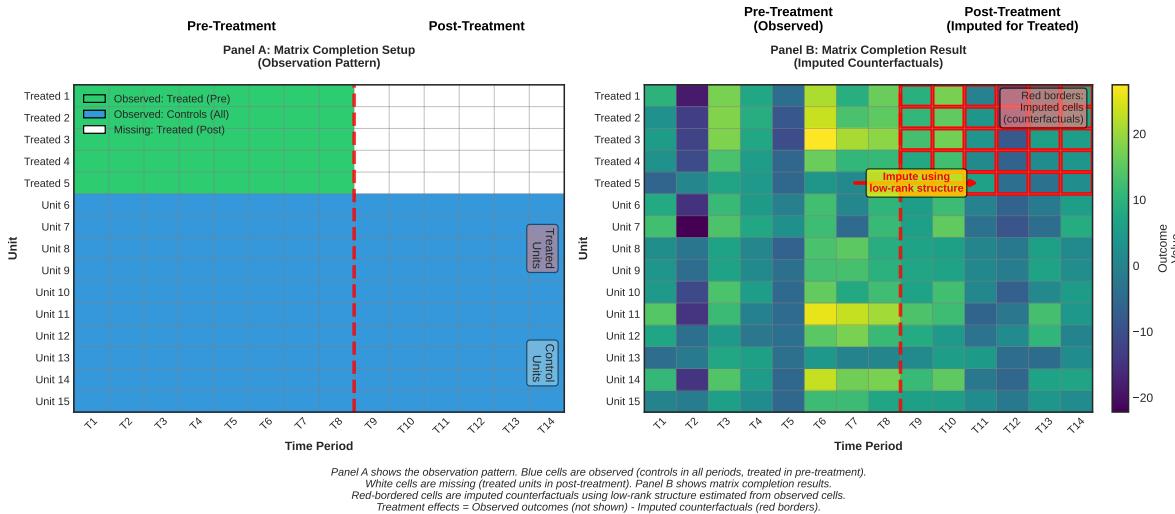


**Fig. 8.3** Factor Model Schematic with Loadings and Common Shocks

Note: Panel (a) displays the factor-model decomposition  $\mathbf{Y} = \boldsymbol{\Lambda}\mathbf{F}' + \boldsymbol{\varepsilon}$ , showing how the outcome matrix is decomposed into loadings (unit sensitivities), factors (common shocks), and idiosyncratic errors. This corresponds to the scalar form  $Y_{it}(0) = \lambda_i' f_t + \varepsilon_{it}$  used throughout the chapter. Panel (b) shows the loading matrix as a heatmap, with rows for units and columns for factors. High values indicate that a unit is strongly exposed to a given factor. Panel (c) plots several common factors over time, for example seasonal patterns, upward trends, and higher-frequency variation. Panel (d) displays the observed outcome matrix, which combines the factor structure with idiosyncratic noise.

**Fig. 8.4** Pre-Period Reconstruction Error vs Rank/Penalty (Validation Curve)

Note: Panel (a) shows rank selection for interactive fixed effects via cross-validation. Training error decreases monotonically with rank, while validation error often has a U-shape. The optimal rank (green vertical line) minimises validation error, balancing fit and complexity. Ranks below the optimum underfit, whereas ranks above it overfit idiosyncratic noise. Validation blocks should consist of untreated pre-periods and respect time ordering. Panel (b) shows penalty selection for matrix completion with nuclear-norm regularisation. Low penalties allow very flexible fits that can overfit noise, while high penalties enforce smoother low-rank structure that may underfit signal. The optimal penalty  $\lambda$  balances these extremes.

**Fig. 8.5** Matrix Completion Illustration with Treated and Missing Cells

Note: Panel (a) shows the observation pattern for matrix completion in a causal panel. Green cells are observed for treated units in the pre-treatment period. Blue cells are observed for control units in all periods. White cells, corresponding to treated units in the post-treatment period (where  $D_{it} = 1$ ), are missing and must be imputed. The red dashed line separates pre-treatment from post-treatment. Panel (b) shows the matrix completion result, in which all cells are filled using the low-rank structure estimated from observed cells. Red-bordered cells are the imputed counterfactual outcomes for treated units in the post-treatment period. Treatment effects are the differences between these imputed values and the observed outcomes.

**Box 8.1: Factor Model Workflow Summary**

**Design.** Specify the target estimand, such as ATT, unit-specific, or period-specific effects. Define treated and untreated cells and record treatment timing. Choose between IFE and matrix completion in light of the rank, interpretability, and data quality, subject to the low-rank and stability diagnostics in Sections 8.8 and 8.4.

**Estimation.** Select rank  $R$  or penalty  $\lambda$  using information criteria, scree plots, and cross-validation, with a pre-specified rule. Estimate loadings and factors, or the completed matrix, on untreated cells using the methods in Sections 8.2 and 8.3.

**Imputation and aggregation.** Impute counterfactuals for treated cells, compute cell-level effects as observed minus imputed outcomes, and aggregate into  $\widehat{\text{ATT}}$  or event-time profiles. Plot dynamics and check that effects emerge after treatment.

**Diagnostics.** Report pre-treatment  $R^2$ , factor stability, and residual patterns. Vary the rank and perform leave-one-out checks to understand how sensitive conclusions are to key choices and individual units or periods.

**Inference.** Use bootstrap methods that match the dependence structure and panel size, and complement them with placebo-in-time diagnostics and, where useful, conformal prediction intervals. Report 95% intervals for the main estimands.

**Robustness.** Compare factor-based estimates with SC and SDID when designs permit, document the assumptions that drive differences (low rank, design support, SUTVA/interference), and provide replication materials so others can scrutinise and build on the analysis.

## Chapter 9

# Advanced Matrix Methods for Causal Inference

This chapter extends the matrix completion tools from Chapter 8 to panel structures that push beyond standard low-rank matrices in marketing work. We focus on situations where standard low-rank matrix completion is not enough and show how to address its limitations through five extensions: tensor completion for multi-way panels (product  $\times$  store  $\times$  time), robust methods for outlier-contaminated data, covariate-assisted completion, time-varying rank models for non-stationary environments, and Bayesian approaches for uncertainty quantification. We discuss the computational trade-offs that matter in large-scale commercial panels and develop a framework for choosing an appropriate method given your data structure and research question. For the core low-rank factor model foundations, see Chapter 8. For alternative identification strategies based on synthetic control and SDID, see Chapters 6 and 7. The chapter closes with four marketing applications drawn from retail, advertising, consumer packaged goods, and platform settings. These extensions improve how we approximate untreated potential outcomes (for example,  $Y_{it}(0)$ ) in complex panels, but they do not relax the design assumptions from Chapter 8 that justify treating treated cells as missing  $Y_{it}(0)$ . Throughout the chapter we treat tensor and advanced matrix methods as alternative ways to model  $Y_{it}(0)$ . The causal targets, such as ATT and dynamic effects, and the need for factor stability and design-linked diagnostics remain unchanged.

## 9.1 Introduction: Beyond Standard Matrix Completion

In Chapter 8 we modelled the outcome matrix as approximately low rank, with a fixed number of latent factors and observations that were treated as equally reliable. That structure is powerful for constructing counterfactual outcomes and, in turn, for estimating treatment effects in panel data under the assignment-mechanism assumptions stated in Chapter 8. In many marketing applications, however, these modelling assumptions are stretched or violated. This chapter extends the standard framework to handle richer data structures and more realistic settings.

### Motivating Scenarios

Four scenarios motivate the extensions in this chapter. First, many marketing panels have genuine multi-way structure. A retailer tracks sales by store, product, and time, and flattening this three-way structure into a two-way matrix can discard useful inductive structure about how products and stores co-move, which in turn can weaken imputation when untreated support is limited. Tensor completion preserves the multi-way structure in the imputation model for untreated potential outcomes (for example,  $\mathcal{Y}_{ijt}(0)$ , with  $i$  indexing stores,  $j$  indexing products, and  $t$  indexing time) and can improve counterfactual reconstruction when the design assumptions that justify treating treated cells as missing  $Y_{it}(0)$  are credible. We develop these ideas in Section 9.2.

Second, retail panels often contain large irregularities. Stockouts create artificial zeros, data-entry errors produce extreme values, and promotional spikes create short-lived deviations from underlying demand. Standard matrix completion treats all observations equally, so a small fraction of extreme cells can distort the fitted low-rank structure and contaminate imputations of  $Y_{it}(0)$ . Robust methods separate sparse, extreme deviations from the low-rank component and can stabilise imputations when the sparse component corresponds to data artefacts rather than treatment effects. We introduce these methods in Section 9.3.

Third, many settings come with rich side information alongside the panel. Store characteristics (size, location, demographics) and time-varying covariates (weather, holidays, macroeconomic conditions) can carry predictive signal for  $Y_{it}(0)$ . Covariate-assisted matrix completion incorporates this information to improve imputation and reduce reliance on purely latent factors. We describe these approaches in Section 9.4.

Finally, the factor structure itself may evolve. A recession introduces new common shocks. A major platform algorithm change alters behaviour. The COVID-19 pandemic created structural breaks in demand patterns. Time-varying rank models allow the effective number and composition of factors to change over time. They can improve  $Y_{it}(0)$  approximation in non-stationary environments, but they do not resolve confounding if breaks coincide with treatment adoption. We develop these models in Section 9.5.

Across these scenarios, our causal targets remain the same as in earlier chapters: ATT and, when time is central, dynamic response paths such as event-time effects  $\theta_k$ . What changes is the model we use for untreated potential outcomes  $Y_{it}(0)$ . Tensor and advanced matrix methods impose richer low-rank structure on  $Y_{it}(0)$  in order to borrow strength across products, stores, markets, and time, while leaving the underlying design

assumptions intact (in particular, the factor-stability and missingness restrictions that justify treating treated cells as missing  $Y_{it}(0)$ , as stated in Chapter 8). See Section 9.9 for diagnostics that target these low-rank and assignment assumptions.

## Chapter Roadmap

Table 9.1 maps the motivating scenarios to chapter sections and summarises the key methods.

**Table 9.1** Chapter 9 roadmap: advanced matrix methods for counterfactual imputation

Scenario	Section	Method	Marketing example
Multi-way structure	9.2	Tensor completion, CP/Tucker	Store $\times$ product $\times$ time sales
Outliers	9.3	Robust PCA, sparse + low-rank	Stockouts, promotions, data errors
Side information	9.4	Covariate-assisted MC	Store demographics, weather
Structural breaks	9.5	Time-varying rank, changepoint	COVID shock, algorithm changes
Uncertainty	9.6	Bayesian matrix completion	Posterior intervals for ATT
Scalability	9.7	Stochastic, distributed methods	Large retail panels
Method selection	9.8	Decision framework	Choosing the right method
Diagnostics	9.9	Validation, residual checks	Assessing model fit
Applications	9.10	Four case studies	Retail, advertising, platform, CPG
Workflow	9.11	Step-by-step protocol	Reproducible analysis

In every row, the method is used to impute untreated potential outcomes (for example,  $Y_{it}(0)$ ) for treated cells. Causal interpretation then depends on the assignment-mechanism assumptions stated in Chapters 2 and 8, and the resulting imputations feed into ATT and dynamic estimands such as  $\theta_k$  as defined in Chapters 2 and 5.

The methods in this chapter build directly on the factor model foundations in Chapter 8. Section 9.8 explains when standard low-rank matrix completion is sufficient, when one of the advanced methods is warranted, and how these extensions relate to synthetic control (Chapter 6) and SDID (Chapter 7), which rely on different identifying restrictions than low-rank completion.

## 9.2 Tensor Completion for Multi-Way Panels

### Motivation: three-way and higher-order data

Marketing panels often have natural multi-way structure. A retailer observes sales for stores indexed by  $i = 1, \dots, N$ , products indexed by  $j = 1, \dots, J$ , and time periods indexed by  $t = 1, \dots, T$ . The outcome is a three-way tensor  $\mathcal{Y} \in \mathbb{R}^{N \times J \times T}$ , where  $\mathcal{Y}_{ijt}$  records sales of store  $i$  for product  $j$  at time  $t$ .

Standard matrix completion flattens this structure. We create a matrix  $\mathbf{Y} \in \mathbb{R}^{(N \times J) \times T}$  by stacking store-product combinations as rows and time as columns. This has two limitations. First, it ignores the natural grouping of products and stores. Products in the same category tend to co-move, and stores in the same region tend to co-move. Flattening can discard useful inductive structure about how products and stores co-move, which in turn can weaken imputation when untreated support is limited. Second, it creates a very large matrix. For example, with  $N = 1000$  and  $J = 200$  we obtain 200,000 rows, which makes both storage and computation demanding.

Tensor completion preserves the multi-way structure. It models the tensor as a sum of low-rank components that capture store effects, product effects, time effects, and their interactions. This representation is more parsimonious, because it uses fewer parameters than a fully general three-way array, and often more interpretable in practice, because factors can be associated with store-, product-, and time-related components, although factor rotations can limit one-to-one interpretation.

Further marketing contexts fit the same pattern. Large e-commerce platforms observe user engagement as a three-way tensor of users by items by contexts such as device, time of day, or location. Advertising data often form a four-way tensor of campaigns by creatives by platforms by time. In all these settings, tensor completion offers a natural way to impute the untreated potential outcomes that feed into causal estimands.

### Tensor decomposition models

We adopt standard CP and Tucker decompositions to model multi-way structure. Let  $\Omega \subseteq \{1, \dots, N\} \times \{1, \dots, J\} \times \{1, \dots, T\}$  denote the set of measured tensor entries, and let  $\mathcal{W} \subseteq \Omega$  denote the treated cells. In causal applications, we use these decompositions as models for the low-rank component of untreated potential outcomes, estimating them from untreated cells in  $\Omega \setminus \mathcal{W}$ .

**Definition 9.1 (Mode- $k$  Product)** For a tensor  $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$  and a matrix  $\mathbf{U} \in \mathbb{R}^{J \times N_k}$ , the mode- $k$  product  $\mathcal{X} \times_k \mathbf{U}$  is a tensor of size  $N_1 \times \dots \times N_{k-1} \times J \times N_{k+1} \times \dots \times N_3$  with entries:

$$(\mathcal{X} \times_k \mathbf{U})_{i_1, \dots, i_{k-1}, j, i_{k+1}, \dots, i_3} = \sum_{i_k=1}^{N_k} \mathcal{X}_{i_1, \dots, i_k, \dots, i_3} \cdot U_{j, i_k}.$$

The mode- $k$  product multiplies each mode- $k$  fibre by the matrix  $\mathbf{U}$ .

**Definition 9.2 (CP Decomposition)** A tensor  $\mathcal{Y} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$  admits a CP decomposition of rank  $R$  if there exist factor matrices  $\mathbf{A} \in \mathbb{R}^{N_1 \times R}$ ,  $\mathbf{B} \in \mathbb{R}^{N_2 \times R}$ ,  $\mathbf{C} \in \mathbb{R}^{N_3 \times R}$  such that

$$\mathcal{Y}_{ijt} = \sum_{r=1}^R A_{ir} B_{jr} C_{tr} = \sum_{r=1}^R \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r,$$

where  $\otimes$  denotes the outer product and  $\mathbf{a}_r$ ,  $\mathbf{b}_r$ ,  $\mathbf{c}_r$  are the  $r$ -th columns of  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$  respectively.

**Definition 9.3 (Tucker Decomposition)** A tensor  $\mathcal{Y} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$  admits a Tucker decomposition with multilinear rank  $(R_1, R_2, R_3)$  if there exist a core tensor  $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$  and factor matrices  $\mathbf{U}^{(1)} \in \mathbb{R}^{N_1 \times R_1}$ ,  $\mathbf{U}^{(2)} \in \mathbb{R}^{N_2 \times R_2}$ ,  $\mathbf{U}^{(3)} \in \mathbb{R}^{N_3 \times R_3}$  such that

$$\mathcal{Y} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}.$$

Tucker allows different ranks for each mode, offering more flexibility than CP at the cost of a denser core tensor.

Technical Note: CP vs Tucker

CP decomposition writes the tensor as a sum of  $R$  rank-one terms,  $\mathcal{Y}_{ijt} \approx \sum_{r=1}^R A_{ir} B_{jr} C_{tr}$ . CP is simple and often interpretable in practice, but it forces all modes to share the same rank.

Tucker decomposition generalises CP by allowing distinct ranks  $(R_1, R_2, R_3)$  for each mode. The core tensor  $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$  captures interactions. Tucker is more flexible but uses more parameters:  $R_1 R_2 R_3 + N_1 R_1 + N_2 R_2 + N_3 R_3$ .

For causal work, CP is often attractive when the number of untreated cells is limited, because it is parsimonious. Tucker becomes appealing when modes have very different intrinsic dimensions and there are enough untreated observations to estimate a richer core.

These decompositions are standard. See, for example, Kolda and Bader [2009] for an overview of tensor methods.

For causal inference, we assume that untreated potential outcomes admit a low-rank tensor decomposition. This is the tensor analogue of the low-rank factor structure in Chapter 8. It rules out idiosyncratic patterns that cannot be captured by a small number of product, store, and time factors.

**Assumption 30 (Tensor low-rank structure)** The untreated potential outcomes admit a low-rank tensor decomposition:

$$\mathcal{Y}_{ijt}(0) = \sum_{r=1}^R \alpha_{ir} \beta_{jr} \gamma_{tr} + \varepsilon_{ijt},$$

where:

- (i) The rank  $R$  satisfies  $R \ll \min(N, J, T)$ .
- (ii) The factor matrices  $\mathbf{A} = (\alpha_{ir})$ ,  $\mathbf{B} = (\beta_{jr})$ ,  $\mathbf{C} = (\gamma_{tr})$  have bounded entries. In particular,  $\max_{i,r} |\alpha_{ir}| \leq \bar{\alpha}$ , and similarly for  $\mathbf{B}$  and  $\mathbf{C}$ .

- (iii) The idiosyncratic errors  $\varepsilon_{ijt}$  have mean zero and bounded variance, and satisfy weak dependence over time and across products and stores in the sense used for matrix completion in Chapter 8: serial and cross-sectional correlations decay sufficiently fast that law-of-large-numbers and central-limit-theorem arguments apply to averages over  $(i, j, t)$ .

## Causal inference with tensors

Here  $\mathcal{Y}_{ijt}(d)$  plays the same role as  $Y_{it}(d)$  in Chapter 2, but with an additional product index  $j$  to reflect the three-way panel structure.

For causal questions we focus on the untreated potential outcomes. A promotion applied to a subset of products  $\mathcal{P}$  in a subset of stores  $\mathcal{S}$  during periods  $\mathcal{T}$  defines a set of treated cells  $\mathcal{W} = \{(i, j, t) : i \in \mathcal{S}, j \in \mathcal{P}, t \in \mathcal{T}\}$ . For each measured cell  $(i, j, t) \in \Omega$ , observed outcomes satisfy  $\mathcal{Y}_{ijt} = \mathcal{Y}_{ijt}(1)$  if  $(i, j, t) \in \mathcal{W}$  and  $\mathcal{Y}_{ijt} = \mathcal{Y}_{ijt}(0)$  otherwise. The counterfactual  $\mathcal{Y}_{ijt}(0)$  is unobserved by design on treated cells.

As in Chapter 5, the average treatment effect on the treated conditions on treated cells,

$$\text{ATT} = \mathbb{E}[\mathcal{Y}_{ijt}(1) - \mathcal{Y}_{ijt}(0) \mid (i, j, t) \in \mathcal{W}].$$

In finite samples, a natural analogue aggregates over treated cells,

$$\widehat{\text{ATT}} = \frac{1}{|\mathcal{W}|} \sum_{(i, j, t) \in \mathcal{W}} (\mathcal{Y}_{ijt} - \hat{\mathcal{Y}}_{ijt}(0)).$$

We obtain  $\hat{\mathcal{Y}}_{ijt}(0)$  by fitting a low-rank tensor model on  $\Omega \setminus \mathcal{W}$  and setting  $\hat{\mathcal{Y}}_{ijt}(0) = \hat{\mathcal{M}}_{ijt}$ . Tensor completion imputes  $\mathcal{Y}_{ijt}(0)$  for treated cells by estimating the low-rank structure from untreated cells.

**Assumption 31 (Tensor factor stability)** The factor structure is stable across treated and untreated cells. In particular, the joint distribution of the factors and idiosyncratic errors is the same for treated and untreated cells, so that

$$(\alpha_{i\cdot}, \beta_{j\cdot}, \gamma_{t\cdot}, \varepsilon_{ijt}) \mid (i, j, t) \in \mathcal{W} \stackrel{d}{=} (\alpha_{i\cdot}, \beta_{j\cdot}, \gamma_{t\cdot}, \varepsilon_{ijt}) \mid (i, j, t) \in \Omega \setminus \mathcal{W}.$$

Equivalently, conditional on the low-rank component determined by  $(\alpha_{i\cdot}, \beta_{j\cdot}, \gamma_{t\cdot})$ , treatment assignment does not depend on idiosyncratic shocks. This restriction takes the form

$$\mathbf{1}\{(i, j, t) \in \mathcal{W}\} \perp \varepsilon_{ijt} \mid (\alpha_{i\cdot}, \beta_{j\cdot}, \gamma_{t\cdot}).$$

The promotion pattern may vary with long-run product, store, or time factors, but it cannot systematically target short-run shocks that are not already captured by the factors.

This assumption can fail in marketing settings where promotions are triggered by recent unexplained sales movements (for example, managers discount products with unexpectedly weak sales even after controlling for

product, store, and time factors). In that case, tensor completion still helps with prediction, but the resulting ATT estimate is not causally identified.

**Definition 9.4 (Tensor Nuclear Norm)** For computational tractability, we use the sum of nuclear norms of matricisations:

$$\|\mathcal{M}\|_{*,\text{sum}} = \sum_{k=1}^3 \|\mathbf{M}_{(k)}\|_*,$$

where  $\mathbf{M}_{(k)}$  denotes the mode- $k$  unfolding of  $\mathcal{M}$  and  $\|\cdot\|_*$  is the matrix nuclear norm.

The estimator solves the regularised problem:

$$\min_{\mathcal{M}} \sum_{(i,j,t) \in \Omega \setminus \mathcal{W}} (\mathcal{Y}_{ijt} - \mathcal{M}_{ijt})^2 + \lambda \|\mathcal{M}\|_{*,\text{sum}},$$

where  $\lambda > 0$  is the regularisation parameter (selected by cross-validation as in Section 8.6).

The following result adapts tensor-completion error bounds in Yuan and Zhang [2016] to our setting. Under Assumptions 30 and 31, tensor completion can recover  $\mathcal{Y}_{ijt}(0)$  accurately enough on treated cells for a plug-in estimator of ATT to be consistent.

**Theorem 9.1 (Consistency of ATT under tensor completion)** *Under Assumptions 30 and 31, and assuming  $|\Omega| \geq C \cdot R \cdot (N+J+T) \log^2(\max(N, J, T))$  for a universal constant  $C > 0$ , the ATT is consistently estimated by*

$$\widehat{\text{ATT}} = \frac{1}{|\mathcal{W}|} \sum_{(i,j,t) \in \mathcal{W}} (\mathcal{Y}_{ijt} - \hat{\mathcal{M}}_{ijt}),$$

where  $\hat{\mathcal{M}}$  is the tensor completion estimator. A full proof follows from the tensor-completion error bounds in Yuan and Zhang [2016], combined with the mapping from prediction error on untreated potential outcomes to bias in the ATT estimand. See also Athey et al. [2021] for the matrix case. The key step is that the ATT estimation error is bounded by the average prediction error on treated cells,

$$|\widehat{\text{ATT}} - \text{ATT}| \leq \frac{1}{|\mathcal{W}|} \sum_{(i,j,t) \in \mathcal{W}} |\hat{\mathcal{M}}_{ijt} - \mathcal{M}_{ijt}^*|,$$

so tensor-completion error bounds translate directly into bounds on bias in ATT.

**Proposition 9.1 (Estimation Error Bound)** *Under Assumptions 30–31, the nuclear-norm regularised estimator satisfies*

$$\frac{1}{NJT} \|\hat{\mathcal{M}} - \mathcal{M}^*\|_F^2 = O_P \left( \sigma^2 R \cdot \frac{N+J+T}{|\Omega|} \right),$$

where  $\mathcal{M}^*$  is the true low-rank component. See Yuan and Zhang [2016] for proof details.

## Algorithms and software

Alternating least squares (ALS) extends naturally from matrices to tensors. As in Chapter 8, ALS alternates between updating factor matrices while holding the others fixed, but now there are three sets of factors (products, stores, times) and the loss is computed over untreated entries in  $\Omega \setminus \mathcal{W}$ . Each update step reduces to a least-squares problem and can be implemented efficiently using sparse linear algebra. In practice ALS often converges in a few dozen iterations in well-conditioned problems, but iteration counts can be much higher when missingness is severe or factors are ill-conditioned. You should monitor objective decrease, held-out prediction error on untreated cells, and the stability of ATT across iterations.

Several software packages implement tensor decompositions with missing data. In Python, `tensorly` provides CP and Tucker decompositions with backends such as NumPy, PyTorch, and TensorFlow, and supports missing entries via masking. Packages such as `sktensor` focus on sparse tensors and are useful when only a small fraction of the full tensor is observed. In R, `rTensor` implements core tensor operations and decompositions, and Bayesian tensor factorisation packages such as `tensorBF` extend these ideas to uncertainty quantification. MATLAB’s Tensor Toolbox offers mature implementations for both CP and Tucker decompositions.

These tools can be plugged directly into the causal workflow from Chapters 8 and 16: estimate a low-rank tensor model for untreated outcomes on  $\Omega \setminus \mathcal{W}$ , impute  $\hat{\mathcal{Y}}_{ijt}(0)$  for treated cells, compute ATT and dynamic response paths, and then apply the diagnostics from Section 9.9 and inference procedures from Chapter 16 that match your dependence structure (serial and cross-sectional) and account for tuning and re-fitting uncertainty. These packages handle the numerical optimisation, but identification and uncertainty quantification still follow the design and inference principles in Chapters 2, 8, and 16. Tensor completion only changes how we approximate  $Y_{it}(0)$ .

## Application: multi-market product launches

Consider a consumer packaged goods company launching dozens of new products across multiple geographic markets over the course of a year, with staggered launch dates. Weekly sales form a product  $\times$  market  $\times$  week tensor. Launch periods for each product–market pair define treated cells, and the remaining cells define untreated observations for learning the low-rank structure.

The main challenge is that traditional difference-in-differences struggles with this many staggered launch paths, and synthetic control is awkward because each product–market combination is effectively unique. Tensor completion offers a unified way to borrow strength across products, markets, and time. We treat post-launch outcomes for treated product–market–week cells as unobserved counterfactuals  $\mathcal{Y}_{ijt}(0)$  and estimate a CP decomposition with a modest rank using ALS on pre-launch and never-treated cells.

The estimated factors have natural interpretations. One factor captures seasonality, with high loadings in summer and lower loadings in winter. Another tracks product-category trends, such as growing demand for health-oriented products and declining demand for indulgence categories. A third reflects market size, with

consistently higher loadings for large markets. Additional factors pick up regional preferences and differences in launch timing.

The completed tensor yields counterfactual sales paths for treated product–market–week cells. A natural summary is ATT aggregated over treated cells,  $\frac{1}{|\mathcal{W}|} \sum_{(i,j,t) \in \mathcal{W}} (\mathcal{Y}_{ijt}(1) - \hat{\mathcal{Y}}_{ijt}(0))$ , and, when launches are staggered, dynamic effects can be summarised as event-time effects  $\theta_k$ . In plausible calibrations, effects build over several weeks as awareness grows, and larger markets tend to show stronger responses.

Diagnostics on pre-launch reconstruction error and placebo exercises provide limited evidence on whether the low-rank approximation is adequate on untreated cells. They cannot by themselves validate the assignment assumptions. Interval estimates and formal inference are developed in Section 9.6. To assess the credibility of Assumption 31, you should check whether launch timing is predictable from recent residuals after fitting a baseline tensor model on pre-launch data. If products or markets with unusually low residual sales are systematically more likely to launch next, then managers are reacting to short-run shocks that the factors do not capture, and the resulting ATT estimate should be interpreted as descriptive rather than causal. Placebo launches and residual-based balance checks, developed in Section 9.9, make these patterns visible.

### 9.3 Robust Matrix Completion with Outliers

#### The Outlier Problem in Marketing Data

Marketing data are often contaminated by large irregularities. Stockouts create artificial zeros. These are not missing values but mismeasured outcomes that reflect a supply constraint rather than underlying demand, and they can violate the additive-noise assumptions behind low-rank completion. Data entry errors can produce extreme values through extra digits or misplaced decimal points. Promotional spikes generate temporary deviations when, for example, a flash sale produces a tenfold increase in sales for a day and then sales return to their previous level.

Standard matrix completion treats all observations equally. It estimates a low-rank structure that fits every cell, including outliers. Large, sparse deviations then pull the low-rank estimate away from the underlying demand structure, degrading the imputation of  $Y_{it}(0)$  and, in turn, potentially biasing estimates of ATT that plug these imputations into treated cells.

Robust matrix completion addresses this problem by decomposing the outcome matrix into three components.

**Definition 9.5 (Low-rank plus sparse decomposition)** The observed outcome matrix  $\mathbf{Y} \in \mathbb{R}^{N \times T}$  admits the decomposition

$$\mathbf{Y} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{E},$$

where:

- (i)  $\mathbf{L}^* \in \mathbb{R}^{N \times T}$  is the low-rank component with  $\text{rank}(\mathbf{L}^*) = R \ll \min(N, T)$ .
- (ii)  $\mathbf{S}^* \in \mathbb{R}^{N \times T}$  is the sparse outlier component with  $\|\mathbf{S}^*\|_0 \leq \rho NT$  for some  $\rho \in (0, 1)$ .
- (iii)  $\mathbf{E} \in \mathbb{R}^{N \times T}$  is a noise matrix with mean-zero entries and bounded variance, satisfying weak dependence over time and across units so that law-of-large-numbers and central-limit-theorem arguments apply to averages over  $(i, t)$ .

This generalises the i.i.d. noise conditions in Candès et al. [2011] to the weak-dependence setting used in Chapter 8.

The low-rank component captures systematic co-movement across units and time, in the spirit of the factor structure in Chapter 8. The sparse component captures cell-specific deviations that are large in magnitude but affect only a small fraction of cells.

#### Robust PCA and Matrix Completion

Principal Component Pursuit (PCP) provides a convex formulation for separating low-rank and sparse components:

$$\min_{\mathbf{L}, \mathbf{S}} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \quad \text{subject to} \quad \mathbf{Y} = \mathbf{L} + \mathbf{S},$$

where  $\|\mathbf{L}\|_*$  is the nuclear norm, encouraging low rank,  $\|\mathbf{S}\|_1$  is the  $\ell_1$  norm, encouraging sparsity, and  $\lambda > 0$  is a regularisation parameter.

In the noiseless exact-recovery theory, work such as Candès et al. [2011] suggests  $\lambda = 1/\sqrt{\max(N, T)}$ . In practice, cross-validation on held-out untreated cells is preferable. Small values of  $\lambda$  prioritise low rank, while large values prioritise sparsity.

When outcomes are missing, we observe  $Y_{it}$  only for  $(i, t) \in \Omega$ . A common noisy, incomplete-data formulation replaces the hard constraint with penalised least squares on observed entries, for example

$$\min_{\mathbf{L}, \mathbf{S}} \frac{1}{2} \sum_{(i,t) \in \Omega} (Y_{it} - L_{it} - S_{it})^2 + \mu \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1,$$

where  $\mu > 0$  and  $\lambda > 0$  control the low-rank and sparse penalties, respectively. In this formulation,  $\mu$  plays the role of the nuclear-norm regularisation parameter from standard matrix completion, while  $\lambda$  governs how aggressively we label large residuals as sparse irregularities. Both PCP and SPCP are convex and can be solved efficiently using Alternating Direction Methods of Multipliers (ADMM). See Section 9.7 for computational details and their relation to the matrix-completion algorithms in Chapter 8.

## Identification and Assumptions

Identification of the low-rank and sparse components requires regularity conditions that mirror those in the matrix-completion literature.

**Assumption 32 (Incoherence)** Let  $\mathbf{L}^* = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$  be the SVD of the low-rank component. The matrix  $\mathbf{L}^*$  is  $\mu$ -incoherent if:

- (i)  $\max_{i \in \{1, \dots, N\}} \|\mathbf{U}_i\|_2^2 \leq \mu R/N$ .
- (ii)  $\max_{t \in \{1, \dots, T\}} \|\mathbf{V}_t\|_2^2 \leq \mu R/T$ .
- (iii)  $\|\mathbf{U}\mathbf{V}^\top\|_\infty \leq \sqrt{\mu R/(NT)}$ .

Incoherence ensures that the singular vectors are spread out rather than concentrated in a few rows or columns. The case  $\mu = 1$  corresponds to maximally spread singular vectors.

**Assumption 33 (Random sparsity pattern)** The support of  $\mathbf{S}^*$  is uniformly distributed among all subsets of  $\{1, \dots, N\} \times \{1, \dots, T\}$  of cardinality at most  $\rho NT$ , independent of  $\mathbf{L}^*$  and  $\mathbf{E}$ .

This assumption is a tractable theoretical idealisation. In retail panels, stockouts and other irregularities are often systematic, not uniformly scattered across cells. In practice we treat the random-support condition as a heuristic and rely on diagnostics and sensitivity analysis when sparsity correlates with unit or time characteristics.

Under these conditions, PCP recovers the low-rank and sparse components with high probability.

**Theorem 9.2 (Exact recovery)** Suppose Assumptions 32 and 33 hold with  $\rho \leq c_0/(\mu R)$  for a universal constant  $c_0 > 0$ . With  $\lambda = 1/\sqrt{\max(N, T)}$  and  $\sigma = 0$  (noiseless), PCP recovers  $\hat{\mathbf{L}} = \mathbf{L}^*$  and  $\hat{\mathbf{S}} = \mathbf{S}^*$  with probability at least  $1 - c_1(NT)^{-10}$ . See Candès et al. [2011] for the proof.

**Proposition 9.2 (Noisy recovery)** *Under the conditions of Theorem 9.2 with noise  $\|\mathbf{E}\|_F \leq \delta$ , Stable PCP satisfies*

$$\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 + \|\hat{\mathbf{S}} - \mathbf{S}^*\|_F^2 \leq C\delta^2$$

for a constant  $C$  depending on  $\mu$ ,  $R$ , and  $\rho$ .

For causal inference, these results matter because they control how accurately we can recover the untreated potential outcomes  $Y_{it}(0)$  that enter estimands such as ATT (defined in Chapter 5). Robust matrix completion improves causal estimates to the extent that it improves recovery of the low-rank component underlying  $Y_{it}(0)$  without misclassifying treatment effects as outliers. These guarantees do not relax the factor-stability and assignment restrictions from Chapter 8. If treatment assignment is correlated with idiosyncratic shocks that are not captured by the low-rank component, then both standard and robust matrix completion will deliver biased causal estimates, regardless of how well they separate outliers from  $\mathbf{L}^*$ .

## Treatment–Outlier Separation

Let  $\Omega \subseteq \{1, \dots, N\} \times \{1, \dots, T\}$  denote the set of observed cells, and let  $\mathcal{W} \subseteq \Omega$  denote the set of treated cells, as in Chapter 8. For causal work we treat entries in  $\mathcal{W}$  as missing untreated potential outcomes  $Y_{it}(0)$  and fit the robust decomposition on  $\Omega \setminus \mathcal{W}$ . We then target ATT by comparing treated outcomes to completed counterfactuals.

To use robust matrix completion for causal inference, we must guard against mistaking genuine treatment effects for outliers.

**Assumption 34 (Treatment–outlier separation)** Let  $\mathcal{W}$  denote the set of treated cells. Robust matrix completion separates treatment effects from outliers if the following hold:

- (i) The robust decomposition is learned from untreated cells. In practice we fit PCP or SPCP on  $\Omega \setminus \mathcal{W}$ , treating entries in  $\mathcal{W}$  as missing when estimating  $\mathbf{L}$  and  $\mathbf{S}$ .
- (ii) Cell-level treated-minus-untreated contrasts  $\Delta_{it} = Y_{it}(1) - Y_{it}(0)$  for  $(i, t) \in \mathcal{W}$  are moderate relative to the noise level, with  $|\Delta_{it}| \leq C_\tau \sigma$ .
- (iii) Outlier magnitudes satisfy  $|S_{it}^*| > C_S \sigma$  with  $C_S \gg C_\tau$ .

The first condition is a design choice: we deliberately exclude treated cells from the outlier-detection step. The magnitude conditions formalise the idea that outliers are both sparse and much larger than typical treatment effects. Together they ensure that the sparse component captures irregularities such as stockouts and data errors, rather than soaking up genuine treatment-induced changes.

**Caution: Large treatment effects**

Assumption 34 is strong. It requires that outliers be rare and large, and that treatment effects be moderate relative to typical noise. In many marketing settings this will fail. Deep promotions, major product relaunches, or algorithm changes on a platform can create large, sparse spikes in the outcome that look very similar to  $\mathbf{S}^*$ . In those cases, a naive robust matrix-completion procedure may classify part of the treatment effect as an outlier and attenuate ATT.

To reduce this risk, you can learn the robust decomposition using only untreated cells, compare robust and standard matrix-completion estimates to detect large discrepancies, and conduct sensitivity analysis by varying the robustness penalties and checking how much ATT changes. If robust and standard methods produce very different ATT estimates and the difference is highly sensitive to tuning, you should treat the results with caution and revisit both the low-rank specification and the treatment–outlier separation strategy. A practical diagnostic is to examine whether large fitted sparse components  $\hat{S}_{it}$  are disproportionately concentrated in treated cells and line up with known promotion windows. If so, the robust procedure is likely misclassifying genuine treatment effects as outliers, and the resulting ATT estimates should be treated as descriptive rather than causal.

### Comparison: Standard vs Robust Matrix Completion

Table 9.2 compares standard and robust matrix completion. Standard MC refers to the low-rank nuclear-norm estimator introduced in Chapter 8.

**Table 9.2** Standard vs robust matrix completion

Aspect	Standard MC	Robust MC (PCP/SPCP)
Model	$\mathbf{Y} = \mathbf{L}^* + \mathbf{E}$	$\mathbf{Y} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{E}$
Outliers	Absorbed into $\mathbf{L}$ (bias)	Separated into $\mathbf{S}$
Assumptions	Low-rank structure	Low rank + sparsity + incoherence
When to use	Relatively clean data	Stockouts, data errors, strong promotions
Regularisation	$\lambda$ (nuclear norm)	$\lambda$ (trade-off $\mathbf{L}$ vs $\mathbf{S}$ )
Computation	SVD-based	ADMM (iterative)
Interpretability	Factors and loadings	Factors plus identified outliers

## Software Implementation

Several software packages implement robust matrix completion or closely related robust PCA methods. In Python, dedicated packages such as `rpca` implement PCP using ADMM and are convenient for separating low-rank and sparse components in panels with missing data. General tensor and matrix libraries like `tensorly` include robust variants that extend these ideas to higher-order arrays. In R, the `rpca` package provides nuclear-norm-plus- $\ell_1$  formulations for robust PCA, and `pcaMethods` offers robust PCA variants that can serve as building blocks in a causal workflow. MATLAB users can access Augmented Lagrange Multiplier implementations from the authors of Candès et al. [2011]. In all cases, these tools are used in combination with the causal designs and estimands from earlier chapters: robust MC delivers cleaner imputations of  $Y_{it}(0)$ , which you then plug into ATT or dynamic treatment-effect estimands as in Chapters 5 and 10.

## Application: Retail Sales with Stockouts

Consider a grocery retailer with 500 products across 100 stores over 104 weeks. Stockouts affect about 5% of product–store–week combinations. They are not random: popular products are more likely to stock out because high demand depletes inventory, and small stores experience more stockouts because of limited shelf space.

A promotional campaign is applied to 50 products in 20 stores for 4 weeks. As in Chapter 5, we target the average treatment effect on the treated (ATT) for this campaign, where treated cells are the product–store–week combinations that receive the promotion and untreated cells define the comparison surface. The untreated potential outcomes  $Y_{it}(0)$  represent what sales would have been for each treated product–store–week cell in the absence of the promotion.

Standard matrix completion applied directly to  $\mathbf{Y}$  produces biased counterfactuals because stockouts create artificial zeros that violate the low-rank model. Robust MC tackles this by solving SPCP on untreated cells with a regularisation parameter  $\lambda$  chosen by cross-validation. The low-rank component  $\hat{\mathbf{L}}$  then captures underlying demand, while the sparse component  $\hat{\mathbf{S}}$  captures stockouts and other irregularities.

Inspecting  $\hat{\mathbf{S}}$  shows that most non-zero entries are negative, consistent with stockouts creating observed sales below underlying demand. A smaller fraction of large positive entries correspond to data errors or exceptional promotions. The completed low-rank matrix  $\hat{\mathbf{L}}$  provides counterfactual paths  $\hat{L}_{it}$  for treated cells. Comparing observed outcomes to these imputations yields an ATT of around 20% higher sales, whereas standard MC without robustness produces an ATT of about 15%. The difference reflects downward bias from stockouts in the standard approach.

Diagnostics support the robust specification. The distribution of  $\hat{\mathbf{S}}$  is highly sparse, with non-zero entries having mean absolute value roughly half of average weekly sales, confirming that the method is isolating substantial deviations rather than noise. A placebo experiment that randomly designates 5% of untreated cells as artificial outliers shows that the robust procedure correctly flags most of them. Sensitivity analysis indicates that ATT estimates are stable as  $\lambda$  varies over a reasonable range.

Formal inference for the robust ATT proceeds as in the inference chapter. Bootstrap procedures that resample units or clustered time blocks, combined with re-solving the SPCP problem in each bootstrap draw, produce empirical confidence intervals. These intervals can be mapped to the ATT estimand defined in Chapter 5 and to dynamic treatment-effect paths when the promotion has staggered timing.

## 9.4 Matrix Completion with Side Information

### Incorporating Covariates

Marketing panels often come with rich side information. Store characteristics such as size, location, demographics, and format vary across units. Time-varying covariates such as weather, holidays, and macroeconomic conditions vary across periods. Network structure, capturing geographic proximity or competitive relationships, links units whose outcomes move together. This side information provides additional signal that can improve imputation of untreated potential outcomes when combined with the low-rank structure from Chapter 8.

We extend the factor model for untreated potential outcomes by adding an observed-covariate component.

**Definition 9.6 (Covariate-assisted matrix completion model)** The untreated potential outcomes follow

$$Y_{it}(0) = \underbrace{X_{it}^\top \beta}_{\text{covariate effect}} + \underbrace{\lambda_i^\top f_t}_{\text{low-rank component}} + \varepsilon_{it},^1$$

where

- (i)  $X_{it} \in \mathbb{R}^p$  is a vector of observed covariates for unit  $i$  at time  $t$ .
- (ii)  $\beta \in \mathbb{R}^p$  is a coefficient vector.
- (iii)  $\lambda_i \in \mathbb{R}^R$  are unit-specific factor loadings.
- (iv)  $f_t \in \mathbb{R}^R$  are time-specific factors.
- (v)  $\varepsilon_{it}$  are idiosyncratic errors with  $\mathbb{E}[\varepsilon_{it} | X_{it}, \lambda_i, f_t] = 0$ .

The covariate term captures systematic variation explained by observed characteristics, while the low-rank component captures residual co-movement after controlling for covariates.

Covariates in marketing panels naturally fall into three categories. Unit-specific covariates are time-invariant characteristics such as store size, urban or rural location, local demographics like median income and population density, and format (for example, supermarket or convenience store). These enter as  $X_i^\top \beta$ . Time-varying covariates include weather variables such as temperature and precipitation, calendar dummies for holidays, and macroeconomic indicators such as unemployment and consumer confidence. These enter as  $X_{it}^\top \beta$ . Network information, such as geographic proximity or competitive relationships between stores, is best handled through graph-regularised methods that we describe below.

---

<sup>1</sup> As in Chapter 10, where we used  $Y_{it}(\infty)$  to denote the potential outcome under never treated in staggered-adoption designs, we use  $Y_{it}(0)$  here to denote untreated potential outcomes for generic treatment patterns. Matrix-completion methods allow arbitrary treatment paths, but the  $Y_{it}(0)$  shorthand for “no treatment” is standard in this literature; see, for example, Athey et al. [2021].

## Graph-Regularised Matrix Completion

Network structure links units whose outcomes are likely to move together. Graph-regularised matrix completion builds this information into the imputation of  $Y_{it}(0)$  by penalising differences between connected units.

**Definition 9.7 (Graph Laplacian)** Let  $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$  be an undirected graph on units with adjacency matrix  $\mathbf{A}$ . The graph Laplacian is  $\mathbf{L}_G = \mathbf{D}^{\text{deg}} - \mathbf{A}$ , where  $\mathbf{D}^{\text{deg}}$  is the degree matrix with  $D_{ii}^{\text{deg}} = \sum_j A_{ij}$ . The quadratic form  $\mathbf{x}^\top \mathbf{L}_G \mathbf{x} = \sum_{(i,j) \in \mathcal{E}} (x_i - x_j)^2$  measures total squared differences between connected nodes, so penalising this term encourages similar values for connected units.

**Definition 9.8 (Graph-regularised estimator)** The graph-regularised matrix completion estimator solves

$$\min_{\mathbf{M}} \sum_{(i,t) \in \Omega \setminus \mathcal{W}} (Y_{it} - M_{it})^2 + \lambda \|\mathbf{M}\|_* + \gamma \text{tr}(\mathbf{M} \mathbf{L}_G \mathbf{M}^\top),$$

where  $\lambda > 0$  controls low-rank regularisation, as in Section 8.6, and  $\gamma > 0$  controls graph smoothness. The trace term expands to

$$\text{tr}(\mathbf{M} \mathbf{L}_G \mathbf{M}^\top) = \sum_{(i,j) \in \mathcal{E}} \|\mathbf{M}_{i\cdot} - \mathbf{M}_{j\cdot}\|_2^2,$$

which penalises differences in the outcome paths of connected units. When the graph captures geographic or competitive proximity, this encourages similar counterfactual paths for nearby or closely competing stores, improving imputation of  $Y_{it}(0)$  when information is sparse. For causal analysis, the graph  $\mathcal{G}$  must be constructed from pre-treatment information (such as geography or long-run competitive relationships) rather than from realised outcomes, to avoid leaking post-treatment information into the imputation of  $Y_{it}(0)$ .

## Estimation: Two-Step Procedure

A simple and computationally efficient approach is to residualise outcomes on covariates and then apply matrix completion to the residuals.

This two-step estimator is attractive when covariates explain a large share of outcome variation. Jointly estimating  $\beta$  and  $\mathbf{L}$  in a single nuclear-norm-penalised objective, as in Chapter 8, can be more efficient when the low-rank component plays a large role but is computationally more demanding.

## Identification with Covariates

Covariates help primarily with efficiency. They can also reduce bias when they capture variation that would otherwise be absorbed into the low-rank component, but they do not repair violations of the underlying factor-

**Algorithm 1** Two-step covariate-assisted matrix completion

---

**Require:** Outcomes  $\{Y_{it}\}$ , covariates  $\{X_{it}\}$ , observed cells  $\Omega$ , treated cells  $\mathcal{W}$   
**Ensure:** Counterfactuals  $\{\hat{Y}_{it}(0)\}_{(i,t) \in \mathcal{W}}$  and treated-minus-untreated contrasts  $\{\hat{\Delta}_{it}\}$

**Step 1: Estimate covariate effects**

- 1: Regress  $Y_{it}$  on  $X_{it}$  using untreated cells:  $\hat{\beta} = \arg \min_{\beta} \sum_{(i,t) \in \Omega \setminus \mathcal{W}} (Y_{it} - X_{it}^{\top} \beta)^2$ .
- 2: Compute residuals  $\tilde{Y}_{it} = Y_{it} - X_{it}^{\top} \hat{\beta}$  for all  $(i, t) \in \Omega \setminus \mathcal{W}$ .

**Step 2: Matrix completion on residuals**

- 3: Solve  $\hat{\mathbf{L}} = \arg \min_{\mathbf{L}} \sum_{(i,t) \in \Omega \setminus \mathcal{W}} (\tilde{Y}_{it} - L_{it})^2 + \lambda \|\mathbf{L}\|_*$ .
- 4: Select  $\lambda$  by cross-validation (Section 8.6).

**Step 3: Impute counterfactuals and treatment effects**

- 5: **for** each treated cell  $(i, t) \in \mathcal{W}$  **do**
  - 6:     Impute  $\hat{Y}_{it}(0) = X_{it}^{\top} \hat{\beta} + \hat{L}_{it}$ .
  - 7:     Compute  $\hat{\Delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ .
  - 8: **end for**
- 

structure assumptions or selection on unobservables. The key condition is that covariates are exogenous once we account for the factors.

**Assumption 35 (Covariate exogeneity)** Covariates  $X_{it}$  satisfy:

- (i) Strict exogeneity.  $\mathbb{E}[\varepsilon_{it} | X_{i1}, \dots, X_{iT}, \lambda_i, f_1, \dots, f_T] = 0$  for all  $i, t$ .
- (ii) Rank condition.  $\mathbb{E}[X_{it} X_{it}^{\top}]$  is positive definite.
- (iii) No incremental selection on covariates after conditioning on factors. Conditional on  $(\lambda_i, f_t)$ , covariates are not predictive of treatment.

Condition (i) is the usual strict exogeneity requirement from panel regression. Condition (iii) formalises the idea that, after controlling for the latent factors, covariates do not pick up differences between treated and control units that would induce additional selection bias. When this condition fails, including covariates can introduce a form of bad-control bias, as in the general discussion in Chapter 2.

Condition (iii) should be read as a sufficient high-level restriction rather than a literal description of most marketing panels. In practice, store size, demographics, or prior sales are often predictive of treatment even after controlling for latent factors. The key requirement is that any remaining dependence of treatment on  $X_{it}$  does not operate through channels that also drive the counterfactual path  $Y_{it}(0)$  once  $(\lambda_i, f_t)$  are accounted for. When this is doubtful, you should treat covariates as potential bad controls and revert to the design-based guidance in Chapter 2.

A convenient way to summarise the efficiency gain from covariates is to compare the variance of the outcome explained by  $X_{it}$  to the total variance of  $Y_{it}(0)$ .

**Proposition 9.3 (Efficiency gain from covariates)** Suppose Assumption 35 holds and that the linear specification for  $Y_{it}(0)$  is correctly specified. In an idealised setting where matrix-completion error is negligible and ATT is estimated by averaging cell-level treated-minus-untreated contrasts based on the completed panel, the asymptotic mean squared error of the covariate-assisted ATT estimator  $\widehat{\text{ATT}}^{\text{cov}}$  is reduced relative to the standard estimator  $\widehat{\text{ATT}}^{\text{std}}$  roughly in proportion to

$$R_X^2 = \frac{\text{Var}(X_{it}^\top \beta)}{\text{Var}(Y_{it}(0))},$$

which measures the fraction of outcome variance explained by covariates. Higher  $R_X^2$  implies larger potential efficiency gains.

These gains are purely efficiency gains under correct specification and Assumption 35. They do not create identification in settings where selection on unobservables already violates the factor-structure assumptions underpinning matrix completion.

Here  $\widehat{\text{ATT}}^{\text{cov}}$  and  $\widehat{\text{ATT}}^{\text{std}}$  denote the ATT estimators with and without covariate adjustment, respectively, both targeting ATT as defined in Chapter 5. The proposition should be read as an idealised efficiency heuristic rather than a precise finite-sample equality.

#### Caution: Endogenous covariates

When covariates are correlated with treatment beyond what is captured by the factors, they can introduce confounding instead of removing it. If treatment is more likely in large stores and store size enters the model as a covariate, the regression term  $X_{it}^\top \beta$  may soak up part of the true treatment effect and push ATT estimates downward. The same concern applies to variables such as prices or advertising that may themselves be affected by treatment or share the same unobserved drivers. In practice you should include covariates that are known to affect outcomes but are plausibly exogenous with respect to treatment assignment, such as weather, holidays, and local demographics. Covariates that may be affected by treatment or that capture managerial decisions on the treatment path belong in the “bad control” category discussed in Chapter 2 and should be excluded or handled with explicit identification strategies.

### Comparison: Standard vs Covariate-Assisted Matrix Completion

Table 9.3 compares standard and covariate-assisted approaches. Standard MC refers to the low-rank nuclear-norm estimator without covariates, while covariate-assisted MC augments this with regression on observed covariates.

### Software Implementation

Several software packages support covariate-assisted matrix completion or closely related models. In Python, libraries such as `fancyimpute` implement matrix completion routines that can be combined with separate regression steps on covariates, while recommender-system libraries like `surprise` support side information for users and items in joint factorisation models. In R, `softImpute` provides nuclear-norm-regularised matrix completion that can be paired with standard regression tools such as `lm()` in the two-step procedure, and

**Table 9.3** Standard vs covariate-assisted matrix completion

Aspect	Standard MC	Covariate-assisted MC
Model	$\mathbf{Y} = \mathbf{L} + \mathbf{E}$	$\mathbf{Y} = X\beta + \mathbf{L} + \mathbf{E}$
Side information	Ignored	Incorporated
Efficiency	Baseline	Potentially improved when $R_X^2$ is large
When to use	No relevant covariates	Rich unit/time covariates
Heterogeneity	Implicit in loadings	Explicit via covariate interactions
Computation	Single-step MC	Two-step (regression plus MC)
Risk	None from covariates	Bad-control bias if covariates endogenous

`cmfrec` implements collective matrix factorisation with side information. For graph-regularised matrix completion, Python packages such as `pygsp` help construct and manipulate graphs whose Laplacians can be fed into custom optimisation routines, and R packages like `igraph` serve a similar role. In all cases, these tools are numerical building blocks. They do not change the underlying causal estimands or assumptions, which continue to follow Chapters 5 and 16; covariate-assisted matrix completion only alters how we approximate  $Y_{it}(0)$ .

### Application: Store Characteristics and Demand

Consider a retailer with 200 stores observing weekly sales of 100 products over 52 weeks. Store characteristics include size measured in square footage, an urban or rural location flag, and local demographics such as median income. A promotional campaign is applied to 20 products in 40 stores for 4 weeks, and we are interested in the ATT for this campaign.

A baseline analysis using standard matrix completion ignores store characteristics. It estimates a low-rank structure that captures co-movement across stores and products, but it does not account for systematic level differences driven by size and demographics. As a result, part of the cross-sectional variation that could be explained by observable store features is left to the low-rank component.

The covariate-assisted approach first regresses sales on store size, location, and demographics using untreated cells. Suppose this first-step regression achieves an  $R^2$  of about 0.60, meaning covariates explain roughly 60 per cent of the variation in untreated outcomes. We then apply matrix completion to the residuals, letting the low-rank component capture remaining co-movement that is not explained by observables, and combine the covariate predictions with the completed residuals to form counterfactuals  $\hat{Y}_{it}(0)$  for treated cells.

In this scenario, the covariate-assisted estimator might yield an ATT of around 18 per cent with a 95 per cent confidence interval from roughly 14 to 22 per cent, compared with about 15 per cent (95 per cent interval from roughly 10 to 20 per cent) from standard matrix completion. The difference reflects store heterogeneity

that the covariate-assisted approach accounts for explicitly. Treatment effects can vary substantially by store size: large stores may see increases near 25 per cent, medium stores around 15 per cent, and small stores around 10 per cent. Standard matrix completion, which averages over these differences implicitly, tends to report a single effect that masks this heterogeneity.

Diagnostics support the value of covariates. Out-of-sample  $R^2$  for predictions of untreated outcomes might rise from about 0.65 for standard MC to about 0.75 for covariate-assisted MC. This ten-percentage-point improvement signals that covariates carry information that the low-rank structure alone was not capturing. Inference for the ATT and heterogeneous effects follows the bootstrap or other procedures developed in Section 9.6 and Chapter 16, applied to the completed panels.

## 9.5 Time-Varying Rank and Non-Stationary Panels

### When Rank Changes Over Time

The low-rank assumption in Chapter 8 requires that the factor structure for untreated potential outcomes is stable over time. The number of factors  $R$  and the loadings  $\lambda_i$  are constant, and the factors  $f_t$  follow a stationary process. This is often violated when structural breaks occur. A recession can introduce new common shocks linked to unemployment and credit constraints. A platform algorithm change can alter user behaviour by creating new engagement patterns. The COVID-19 pandemic created a major break in demand through remote work, supply-chain disruptions, and shifts in consumption patterns.

When the effective rank changes over time, standard matrix completion extrapolates the pre-break factor structure into the post-break period and can produce biased imputations of  $Y_{it}(0)$ . Time-varying rank models adapt to non-stationary environments by allowing the factor structure, and in particular the rank, to change over time while still exploiting low-rank structure within relatively stable regimes.

**Definition 9.9 (Time-varying factor model)** The untreated potential outcomes follow a time-varying factor structure

$$Y_{it}(0) = \lambda_i(t)^\top f_t + \varepsilon_{it},$$

where

- (i)  $\lambda_i(t) \in \mathbb{R}^{R_t}$  are time-varying factor loadings for unit  $i$ .
- (ii)  $f_t \in \mathbb{R}^{R_t}$  are factors at time  $t$ .
- (iii) the rank  $R_t$  may vary with time, taking values in  $\{R_{\min}, \dots, R_{\max}\}$ .

See Su and Wang [2017] for theoretical properties of time-varying factor models.

### Identification Under Non-Stationarity

For causal inference we still require the basic potential-outcomes structure from earlier chapters: observed outcomes satisfy  $Y_{it} = Y_{it}(1)$  for treated cells and  $Y_{it} = Y_{it}(0)$  for untreated cells, and ATT and related estimands are defined as in Chapter 5. Time variation in rank does not change the estimand, but it complicates how we recover  $Y_{it}(0)$  from incomplete panels. These models do not relax the identification assumptions from Chapter 8. We still require that, conditional on the evolving low-rank component, the pattern of treated and untreated cells is compatible with treating treated entries as missing  $Y_{it}(0)$ . Time-varying rank changes how we approximate  $Y_{it}(0)$  in non-stationary environments, not the underlying design assumptions.

Identification under non-stationarity requires that the factor structure evolves smoothly between structural breaks and that the pattern of treated and untreated cells is compatible with learning the evolving low-rank component. Within each regime between breaks, the factor model behaves like the static model in Chapter 8. Across regimes, loadings, factors, and rank drift.

**Assumption 36 (Smooth factor evolution)** The factor structure evolves smoothly over time, apart from a small number of structural breaks:

- (i) Loading stability. When  $R_{t-1} = R_t$ , loadings satisfy  $\|\lambda_i(t) - \lambda_i(t-1)\| \leq \delta_\lambda$  for all  $i$ .
- (ii) Factor stability. Factors satisfy  $\|f_t - f_{t-1}\| \leq \delta_f$ .
- (iii) Rank changes. Rank changes occur at isolated times  $t$  in a set  $\mathcal{B} = \{t_1, \dots, t_K\}$  with  $|\mathcal{B}| \ll T$ .

The smoothness parameters  $\delta_\lambda$  and  $\delta_f$  control the rate at which the factor structure drifts between breaks.

Combined with the missing-at-random assumptions for untreated entries from Chapter 8, this condition allows us to use nearby periods to learn the evolving low-rank structure for  $Y_{it}(0)$ , and to update that structure when evidence of a structural break appears.

## Subspace Tracking Methods

Let  $D_{it}$  denote the treatment indicator as in earlier chapters, with  $D_{it} = 1$  for treated cells and  $D_{it} = 0$  otherwise. Let  $\Omega \subseteq \{1, \dots, N\} \times \{1, \dots, T\}$  denote the set of observed cells and let  $\mathcal{W} \subseteq \Omega$  denote the treated cells. For causal work we treat  $Y_{it}(0)$  as missing on  $\mathcal{W}$  and learn the evolving low-rank structure from  $\Omega \setminus \mathcal{W}$ .

Online matrix-completion methods update the low-rank estimate as new periods arrive, using temporal smoothing to encourage stability while remaining flexible enough to adapt at breaks.

**Definition 9.10 (Online matrix-completion objective)** At time  $t$ , an online estimator for the low-rank component  $\mathbf{L}_t$  may solve

$$\hat{\mathbf{L}}_t = \arg \min_{\mathbf{L}} \sum_{(i,s) \in \Omega \setminus \mathcal{W}, s \leq t} (Y_{is} - L_{is})^2 + \lambda_t \|\mathbf{L}\|_* + \alpha \|\mathbf{L} - \hat{\mathbf{L}}_{t-1}\|_F^2,$$

where

- (i)  $\lambda_t > 0$  is a time-varying regularisation parameter, selected as in Section 8.6.
- (ii)  $\alpha > 0$  is a temporal smoothing parameter.
- (iii)  $\hat{\mathbf{L}}_{t-1}$  is the previous-period estimate, padded as needed when the effective rank changes.

The objective restricts the fit to untreated observed cells, so that the evolving low-rank structure is learned from  $Y_{is}(0)$  on  $\Omega \setminus \mathcal{W}$ . Here  $\mathbf{L}_t$  denotes the current low-rank estimate based on data up to time  $t$ , so the objective pools reconstruction error from periods 1 through  $t$  while penalising deviations from the previous estimate  $\hat{\mathbf{L}}_{t-1}$ .

Large values of  $\alpha$  enforce smooth evolution but may make the estimator slow to adapt to breaks. Small values allow rapid adaptation but risk overfitting noise. In practice, rolling-window cross-validation or information criteria applied to reconstruction error provide guidance on the choice of  $\alpha$ .

The proximal operator for the nuclear norm is discussed in Section 9.7, where we describe singular-value thresholding and related algorithms. Here it provides a convenient shorthand for updating the low-rank estimate at each step.

**Algorithm 2** Online matrix completion with subspace tracking

---

**Require:** Panel data  $\{Y_{it}\}$ , treatment indicators  $\{D_{it}\}$ , initial rank  $R_0$ , regularisation schedule  $\{\lambda_t\}$ , smoothing parameter  $\alpha$

**Ensure:** Time-varying low-rank estimates  $\{\hat{\mathbf{L}}_t\}$ , detected change points  $\hat{\mathcal{B}}$ , and counterfactuals for treated cells

- 1: Estimate an initial low-rank matrix  $\hat{\mathbf{L}}_1$  from the first  $T_0$  periods using standard matrix completion on untreated cells.
- 2: Set  $R_1 = R_0$ .
- 3: **for**  $t = T_0 + 1, \dots, T$  **do**
- 4:     Update the low-rank estimate using a proximal gradient step of the online objective, for example

$$\hat{\mathbf{L}}_t = \text{prox}_{\lambda_t \|\cdot\|_*}(\hat{\mathbf{L}}_{t-1} - \eta \nabla \ell_t(\hat{\mathbf{L}}_{t-1})),$$

where  $\ell_t$  is the squared loss on untreated cells at time  $t$  and  $\eta$  is a step size.

- 5:     Monitor reconstruction error and compute a statistic  $\Delta_t$  based on the change in fit between periods  $t-1$  and  $t$ .
  - 6:     **if**  $\Delta_t$  exceeds a chosen threshold **then**
  - 7:         Flag  $t$  as a potential change point and add it to  $\hat{\mathcal{B}}$ .
  - 8:         Re-estimate the effective rank  $R_t$  using cross-validation or information criteria on a window around  $t$ .
  - 9:     **end if**
  - 10:    For treated cells  $(i, t)$ , impute  $\hat{Y}_{it}(0) = \hat{L}_{it}$  and compute treated-minus-untreated contrasts  $\hat{\Delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ .
  - 11: **end for**
- 

**Proposition 9.4 (Change-point detection via reconstruction error)** Suppose Assumption 36 holds with true change points  $\mathcal{B} = \{t_1^*, \dots, t_K^*\}$ , and that the low-rank component is consistently estimated within each regime. Standard change-point detection methods applied to a reconstruction-error series built from  $\hat{\mathbf{L}}_t$  can consistently recover the number and approximate locations of the structural breaks under suitable regularity conditions. See, for example, Su and Wang [2017] and the general change-point literature.

A practical complication arises when major structural breaks coincide with treatment adoption. If treated units adopt a new advertising strategy at the same time that a recession or platform change shifts the factor structure, then even a time-varying low-rank model cannot fully disentangle treatment effects from break-induced changes in  $Y_{it}(0)$ . In that case, credible causal interpretation requires either external information about the break or a design that provides untreated units exposed to the same break but not to the treatment.

The proposition should be read as a qualitative guide rather than a complete theory: the main message is that large, persistent jumps in reconstruction error are evidence of breaks in the factor structure and should trigger re-estimation of the low-rank model.

### Comparison: Standard vs Time-Varying Matrix Completion

Table 9.4 contrasts standard and time-varying approaches. Standard MC assumes fixed  $R$ ,  $\lambda_i$ , and  $f_t$ , and can be seriously biased after a structural break. Time-varying MC allows  $R_t$ ,  $\lambda_i(t)$ , and  $f_t$  to evolve and uses change-point detection to adapt when major shocks, such as recessions or pandemics, occur.

**Table 9.4** Standard vs time-varying matrix completion

Aspect	Standard MC	Time-varying MC
Factor structure	Fixed $R$ , $\lambda_i$ , $f_t$	$R_t$ , $\lambda_i(t)$ , $f_t$ vary over time
Structural breaks	Not handled (bias)	Detected and adapted to
Assumptions	Stable factors	Smooth evolution plus isolated breaks
When to use	Stable environments	Recessions, pandemics, policy changes
Computation	Single estimation	Online or sequential updates
Parameters	$\lambda$ (regularisation)	$\lambda_t$ , $\alpha$ (smoothing and adaptation)
Risk	Bias after breaks	Overfitting if $\alpha$ too small

### Software Implementation

Several software packages support time-varying factor models and change-point detection that can be combined with matrix completion. In Python, `ruptures` provides a flexible change-point detection library with algorithms such as PELT and binary segmentation that can be applied to reconstruction-error series from matrix completion. Packages like `dynfact` implement dynamic factor models with time-varying loadings in a state-space framework, which can be used to model evolving factors in smaller panels. Custom implementations using `scipy.optimize` or `pytorch` can handle the online matrix-completion objective directly.

In R, `changepoint` and `bfast` offer change-point detection for time series that can be applied to aggregate or factor-level series, `tseries` includes structural-break tests, and `dfms` implements dynamic factor models in state space. MATLAB users can employ algorithms such as GRASTA (Grassmannian Rank-One Update Subspace Tracking) for online subspace tracking, or build custom solvers using the Optimization Toolbox. In all cases, these time-varying tools are layered on top of the causal designs and estimands from earlier chapters: they provide better imputations of  $Y_{it}(0)$  in non-stationary environments, which you then plug into ATT and dynamic-treatment estimators as in Chapters 5 and 10.

## Application: COVID-19 Structural Break

Consider an e-commerce platform that tracks daily user engagement, such as clicks and purchases, for 1,000 users over 365 days. The COVID-19 pandemic creates a clear structural break in mid-March 2020. Before the pandemic, a low-rank factor model with three factors captures most of the variation in engagement: one factor drives day-of-week patterns, another captures seasonality across the year, and a third reflects persistent differences in user preferences.

After the onset of COVID-19, two new forces become important. Remote work shifts activity towards daytime, changing the loadings on days and times of day. Supply-chain disruptions and changes in product availability alter browsing and purchasing patterns. The effective rank of the engagement process increases as these new common shocks appear. In April 2020 the platform rolls out a personalised recommendation feature for a subset of users. We are interested in the ATT of this feature on engagement, defined as the average difference between  $\hat{Y}_{it}(1)$  and the untreated potential outcomes  $\hat{Y}_{it}(0)$  for treated users in the post-launch period.

Standard matrix completion that fits a static three-factor model on the full sample struggles in this setting. It tries to explain post-COVID variation using the pre-COVID factors, so the imputed counterfactuals for treated users in April and beyond are contaminated by misspecification. A time-varying matrix-completion approach instead learns the low-rank structure on pre-COVID data, monitors reconstruction error over time, and detects a sharp increase in error around mid-March 2020. This triggers a re-estimation of the factor model with a higher effective rank in the post-COVID regime.

With the time-varying model, the completed panel  $\hat{Y}_{it}(0)$  for treated users in April and later incorporates the new remote-work and supply-chain factors, so the counterfactual paths better reflect what engagement would have looked like without the recommendation feature but with COVID-19 dynamics. In a representative application, the time-varying estimator might yield an ATT of about 12 per cent higher engagement with a 95 per cent confidence interval from roughly 8 to 16 per cent, compared with around 8 per cent (95 per cent interval from roughly 4 to 12 per cent) from a static three-factor model. The static estimate is biased downward because it forces the pre-COVID factor structure onto the post-COVID period.

Diagnostics support the time-varying specification. Change-point detection applied to the reconstruction-error series flags mid-March 2020 as a break. Model fit improves once the new regime is in place: the pre-break factor model explains, say, around 80 per cent of variation, while the post-break model with five factors explains closer to 85 per cent. Sensitivity analysis shows that ATT estimates are stable over a range of smoothing parameters, for example  $\alpha$  between 0.1 and 1.0. Confidence intervals can be constructed using block bootstrap methods with weekly blocks, as described in Section 9.6 and Chapter 16, applied to the completed panels.

## 9.6 Bayesian Matrix Completion

### Bayesian Framework for Uncertainty Quantification

Bayesian matrix completion provides a coherent way to quantify uncertainty about the low-rank structure that underpins our counterfactuals. We place priors on the rank, loadings, and factors for the untreated potential outcomes, then use the observed panel to update these priors to a posterior distribution. This yields posterior distributions and credible intervals for counterfactual paths and treatment effects that reflect estimation uncertainty under the same identification assumptions used in the factor-model chapter.

**Definition 9.11 (Bayesian factor model)** A basic Bayesian specification for untreated potential outcomes is

$$\begin{aligned} Y_{it}(0) \mid \lambda_i, f_t, \sigma^2 &\sim \mathcal{N}(\lambda_i^\top f_t, \sigma^2), \\ \lambda_i \mid \Sigma_\lambda &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_\lambda), \quad i = 1, \dots, N, \\ f_t \mid \Sigma_f &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_f), \quad t = 1, \dots, T, \\ \Sigma_\lambda &\sim \text{IW}(\nu_\lambda, \Psi_\lambda), \\ \Sigma_f &\sim \text{IW}(\nu_f, \Psi_f), \\ \sigma^2 &\sim \text{IG}(a_\sigma, b_\sigma), \end{aligned}$$

where IW denotes the inverse-Wishart distribution and IG denotes the inverse-gamma distribution.

This prior structure encodes the same low-rank factor model for  $Y_{it}(0)$  as in Chapter 8, but treats the factors, loadings, and noise variance as random rather than fixed unknowns.

**Definition 9.12 (ARD prior for rank selection)** The inverse-Wishart specification treats  $\Sigma_\lambda$  and  $\Sigma_f$  as full covariance matrices. An alternative, more parsimonious approach is to use Automatic Relevance Determination (ARD) priors that operate factor by factor. For a maximum rank  $R_{\max}$ ,

$$\lambda_{ir} \mid \tau_r \stackrel{\text{ind}}{\sim} \mathcal{N}(0, \tau_r^{-1}), \quad \tau_r \stackrel{\text{iid}}{\sim} \text{Gamma}(a_\tau, b_\tau), \quad r = 1, \dots, R_{\max},$$

where the precision parameters  $\tau_r$  are shared across units. Large values of  $\tau_r$  shrink the  $r$ th factor towards zero across all units, effectively reducing the rank.

A simple way to summarise the effective rank is to count how many factors have posterior variance above a small threshold. If  $\bar{\tau}_r$  denotes the posterior mean of  $\tau_r$ , a convenient summary is

$$\hat{R}_{\text{eff}} = \sum_{r=1}^{R_{\max}} \mathbf{1} \left\{ \frac{1}{\bar{\tau}_r} > \epsilon \right\},$$

with  $\epsilon$  chosen as a small fraction of a typical loading variance, for example one per cent of  $\text{Var}(\hat{\lambda})$ . This threshold is an implementation choice rather than a theoretical constant and can be tuned using held-out prediction error.

Under standard regularity conditions on the low-rank structure and priors, Bayesian factor models with ARD priors concentrate posterior mass near the true rank and loading structure as the panel grows. The precise conditions and proofs are technical. Here the main message is that, with a well-specified prior and sufficient data, the posterior focuses on factor structures that reconstruct  $Y_{it}(0)$  well.

**Proposition 9.5 (Posterior concentration and counterfactuals)** *Suppose the untreated potential outcomes satisfy a low-rank factor model as in Chapter 8 with true rank  $R^*$ , and that the Bayesian specification with ARD prior is correctly specified with  $R_{\max} \geq R^*$ . As  $N$  and  $T$  grow proportionally, the posterior distribution concentrates around factor structures that recover the low-rank component of  $Y_{it}(0)$  with vanishing error. In particular, posterior means of counterfactuals  $Y_{it}(0)$  for untreated cells converge to their true values, and posterior credible intervals for smooth functionals such as ATT have approximate frequentist coverage under the model.*

This proposition is intentionally high level. It signals that Bayesian matrix completion can match the large-sample performance of frequentist methods when the model is correctly specified, without over-claiming specific rates or exact coverage. These results all hold under the maintained low-rank factor model and the same factor-stability and sampling assumptions as in Chapter 8. Bayesian methods do not repair violations of these assumptions. They only provide a coherent way to propagate estimation uncertainty for  $Y_{it}(0)$  and ATT when the model is well specified.

**Proposition 9.6 (Credible intervals for treatment effects)** *Let  $\{\text{ATT}^{(s)}\}_{s=1}^S$  denote posterior draws of ATT from a Markov chain Monte Carlo (MCMC) run based on the Bayesian factor model. The equal-tailed  $(1 - \alpha)$  credible interval is*

$$\text{CI}_{1-\alpha}(\text{ATT}) = [\text{ATT}^{(\lfloor S\alpha/2 \rfloor)}, \text{ATT}^{(\lceil S(1-\alpha/2) \rceil)}],$$

where  $\text{ATT}^{(k)}$  is the  $k$ th order statistic of the draws. When the model is correctly specified and the posterior concentrates as in Proposition 9.5, these credible intervals provide a good approximation to frequentist confidence intervals for ATT. Their coverage properties degrade when the factor model is misspecified.

Here ATT is the average treatment effect on the treated as defined in Chapter 5, and posterior draws  $\text{ATT}^{(s)}$  are obtained by imputing untreated potential outcomes  $Y_{it}(0)$  for treated cells in each MCMC draw and averaging the resulting treatment effects.

## Gibbs Sampler Algorithm

Posterior inference typically proceeds via MCMC. A Gibbs sampler exploits the conditional conjugacy of the Gaussian factor model and ARD priors.

Let  $\Omega \subseteq \{1, \dots, N\} \times \{1, \dots, T\}$  denote the set of observed cells and let  $\mathcal{W} \subseteq \Omega$  denote the treated cells. As in Chapter 8, the Bayesian matrix-completion step conditions on  $\Omega \setminus \mathcal{W}$  to learn the factor structure for  $Y_{it}(0)$  and imputes counterfactuals for cells in  $\mathcal{W}$ .

---

**Algorithm 3** Gibbs sampler for Bayesian matrix completion

---

**Require:** Observed outcomes  $\{Y_{it}\}_{(i,t) \in \Omega \setminus \mathcal{W}}$ , maximum rank  $R_{\max}$ , hyperparameters  $(\nu_\lambda, \Psi_\lambda, \nu_f, \Psi_f, a_\sigma, b_\sigma, a_\tau, b_\tau)$ , total iterations  $S$ , burn-in  $B$

**Ensure:** Posterior draws of ATT, credible intervals, and effective rank summary  $\hat{R}_{\text{eff}}$

- 1: Initialise loadings, factors, precisions, and variance  $(\lambda_i^{(0)}, f_t^{(0)}, \tau_r^{(0)}, \sigma^{2(0)})$  using a low-rank SVD on untreated cells.
- 2: **for**  $s = 1, \dots, S$  **do**
- 3:     For each unit  $i$ , sample  $\lambda_i^{(s)}$  from its Gaussian full conditional given  $\mathbf{Y}$ ,  $\mathbf{f}^{(s-1)}$ ,  $\tau^{(s-1)}$ , and  $\sigma^{2(s-1)}$ .
- 4:     For each time  $t$ , sample  $f_t^{(s)}$  from its Gaussian full conditional given  $\mathbf{Y}$ ,  $\boldsymbol{\lambda}^{(s)}$ , and  $\sigma^{2(s-1)}$ .
- 5:     For each factor  $r$ , sample the ARD precision  $\tau_r^{(s)}$  from its Gamma full conditional based on the current loadings.
- 6:     Sample the noise variance  $\sigma^{2(s)}$  from its inverse-gamma full conditional.
- 7:     For treated cells  $(i, t)$  in  $\mathcal{W}$ , impute untreated potential outcomes by drawing  $\hat{Y}_{it}^{(s)}(0)$  from  $\mathcal{N}(\lambda_i^{(s)\top} f_t^{(s)}, \sigma^{2(s)})$ .
- 8:     Compute the draw of ATT as

$$\text{ATT}^{(s)} = \frac{1}{|\mathcal{W}|} \sum_{(i,t) \in \mathcal{W}} (Y_{it} - \hat{Y}_{it}^{(s)}(0)).$$

- 9: **end for**
- 

After discarding an initial burn-in period and thinning if necessary, the retained draws  $\{\text{ATT}^{(s)}\}$  provide a Monte Carlo approximation to the posterior distribution of ATT.

Variational inference offers a faster but approximate alternative. It replaces the exact posterior with a factorised Gaussian approximation and chooses its parameters to minimise Kullback–Leibler divergence. Variational methods are attractive for exploratory analysis or very large panels but tend to underestimate posterior uncertainty. For final reporting of treatment effects we recommend MCMC when computationally feasible.

## Hierarchical Models

Hierarchical Bayesian models share information across related panels. In retail settings, this is natural when multiple product categories display similar seasonal patterns and macro shocks but differ in category-specific dynamics. A hierarchical factor model lets us borrow strength across categories while still allowing each category to deviate from the chain-wide average.

**Definition 9.13 (Hierarchical factor model)** Suppose we observe untreated potential outcomes  $Y_{ijt}(0)$  for product  $i$  in category  $j$  at time  $t$ . A hierarchical specification is

$$Y_{ijt}(0) = \lambda_{ij}^\top f_{jt} + \varepsilon_{ijt}, \quad f_{jt} \sim \mathcal{N}(\mu_t, \Sigma), \quad \mu_t \sim \mathcal{N}(0, \Sigma_0),$$

where  $f_{jt}$  is a category-level factor and  $\mu_t$  is a population-level factor capturing chain-wide shocks.

Here  $i$  indexes products within category  $j$  and  $t$  indexes time, so  $Y_{ijt}(0)$  is a three-way panel in the sense of Section 9.2, but we model it as a collection of category-specific matrices to keep the hierarchical structure simple.

This specification shrinks category-level factors towards the population mean. Categories with limited data, such as niche or newly launched lines, then borrow information from larger categories whose factor structures are better identified.

In treatment-effect applications, we typically treat  $Y_{ijt}(0)$  as the model for untreated potential outcomes and use the hierarchical structure to improve imputation for categories with few treated products or short pre-treatment histories.

## Model Selection and Comparison

Bayesian model selection uses the marginal likelihood, which penalises complexity and favours parsimonious models. Bayes factors compare marginal likelihoods,  $\text{BF}_{12} = p(\mathbf{Y} | M_1)/p(\mathbf{Y} | M_2)$ , with values above about 10 often taken as suggestive evidence for one model over another when priors are well calibrated.

In practice, exact marginal-likelihood computation for matrix-completion models can be demanding. Approximations based on bridge sampling, harmonic-mean estimators, or information criteria are often used instead. Posterior predictive checks provide a complementary diagnostic by comparing observed data to data simulated from the posterior predictive distribution. Large discrepancies in simple summaries (means, variances, autocorrelations) indicate model misspecification.

## Frequentist vs Bayesian Inference in This Setting

Chapter 16 compares frequentist and Bayesian inference in detail. For matrix completion, the main distinction is how we construct uncertainty intervals for treatment effects.

Frequentist approaches treat the low-rank component as a fixed but unknown object and use bootstrap or asymptotic approximations to quantify uncertainty in ATT, for example by resampling units or time blocks and re-estimating the matrix-completion model in each bootstrap sample. Bayesian approaches place a prior on the factor structure, use MCMC or variational methods to approximate the posterior, and derive credible intervals directly from posterior draws of ATT.

Table 9.5 summarises these differences in the specific context of matrix completion.

**Table 9.5** Frequentist vs Bayesian inference for matrix completion

Aspect	Frequentist (bootstrap)	Bayesian (MCMC)
Uncertainty measure	Confidence intervals for ATT	Credible intervals for ATT
Interpretation	Long-run coverage under resampling scheme	Posterior probability under the model
Rank selection	Cross-validation or information criteria	ARD prior and posterior concentration
Computation	Repeated re-fitting (parallelisable)	Sequential MCMC or variational updates
Small-sample behaviour	Coverage can be distorted	Sensitive to priors and can stabilise estimates
When useful	Very large panels, weak prior information	Smaller panels or when prior knowledge is available
Software	Custom code, <code>boot</code>	<code>Stan</code> , <code>pymc</code> , custom Gibbs samplers

## Software Implementation

Several software frameworks support Bayesian matrix completion or closely related factor models. In Python, `pymc` and `tensorflow-probability` allow you to specify custom factor models with ARD priors and fit them using MCMC or variational inference. Packages such as `bpmf` implement Bayesian probabilistic matrix factorisation with Gibbs samplers designed for low-rank panels. In R, `brms` and `blavaan` provide high-level interfaces to Stan for specifying latent-factor models, and `MCMCpack` includes more general-purpose MCMC routines. Stan itself offers a flexible language for implementing the Bayesian factor model directly, with `rstan` and `pystan` providing interfaces for R and Python.

In all these cases, the Bayesian machinery is used to approximate the posterior distribution of  $Y_{it}(0)$  under a low-rank factor model and to propagate that uncertainty through to ATT and related estimands defined in Chapters 5 and 10.

## Application: Hierarchical Demand Across Categories

Consider a grocery retailer with 10 product categories, each containing about 50 products, observed over 52 weeks. A promotional campaign is applied to 5 products in each category for 4 weeks. We are interested in ATT at the category level and in how uncertainty behaves across large and small categories.

A first approach fits separate factor models to each category using standard matrix completion. Large categories with many products and long histories yield reasonably precise estimates, but small categories produce noisy factor estimates and wide confidence intervals for ATT. A second approach pools all categories into a single factor model, which improves precision for small categories but risks misspecifying category-specific dynamics.

A hierarchical Bayesian matrix-completion model strikes a middle ground. It introduces population-level factors that capture chain-wide shocks, such as overall seasonality or macroeconomic shifts, and category-level deviations that capture differences in category-specific demand cycles. The hierarchical prior shrinks category-level factors towards the population mean, with the strength of shrinkage depending on how much data each category contributes.

In a representative analysis, ATT estimates for large categories (more than 40 products) might be around 15 per cent under both standard and hierarchical models, but the hierarchical model reduces standard errors slightly by pooling information, for example from 2.5 percentage points to roughly 2.3. For medium-sized categories (20 to 40 products), hierarchical shrinkage can reduce standard errors more substantially, say from about 4.0 to 3.2. For small categories (fewer than 20 products), where separate factor models are very noisy, the hierarchical model can cut standard errors by roughly a third, for example from 6.5 to 4.5 percentage points, by borrowing strength from larger categories that share similar factor structure.

The posterior also yields 95 per cent credible intervals that reflect this category-size gradient. A large category with around 50 products might have an ATT near 15 per cent with a 95 per cent credible interval from roughly 12 to 18 per cent. A small category with about 15 products might have an ATT near 12 per cent but with a wider 95 per cent credible interval from roughly 8 to 16 per cent. Bayes factors or marginal-likelihood approximations often favour the hierarchical model over a fully pooled alternative, reflecting its ability to capture both common and category-specific components without overfitting.

## 9.7 Computational Methods for Large-Scale Problems

### Scalability Challenges

Large-scale marketing panels pose serious computational challenges. A panel with 10,000 products, 1,000 stores, and 100 weeks contains one billion potential product–store–week cells. Even if only a fraction are observed, naive matrix-completion algorithms based on dense nuclear-norm minimisation and full singular-value decompositions have per-iteration complexity on the order of  $\min(N^2T, NT^2)$ , which is infeasible at this scale.

Memory constraints also bind. Storing a dense matrix of size  $10,000 \times 100,000$  (products by store–weeks) in double precision requires about eight gigabytes. Most marketing panels are highly sparse, because many product–store combinations are never observed or are observed only intermittently, but basic algorithms do not automatically exploit that sparsity. Using sparse storage formats, such as compressed sparse row, reduces memory to order  $|\Omega|$ , where  $\Omega$  is the set of observed cells. In causal applications we fit the low-rank model on  $\Omega \setminus \mathcal{W}$ , where  $\mathcal{W}$  denotes treated cells. Scalable algorithms must exploit both sparsity and low rank.

In what follows we compare algorithms that all target the same low-rank solution described in Chapter 8. They differ only in how they approximate that solution and how their computational cost scales with the panel size. When you use any of these methods to impute  $Y_{it}(0)$ , the causal interpretation of ATT and related estimands remains the same. The main question is whether numerical approximation error is negligible relative to sampling variability.

### Efficient Algorithms

We summarise rough per-iteration complexity for commonly used algorithms. These orders suppress constants and implementation details but provide guidance on how each method scales with panel size, rank, and sparsity.

**Proposition 9.7 (Computational complexity, rough orders)** *For an  $N \times T$  panel with rank  $R$  and  $|\Omega|$  observed entries, typical per-iteration computational costs are of the following orders:*

- (i) nuclear-norm minimisation via dense SVD: about  $\min(N^2T, NT^2)$ .
- (ii) Soft-Impute with truncated SVD: about  $|\Omega|R + (N + T)R^2$ .
- (iii) alternating least squares (ALS): about  $|\Omega|R^2 + (N + T)R^3$ .
- (iv) stochastic gradient descent (SGD) on mini-batches of size  $B$ : about  $BR$  per update.
- (v) MCMC (Gibbs sampling) for Bayesian matrix completion: about  $(N + T)R^2$  per posterior draw.

*These expressions describe orders of magnitude rather than exact equalities and must be interpreted together with the number of iterations required for convergence. These orders assume that implementations exploit sparse storage for the observed set  $\Omega$  rather than forming dense  $N \times T$  matrices in memory. From a causal*

*perspective, the relevant diagnostic is that ATT and related estimands remain stable as you tighten convergence tolerances or increase iteration counts, so that optimisation error is negligible relative to sampling uncertainty.*

---

**Algorithm 4** Soft-Impute for matrix completion

---

**Require:** Observed entries  $\{Y_{it}\}_{(i,t) \in \Omega}$ , regularisation parameter  $\lambda$ , target rank  $R$ , convergence tolerance  $\epsilon$

**Ensure:** Low-rank estimate  $\hat{\mathbf{L}}$

- 1: Initialise  $\hat{\mathbf{L}}^{(0)}$  as the zero matrix.
  - 2: **for**  $k = 1, 2, \dots$  until convergence **do**
  - 3:   Fill in missing entries using the current estimate:  $\tilde{\mathbf{Y}}^{(k)} = P_\Omega(\mathbf{Y}) + P_{\Omega^c}(\hat{\mathbf{L}}^{(k-1)})$ , where  $P_\Omega$  and  $P_{\Omega^c}$  project onto observed and missing cells.
  - 4:   Compute a rank- $R$  truncated SVD of  $\tilde{\mathbf{Y}}^{(k)}$ :  $\tilde{\mathbf{Y}}^{(k)} \approx \mathbf{U}\Sigma\mathbf{V}^\top$ .
  - 5:   Apply soft-thresholding to the singular values: set  $\hat{\mathbf{L}}^{(k)} = \mathbf{U} \operatorname{diag}((\sigma_r - \lambda)_+) \mathbf{V}^\top$ , where  $(x)_+ = \max(0, x)$ .
  - 6:   **if**  $\|\hat{\mathbf{L}}^{(k)} - \hat{\mathbf{L}}^{(k-1)}\|_F / \|\hat{\mathbf{L}}^{(k-1)}\|_F < \epsilon$  **then**
  - 7:     Stop and return  $\hat{\mathbf{L}}^{(k)}$ .
  - 8:   **end if**
  - 9: **end for**
- 

Soft-Impute exploits sparsity by operating only on observed entries and using truncated SVD, computing just the leading  $R$  singular values and vectors. For sparse panels with  $|\Omega| \ll NT$ , this is much faster than dense SVD while approximating the same low-rank solution.

Alternating least squares (ALS), introduced in Section 9.2, alternates between updating loadings and factors. Each step solves a least-squares problem of the form

$$\hat{\boldsymbol{\lambda}}_i = \left( \sum_{t:(i,t) \in \Omega} f_t f_t^\top + \lambda \mathbf{I} \right)^{-1} \sum_{t:(i,t) \in \Omega} Y_{it} f_t,$$

with analogous updates for  $f_t$ . ALS is particularly effective when the rank is known and the panel is moderately sized or when we work with tensor structures as in Section 9.2.

Stochastic gradient descent (SGD) updates loadings and factors using mini-batches of observed entries. A typical update for  $\lambda_i$  with mini-batch  $B_k$  is

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \eta_k \sum_{t \in B_k} (Y_{it} - \lambda_i^{(k)\top} f_t^{(k)}) f_t^{(k)},$$

where the step size sequence  $\eta_k$  controls convergence. SGD handles datasets that do not fit in memory by streaming mini-batches, at the cost of slower and noisier convergence.

Convergence guarantees for ALS and SGD come from the optimisation and numerical-linear-algebra literature. For example, analyses of CP tensor decomposition show that, under mild regularity conditions, the ALS objective decreases monotonically and converges to a stationary point, with iteration counts that depend on condition numbers. Rather than restating full theorems here, we refer interested readers to the specialised literature and focus on practical consequences for marketing panels: well-tuned ALS and SGD

implementations converge reliably on many real-world problems, but can stall or converge slowly when the factor structure is ill-conditioned.

## Algorithm Selection Guide

Table 9.6 compares algorithms across key dimensions relevant for large marketing panels.

**Table 9.6** Algorithm comparison for matrix completion

Algorithm	Complexity (per iter.)	Memory	Sparsity support	When to use
Nuclear norm (dense)	$\min(N^2T, NT^2)$	order $NT$	No	Small dense panels
Soft-Impute	$ \Omega R$	order $ \Omega $	Yes	Large sparse panels
ALS	$ \Omega R^2$	order $ \Omega $	Yes	Moderate size, known rank, tensors
SGD	$BR$	order $B$	Yes	Very large panels, streaming data
MCMC	$(N + T)R^2$ per draw	order $NT$	Limited	When full uncertainty quantification is needed

In a large retail chain with millions of product–store–week observations but extreme sparsity, Soft-Impute or ALS on sparse matrices will typically be the workhorses. For impression-level advertising data or platform engagement logs that arrive continuously, SGD-style updates are more natural. When panels are smaller but uncertainty quantification is central—for example, in regulatory submissions or high-stakes budget allocations—Bayesian matrix completion with MCMC, as in Section 9.6, becomes attractive despite its higher computational cost.

From a causal perspective, all of these algorithms are tools for approximating the same low-rank estimate of  $Y_{it}(0)$ . Provided you drive the optimisation error to a level that is small relative to sampling variability, your ATT and related treatment-effect estimates will be dominated by design and sampling uncertainty rather than numerical approximation.

## Approximation Methods

Sketching and randomised algorithms provide approximate solutions with provable error bounds, which can be valuable when exact SVDs are too costly.

Randomised SVD replaces a full SVD with a low-dimensional projection. One simple scheme is to draw a random matrix  $\Omega$  of size  $T \times (R + p)$ , with  $p$  an oversampling parameter, form  $\mathbf{Q} = \text{orth}(\mathbf{Y}\Omega)$  as an

orthonormal basis for the range of  $\mathbf{Y}$ , and then compute the SVD of the smaller matrix  $\mathbf{Q}^\top \mathbf{Y}$ . This reduces the cost from order  $NT^2$  to roughly  $NTR$ .

Under standard conditions on the random projection and oversampling, the approximation error satisfies bounds of the form

$$\|\mathbf{Y} - \hat{\mathbf{Y}}\|_F \leq (1 + \epsilon)\|\mathbf{Y} - \mathbf{Y}_R\|_F$$

with high probability, where  $\mathbf{Y}_R$  is the best rank- $R$  approximation and  $\epsilon$  depends on the amount of oversampling. See the randomised linear algebra literature for details. See, for example, Halko et al. [2011] for a detailed treatment of randomised SVD methods and error bounds. In practice, randomised SVD trades a small, controllable increase in reconstruction error for a substantial reduction in runtime.

Low-rank approximation more generally trades accuracy for speed by intentionally solving with a smaller rank than the true effective rank. This can be useful when rank is only approximately low and higher-order factors contribute little to treatment-effect estimates. The key discipline is to check that ATT and other causal estimands are stable as you increase  $R$ . If estimates drift substantially, you are underfitting the factor structure.

## Software and Implementation

Several software packages expose scalable matrix-completion algorithms suitable for large marketing panels. In R, `softImpute` implements the Soft-Impute algorithm and handles matrices with millions of entries using sparse storage, `MatrixCompletion` provides nuclear-norm minimisation routines, and `tensorBF` offers Bayesian tensor factorisation for multi-way panels. In Python, `fancyimpute` provides multiple imputation algorithms, including Soft-Impute, and can be combined with SciPy's sparse matrices, `implicit` implements ALS for large sparse matrices in recommender-system contexts, and `tensorly` implements CP and Tucker decompositions for tensor data. Julia users can employ `LowRankModels.jl` for generalised low-rank models and `TensorToolbox.jl` for tensor decomposition.

In all cases, the workflow is the same: choose an algorithm that matches your panel's size and sparsity, fit a low-rank model for untreated potential outcomes  $Y_{it}(0)$ , and then compute ATT and related estimands exactly as in the earlier chapters. The computational choices affect how fast and how accurately you reach the low-rank solution, not the definition of the causal estimand.

## Benchmarking

Benchmarking studies compare the speed and accuracy of different algorithms on synthetic and real panels. Table 9.7 reports illustrative relative performance numbers for a representative sparse panel on standard hardware. Actual performance will depend on sparsity, rank, and implementation.

These numbers illustrate typical trade-offs. For very sparse matrices Soft-Impute can be an order of magnitude faster than dense nuclear-norm solvers with only a small loss in reconstruction accuracy, while

**Table 9.7** Illustrative benchmark results (Intel i7, 32GB RAM, synthetic sparse panel)

Algorithm	Sparsity	Time (relative)	MSE (relative)
Nuclear norm (dense)	1% observed	1.0× (baseline)	1.0×
Soft-Impute	1% observed	0.1× (about ten times faster)	1.05×
ALS	1% observed	0.15×	1.02×
Soft-Impute	50% observed	0.5×	1.0×
ALS	50% observed	0.2×	1.0×

for moderately dense matrices ALS often dominates. For tensor data, CP-ALS is usually substantially faster than Tucker-based methods at a given rank. In applied work you should run small-scale benchmarks on your own data, record both runtime and the stability of ATT and related estimands across algorithms, and choose the method that delivers robust causal estimates within your computational budget. Even when approximate algorithms increase reconstruction MSE by ten per cent relative to dense solvers, the induced changes in ATT are often much smaller than design-driven sampling variability. The right benchmark is not only matrix-level MSE but the stability of treatment-effect estimates across algorithms and tuning choices.

## 9.8 Connections and Comparisons

This section synthesises the advanced methods presented in this chapter and provides guidance on when to use each approach in marketing applications.

### Tensor vs Matrix Completion

Tensor completion (Section 9.2) preserves multi-way structure, while matrix completion flattens it into two dimensions. In a store–product–time panel, tensor completion models sales as a three-way array with separate factors for stores, products, and time. Matrix completion instead stacks store–product combinations as rows and takes time as the column dimension.

**Table 9.8** Tensor vs matrix completion

Aspect	Tensor completion	Matrix completion
Structure	Preserves multi-way structure (store $\times$ product $\times$ time)	Flattens to two-way panel
Mode interactions	Models interactions across modes explicitly	Ignores mode-specific structure
Computation	Typically higher, often scaling like $ \Omega R^K$ for $K$ modes	Roughly order $ \Omega R$
Data pattern	Exploits structure within each mode	Treats unit–time cells as exchangeable given factors
When attractive	Informative multi-way structure, moderate size	Very large panels, limited resources

The exact complexity for tensor methods depends on the chosen decomposition (CP vs Tucker) and optimisation routine, but a useful rule of thumb is that each additional mode increases the effective cost by a factor on the order of the target rank.

From a causal perspective, both approaches target the same underlying object: the untreated potential outcomes  $Y_{it}(0)$  for treated cells. Tensor completion can be particularly useful when treatment varies along one mode but there remains rich, untreated variation in the other modes. For example, if a promotion is rolled out to all products in a subset of stores, then product and time dimensions still carry information about  $Y_{ijt}(0)$  within treated stores. Tensor methods exploit this extra structure more directly than a flattened matrix that treats each product–store pair as a separate unit. Matrix completion is often preferable when the panel is extremely large or when the multi-way structure is weak or incidental, because it is computationally more tractable.

## Robust vs Standard Methods

Standard matrix completion assumes that idiosyncratic noise is light tailed. When retail data contain frequent stockouts, data-entry errors, or extreme promotional spikes, a small fraction of cells can be far from the low-rank structure. Robust methods (Section 9.3) explicitly decompose the outcome matrix into a low-rank component and a sparse outlier component, which can substantially improve the imputation of  $Y_{it}(0)$  for treated cells.

Rather than re-tabling all differences, it is helpful to think in terms of data diagnostics. Fit a standard matrix-completion model and examine residuals on untreated cells. If you see heavy-tailed residuals, many observations more than three standard deviations from zero, or clear patterns associated with stockouts or data glitches, robust methods are warranted. When outliers are rare and small relative to typical variation, standard matrix completion usually suffices and is computationally cheaper.

For causal work, the main risk is misclassifying treatment effects as outliers when treatment induces large, sparse spikes that resemble stockouts or glitches. Section 9.3 discussed this treatment–outlier separation problem in detail. Robust methods are most attractive when outliers are common and clearly distinct in magnitude from plausible treatment effects, and when you can learn the sparse component from untreated cells. In practice you should compare robust and standard matrix-completion estimates of ATT and inspect where the sparse component places mass. If large fitted outliers are concentrated in treated cells and align with known promotion windows, Assumption 34 is likely violated and robust methods may underestimate true treatment effects.

## Frequentist vs Bayesian

Frequentist and Bayesian approaches to matrix completion share the same identification logic: both rely on the low-rank factor structure for  $Y_{it}(0)$  and on treatment assignment assumptions developed earlier in the book. They differ in how they quantify uncertainty. Frequentist methods use bootstrap or asymptotic approximations applied to point estimators such as nuclear-norm minimisers or ALS solutions. Bayesian methods, as in Section 9.6, place priors on factors and loadings and produce posterior distributions for counterfactuals and treatment effects.

In large marketing panels where speed is critical and prior information is weak, frequentist procedures with robust bootstrap or asymptotic standard errors are often the default. In settings with smaller panels, strong prior beliefs, or a need for richer uncertainty statements—for example when evaluating loyalty-program changes in a small set of key markets—Bayesian matrix completion can be valuable. Chapter 16 and Section 9.6 provide more detailed comparisons. Here we simply emphasise that both approaches operate on the same completed panel and target the same estimands such as ATT.

## Nonlinear Panels with Interactive Fixed Effects

Nonlinear panel models with interactive fixed effects extend the low-rank factor structure from Chapter 8 to non-Gaussian outcomes such as binary adoption or churn and count data. The idea is to retain a low-rank interactive effect as a high-dimensional nuisance while modelling the outcome through a nonlinear link function.

The computational challenge is that fixed-effects estimators for these models require optimising a non-convex likelihood over both the structural parameters and the high-dimensional interactive fixed effects. For realistic  $N$  and  $T$  this is numerically burdensome. Recent work by Zeleneev and Zhang [2025] proposes a two-step approach. First, a convex nuclear-norm-regularised problem delivers a preliminary estimate of the interactive-effects matrix and structural parameters. Second, gradient descent on the original likelihood, initialised at that solution, refines the estimate and can converge to the fixed-effects estimator under local convexity conditions.

These models have natural marketing applications. For binary outcomes such as subscription or churn decisions in streaming video or mobile apps, a logit model with interactive fixed effects can capture latent demand factors that vary across users and time. For counts such as weekly purchase frequencies within product categories, a Poisson model with interactive fixed effects can capture category–time shocks that are not explained by observables. In network settings where peer effects depend on latent similarity—such as social sharing of campaigns or referral programs—a low-rank interaction structure can capture homophily in unobservables.

The identification story in these nonlinear models mirrors the linear case: we still require that, after conditioning on observed covariates and the interactive fixed effects, treatment assignment is as good as random, and that the low-rank structure for the latent index or mean function is correctly specified. This chapter focuses on the computational regularisation intuition and does not attempt to catalogue the additional identification conditions these nonlinear models require. For credible causal use, you must still supply a design in which, after conditioning on observed covariates and interactive fixed effects, treatment assignment is as good as random, and you must verify that the low-rank structure for the latent index is a plausible approximation in your marketing context.

## Integration with Other Methods

Matrix completion is not a competitor to synthetic control or difference-in-differences. It is a complementary tool for constructing counterfactuals. In many applications, hybrid or ensemble approaches provide robustness to misspecification by combining strengths of different methods.

One hybrid combines matrix completion with synthetic control. Synthetic control constructs weights on donors to match a treated unit's pre-treatment path. Matrix completion recovers a low-rank structure for untreated outcomes. A simple hybrid treats the synthetic-control prediction as a covariate and lets matrix completion model the residual structure,

$$\hat{Y}_{it}(0) = \sum_{j \in \text{donor}} w_j Y_{jt} + \hat{L}_{it},$$

where  $w_j$  are synthetic-control weights and  $\hat{L}_{it}$  is the low-rank residual from matrix completion. The first term captures similarities to specific donors, while the second term captures broader factor-driven co-movement. The estimand remains ATT as defined in Chapter 5, using these hybrid counterfactuals.

Another hybrid combines difference-in-differences (DiD) with matrix completion. DiD removes additive unit and time fixed effects,  $\hat{\alpha}_i$  and  $\hat{\delta}_t$ , and matrix completion is applied to the residuals,

$$\tilde{Y}_{it} = Y_{it} - \hat{\alpha}_i - \hat{\delta}_t, \quad \hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\delta}_t + \hat{L}_{it}.$$

Here DiD soaks up large additive heterogeneity, while matrix completion captures remaining interactive structure. The causal estimand remains ATT or a dynamic treatment-effect path based on  $\hat{Y}_{it}(0)$ . The hybrid merely changes how we estimate that counterfactual surface.

Ensemble approaches average treatment-effect estimates from multiple methods to reduce sensitivity to any single specification. If  $\widehat{\text{ATT}}^{(1)}, \dots, \widehat{\text{ATT}}^{(M)}$  are ATT estimates from  $M$  different methods (for example, matrix completion, synthetic control, DiD with high-dimensional controls), an ensemble estimator takes the form

$$\widehat{\text{ATT}}^{\text{ens}} = \sum_{m=1}^M w_m \widehat{\text{ATT}}^{(m)}, \quad \text{with} \quad \sum_{m=1}^M w_m = 1.$$

Ensemble estimators reduce sensitivity to any single specification but do not create identification on their own. They are most useful when you believe that at least one of the underlying methods is approximately correctly specified and that others fail in different ways, so that averaging shrinks idiosyncratic biases.

Weights can be chosen by cross-validation on untreated cells, by inverse-variance weighting using estimated standard errors, or simply set to equal weights when methods are sufficiently diverse. In all cases the estimand is still ATT. The ensemble aims to stabilise estimates when each individual method is only approximately correctly specified.

## Practical Guidance: Method Selection

In practice you can think of method selection as a sequence of questions about your data and objectives rather than as a rigid decision tree. If the panel has genuinely multi-way structure—such as users by items by context in a platform setting, or products by stores by regions in global retail—tensor completion is attractive when the panel is of moderate size and when treatment varies along one mode but there remains rich untreated variation along the others. When the panel is extremely large or the multi-way structure is weak, flattening to a matrix and using matrix completion is usually more practical.

Data quality is the next consideration. If stockouts, data-entry errors, or extreme promotions generate many large residuals relative to the low-rank structure, robust matrix completion is more reliable; when such outliers are rare and modest in size, standard low-rank methods are typically sufficient. Rich side

information—store demographics, weather, macro variables, or known network structure—pushes you towards covariate-assisted and graph-regularised methods (Section 9.4), especially when covariates explain a large share of outcome variance and are plausibly exogenous.

Stability and dynamics matter as well. When the environment is relatively stable, standard low-rank matrix completion can safely pool long time spans. When recessions, pandemics, or major policy changes create obvious structural breaks, time-varying rank models (Section 9.5) combined with change-point detection provide better imputations of  $Y_{it}(0)$  in post-break periods. Outcome type also guides choices: continuous outcomes fit naturally into the Gaussian factor models used throughout this chapter; binary or count outcomes with important nonlinearities motivate nonlinear interactive fixed-effect models along the lines of Zeleneev and Zhang [2025].

Finally, your need for uncertainty quantification and your computational budget determine whether you lean on frequentist or Bayesian estimators and which algorithms you choose. When you require only point estimates and approximate standard errors in very large panels, fast frequentist algorithms such as Soft-Impute, ALS, or SGD (Section 9.7) are natural. When full posterior inference for ATT and related estimands is essential, and panel sizes are moderate, Bayesian matrix completion (Section 9.6) offers a principled route at higher computational cost. Throughout, the key discipline is to treat matrix and tensor methods as tools for constructing  $Y_{it}(0)$  under clear identification assumptions, not as black boxes that magically deliver causal estimates. Across all these choices, method selection cannot compensate for a weak or violated design. If the assumptions that justify using matrix or tensor methods for  $Y_{it}(0)$  do not hold, no amount of algorithmic sophistication will deliver credible causal estimates.

## 9.9 Diagnostics and Validation

When we use matrix or tensor completion to construct counterfactuals, the quality of our causal estimates depends on three ingredients: the low-rank specification for untreated potential outcomes, the way treatment induces missing cells, and the numerical behaviour of the algorithms. Diagnostics and validation are therefore about stress-testing these three ingredients. They do not create identification by themselves, but they help detect when the assumptions underpinning ATT and related estimands are implausible or when numerical approximations are too crude.

For complementary inference and design tools when using SC/SDID rather than factor-based identification, see Chapters 6 and 7, and the design-diagnostics protocol in Chapter 17.

### Rank Diagnostics: Eigenvalues and Stability

Rank choice is central to factor-based identification. If the rank is set too low, we underfit the untreated potential outcomes  $Y_{it}(0)$  and bias ATT. If it is set too high, we overfit noise and inflate variance. Scree plots offer a useful first look but lack formal statistical properties.

For matrix factor models, Babii et al. [2024] propose an eigenvalue-ratio test that formalises the search for the “cliff” in the spectrum. In our setting we often work with multi-way arrays  $\mathcal{Y}$  constructed from outcomes or untreated outcomes. A practical approach is to matricise the tensor along each mode  $j$  into a matrix  $\mathbf{Y}_{(j)}$  and apply the eigenvalue-ratio logic to each unfolding. In practice we apply this logic to matricisations built from untreated observed cells only, using  $\Omega \setminus \mathcal{W}$ , so that the spectrum reflects the low-rank structure of untreated potential outcomes rather than a mixture of treated and untreated paths. Let  $\hat{\sigma}_{r,j}^2$  denote the  $r$ th largest eigenvalue of  $\mathbf{Y}_{(j)} \mathbf{Y}_{(j)}^\top$ . For a given upper bound  $K$  on the number of factors in mode  $j$ —typically a modest value such as 10 to 20—we compute the ratio statistic

$$S_j = \max_{2 \leq r \leq K-2} \frac{\hat{\sigma}_{r,j}^2 - \hat{\sigma}_{r+1,j}^2}{\hat{\sigma}_{r+1,j}^2 - \hat{\sigma}_{r+2,j}^2}$$

for  $r$  in a range such as  $2 \leq r \leq K-2$  so that both numerator and denominator are defined. This ratio highlights points where the spectrum drops sharply and then flattens. Large values arise when the denominator approaches zero, which typically happens at the cut-off between signal and noise.

In practice, you generate a reference distribution for  $S_j$  by simulating matrices with pure noise and no factors, then compute p-values by comparing the observed statistic with this reference. Combining evidence across modes—for example, by averaging or taking the maximum of mode-specific p-values—yields a consensus view of how many factors are needed to capture most of the structure in  $Y_{it}(0)$ . This procedure should be seen as a principled refinement of the scree plot rather than as an oracle. It complements cross-validation and substantive judgement.

Cross-validation for rank selection, discussed in Section 8.6, remains essential. You randomly hold out cells from the untreated set, estimate the low-rank approximation on the remaining cells, and compute prediction

error on the held-out cells. Repeating this over multiple splits and comparing errors across candidate ranks gives a data-driven sense of how many factors are useful for predicting  $Y_{it}(0)$ . Stability across subsamples—for example, comparing estimated factors on early vs late pre-treatment periods—provides an additional check: if factor loadings and time patterns are highly correlated across subsamples, the rank choice is likely robust.

## Outlier Detection and Influence

Outliers and influential cells can distort counterfactuals and hence treatment effects. Robust matrix completion (Section 9.3) explicitly separates a sparse outlier component from the low-rank structure. Diagnostics help you decide when robustness is needed and which cells deserve closer inspection.

Residual analysis is the starting point. Plot the residuals from a standard matrix-completion fit on untreated cells. If the distribution is roughly symmetric and light-tailed, standard methods are usually adequate. Heavy tails, clusters of very large residuals, or systematic patterns by unit or time suggest that stockouts, data glitches, or mis-coded promotions are present and that robust methods are warranted.

Influence diagnostics adapt ideas from regression. One natural measure is the change in the low-rank estimate when a cell is removed. Let  $\hat{\mathbf{L}}$  denote the fitted low-rank matrix and  $\hat{\mathbf{L}}_{(-it)}$  the estimate obtained when cell  $(i, t)$  is omitted. A global influence measure compares these two objects, scaled by noise and rank. A cell is influential if removing it changes many fitted values by a non-trivial amount. Classical quantities such as Cook's distance and DFFITS provide intuition, but in the high-dimensional, low-rank setting it is safest to treat their thresholds as rough guides rather than as hard cut-offs. Cells that show up as extreme under several influence metrics, or that are clear outliers in residual plots, merit business review: they may correspond to mis-coded prices, mega-promotions, system outages, or extraordinary events like panic buying.

A simple and effective global diagnostic is to compare ATT estimates from standard and robust matrix completion. Large discrepancies, especially when robust and non-robust ATT disagree in sign or differ by more than, say, 20 per cent, signal that outliers are materially affecting your conclusions. Sensitivity analysis over a range of robustness tuning parameters strengthens this check. When robust and standard ATT estimates differ sharply, it is critical to check where the robust procedure is placing the sparse component. If large fitted outliers are disproportionately in treated cells and line up with known promotion windows, this is evidence against the treatment–outlier separation condition in Assumption 34. Robust methods may then be attenuating genuine treatment effects rather than correcting data glitches.

## Covariate Relevance and Sensitivity

Covariate-assisted matrix completion (Section 9.4) introduces observed characteristics into the model for  $Y_{it}(0)$ . Diagnostics here serve two roles: they indicate whether covariates are actually adding predictive signal for untreated outcomes, and they flag potential bad-control problems when covariates are tied to treatment assignment.

Variable-importance analysis starts from the share of variance explained by covariates. One practical approach is to re-fit the model while dropping each covariate or group of covariates in turn and to record the change in out-of-sample prediction error on untreated cells. Covariates whose removal substantially worsens prediction of  $Y_{it}(0)$  are carrying real signal. Those that barely move the error can often be dropped for parsimony.

Partial dependence plots and related tools from machine learning can reveal whether relationships between covariates and outcomes are roughly linear or strongly nonlinear. When plots show clear curvature or interactions—say, weather effects that differ sharply by region—it may be worth enriching the covariate specification with transformations or interactions, always bearing in mind the risk of overfitting.

Most importantly, you should treat covariate sensitivity as a causal diagnostic. Estimating ATT with broad covariate sets, with only clearly exogenous variables (such as weather and holidays), and with no covariates at all provides a three-way comparison. Stable ATT estimates across these specifications suggest that covariates are mostly improving efficiency. Large shifts, especially when including variables that could be affected by treatment (prices, advertising budgets), are a red flag for bad-control bias and should be reported transparently. Large, systematic shifts in ATT when you add or remove covariates that are plausibly affected by treatment or share unobserved drivers with treatment should be interpreted as evidence against the identification conditions in Assumption 35, not as a tuning issue. In those cases you should revert to designs that avoid such bad controls rather than trying to “fix” the problem with richer matrix-completion models.

## Computational Validation

Because we rely on numerical optimisation to recover the low-rank structure, we must also check that the algorithms have converged to a stable solution and that approximation shortcuts do not dominate uncertainty.

Convergence diagnostics begin with the objective function. For nuclear-norm and Soft-Impute-type algorithms, the reconstruction error on untreated cells should decrease steadily and then flatten. Sudden jumps or oscillations suggest poor step-size choices or numerical instability. Running the algorithm from several random initialisations and comparing the resulting low-rank estimates gives an informal sense of whether local minima are an issue. If different runs produce very similar imputed  $Y_{it}(0)$  and ATT, you can be more confident in the numerical side of the analysis.

Approximation methods such as randomised SVD and intentional low-rank truncation introduce additional error. You can quantify this by comparing fits from approximate and more exact methods on a subset of the data small enough to permit dense SVD or higher-rank estimation. If the differences in reconstructed  $Y_{it}(0)$  and in ATT are small relative to the estimated standard errors, the approximations are unlikely to change substantive conclusions. If not, you may need to increase the approximation fidelity or adjust the panel size, rank, or sparsity threshold.

## Software for Diagnostics

General matrix and regression diagnostic tools are valuable here. In R, packages such as `RMTstat` support random matrix theory calculations useful for eigenvalue-based rank diagnostics. `car` and related packages provide influence and leverage diagnostics that can be adapted to residuals from matrix-completion fits. `ggplot2` makes it straightforward to produce scree plots, residual histograms, and partial dependence plots. In Python, `statsmodels` and `sklearn.inspection` offer influence measures and partial dependence tools, while `matplotlib` and `seaborn` cover visualisations. The key is not the specific package but a disciplined workflow. Apply these tools to residuals and fitted values from your matrix or tensor model, not just to simple regressions.

## A Practical Diagnostic Workflow

In applied marketing panels, a practical workflow for matrix-completion diagnostics can be summarised in a few steps. Before estimation, inspect the raw data for obvious outliers and missingness patterns, and generate scree plots for the main matrixisations to get a first sense of rank. After fitting initial low-rank models, use eigenvalue-ratio ideas and cross-validation to refine rank choices and check that factor structures are stable across subsamples. Examine residuals and influence measures to detect outliers or influential cells, and compare ATT estimates from standard and robust matrix completion when serious outliers are plausible.

Next, evaluate the contribution and exogeneity of covariates by looking at variable-importance measures and by comparing treatment-effect estimates across different covariate sets. Throughout, monitor convergence diagnostics and run a few alternative numerical configurations to ensure that solutions are not artefacts of a particular initialisation or approximation. Finally, document the chosen rank, any influential observations you identified and how you handled them, the range of ATT estimates across reasonable specifications, and the main diagnostic plots. This documentation is as important for internal decision-making as it is for external scrutiny. It shows that your causal conclusions rest on a carefully stress-tested low-rank model rather than on a single opaque algorithm run. None of these diagnostics can turn a weak or violated design into a credible one. Their role is to expose tensions between the low-rank assumptions and the data so that you can revisit the design or the use of matrix and tensor methods where necessary.

## 9.10 Marketing Applications

This section illustrates how the advanced matrix and tensor methods in this chapter can be deployed in realistic marketing settings. In each case, the goal is to estimate treatment effects such as ATT or dynamic response paths, with matrix or tensor completion providing the counterfactual untreated outcomes  $Y_{it}(0)$  as defined in earlier chapters. In each case, we assume a design that justifies using matrix or tensor completion for causal interpretation: randomised rollouts, staggered adoption with credible parallel-trends assumptions, or other assignment mechanisms introduced in earlier chapters. The advanced methods here change how we approximate  $Y_{it}(0)$  within those designs. They do not, by themselves, weaken the underlying identification assumptions.

**Table 9.9** Summary of marketing applications

Case	Method	Data structure	Challenge	Illustrative finding
E-commerce	Tensor completion	Users $\times$ items $\times$ context	Multi-way structure	Higher CTR and purchase rates
Retail stockouts	Robust MC	Products $\times$ store-weeks	Outliers (stockouts)	Larger ATT once stockouts corrected
Product launch	Covariate-assisted tensor completion	Products $\times$ markets $\times$ weeks	Market heterogeneity	Strong average effect, heterogeneous across markets
Advertising	Time-varying + Bayesian	Platforms $\times$ creatives $\times$ weeks	Algorithm changes	Positive platform-specific effects with credible intervals

### Case Study 1: E-Commerce Recommendation Systems

An e-commerce platform observes user engagement for tens of thousands of users across thousands of items and multiple contexts, such as mobile vs desktop and different times of day. The data form a three-way tensor of users by items by contexts. The platform rolls out a new recommendation algorithm to a subset of users in the mobile context for four weeks and wants to estimate the effect on engagement measures such as click-through rate (CTR) and purchase rate.

The cleanest design is a randomised A/B test in which eligible users are randomly assigned to the new algorithm or to business-as-usual. If instead the platform targets heavy users or specific segments, tensor completion still improves prediction of engagement but no longer guarantees causal identification without stronger assumptions about how assignment depends on past behaviour and unobservables.

Using tensor completion (Section 9.2), we model engagement as a low-rank CP decomposition with a modest rank, for example  $R = 10$ , estimated by ALS. One factor captures broad user preferences (for instance, electronics vs fashion), another captures item popularity, a third picks up context effects (mobile

users browse more but buy less), and additional factors capture finer-grained interactions. Under the low-rank structure for untreated outcomes and the treatment assignment assumptions set out earlier in the book, this tensor model yields imputations of untreated potential outcomes, for example  $\mathcal{Y}_{ijq}(0)$  with  $i$  indexing users,  $j$  indexing items, and  $q$  indexing contexts.

ATT for CTR and purchases is then defined as in Chapter 5, averaging differences between observed and imputed untreated outcomes over treated user-item-context cells during the evaluation period. In a representative application, the estimated ATT for CTR might correspond to a mid-single-digit percentage increase, with purchase-rate effects somewhat smaller but still economically meaningful. Tensor completion is particularly valuable here because it leverages similarities across users, items, and contexts simultaneously. Diagnostics based on cross-validation, placebo tests on pre-treatment periods, and residual analysis (Section 9.9) support the chosen rank and confirm that the tensor model captures most of the systematic variation in engagement.

### Case Study 2: Retail Chain with Stockouts

Consider a grocery retailer with several hundred products across around 150 stores observed weekly for two years. Stockouts affect a non-trivial share of product-store-week cells and are more common for popular products and smaller stores. The retailer implements a new inventory-management system in a subset of stores for 12 weeks and seeks to measure its causal effect on sales.

Standard matrix completion treats observed zeros from stockouts as genuine low sales and therefore underestimates underlying demand. Robust matrix completion (Section 9.3) instead decomposes the outcome matrix into a low-rank demand component and a sparse outlier component. Fitting a stable principal-component pursuit model on untreated cells, with the regularisation parameter chosen by cross-validation, yields a low-rank estimate  $\hat{\mathbf{L}}$  that reflects underlying demand and a sparse component  $\hat{\mathbf{S}}$  that captures stockouts and other large irregularities.

For causal interpretation we need store-level assignment of the new system to be plausibly exogenous, for example via a phased rollout that is unrelated to recent unexplained demand shocks after controlling for observed covariates and factors. If early-adopting stores are chosen precisely because of recent stockout problems, then even a robust low-rank model cannot fully separate treatment effects from evolving inventory policies.

Imputing  $\hat{Y}_{it}(0)$  for treated store–product–week cells from  $\hat{\mathbf{L}}$  gives an ATT for sales that is typically larger than what standard matrix completion would suggest, because it attributes part of the observed increase to improved availability rather than purely to demand shifts. In a stylised panel calibrated to a mid-sized retailer, robust estimates of the sales lift from better inventory management can be materially higher than standard estimates that do not correct for stockouts. Residual diagnostics and sensitivity of ATT to robustness tuning, as described in Section 9.9, help assess whether outliers are driving the difference.

### Case Study 3: Multi-Market Product Launch

A consumer packaged goods company launches a new line of products across multiple geographic markets over the course of a year. Launch dates are staggered across markets. Markets differ in size, income, and competitive intensity. Management wants to understand not only the average launch effect but also how it varies across markets and over time.

The staggered launch plays the same role as in the event-study designs of Chapters 4 and 5: untreated markets in each period provide a comparison surface for treated markets under a parallel-trends assumption once we control for observed covariates and latent factors. Matrix and tensor completion refine that comparison by exploiting cross-market and cross-product structure, but they still rely on the underlying staggered-adoption design for identification.

Covariate-assisted tensor completion combines market characteristics with a tensor representation of product  $\times$  market  $\times$  week outcomes. As in Section 9.4, we first regress sales on market-level covariates such as store density, median income, and competitor presence to remove variation explained by observables. We then apply a low-rank CP decomposition to the residual tensor to capture remaining co-movement across products, markets, and weeks.

This structure yields imputations of untreated potential outcomes for market–product–week cells affected by the launch, for example  $\mathcal{Y}_{ijt}(0)$  with  $i$  indexing markets and  $j$  indexing products. Aggregating differences between observed and imputed untreated outcomes over treated cells produces an estimate of ATT for the launch, with further aggregation by market segment revealing heterogeneity. In plausible calibrations, larger and higher-income markets tend to show stronger and earlier responses than smaller or lower-income ones, with effects building over several weeks as awareness grows. Covariate-assisted tensor completion makes these patterns visible by combining observed heterogeneity with latent factor structure, while cross-validation and covariate-sensitivity checks ensure that results do not hinge on a particular covariate specification.

### Case Study 4: Advertising Across Platforms

An advertiser runs campaigns across several digital platforms—search, social, display, and video—using a portfolio of creatives over half a year. The outcome is a three-way tensor of platforms by creatives by weeks. Midway through the period, one platform changes its ranking algorithm in a way that affects organic and paid reach. Later, the advertiser introduces a new bidding strategy on two platforms. The aim is to estimate platform-specific treatment effects on CTR and conversions while accounting for both structural breaks and cross-platform information sharing.

For the bidding-strategy effect to be identified, we need variation across platform–creative–week cells that experience the same platform algorithm regime but differ in whether the new bidding strategy is applied. The time-varying rank component soaks up the common algorithm change, while the bidding-strategy indicator provides the treatment variation within each regime. If adoption of the new bidding strategy is itself targeted based on recent unexplained performance, the same selection concerns as in Chapter 2 apply.

Time-varying rank models (Section 9.5) allow the effective number of factors to increase after the algorithm change, so that the low-rank representation of untreated outcomes reflects the new environment. Bayesian hierarchical matrix or tensor completion (Section 9.6) then shares information across platforms by shrinking platform-level factors toward a population-level set of factors. This is especially helpful for video or smaller platforms with fewer observations.

Within this framework, ATT for the bidding strategy is defined as the average difference between observed and imputed untreated outcomes on treated platform–creative–week cells in the post-treatment window, with posterior draws providing credible intervals. In a stylised example, search and social may both exhibit positive CTR lifts, with search effects somewhat larger and more precisely estimated. The hierarchical structure reduces uncertainty for platforms with less data by borrowing strength from better-measured platforms. Change-point diagnostics on reconstruction error and Bayes factors comparing static and time-varying models help justify the additional rank introduced after the algorithm change.

## Lessons Learned

Across these applications, several themes recur. First, the choice of method should follow the data structure. Multi-way panels with rich interactions across dimensions naturally call for tensor completion; panels dominated by outliers benefit from robust matrix completion; and panels with strong observed heterogeneity or network structure are good candidates for covariate-assisted and graph-regularised methods. When environments change sharply over time, time-varying rank models better capture the new regime than static factors.

Second, treatment effects in marketing are rarely homogeneous. The tensor and matrix methods in this chapter are most valuable when they reveal systematic variation across users, markets, products, platforms, and time, rather than when they return a single headline ATT. In practice, reporting subgroup effects and dynamic responses alongside overall averages gives a more faithful picture of how interventions work.

Third, diagnostics and uncertainty quantification are not optional extras. Cross-validation, placebo tests on pre-treatment periods, residual analysis, and sensitivity to tuning parameters and covariate sets all feed into a judgement about whether the low-rank model is an adequate description of untreated outcomes. Bootstrap confidence intervals for frequentist methods and credible intervals for Bayesian methods make clear how much uncertainty remains around estimated effects.

Finally, robust methods matter in marketing data sets riddled with stockouts, data-entry errors, and occasional shocks. Comparing robust and standard matrix-completion estimates is a low-cost way to gauge how sensitive your conclusions are to such irregularities. When combined with the broader causal-design tools in this book, advanced matrix and tensor methods become a flexible, diagnostically grounded machinery for learning about treatment effects in complex marketing environments.

Across all four cases, the common pattern is that advanced matrix and tensor methods amplify a good design but cannot rescue a bad one. When assignment mechanisms violate the identification conditions laid out

in the early chapters, the estimates in this section should be read as descriptive summaries of counterfactual models rather than as causal effects.

## 9.11 Practical Workflow and Software

This section summarises a practical workflow for implementing the advanced matrix and tensor methods from this chapter in real marketing applications. It draws together the method comparisons in Section 9.8, the diagnostics in Section 9.9, and the applications in Section 9.10, and links them to concrete software choices.

### Overview of the Workflow

In all of the case studies in Section 9.10, the workflow followed the same broad pattern. First, we organised outcomes into a matrix or tensor and made the treatment design explicit, separating treated cells where we observe  $Y_{it}(1)$  from untreated cells that inform the model for  $Y_{it}(0)$ . Second, we chose an appropriate matrix or tensor completion method based on the data structure, data quality, and covariate information, using standard matrix completion as a baseline and layering on tensor, robust, or covariate-assisted extensions when warranted. Third, we tuned key hyperparameters such as rank and regularisation strength using cross-validation and eigenvalue-based diagnostics. Fourth, we estimated the model and generated imputations of  $Y_{it}(0)$  for treated cells, translating these into treatment-effect estimators such as ATT as defined in Chapter 5. Finally, we subjected the results to a battery of diagnostics and uncertainty quantification checks before reporting effects and heterogeneity.

### Method Selection in Practice

Method selection is largely driven by data structure, data quality, and the nature of the research question. Section 9.8 already set out the main trade-offs. Here we just highlight how they play out operationally.

When the panel has a genuinely multi-way structure, as in users by items by context or products by stores by regions, and treatment varies along only one of these modes, tensor completion is attractive because it can borrow strength across the remaining modes. When the panel is effectively two-dimensional, or when the third dimension is thin or noisy, flattening to a matrix and using matrix completion is usually more robust and computationally efficient.

Data quality considerations often push you towards robust methods. If residuals from standard matrix completion show heavy tails or clear patterns associated with stockouts, data-entry errors, or extreme promotions, robust matrix completion separates a sparse outlier component from the low-rank structure and yields more reliable imputations of  $Y_{it}(0)$ . When covariates such as demographics, weather, or network structure explain a large share of outcome variation and are plausibly exogenous, covariate-assisted and graph-regularised methods help by removing predictable variation before low-rank modelling.

Finally, dynamics and inference needs shape the choice among static, time-varying, frequentist, and Bayesian approaches. In relatively stable environments, static low-rank models estimated by fast frequentist algorithms often suffice. When structural breaks such as recessions or algorithm changes are present,

time-varying rank models with change-point detection provide better counterfactuals post-break. When full posterior uncertainty for ATT and related estimands is required—for instance in high-stakes budget or policy decisions—Bayesian matrix or tensor completion is appropriate despite higher computational cost.

## Step-by-Step Implementation Guide

**Step 1: Prepare the data.** Organise outcomes into a matrix or tensor and define the treatment indicators. Make the assignment mechanism explicit by stating what you assume about  $\Pr(\mathbf{D} \mid \{Y_{it}(d_i^t)\}_{i,t}, \mathbf{X})$  in your setting and why, under those assumptions, treated and untreated cells can be compared. Explicitly separate treated cells  $\mathcal{W}$ , where you observe  $Y_{it}(1)$ , from untreated observed cells  $\Omega \setminus \mathcal{W}$  that will inform the model for  $Y_{it}(0)$ . Assemble covariates at the unit and time level, and visualise the panel using simple plots—time series by unit, histograms of missingness, and basic heatmaps—to understand the scale, sparsity, and presence of obvious anomalies.

**Step 2: Select the method.** Start from the guidance in Section 9.8. As a baseline, fit a standard low-rank matrix-completion model. Then add complexity only where the data justify it: use tensor methods when there is clear multi-way structure and enough observations per mode, bring in robust methods when outliers are a concern, and incorporate covariates or graph structure when they demonstrably improve prediction of untreated outcomes.

**Step 3: Choose tuning parameters.** Choose candidate ranks using a combination of scree plots, eigenvalue-based diagnostics, and substantive judgement about how many factors are plausible. Refine rank and regularisation parameters through cross-validation on untreated observed cells  $\Omega \setminus \mathcal{W}$ , comparing prediction performance across candidate values and checking stability across subsamples. For time-varying models, tune smoothing and change-point thresholds by examining reconstruction error over time. For Bayesian models, calibrate priors to be weakly informative unless strong prior knowledge is available, and check that posterior inferences are not unduly sensitive to prior hyperparameters.

**Step 4: Estimate the model.** Run the chosen algorithm—ALS, nuclear-norm minimisation, Soft-Impute, stochastic gradient descent, or MCMC—using implementations that exploit matrix sparsity where possible. Monitor convergence by tracking the objective function and key parameter summaries over iterations and by comparing results across multiple initialisations. Once convergence is achieved, compute the imputed untreated outcomes  $\hat{Y}_{it}(0)$  for treated cells and form treatment effects as differences  $Y_{it}(1) - \hat{Y}_{it}(0)$ , aggregating as needed to form ATT or dynamic response profiles.

**Step 5: Conduct diagnostics.** Apply the diagnostic tools from Section 9.9 to untreated cells and fitted  $Y_{it}(0)$ : revisit rank choice using cross-validation and eigenvalue-based tests, inspect residuals for heavy tails and structure, compare standard and robust fits where outliers are plausible, and examine the contribution and exogeneity of covariates by looking at variable-importance measures and covariate-sensitivity of ATT.

In panels with potential structural breaks, inspect reconstruction error over time and apply change-point methods to detect regime shifts.

**Step 6: Perform inference.** Quantify uncertainty in treatment-effect estimates using the inference tools from Chapter 16. Start by stating the independent sampling unit. In many marketing panels this is a cluster  $c = 1, \dots, G$  (markets, regions, geo cells), so the effective sample size is  $G$  rather than  $NT$ . For frequentist estimators, this may involve CRVE or a wild cluster bootstrap, with re-fitting inside each resample when tuning or regularisation is part of the estimator. For Bayesian estimators, use posterior draws of ATT and dynamic treatment effects to construct credible intervals, and remember that their interpretation is model-based. Randomisation inference is exact under the sharp null and must be conditioned on the randomisation or re-randomisation rule when such a rule is used.

**Step 7: Report results.** Present treatment-effect estimates with their uncertainty measures and highlight heterogeneity across units, markets, user segments, and time where it is substantively meaningful. Compare results to simpler alternatives such as difference-in-differences or synthetic control, explaining where matrix and tensor methods change the conclusion and why. Be explicit about tuning choices, diagnostics performed, and any influential observations or design features that materially affect the results. Where possible, provide replication materials—code, summary data, and diagnostic plots—to make the analysis transparent.

## Software Recommendations

Table 9.10 lists representative software packages that implement the main classes of methods discussed in this chapter. These are not exhaustive, but they provide starting points in the most common languages.

**Table 9.10** Software packages by task and language

Task	R	Python	Julia
Standard matrix completion	<code>softImpute</code>	<code>fancyimpute</code>	<code>LowRankModels.jl</code>
Tensor decomposition	<code>rTensor</code>	<code>tensorly</code>	<code>TensorToolbox.jl</code>
Robust PCA / matrix completion	Packages such as <code>rrcov</code> , custom PCP implementations	Dedicated RPCA libraries (for example <code>rpca</code> ), custom ADMM solvers	Custom implementations
Bayesian factor models	<code>rstan</code> , <code>brms</code> (for custom factor structures)	<code>pymc</code> , <code>numpyro</code>	<code>Turing.jl</code>

For sparse matrices, `softImpute` in R and matrix-completion routines in `fancyimpute` provide efficient implementations of nuclear-norm and related estimators. For tensors, `tensorly` in Python and `rTensor` in R implement CP and Tucker decompositions with missing data support. Robust matrix completion often requires either dedicated RPCA libraries or custom ADMM implementations built on top of general opti-

misation toolkits. Bayesian factor models can be implemented in general-purpose frameworks such as Stan, PyMC, NumPyro, or Turing by encoding the models of Section 9.6 directly.

## Common Pitfalls and Solutions

Despite the flexibility of matrix and tensor methods, several pitfalls recur in practice. Overfitting is a first concern: choosing a rank that is too large or using too weak regularisation leads to excellent in-sample fit but poor out-of-sample performance and unstable treatment effects. Cross-validation and a bias towards parsimony mitigate this risk. Computational bottlenecks arise when attempting to run dense algorithms on very large panels. Exploiting sparsity and using algorithms such as Soft-Impute, SGD, or sparse ADMM are the main remedies.

Interpretation can also be challenging when many factors are included. Plotting factor loadings over time or across units and relating them to substantive features—such as known seasonality, product categories, or platform characteristics—helps turn abstract factors into interpretable components. Finally, sensitivity to tuning parameters is unavoidable in flexible models. Rather than hiding this, you should report how treatment-effect estimates change across reasonable ranges of rank and regularisation parameters, and use data-driven tuning to anchor your preferred specification. As a rule of thumb, if reasonable changes in rank or regularisation move ATT estimates by more than their reported standard errors, you should treat the design or low-rank specification as fragile and report that fragility explicitly.

These workflow, software, and pitfall guidelines are not a substitute for careful design and diagnostics, but they provide a template for implementing the methods in this chapter on real marketing panels in a transparent and reproducible way.

## 9.12 Conclusion and Future Directions

This chapter has extended the matrix-completion framework from Chapter 8 to a family of advanced methods for complex panel structures. The common thread is the use of low-rank structure for untreated potential outcomes  $Y_{it}(0)$ , combined with clear treatment designs and disciplined diagnostics, to construct credible counterfactuals in settings where simple difference-in-differences or synthetic control would struggle.

Tensor completion preserves genuine multi-way structure when we observe outcomes across several dimensions, such as users, items, and contexts. Robust matrix completion separates sparse, extreme deviations such as stockouts from the underlying demand surface. Covariate-assisted and graph-regularised methods incorporate observed heterogeneity and network links into the low-rank model. Time-varying rank models allow the factor structure to evolve with structural breaks or algorithm changes. Bayesian versions of these models treat factors and loadings as random and provide posterior distributions for counterfactuals and treatment effects.

These tools are not competing estimators with different causal estimands. They are alternative ways to approximate the same objects—untreated potential outcomes and, in turn, ATT and dynamic treatment effects—under the same potential-outcomes framework developed in earlier chapters. Method choice is driven by data structure, data quality, the richness of side information, and the need for uncertainty quantification, as summarised in Sections 9.8 and 9.11.

### When to Use Advanced Methods

Advanced matrix and tensor methods are most valuable when the data call for them. When panels are small, tidy, and close to the textbook staggered-adoption designs of Chapters 5 and 10, simpler approaches may suffice. Low-rank completion becomes a natural starting point when panels are high-dimensional, sparse, and riddled with missing cells or outliers. It is also attractive when outcomes vary along several meaningful modes or when strong, plausibly exogenous covariates explain most of the variation in untreated outcomes.

Tensor methods are particularly attractive when the multi-way structure is not an artefact of coding but reflects real economic interactions—for example, users by items by device types in digital platforms, or products by stores by regions in global retail. Robust methods matter when the data contain a non-trivial fraction of large irregularities that standard models would absorb into the low-rank component. Covariate-assisted and graph-regularised methods help when demographic, behavioural, or spatial variables explain substantial variation in  $Y_{it}(0)$  and can be treated as exogenous. Time-varying models earn their keep when reconstructing untreated potential outcomes across sharp regime changes, such as recessions or major platform algorithm shifts. Bayesian approaches make sense when practitioners need full posterior distributions for effects—for example, in budget or policy choices where risk attitudes matter and panels are not so large that MCMC becomes intractable.

Rather than memorising thresholds, the practical rule is to start from the simplest adequate model, layer in complexity only when diagnostics or domain knowledge demand it, and always cross-check results across a small menu of reasonable specifications.

## Open Research Questions

Several methodological questions remain open and are particularly relevant for marketing applications.

One direction is to extend tensor completion to settings where the effective rank changes over time across multiple modes. Section 9.5 developed time-varying rank in two-way panels. Analogous results for multi-way panels would be valuable for platforms and retailers whose environments shift across both units and contexts.

Another direction is to integrate robustness and Bayesian uncertainty quantification more tightly. Robust matrix completion handles outliers well but is typically framed in a frequentist optimisation language. Fully Bayesian low-rank-plus-sparse models with scalable inference would allow practitioners to propagate both factor and outlier uncertainty into treatment-effect distributions.

Scalability of Bayesian methods to truly massive panels is also an active area. The models in Section 9.6 are currently best suited to moderate panels. Advances in stochastic variational inference, subsampling-based MCMC, and hardware-aware implementations could make Bayesian matrix and tensor completion more accessible for panels with hundreds of millions or billions of cells.

A further research frontier is incorporating richer network structure—such as spatial correlation, competitive relationships, or social networks—directly into tensor and matrix models for  $Y_{it}(0)$ . Graph-regularised approaches in Section 9.4 provide a first step. More expressive models that combine low rank with flexible graph dynamics would be highly relevant for geographic retail and platform settings.

Finally, there is ongoing work on non-convex regularisers and explicit rank constraints that may deliver better statistical properties than nuclear norms, for example lower bias in estimated singular values and improved recovery of weak factors, provided they can be optimised reliably at scale.

## Emerging Methods

Recent work at the interface of deep learning and matrix completion explores neural architectures that can capture nonlinear structure in panels. Autoencoder-based models learn low-dimensional representations of units and times that need not be linear. Generative models can approximate complex outcome distributions and offer flexible imputations. These approaches hold promise in extremely large panels where linear low-rank models underfit important patterns.

From a causal perspective, however, deep models are still primarily prediction tools. To use them in treatment-effect analysis, we must embed them in the same potential-outcomes framework as the linear methods in this book. That means training them on untreated data to predict  $Y_{it}(0)$ , validating them with the diagnostics from Section 9.9, and plugging their predictions into the same ATT estimands as before. The

appeal is their flexibility. The risk is that their complexity makes them harder to diagnose and interpret, reinforcing the need for strong design and transparent diagnostics.

Advanced matrix and tensor methods also interact naturally with the causal machine-learning tools in Chapters 12 and 13. In double machine learning, for example, one can use matrix or tensor completion as one of the nuisance learners for outcome or treatment models, alongside lasso, random forests, or boosting. The orthogonality conditions in those chapters carry over unchanged: matrix or tensor completion can serve as one of the nuisance learners, but identification of treatment effects still depends on unconfoundedness or parallel-trends assumptions rather than on the flexibility of the learner. Cross-fitting then helps avoid overfitting when high-dimensional covariates coexist with low-rank structure. Group-lasso or related regularisers can be used to select covariates in the covariate-assisted part of the model, while low-rank structure captures remaining dependence.

In all these hybrids, matrix completion retains its role as a way to estimate nuisance components—conditional expectations or residuals—under low-rank assumptions. Identification of treatment effects still hinges on the same unconfoundedness or parallel-trends assumptions, now enforced in a richer function class.

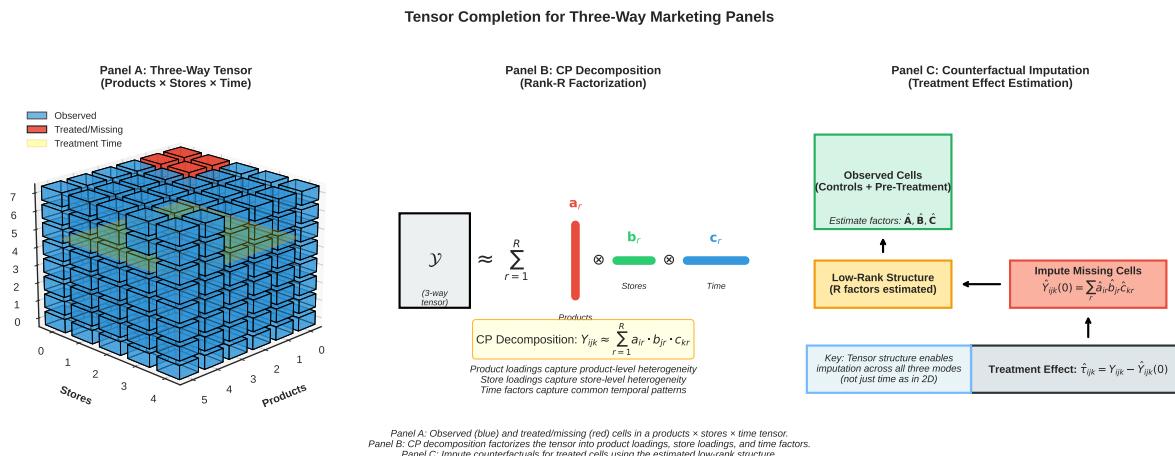
## Summary

The workflow and software recommendations in Section 9.11 provide a template for implementing these methods in practice, and the marketing case studies in Section 9.10 show how they work in concrete settings. Figures 9.1 and 9.2 visualise two core ideas: tensor completion as a way to borrow strength across multiple dimensions, and robust matrix completion as a way to separate systematic structure from outliers.

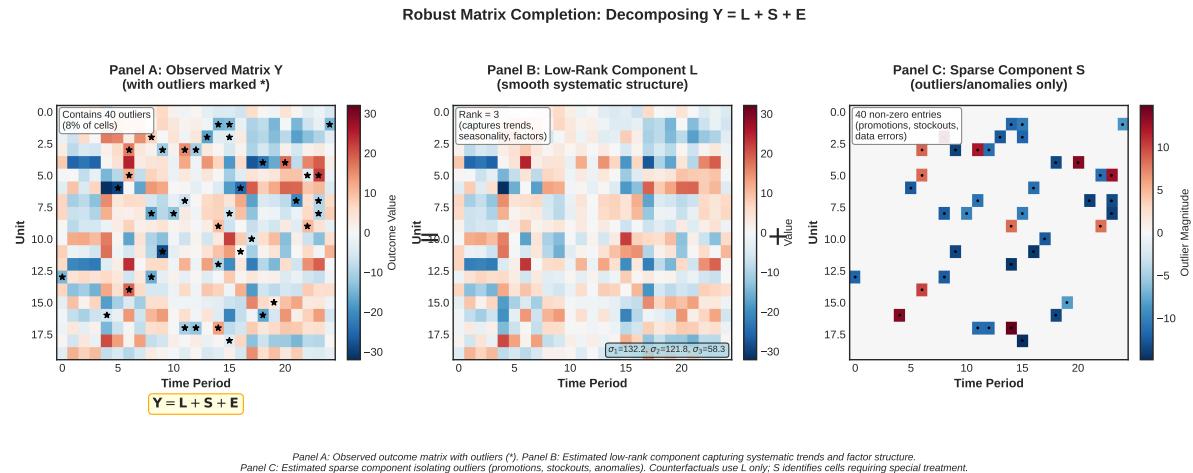
Used thoughtfully, with clear treatment designs, careful rank and regularisation choices, and rigorous diagnostics, advanced matrix and tensor methods give marketers a powerful set of tools for constructing counterfactuals in complex panels. They do not replace the design-based reasoning developed in earlier chapters. They extend it to the high-dimensional, sparse, and noisy data environments that increasingly characterise modern marketing. When treatment assignment remains correlated with shocks that are not captured by observed covariates or the low-rank structure, these methods still produce useful predictive counterfactuals but they no longer identify causal effects. In that case, their estimates should be interpreted as descriptive projections rather than as ATT or dynamic responses.

**Table 9.11** Comparison of advanced matrix and tensor methods

Method	Key feature	Best suited for	Computational profile
Tensor completion	Preserves multi-way structure	Multi-way panels with informative modes	High, scales with number of modes and rank
Robust MC	Separates outliers from low rank	Panels with frequent, large irregularities	Medium, convex but heavier than standard MC
Covariate-assisted	Incorporates side information	Panels with strong, plausibly exogenous covariates	Low to medium, regression plus MC
Time-varying rank	Adapts to structural breaks	Non-stationary environments with regime shifts	Medium, online or piecewise estimation
Bayesian	Provides posterior uncertainty	Settings where full uncertainty quantification is needed	High, MCMC or variational inference
Soft-Impute and related	Efficient for sparse matrices	Very large, sparse two-way panels	Low, scalable to millions of cells

**Fig. 9.1** Tensor completion schematic for three-way panels

Note: Panel A shows a three-way tensor ( $\text{products} \times \text{stores} \times \text{time}$ ) with observed cells (blue) and treated or missing cells (red). Panel B illustrates the CP decomposition into product loadings, store loadings, and time factors. Panel C shows the workflow for imputing counterfactual untreated outcomes and constructing treatment effects.



**Fig. 9.2** Robust matrix completion: separating low-rank structure from outliers

*Note:* The decomposition  $\mathbf{Y} = \mathbf{L} + \mathbf{S} + \mathbf{E}$  separates the observed matrix into a low-rank component capturing systematic structure, a sparse component isolating outliers, and a noise term. This separation improves imputation of untreated potential outcomes in the presence of stockouts, data errors, or other extreme observations.

## **Part V**

### **Dynamics, Heterogeneity, and Spillovers**



## Chapter 10

# Dynamic Treatment Effects

This chapter extends event-study estimands to dynamic, path-dependent settings. We define dynamic treatment paths, event-time effects, and summary measures such as long-run multipliers and half-lives. Throughout the chapter we follow the notation from Chapter 2: dynamic treatment paths are written as  $\underline{d}_i^t$ , path-dependent potential outcomes as  $Y_{it}(\underline{d}_i^t)$ , and event-time effects as  $\theta_k$ . When  $\theta_0$  is substantively non-zero, we summarise persistence using  $LRM = \sum_{k=0}^{\bar{K}} \theta_k / \theta_0$ , with  $\bar{K}$  chosen so that  $\theta_k$  has effectively dissipated. When  $\theta_0$  is close to zero, we report the long-run effect  $\sum_{k=0}^{\bar{K}} \theta_k$  directly.

We show how to apply identification strategies under staggered adoption for dynamic effects, clarifying no-anticipation versus limited anticipation. The dynamic structure sharpens how we summarise effects over time but does not relax the design and identification assumptions from Chapters 4 and 5. Parallel trends or related conditions remain the core requirements. You will implement heterogeneity-robust event studies, distributed-lag models, and dynamic difference-in-differences for continuous and discrete treatments. Finally, you will conduct diagnostics and inference that quantify uncertainty conditional on the identification assumptions. Inference treats units (or clusters) as independent and allows serial dependence within them.

## 10.1 Motivation and Setup

Many marketing interventions shift outcomes over multiple periods, so a single post-treatment average obscures the timing of effects. In this chapter we make the target dynamic estimand explicit before discussing identification and estimation. A television advertising campaign builds awareness over weeks and generates incremental sales over months. Effects decay gradually as memory fades and competitive messages crowd the consumer's mind. A promotional discount accelerates buying in the promotion period, creating stockpiling and purchase acceleration, followed by post-promotion dips as consumers draw down inventories. A loyalty programme enrols customers gradually, generates habit formation and switching costs over quarters or years, and exhibits persistence long after the programme ends.

These dynamic patterns are central to marketing decision-making. Managers must forecast the cumulative impact of campaigns, understand the time profile of returns, and decide whether to sustain, refresh, or terminate investments based on the speed of ramp-up and decay.

### Key Concepts: Dynamic Treatment Effects

An impulse response describes how a one-period change in treatment today affects outcomes over future periods. In staggered adoption with absorbing treatment, the event-time effects  $\theta_k$  provide a closely related dynamic summary (Section 10.2).

The long-run multiplier cumulates effects across lags and normalises by the impact effect. When  $\theta_0$  is substantively non-zero, we compute  $\text{LRM} = \sum_{k=0}^{\bar{K}} \theta_k / \theta_0$ , with  $\bar{K}$  chosen so that  $\theta_k$  has effectively dissipated. When  $\theta_0$  is close to zero or changes sign, we report the long-run effect  $\sum_{k=0}^{\bar{K}} \theta_k$  directly. Half-life measures the time required for effects to decay to half their initial magnitude. If effects follow geometric decay  $\beta_s = \beta_0 \delta^s$ , the half-life is  $h^* = \log 2 / (-\log \delta)$ .

Event time indexes periods relative to treatment adoption, with  $k = t - G_i$  where  $G_i$  denotes the first treated period for unit  $i$ . Pre-treatment periods have  $k < 0$ , adoption occurs at  $k = 0$ , and post-treatment periods have  $k > 0$ .

Carryover describes how past treatments influence current outcomes. Anticipation describes how expected future treatments influence current outcomes. When outcomes depend on the treatment history, we represent this dependence using path-dependent potential outcomes  $Y_{it}(\underline{d}_i^t)$ . This notation is a bookkeeping device that lets us state no-anticipation and carryover restrictions precisely (see Section 10.2).

Dynamic treatment effects formalise the time dimension of causal inference in panels. Rather than estimating a single average treatment effect that averages over all post-treatment periods, dynamic methods trace out the full trajectory of effects over event time, estimating impulse responses that quantify how outcomes respond to treatment shocks over lags and summarising dynamics through the long-run multiplier and half-life. These estimands connect directly to substantive questions in marketing: How long does advertising continue to generate incremental sales? When does a promotion's stockpiling effect dissipate? How persistent are loyalty programme effects?

## Design-Based Perspective

The design-based perspective emphasised by Angrist and Pischke [2010] frames dynamics naturally. Chapters 4 and 5 provide the main estimands. This chapter adds explicit dynamic structure and connects to distributed lags and ad-stock models. Event-study designs visualise treatment effects over leads and lags, using pre-trend diagnostics and tracing post-treatment dynamics transparently (see Chapter 5 for specification and plotting). Staggered difference-in-differences designs aggregate cohort-time effects into event-time profiles, averaging effects across cohorts at each lag while avoiding contamination from already-treated controls (see Chapter 4 for estimands and weights).

Synthetic control methods impute counterfactuals that evolve dynamically, permitting estimation of time-varying treatment effects for single treated units (see Chapters 6 and 7). We prefer design-based methods because they minimise functional-form restrictions. Conditional on a credible assignment mechanism (for example, parallel trends plus no anticipation in staggered adoption), they deliver interpretable estimates of event-time effects. Sufficient pre-treatment data are needed for diagnostics and credibility assessment. For design diagnostics and inference under serial dependence, see Chapters 17 and 16.

## Model-Based Ad-Stock Tradition

The model-based ad-stock tradition, prevalent in marketing science since Clarke [1976], complements the design-based approach by imposing economic structure on dynamic relationships. Identification still ultimately relies on the assignment mechanism rather than on functional-form assumptions. Structural restrictions can sharpen estimation, but they do not substitute for a credible design. For measure-theoretic foundations, see Appendix A.

The ad-stock approach posits that advertising effects accumulate through a stock variable that depreciates geometrically, generating carryover that decays exponentially. Distributed-lag models generalise this structure by allowing flexible lag specifications—polynomial, unrestricted finite lags—that accommodate non-monotonic dynamics such as wear-in or delayed effects. Habit formation models embed dynamics in utility functions, generating state dependence where past choices influence current preferences.

These structural models yield interpretable parameters (depreciation rates, habit coefficients) and enable counterfactual simulations. However, they impose strong functional form assumptions that may not hold in the data.

## Synthesis: Design-Based and Model-Based

### When to Use Design-Based vs Model-Based Methods

Design-based methods such as event studies, difference-in-differences, and synthetic control are appropriate when the goal is to document the presence and shape of dynamics transparently, when functional form assumptions are uncertain or unverifiable, when pre-treatment data are available for parallel-trends diagnostics, and when credibility of identification is paramount.

Model-based methods such as ad-stock, distributed-lag, and habit formation models are appropriate when the goal is to forecast effects under untried policies, when structural parameters (depreciation rates, habit coefficients) are of substantive interest, when economic theory provides credible restrictions, or when data are insufficient for flexible design-based estimation.

A synthesis approach proceeds in four stages. First, estimate a design-based event study to document the dynamics transparently. Second, fit a model-based specification to impose structure and sharpen estimates. Third, compare the two and assess sensitivity to functional form. Fourth, report both sets of results and treat the model-based specification as sensitivity analysis when it disagrees with the design-based estimates.

Design-based methods deliver transparent, assumption-lean estimates. Model-based methods impose structure that sharpens estimation, permits extrapolation, and facilitates interpretation. The appropriate choice depends on the substantive question, data richness, and credibility of structural assumptions.

## Connection to Causal Frameworks

The connection to potential outcomes and causal frameworks (Chapter 2) sharpens the estimands. When outcomes depend on the treatment history, we represent this dependence using path-dependent potential outcomes  $Y_{it}(\underline{d}_i^t)$ . This notation is a bookkeeping device that lets us state no-anticipation and carryover restrictions precisely (see Section 10.2). Formally, we write treatment paths as  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$  and potential outcomes as  $Y_{it}(\underline{d}_i^t)$ , with adoption-time shorthand  $Y_{it}(g)$  and  $Y_{it}(\infty)$  defined in Chapter 2. The observed outcome is a function of the entire treatment path, and the treatment effect at time  $t$  contrasts outcomes under the observed path with an alternative path (for example, no treatment). Estimating such effects requires comparing units with different treatment paths while holding constant other determinants—motivating panel data where multiple paths are observed.

### Example: Promotional Calendar Dynamics

Consider a retailer that implements discounts every four weeks over a year. Sales spike in promotion weeks, dip in weeks immediately following (as consumers draw down stockpiled inventories), and return to baseline before the next promotion. A single ATT that pools all periods obscures this rich dynamic structure and provides little guidance for optimising the promotion calendar.

An event-study design traces sales over leads and lags relative to promotion week. The design reveals the spike (immediate effect), the dip (post-promotion stockpiling effect), and the return to baseline (effect duration). It quantifies the net cumulative effect accounting for both spike and dip, informing the decision of whether to increase or decrease promotion frequency.

See Section 10.8 for an illustrative worked example.

## Chapter Roadmap

This chapter develops dynamic treatment effect methods with a focus on practical implementation. Table 10.1 provides the chapter structure.

**Table 10.1** Chapter 10 Roadmap

Section	Topic	Key Content
10.2	Estimands	Path-dependent potential outcomes, event-time effects, impulse responses, long-run multipliers
10.3	Identification	Parallel trends, no-anticipation, control group choices
10.4	Estimation	Heterogeneity-robust event studies, distributed-lag models, dynamic DiD
10.5	Anticipation & Carryover	Modelling anticipation, separating carryover, mediation
10.6	Inference	Serial dependence, multiple testing, uniform confidence bands
10.7	Diagnostics	Pre-trend diagnostics, lag selection, stability checks
10.8	Marketing Applications	Advertising dynamics, promotional calendars, loyalty programmes
10.9	Workflow	Checklist for practitioners

Together, these sections show you how to estimate, diagnose, and report dynamic treatment effects, and how to communicate uncertainty conditional on the identification assumptions.

## 10.2 Potential Outcomes and Dynamic Estimands

Dynamic treatment effects require careful specification of the estimand: what causal quantity is being estimated? This section formalises path-dependent potential outcomes, defines event-time effects and distributed-lag impulse responses, and introduces summary metrics including the long-run multiplier and the half-life. These definitions refine the event-time estimands introduced in Chapters 4 and 5, now expressed explicitly in path-dependent potential-outcome notation. Clear estimand definitions are essential because different estimands require different identification assumptions and estimation methods, and because policy-relevant questions (How long do effects last? What is the cumulative impact?) map to specific estimands. Section 10.1 provides intuition and marketing motivation.

### Path-Dependent Potential Outcomes

Path-dependent potential outcomes extend the standard potential outcomes framework to allow outcomes at time  $t$  to depend on the full history of treatment assignments.

**Definition 10.1 (Treatment Path)** For unit  $i$  observed over periods  $t = 1, \dots, T$ , define:

- (i) The **contemporaneous treatment**  $D_{it} \in \{0, 1\}$  (or  $D_{it} \in \mathcal{D} \subseteq \mathbb{R}$  for continuous treatments).
- (ii) The **treatment history** up to period  $t$ :  $\underline{d}_i^t = (d_{i1}, d_{i2}, \dots, d_{it}) \in \{0, 1\}^t$ .
- (iii) The **full treatment path**:  $\underline{d}_i = (d_{i1}, \dots, d_{iT}) \in \{0, 1\}^T$ .
- (iv) The **adoption time** for absorbing treatments:  $G_i = \min\{t : D_{it} = 1\}$ , with  $G_i = \infty$  for never-treated units.

**Definition 10.2 (Path-Dependent Potential Outcomes)** The potential outcome  $Y_{it}(\underline{d}_i^t)$  is the outcome unit  $i$  would exhibit at time  $t$  under treatment history  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$ . Write the realised outcome as  $Y_{it} = Y_{it}(\underline{D}_i^t)$ , where  $\underline{D}_i^t$  is the realised history up to  $t$ .

With  $T$  periods and binary treatment, the number of possible histories grows exponentially. At time  $t$  there are  $2^t$  possible histories, so by the final period there are  $2^T$  possible histories. For a panel with  $T = 52$  weeks, this is  $2^{52} \approx 4.5 \times 10^{15}$ . This exponential growth explains why restrictions (absorbing treatment, no anticipation, lag truncation) are essential for identification.

**Definition 10.3 (Canonical Treatment Paths)** Define the following canonical paths:

- (i) **Never-treated path**:  $\underline{d}^{\text{never}} = (0, 0, \dots, 0) \in \{0\}^T$ .
- (ii) **Always-treated path**:  $\underline{d}^{\text{always}} = (1, 1, \dots, 1) \in \{1\}^T$ .
- (iii) **Adoption-at- $g$  path**:  $\underline{d}^{\text{adopt}}(g) = (\underbrace{0, \dots, 0}_{g-1}, \underbrace{1, \dots, 1}_{T-g+1})$ , switching from 0 to 1 at period  $g$ .

The untreated potential outcome is  $Y_{it}(\infty) \equiv Y_{it}(\underline{d}^{\text{never}})$ , and the treated potential outcome under adoption at  $g$  is  $Y_{it}(g) \equiv Y_{it}(\underline{d}^{\text{adopt}}(g))$ .

### Event-Time Effects

Event-time effects  $\theta_k$  average treatment effects across units at a fixed lag  $k$  relative to treatment adoption.

**Definition 10.4 (Event-Time Effects)** For cohort  $g$  (units first treated in period  $g$ ) and event time  $k = t - g$  (periods relative to adoption):

- (i) The **cohort-specific event-time effect** is:

$$\theta_k(g) = \mathbb{E} [Y_{i,g+k}(g) - Y_{i,g+k}(\infty) \mid G_i = g]$$

Note that  $\theta_k(g)$  corresponds to the cohort-time average treatment effect  $\tau(g, t)$  defined in Chapter 4, where  $t = g + k$ .

- (ii) The **aggregate event-time effect** is:

$$\theta_k = \sum_{g \in \mathcal{G}_k} w_g^k \theta_k(g),$$

where  $\mathcal{G}_k = \{g : 1 \leq g \leq T - k\}$  is the set of cohorts observed at event time  $k$ , and  $w_g^k$  are cohort weights satisfying  $\sum_g w_g^k = 1$ .

Event-time effects trace the dynamic trajectory of treatment responses. For advertising,  $\theta_0$  is the immediate effect when the campaign airs,  $\theta_1$  is the one-period-ahead effect (carryover),  $\theta_2$  is the two-period-ahead effect, and so on. Plotting  $\theta_k$  against  $k$  visualises whether effects build (ramp-up), peak and decay (carryover), or exhibit non-monotonic patterns (wear-in then wear-out).

### Distributed-Lag Impulse Responses

Distributed-lag impulse responses provide an alternative representation of dynamics.

**Definition 10.5 (Distributed-Lag Impulse Response)** Consider a linear distributed-lag model:

$$Y_{it} = \sum_{s=0}^{\bar{L}} \beta_s D_{i,t-s} + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $D_{it} \in \mathbb{R}$  is treatment intensity,  $\bar{L}$  is the maximum lag,  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects, and  $\varepsilon_{it}$  are idiosyncratic errors.

- (i) The **impulse response function** is the sequence  $\{\beta_s\}_{s=0}^{\bar{L}}$ .
- (ii) The **contemporaneous effect** is  $\beta_0$ .
- (iii) The **cumulative effect at horizon  $h$**  is  $\sum_{s=0}^h \beta_s$ .

The impulse response captures the total effect of a transient treatment shock (a pulse lasting one period). A TV campaign airing for one week generates sales in the campaign week ( $\beta_0$ ) and incremental sales in subsequent weeks ( $\beta_1, \beta_2, \dots$ ) as consumers respond with lags.

## Long-Run Multiplier

**Definition 10.6 (Long-Run Effect and Long-Run Multiplier)** The long-run effect aggregates dynamic impacts over lags, and the long-run multiplier scales this effect by the impact response.

- (i) For distributed-lag models with impulse response coefficients  $\{\beta_s\}$ , the long-run *effect* is  $\sum_{s=0}^{\bar{L}} \beta_s$  and the long-run *multiplier* is

$$\text{LRM} = \frac{\sum_{s=0}^{\bar{L}} \beta_s}{\beta_0}.$$

- (ii) For event-time effects with absorbing binary treatment and flow outcomes, the long-run *effect* is  $\sum_{k=0}^{\bar{K}} \theta_k$  and the long-run *multiplier*, consistent with the notation guide, is

$$\text{LRM} = \frac{\sum_{k=0}^{\bar{K}} \theta_k}{\theta_0}.$$

For stock outcomes (for example, brand equity), the long-run level effect at horizon  $\bar{K}$  is  $\theta_{\bar{K}}$  rather than a sum.

- (iii) Under geometric decay with rate  $\delta \in (0, 1)$  and  $\beta_s = \beta_0 \delta^s$ , the long-run effect as  $\bar{L} \rightarrow \infty$  is  $\beta_0/(1 - \delta)$  and the corresponding multiplier is  $\text{LRM} = 1/(1 - \delta)$ .

When  $\beta_0$  (or  $\theta_0$ ) is close to zero, report the long-run effect directly and treat the multiplier as secondary. The long-run effect exists and is finite if the impulse response is absolutely summable,  $\sum_{s=0}^{\infty} |\beta_s| < \infty$ .

The absolute summability condition ensures that effects decay fast enough that the cumulative sum converges. Geometric decay with  $\delta < 1$  satisfies this condition (a convergent geometric series), while polynomial decay  $\beta_s \propto s^{-\alpha}$  requires  $\alpha > 1$ .

## Half-Life

**Definition 10.7 (Half-Life)** The half-life  $h^*$  is the time required for the treatment effect to decay to half its initial value:

- (i) For geometric decay  $\beta_s = \beta_0 \delta^s$ :

$$h^* = \frac{\log(0.5)}{\log(\delta)} = \frac{\log 2}{-\log \delta}$$

- (ii) For general impulse responses, two definitions are common: the effect half-life, where  $h^*$  solves  $\beta_{h^*} = \beta_0/2$  (time until instantaneous effect halves), and the cumulative half-life, where  $h^*$  solves  $\sum_{s=0}^{h^*} \beta_s = \text{Long-Run Effect}/2$  (time until half the total effect has occurred).

The half-life is inversely related to the decay rate: faster decay (smaller  $\delta$ ) implies shorter half-life.

#### Example: Computing LRM and Half-Life

Consider a weekly advertising campaign with estimated impulse response  $\beta_s = 100 \times 0.7^s$  (geometric decay with  $\delta = 0.7$  and  $\beta_0 = 100$  units of incremental sales). The long-run effect is  $\sum_{s=0}^{\infty} \beta_s = \beta_0/(1 - \delta) = 100/0.3 = 333.3$  units, meaning a one-week campaign generates 333 units of cumulative incremental sales over all time. The corresponding long-run multiplier is  $1/(1 - \delta) = 3.33$ , which expresses the total effect in units of the impact effect  $\beta_0$ . The half-life is  $h^* = \log 2 / (-\log 0.7) = 0.693/0.357 = 1.94$  weeks, meaning effects decay to half their initial magnitude in approximately 2 weeks. With a 2-week half-life, weekly pulsing maintains awareness better than monthly pulsing. If the long-run effect is 333 units and cost-per-campaign is \$10,000, the cost-per-incremental-unit is \$30.

## Correspondence Between Event-Time and Impulse Response

**Proposition 10.1 (Correspondence between Event-Time and Impulse Response)** *Under the following conditions:*

- (i) *Treatment is binary and absorbing (once adopted, remains on).*
- (ii) *Effects are additively separable across lags,*

*the event-time effects  $\theta_k$  and impulse responses  $\beta_s$  are related by:*

$$\theta_k = \sum_{s=0}^k \beta_s \Leftrightarrow \beta_k = \theta_k - \theta_{k-1},$$

*with the convention  $\theta_{-1} = 0$ . Under absorbing treatment, the level effect at lag  $k$ ,  $\theta_k$ , embeds the cumulative impact of all past shocks, so  $\beta_k$  can be recovered as the increment  $\theta_k - \theta_{k-1}$ . Note that for flow outcomes (for example, weekly sales),  $\theta_k$  in Chapters 4 and 5 represents the per-period effect at lag  $k$ , and the long-run effect is  $\sum_k \theta_k$ . The correspondence here shows how to recover impulse responses from that representation.*

*Remark 10.1* The algebraic mapping in Proposition 10.1 is structural and does not require identification assumptions. Parallel trends and related conditions are needed to identify  $\theta_k$  from the data (see Section 10.3), but the relationship between  $\theta_k$  and  $\beta_k$  is purely definitional given the absorbing-treatment structure.

This correspondence enables analysts to move between event-study designs (which estimate  $\theta_k$ ) and distributed-lag models (which estimate  $\beta_s$ ), choosing whichever representation suits the substantive question.

## Summary of Estimands

Table 10.2 summarises the key dynamic estimands.

**Table 10.2** Dynamic Estimands

Estimand	Formula	Interpretation
Cohort-specific event-time effect $\theta_k(g)$	$\mathbb{E}[Y_{i,g+k}(g) - Y_{i,g+k}(\infty) \mid G_i = g]$	Effect $k$ periods after adoption for cohort $g$
Aggregate event-time effect $\theta_k$	$\sum_{g \in \mathcal{G}_k} \omega_g^k \theta_k(g)$	Weighted-average effect $k$ periods after adoption
Impulse response $\beta_s$	Coefficient $\beta_s$ in the distributed-lag model (Definition 10.5)	Effect of a lag- $s$ treatment shock
Long-run effect	$\sum_{s=0}^{\bar{L}} \beta_s$ (impulse) or $\sum_{k=0}^{\bar{K}} \theta_k$ (event-time, flows)	Cumulative effect over all lags
Long-run multiplier	$(\sum_{s=0}^{\bar{L}} \beta_s)/\beta_0$ or $(\sum_{k=0}^{\bar{K}} \theta_k)/\theta_0$	Long-run effect in units of the impact effect
Half-life $h^*$	$\log 2 / (-\log \delta)$	Time to decay to 50%
Cumulative effect at $h$	$\sum_{s=0}^h \beta_s$	Total effect through horizon $h$

### Practical Guidance: Defining Dynamic Estimands

Pre-specify the event-time window based on substantive knowledge about expected effect duration. For advertising effects, a window of 0–12 weeks is typical. For loyalty programmes, 0–8 quarters is appropriate. Report all lags within the window and bin distant lags when support is sparse.

Define the long-run multiplier carefully based on the estimand type and outcome nature. For **flow outcomes** (for example, weekly sales), the long-run effect  $\sum_{k=0}^{\bar{K}} \hat{\theta}_k$  sums the period-specific event-time effects, while the multiplier LRM scales this by  $\hat{\theta}_0$ . For **stock outcomes** (for example, brand equity), the long-run effect is  $\hat{\theta}_{\bar{K}}$ , the level at the longest observed lag. For impulse responses, the long-run effect  $\sum_{s=0}^{\bar{L}} \hat{\beta}_s$  sums the lag-specific effects. If geometric decay is plausible, extrapolate the long-run effect using  $\hat{\theta}_0/(1 - \hat{\delta})$ .

Estimate half-life with uncertainty. Compute impulse responses as  $\hat{\beta}_k = \hat{\theta}_k - \hat{\theta}_{k-1}$ , fit geometric decay  $\hat{\beta}_k = \hat{\beta}_0 \hat{\delta}^k$  to the impulse response profile, compute  $\hat{h}^* = \log 2 / (-\log \hat{\delta})$ , and report uncertainty using the delta method or bootstrap.

Report both event-time effects and impulse responses. Event-time effects  $\hat{\theta}_k$  provide direct interpretation of cumulative impacts at each lag. Impulse responses  $\hat{\beta}_k = \hat{\theta}_k - \hat{\theta}_{k-1}$  reveal the lag structure and decay pattern. Use the correspondence proposition to check internal consistency between the two representations.

### 10.3 Identification

Causal identification of dynamic treatment effects requires assumptions about how treated and control units would have evolved in the absence of treatment, about whether agents anticipate future treatments, and about the structure of the comparison group. This section articulates these assumptions, clarifies the role of parallel trends and no-anticipation, and discusses control group choices under staggered adoption.

Table 10.3 summarises the key assumptions for dynamic treatment effects.

**Table 10.3** Identification Assumptions for Dynamic Effects

Assumption	Requirement	When Required
Parallel trends	Treated and control share time shocks	Event-study, DiD
No anticipation	Pre-treatment outcomes unaffected	Event-study, DiD
Limited anticipation	Anticipation bounded by $\bar{a}$ periods	If anticipation present
Overlap/support	Comparison group exists at each lag	All dynamic methods
Strict exogeneity	$\mathbb{E}[\varepsilon_{it}   D_{i1}, \dots, D_{iT}, X_{i1}, \dots, X_{iT}, \alpha_i] = 0$	Distributed-lag, short panels
Sequential exogeneity	$\mathbb{E}[\varepsilon_{it}   D_{i1}, \dots, D_{it}, X_{i1}, \dots, X_{it}, \alpha_i] = 0$	Dynamic panels, Arellano–Bond
No carryover	Effects depend only on $D_{it}$	Simplifies to static ATT

#### Parallel Trends for Event-Time Effects

The core identification assumption for event-time effects is parallel trends for dynamic contrasts.

**Assumption 37 (Parallel Trends for Event-Time Identification)** For each cohort  $g$  and event time  $k$ , the average untreated potential outcome for cohort  $g$  evolves parallel to the comparison group:

$$\mathbb{E}[Y_{i,g+k}(\infty) - Y_{i,g-1}(\infty) | i \in \mathcal{C}_{g,k}^{\text{treated}}] = \mathbb{E}[Y_{i,g+k}(\infty) - Y_{i,g-1}(\infty) | i \in \mathcal{C}_{g,k}],$$

where  $\mathcal{C}_{g,k}^{\text{treated}}$  denotes units with  $G_i = g$  and  $\mathcal{C}_{g,k}$  is the comparison group for cohort  $g$  at event time  $k$  (never-treated or not-yet-treated units).

This formulation imposes parallel trends in changes from the baseline period  $g - 1$  to each later period  $g + k$ , rather than parallel levels. It ensures that, absent treatment, cohort  $g$  and its comparison group would experience the same cumulative change between  $g - 1$  and  $g + k$ .

This is the dynamic, event-time analogue of the staggered-parallel-trends assumption in Chapter 4. It requires that, absent treatment, cohort  $g$  and its comparison group would experience the same period-to-period changes between the baseline  $g - 1$  and each later lag  $g + k$ . Parallel trends has diagnostic implications in pre-treatment periods (see Proposition 10.2 and the pre-trend discussions in Chapters 4 and 5).

## Control Group Choices

The comparison group  $\mathcal{C}_{g,k}$  must be chosen carefully to avoid contamination and ensure parallel trends.

**Table 10.4** Control Group Choices

Control Group	Definition	Advantages	Disadvantages
Never-treated	Units never adopting treatment	Stable benchmark, no contamination	May be non-representative, sparse if universal adoption
Not-yet-treated	Units adopting later than cohort $g$	Larger pool, more comparable to treated	Moving composition, shrinks at distant lags
Clean controls	Early adopters excluded from sample	Avoids contamination from already-treated	May be sparse, selection issues

### Control Group Selection

The primary recommendation is to use never-treated units when available and comparable, as they provide a stable benchmark uncontaminated by treatment effects. When never-treated units are sparse, use not-yet-treated units but restrict the event-time window to lags where the control pool remains stable and large, report the number of control units at each lag via a support plot, and bin distant lags when support becomes sparse. Avoid using already-treated units as controls because their outcomes are contaminated by their own treatment effects. See Chapter 4 for comprehensive guidance on staggered adoption.

## Anticipation Assumptions

These assumptions refine the path-dependent potential-outcomes framework in Section 10.2 and the no-anticipation discussion in Chapters 2 and 5.

**Assumption 38 (No Anticipation)** Treatment does not affect outcomes before adoption:

$$Y_{it}(\underline{d}_i) = Y_{it}(\underline{d}'_i) \quad \text{for all } \underline{d}_i, \underline{d}'_i \text{ such that } d_{is} = d'_{is} \text{ for all } s \leq t.$$

Under this restriction, the potential outcome at time  $t$  depends only on the treatment history up to time  $t$ .

$$Y_{it}(\underline{d}_i) = Y_{it}(\underline{d}_i^t),$$

where  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$ .

No anticipation ensures that pre-treatment outcomes provide unbiased estimates of the treated group's untreated trajectory.

**Assumption 39 (Limited Anticipation)** Treatment can be anticipated up to  $\bar{a}$  periods in advance:

$$Y_{it}(\underline{d}_i) = Y_{it}(\underline{d}'_i) \quad \text{for all } \underline{d}_i, \underline{d}'_i \text{ such that } d_{is} = d'_{is} \text{ for all } s \leq t + \bar{a}.$$

For adoption at  $g$ , anticipatory effects may occur in periods  $g - \bar{a}, \dots, g - 1$ , but outcomes in periods  $t < g - \bar{a}$  are unaffected.

Operationally, limited anticipation means that event-time plots should exclude or interpret with caution the pre-treatment lags inside the anticipation window  $\{g - \bar{a}, \dots, g - 1\}$ . Pre-trend diagnostics and placebo checks should focus on earlier lags  $k < -\bar{a}$ , where outcomes are guaranteed to be unaffected by future treatment under this assumption.

Limited anticipation relaxes no-anticipation by allowing a finite anticipation window. See Section 10.5 for estimation under anticipation.

## Overlap and Support

**Assumption 40 (Overlap and Support in Event Time)** For each  $(g, k)$  used in estimation, there are untreated comparison units at time  $g + k$  with overlapping covariates relative to cohort  $g$ . Formally, for units with  $G_i = g$  and characteristics  $X_i$  in the support of the treated distribution, there exist comparison units with similar  $X_i$  in  $\mathcal{C}_{g,k}$ , and the comparison group size satisfies  $|\mathcal{C}_{g,k}| \geq M$  for a minimum threshold  $M$  (for example,  $M = 10$ ).

In practice this means that, for each cohort  $g$  that contributes to event time  $k$ , treated units with  $G_i = g$  have comparable units in  $\mathcal{C}_{g,k}$  with similar  $X_i$  and sufficient sample size. When adoption is nearly deterministic in  $X_i$  or when almost all units adopt by early periods, overlap and support at larger lags break down.

Overlap ensures comparability. Support ensures stable estimates. Violations occur when treatment adoption is deterministic or when the comparison pool shrinks at distant lags. To diagnose support violations, plot the number of treated and control units at each lag (the support plot), flag lags where support is sparse, and bin distant lags (for example, pooling lags 10–12) when necessary. See Section 10.7 for event-study-specific diagnostics and Chapter 17 for comprehensive diagnostic guidance.

## Carryover and Exogeneity

**Assumption 41 (No Carryover)** Treatment effects depend only on contemporaneous treatment:

$$Y_{it}(\underline{d}_i) = Y_{it}(d_{it}) \quad \text{for all } \underline{d}_i \in \{0, 1\}^T.$$

This reduces the potential outcomes to  $Y_{it}(0)$  and  $Y_{it}(1)$ , with treatment effect  $\tau_{it} = Y_{it}(1) - Y_{it}(0)$ .

No carryover is rarely satisfied in marketing (effects persist over time). When carryover is present, use distributed-lag or event-study methods that explicitly model dynamics.

**Assumption 42 (Strict Exogeneity)** In the distributed-lag model, treatment is strictly exogenous:

$$\mathbb{E}[\varepsilon_{it} \mid D_{i1}, \dots, D_{iT}, X_{i1}, \dots, X_{iT}, \alpha_i] = 0 \quad \text{for all } t = 1, \dots, T.$$

This rules out feedback from current shocks to treatment, including through future treatment choices.

**Assumption 43 (Sequential Exogeneity)** Treatment is sequentially exogenous (predetermined):

$$\mathbb{E}[\varepsilon_{it} \mid D_{i1}, \dots, D_{it}, X_{i1}, \dots, X_{it}, \alpha_i] = 0 \quad \text{for all } t = 1, \dots, T.$$

This allows treatment to respond to past outcomes and past shocks, but requires that the current innovation is uncorrelated with the current treatment, conditional on the observed history.

*Remark 10.2 (Hierarchy of Exogeneity Assumptions)* The exogeneity assumptions form a hierarchy from strongest to weakest. Random assignment in the dynamic setting requires that the assignment mechanism for the full panel is independent of the collection of dynamic potential outcomes conditional on observables:

$$\Pr(\mathbf{D} \mid \{Y_{it}(\underline{d}_i^t) : \underline{d}_i^t \in \mathcal{D}^t\}_{i,t}, \mathbf{X}) = \Pr(\mathbf{D} \mid \mathbf{X}),$$

where  $\mathbf{D}$  denotes the full treatment assignment matrix,  $\mathbf{X}$  denotes the full covariate history used by the assignment mechanism,  $\mathcal{D}$  is the treatment space, and  $\mathcal{D}^t$  denotes length- $t$  treatment histories. Strict exogeneity requires

$$\mathbb{E}[\varepsilon_{it} \mid \underline{D}_i, X_{i1}, \dots, X_{iT}, \alpha_i] = 0 \quad \text{for all } t,$$

where  $\underline{D}_i = (D_{i1}, \dots, D_{iT})$ . Sequential exogeneity requires

$$\mathbb{E}[\varepsilon_{it} \mid \underline{D}_{it}, X_{i1}, \dots, X_{it}, \alpha_i] = 0 \quad \text{for all } t,$$

where  $\underline{D}_{it} = (D_{i1}, \dots, D_{it})$ . Contemporaneous exogeneity requires only

$$\mathbb{E}[\varepsilon_{it} \mid D_{it}, \alpha_i] = 0 \quad \text{for all } t.$$

Following Arellano [2003] and Chamberlain [1984], strict exogeneity is required for within-group estimation in short panels. Sequential exogeneity permits dynamic specifications but requires instrumental variables (Arellano–Bond). Design-based identification (randomisation, staggered DiD with clean controls) is generally preferable to leaning on strict or sequential exogeneity in purely observational settings.

## Relation to Factor Designs

Parallel trends requires that treated and control groups share the same time-varying shocks. This fails when groups differ in sensitivity to common shocks (for example, large markets respond more to recessions). Factor models (Chapter 8) relax parallel trends by allowing heterogeneous exposure, estimating unit-specific loadings. Factor-based identification replaces Assumption 37 with a stronger structural restriction on  $Y_{it}(0)$ : it assumes that untreated outcomes lie in a low-rank factor space with stable loadings, and uses that structure to impute dynamic counterfactuals when simple parallel trends is implausible.

## Identification Results

**Theorem 10.1 (Identification of Event-Time Effects)** *Under Assumptions 38 and 37, the cohort-specific event-time effect  $\theta_k(g)$  (Definition 10.4) is identified by:*

$$\theta_k(g) = (\mathbb{E}[Y_{i,g+k} \mid G_i = g] - \mathbb{E}[Y_{i,g-1} \mid G_i = g]) - (\mathbb{E}[Y_{i,g+k} \mid i \in \mathcal{C}_{g,k}] - \mathbb{E}[Y_{i,g-1} \mid i \in \mathcal{C}_{g,k}]).$$

*The aggregate event-time effect  $\theta_k$  is identified by aggregating across cohorts:*

$$\theta_k = \sum_{g \in \mathcal{G}_k} w_g^k \theta_k(g).$$

**Proposition 10.2 (Pre-Trend Diagnostic)** *Under Assumption 38, the event-time effects at negative lags satisfy  $\theta_k(g) = 0$  for all  $k < 0$ . This provides a diagnostic implication: if  $\hat{\theta}_k \neq 0$  for  $k < 0$ , then either anticipation is present (Assumption 38 fails) or parallel trends fails (Assumption 37 fails). Use lead coefficients as a diagnostic: large or patterned leads indicate tension with no-anticipation or with parallel trends. Non-rejection is inconclusive and should be interpreted alongside design evidence on announcement timing and information sets. For inference on the pre-trend statistic and its distribution, see Proposition 10.5 in Section 10.6.*

**Theorem 10.2 (Identification of Impulse Response)** *Under Assumption 42 and the rank condition, the impulse response coefficients  $\{\beta_s\}_{s=0}^{\bar{L}}$  (Definition 10.5) are identified. Specifically, let  $\tilde{\mathbf{D}}_i$  denote the matrix of within-demeaned treatment lags for unit  $i$ , with rows corresponding to periods  $t$  and columns to  $D_{i,t}, D_{i,t-1}, \dots, D_{i,t-\bar{L}}$ . Given:*

$$\text{rank}(\mathbb{E}[\tilde{\mathbf{D}}_i \tilde{\mathbf{D}}_i']) = \bar{L} + 1,$$

*the impulse response coefficients  $\{\beta_s\}_{s=0}^{\bar{L}}$  are identified and consistently estimated by within-group OLS.*

### Summary: Assumptions by Method

For event-study designs, identification requires parallel trends (Assumption 37) and no anticipation (Assumption 38). Use Proposition 10.2 as a diagnostic implication. Check overlap and support at each lag (Assumption 40).

For distributed-lag models, identification requires strict exogeneity (Assumption 42) for within-group estimation. If a lagged dependent variable is present, sequential exogeneity (Assumption 43) is needed instead, along with the rank condition for identification.

When parallel trends is implausible, consider factor models (Chapter 8) that allow heterogeneous loadings on common shocks, or synthetic control methods (Chapter 6) that use data-driven weighting to construct comparable counterfactuals.

## 10.4 Estimation Strategies

Estimating dynamic treatment effects requires choosing an estimation strategy that accommodates the data structure (binary or continuous treatment, staggered or common adoption) and provides unbiased estimates under the identification assumptions. This section presents heterogeneity-robust event studies, distributed-lag models, dynamic difference-in-differences, and continuous dose-response methods. Section 10.2 defines the estimands. Section 10.3 provides identification assumptions.

Table 10.5 summarises the main estimation approaches.

**Table 10.5** Dynamic Estimation Methods

Method	Treatment Type	Key Assumptions	Output
Heterogeneity-robust event study	Binary, staggered	Parallel trends, no anticipation	Event-time effects $\theta_k$
Callaway–Sant’Anna	Binary, staggered	Parallel trends, no anticipation	$\text{ATT}(g, t)$ , aggregated $\theta_k$
Distributed-lag (linear)	Continuous	Strict exogeneity	Impulse response $\beta_s$
Geometric (Koyck) ad-stock	Continuous	Strict exogeneity, geometric decay	$\beta_0$ , $\delta$ , long-run effect and LRM
Almon polynomial	Continuous	Strict exogeneity, polynomial structure	Smooth impulse response
LongBet (tree-based)	Binary/continuous	Sequential exogeneity	Heterogeneous $\psi(X_i, t, s)$
Dynamic DiD	Binary, staggered	Parallel trends, no anticipation	$\text{ATT}(g, t)$ , $\theta_k$
Continuous dose-response	Continuous	Conditional exogeneity	Dose-response $\mu_k(d)$

### Choosing an Estimation Method

The choice of estimation method depends on treatment structure and research goals. For binary treatments with common adoption timing, a standard event study with two-way fixed effects is often a sensible starting point under the usual design assumptions. For binary treatments with staggered adoption, heterogeneity-robust methods such as Callaway and Sant'Anna [2021] and Sun and Abraham [2021] avoid the biases that arise when already-treated units serve as implicit controls [Goodman-Bacon, 2021, de Chaisemartin and d'Haultfoeuille, 2020]. These methods estimate cohort-time effects  $\text{ATT}(g, t)$  and aggregate them into event-time effects  $\theta_k$ .

For continuous treatments such as advertising GRPs or discount depth, distributed-lag models are the natural choice. The geometric (Koyck) specification is appropriate when exponential decay is plausible. Almon polynomials provide smooth, potentially non-monotonic dynamics with fewer parameters. When dynamics are heterogeneous across units, LongBet offers tree-based discovery of differential decay rates and lift patterns, while simpler approaches interact event-time dummies with covariates. When the goal is structural parameters such as half-life and long-run multiplier, geometric ad-stock models provide closed-form expressions. You can also fit a decay curve to flexible event-study estimates.

### Heterogeneity-Robust Event Studies

Heterogeneity-robust event studies estimate event-time effects  $\theta_k$  while accounting for heterogeneity in treatment timing and effects across cohorts. The estimand definitions and identification assumptions follow Chapters 4 and 5. Here we focus on aggregation and dynamic-specific issues. In practice, the heterogeneity-robust event-study estimands in this book are implemented via Callaway and Sant'Anna [2021] or Sun and Abraham [2021], which estimate cohort-time  $\text{ATT}(g, t)$  and then aggregate them into  $\theta_k$ . Chapter 5 provides the regression details.

Under staggered adoption with heterogeneous effects, two-way fixed effects estimates are biased because already-treated units serve as implicit controls. Heterogeneity-robust methods address this by estimating cohort-time effects  $\text{ATT}(g, t)$  separately, then aggregating into  $\theta_k$ .

**Definition 10.8 (Cohort-Time Average Treatment Effect)** The average treatment effect for cohort  $g$  in period  $t \geq g$  is:

$$\text{ATT}(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g].$$

This estimand is identified via the DiD-style formula from Theorem 10.1, which differences treated and control groups both over time (from baseline  $g - 1$  to period  $t$ ) and across groups (cohort  $g$  versus comparison group  $C_{g,k}$  with  $k = t - g$ ). Practical estimation uses Callaway and Sant'Anna [2021] or Sun and Abraham [2021], as described in Chapter 4.

**Proposition 10.3 (Aggregation from Cohort-Time to Event-Time)** *The event-time effect at lag  $k$  is obtained by aggregating along the diagonal  $t = g + k$ :*

$$\hat{\theta}_k = \sum_{g \in \mathcal{G}_k} w_g^k \widehat{\text{ATT}}(g, g+k),$$

where the cohort weights  $w_g^k$  are either cohort-size weights  $w_g^k = N_g / \sum_{g' \in \mathcal{G}_k} N_{g'}$  or uniform weights  $w_g^k = 1/|\mathcal{G}_k|$ . This aggregation matches Definition 10.4:  $\hat{\theta}_k$  is the empirical counterpart of the aggregate event-time effect obtained by weighting cohort-specific event-time effects along the diagonal  $t = g + k$ .

**Theorem 10.3 (Consistency and Asymptotic Normality)** Under Assumptions 38, 37, and regularity conditions,  $\hat{\theta}_k \xrightarrow{P} \theta_k$  as the number of independent units (or clusters) grows. When units are independent,  $\sqrt{N}(\hat{\theta}_k - \theta_k) \xrightarrow{d} \mathcal{N}(0, V_k)$ . When the independent units are clusters, state the asymptotic rate in terms of the number of independent clusters  $G$ . The variance  $V_k$  accounts for estimation error in each  $\widehat{\text{ATT}}(g, g+k)$  and covariance across cohorts.

Support shrinks at extreme lags. Common practice bins distant leads (for example,  $k \leq -5$ ) and lags (for example,  $k \geq 10$ ) into single categories. See Chapter 5 for specification and diagnostic guidance.

## Distributed-Lag Models

Distributed-lag models estimate impulse responses by regressing outcomes on current and lagged treatment:

$$Y_{it} = \sum_{s=0}^{\bar{L}} \beta_s D_{i,t-s} + \alpha_i + \lambda_t + \varepsilon_{it}.$$

Without a credible design or exogeneity assumptions on  $D_{it}$ , distributed-lag estimates describe predictive dynamics rather than causal impulse responses (see Assumption 42).

The geometric (Koyck) ad-stock model specifies a parsimonious structure for the impulse response. These structural assumptions sharpen dynamic interpretation and enable extrapolation, but they do not by themselves create identification. The design-based assumptions on treatment assignment and comparison groups remain critical.

**Definition 10.9 (Geometric (Koyck) Ad-Stock Model)** The geometric ad-stock model specifies:

$$Y_{it} = \beta_0 D_{it} + \beta_1 A_{it} + \alpha_i + \lambda_t + \varepsilon_{it},$$

where the ad-stock variable accumulates past treatment with geometric decay:

$$A_{it} = \sum_{s=1}^{t-1} \delta^{s-1} D_{i,t-s} = \delta A_{i,t-1} + D_{i,t-1}, \quad A_{i1} = 0.$$

The parameters are:  $\beta_0$  (contemporaneous effect),  $\beta_1$  (carryover effect per unit of ad-stock), and  $\delta \in (0, 1)$  (decay rate). This two-parameter structure allows the contemporaneous effect to differ from the carryover pattern. The implied impulse response is  $\beta_s = \beta_0 \mathbf{1}\{s=0\} + \beta_1 \delta^{s-1} \mathbf{1}\{s \geq 1\}$ . The long-run effect of a unit

treatment pulse is

$$\sum_{s=0}^{\infty} \beta_s = \beta_0 + \frac{\beta_1}{1-\delta},$$

provided  $\delta \in (0, 1)$ . The corresponding long-run multiplier is

$$\text{LRM} = \frac{\beta_0 + \beta_1/(1-\delta)}{\beta_0},$$

which expresses the long-run effect in units of the impact effect  $\beta_0$ . When  $\beta_0$  is close to zero, report the long-run effect  $\beta_0 + \beta_1/(1-\delta)$  directly and treat LRM as secondary. The standard Koyck model is the special case  $\beta_0 = \beta_1$ .

**Proposition 10.4 (Nonlinear Least Squares Estimation)** *Estimate  $(\beta_0, \beta_1, \delta)$  by grid search over  $\delta \in \{0.5, 0.55, \dots, 0.95\}$  with OLS for  $(\beta_0, \beta_1)$  at each  $\delta$ , minimising the residual sum of squares.*

The Almon polynomial specification offers an alternative that constrains the impulse response to lie on a polynomial of degree  $q$ . Write  $\beta_s = \sum_{j=0}^q \gamma_j s^j$  for  $s = 0, 1, \dots, \bar{L}$ . This imposes smoothness and reduces the number of free parameters from  $\bar{L}+1$  to  $q+1$ . For maximum flexibility, estimate  $\beta_s$  freely without functional form and use regularisation (ridge, lasso) to stabilise estimates. See Chapter 12 for penalty choices.

#### Lag Length Selection

Choosing the maximum lag  $\bar{L}$  involves a bias-variance trade-off. Too short a window truncates the impulse response and biases the LRM downward. Too long a window includes zero-effect lags, inflates standard errors, and risks overfitting.

When selecting lag length, begin with substantive knowledge: advertising effects typically span 6–12 weeks, while loyalty programme effects span 4–8 quarters. Use information criteria (AIC/BIC) to penalise excess lags. Report results for multiple  $\bar{L}$  as sensitivity analysis, and include lags until coefficients are consistently insignificant.

## Dynamic Panel Models

### Technical Note: Causal Interpretation of Dynamic Panel Models

Traditional dynamic panel models include lagged outcomes:

$$Y_{it} = \beta D_{it} + \rho Y_{it-1} + \alpha_i + \lambda_t + \varepsilon_{it}.$$

Estimation uses GMM (Arellano–Bond) with lagged levels as instruments.

Marx et al. [2024] show that under sequential exogeneity (Assumption 43), the GMM estimand identifies a weighted average of heterogeneous intertemporal treatment effects. The coefficient  $\beta$  captures the contemporaneous effect. The coefficient  $\rho$  captures persistence. However, if state dependence is spurious (driven by serial correlation in unobservables), the dynamic panel model biases treatment effects. See Arellano [2003] for foundations.

## Tree-Based Dynamic Heterogeneity

Standard distributed-lag models assume a constant decay rate  $\delta$  across units. Wang et al. [2024] introduce LongBet, a tree-based method for heterogeneous dynamic treatment effects. The model decomposes outcomes into prognostic and treatment terms:

$$Y_{it} = \mu(X_i, t) + \psi(X_i, t, s)D_{it} + \varepsilon_{it},$$

where  $s$  is time since treatment. Functions  $\mu$  and  $\psi$  are modelled using Bayesian Additive Regression Trees (BART). Crucially,  $\psi(X_i, t, s)$  allows the entire dynamic trajectory to vary with unit characteristics  $X_i$ . As with other observational ML methods, credible causal interpretation requires a sequential-exogeneity or unconfoundedness condition given  $X_i$  and past outcomes: conditional on observed history and covariates, treatment paths must be as good as random. LongBet does not weaken these assumptions. It only flexibly models how dynamic effects vary with  $X_i$ .

LongBet offers three advantages for marketing dynamics. First, it discovers heterogeneous decay—for example, that premium customers have slower decay than price-sensitive switchers. Second, trees learn non-monotonic patterns such as wear-in and delayed peaks without prespecifying functional form. Third, the prognostic term (seasonality, trends) is regularised separately from the treatment term, improving estimation of both. In many marketing panels, simpler interactions between event-time dummies and a few key covariates will suffice; LongBet is most valuable when you expect complex, high-dimensional heterogeneity in decay patterns that cannot be captured with low-dimensional interactions. Implementation is available via the **LongBet** package in R and Python. See Chapter 12 for tree-based methods.

## Dynamic Difference-in-Differences

Dynamic DiD under staggered adoption aggregates  $\text{ATT}(g, t)$  along event time  $k = t - g$  using the same aggregation formula as Proposition 10.3. This is conceptually identical to the staggered-adoption DiD estimators in Chapter 4, expressed in event-time coordinates. Dynamic DiD is DiD with explicit attention to the lag structure of  $\text{ATT}(g, t)$  and aggregation into  $\theta_k$ . Dynamic DiD therefore inherits the same identification assumptions and diagnostics as staggered DiD: parallel trends, no anticipation (or a bounded anticipation window), and adequate overlap in event time. Plotting  $\hat{\theta}_k$  against  $k$  visualises dynamics. Checking whether  $\hat{\theta}_k$  is close to zero for  $k < 0$  is a pre-trend diagnostic. See Chapter 4 for estimators, weights, and inference.

## Continuous Treatments and Dose-Response

For continuous treatment intensity (advertising GRPs, discount depth, points earned), you can either work with impulse responses from distributed-lag models or define a dynamic dose-response at a horizon  $k$ . One convenient definition is

$$\mu_k(d) = \mathbb{E}[Y_{i,t+k}(\underline{d}_i^{t+k}(d))],$$

where  $\underline{d}_i^{t+k}(d)$  is the path that sets  $d_{it} = d$  and  $d_{is} = 0$  for  $s \neq t$  up to time  $t + k$ . Identification requires a conditional-exogeneity and overlap condition for the relevant dose histories, as in Chapter 14. Double machine learning methods orthogonalise treatment effects against nuisance functions and extend directly once you have specified the dynamic dose path.

**Table 10.6** Software for Dynamic Treatment Effect Estimation

Method	R	Python/Stata
Callaway–Sant’Anna	<code>did</code>	<code>csdid</code> (Stata)
Sun–Abraham	<code>fixest</code>	<code>eventstudyinteract</code> (Stata)
Distributed-lag	<code>dynlm</code> , <code>plm</code>	<code>linearmodels</code> (Python)
Geometric ad-stock	Custom NLS	Custom NLS
LongBet	<code>LongBet</code>	<code>longbet</code> (Python)
DML dose-response	<code>DoubleML</code>	<code>econml</code> (Python)

## 10.5 Anticipation, Carryover, and Mediation

Dynamic treatment effects arise from three distinct mechanisms. Anticipation describes forward-looking behaviour where agents adjust current outcomes in response to expected future treatments. Carryover captures the lagged effects of past treatments persisting into current periods. Mediation refers to indirect effects operating through intermediate variables. These mechanisms extend the anticipation and carryover ideas already introduced in Chapters 5 and 4 to the fully dynamic, path-dependent framework of this chapter. Understanding which mechanisms drive dynamic patterns informs both interpretation and design.

Table 10.7 summarises the three mechanisms and their diagnostics.

**Table 10.7** Dynamic Mechanisms in Marketing

Mechanism	Direction	Marketing Examples	Diagnostic
Anticipation	Future → Present	Delay buying before promotion, stockpile before price increase	Event-time leads ( $k < 0$ )
Carryover	Past → Present	Advertising builds awareness, promotion generates trial	Event-time lags ( $k > 0$ )
Mediation	Treatment → Mediator → Outcome	Advertising → Awareness → Sales	Path analysis, front-door

### Anticipation

Anticipation occurs when agents are forward-looking and adjust current behaviour in response to expected future treatments. In marketing, consumers may delay buying in anticipation of a promotion (reducing current sales) or accelerate buying before a price increase (stockpiling). Search intensity often increases before a product launch, generating buzz. Firms may also adjust pricing, advertising, or inventory in anticipation of competitor actions.

Anticipatory responses bias event-time estimates if they occur in the pre-treatment period, because pre-treatment outcomes no longer reflect the untreated potential outcome but rather a mixture of the untreated outcome and the anticipatory response. This violates the no-anticipation assumption (Assumption 38 in Section 10.3). When anticipation is present but bounded, the limited-anticipation restriction (Assumption 39) clarifies which pre-period lags remain valid for diagnostics.

To diagnose anticipation, estimate the event-study specification including leads ( $k < 0$ ) alongside lags ( $k \geq 0$ ) and plot  $\hat{\theta}_k$  for  $k < 0$ . Use lead coefficients as a diagnostic: large or patterned leads indicate tension with no-anticipation or with parallel trends. Non-rejection of a joint diagnostic on leads is inconclusive and should be interpreted alongside design evidence on announcement timing and information sets.

### Distinguishing Anticipation from Pre-Trends

Anticipation and pre-trends both generate non-zero event-time leads but arise from different sources and require different responses. Evidence favouring anticipation includes an effect that begins exactly one period before treatment, absence of effects at earlier leads ( $k = -2, -3, \dots$ ), and institutional knowledge confirming that agents could anticipate (such as pre-announcement of the treatment). Evidence favouring differential pre-trends includes effects present at multiple leads, a linear trend pattern across leads, and circumstances where agents could not have known about treatment.

A design-based diagnostic compares pre-announced versus surprise treatments. If the pre-treatment dip appears only for pre-announced treatments, anticipation is the more likely explanation. Statistical patterns can suggest anticipation or pre-trends, but design documentation (announcement timing, information sets) is decisive. Wherever possible, you should use explicit information on announcement dates, eligibility rules, and targeting policies to distinguish anticipation from pre-trends. Time-series patterns alone are rarely decisive.

When anticipation is detected, analysts face three options. The first is to model anticipation explicitly by including leads in the specification and interpreting them as anticipatory effects, reporting both anticipatory effects (leads) and post-treatment effects (lags), and discussing the total effect as the sum of anticipation and post-treatment components. The second option is to redefine the intervention time: if consumers anticipate one period in advance, redefining treatment to switch on at  $g - 1$  captures the onset of anticipation as the treatment itself. The third option is to exclude the anticipation period by using only periods  $t \leq g - \bar{a}$  as pre-treatment, where  $\bar{a}$  is the anticipation window. This restricts the comparison to periods before anticipation began. The choice depends on the substantive question, the strength of evidence, and whether anticipation is of independent interest.

### Carryover and Decay

Carryover describes how treatment effects persist and dissipate over lags. Positive carryover occurs when past treatments increase current outcomes: advertising builds brand awareness as a cumulative stock, promotions generate trial that leads to repeat buying, and loyalty programmes create habits and routines. Negative carryover occurs when past treatments decrease current outcomes: promotions may exhaust demand through stockpiling, and repeated exposures can cause satiation and wear-out.

### The Attention Paradox in Dynamic Advertising Effects

Consider a video platform that runs a six-week campaign with a high-frequency pre-roll ad for a focal brand. Let  $D_{it}$  indicate exposure to the high-frequency schedule in week  $t$  for market  $i$ , and let  $G_i$  be the week when the schedule is turned on. We estimate cohort-specific event-time effects  $\theta_k(g)$  and aggregated effects  $\theta_k$  as in Definition 10.4, with  $k = t - G_i$ .

Suppose the estimated profile  $\{\hat{\theta}_k\}$  for  $k = 0, 1, \dots, 8$  shows positive effects in early weeks ( $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2 > 0$ ) as high attention increases sales and brand search, neutral effects in the middle period ( $\hat{\theta}_3 \approx 0$ ), and negative effects in later weeks ( $\hat{\theta}_4, \hat{\theta}_5 < 0$ ) as further exposure to the same creative reduces sales relative to a counterfactual with shorter duration.

If  $\theta_k$  are per-period flow effects (as in Chapters 4 and 5), define the cumulative effect through horizon  $H$  as  $\hat{L}_\theta(H) = \sum_{k=0}^H \hat{\theta}_k$ . If you work with increments  $\hat{\beta}_k = \hat{\theta}_k - \hat{\theta}_{k-1}$ , recover  $\hat{\theta}_k$  by cumulating the increments and then compute  $\hat{L}_\theta(H)$ . In this pattern,  $\hat{L}_\theta(3)$  is positive and substantial, but  $\hat{L}_\theta(6)$  may be only slightly larger or even smaller if the late negative effects dominate. More “attention” in the sense of longer exposure to the same ad generates negative carryover.

The measurement implication is that focusing only on early positive lags or on total impressions obscures the attention paradox: past a point, incremental event-time effects turn negative even as exposure accumulates. Reporting the full  $\hat{\theta}_k$  trajectory, cumulative effects  $\hat{L}_\theta(H)$  for multiple horizons, and confidence bands helps identify where attention becomes counterproductive and informs optimal campaign duration and rotation of creatives.

Interpreting lag profiles as carryover requires care. Elevated post-treatment sales could reflect true carryover (consumers who tried during promotion continue buying) or sample selection (promotion attracted high-valuation consumers who would have bought eventually). Distinguishing these requires additional evidence, such as tracking the same consumers over time or comparing exposed versus unexposed consumers.

The most common functional form for decay is geometric:

$$\beta_s = \beta_0 \delta^s,$$

where  $\beta_0$  is the immediate effect and  $\delta \in (0, 1)$  is the decay rate. Under geometric decay  $\beta_s = \beta_0 \delta^s$ , the long-run effect of a one-period treatment pulse is  $\beta_0/(1 - \delta)$  and the corresponding long-run multiplier is  $1/(1 - \delta)$ . The half-life is  $h^* = \log 2 / (-\log \delta)$ . Sensitivity analysis should compare geometric, polynomial, and flexible lag specifications to assess robustness to functional form assumptions.

## Mediation

Mediation refers to indirect effects operating through intermediate variables. Let  $M_{it}$  denote a mediator (for example, awareness or reward redemption) measured after treatment. Mediation channels describe pathways through which treatments affect outcomes. Advertising may operate through awareness, consideration, and

trial to generate sales. Promotions may trigger deal-seeking behaviour, stockpiling, and category expansion. Loyalty programmes may create switching costs, habit formation, and reward accumulation that drive retention and sales.

**Table 10.8** Mediation Channels in Marketing

Treatment	Mediators	Outcome
Advertising	Awareness → Consideration → Trial	Sales
Promotions	Deal-seeking → Stockpiling → Category expansion	Sales
Loyalty programmes	Switching costs → Habit formation → Rewards	Retention, Sales

Understanding mediation channels informs design (which channels to activate), targeting (which customers respond to which channels), and optimisation (how to allocate budget across channels).

#### Warning: Post-Treatment Conditioning Bias

Naïve mediation analysis conditions on post-treatment mediators, for example by regressing sales on treatment and reward redemption. This approach opens backdoor paths through confounders of the mediator-outcome relationship, closes frontdoor paths that represent the causal effect, and produces estimates without causal interpretation. In dynamic panels this problem is amplified because mediators often sit on the same time path as outcomes and treatments, so conditioning on them both blocks genuine carryover channels and opens paths through time-varying confounders.

Credible mediation analysis requires sequential ignorability (no unmeasured confounding of the mediator-outcome relationship) and specialised methods such as front-door adjustment or instrumental variables for the mediator. This is a special case of the general admonition in Chapter 2 against conditioning on post-treatment variables without a causal graph and explicit assumptions. See Pearl [2009] for formal treatment.

### Practical Guidance: Mediation in Marketing Panels

Report total effects as the primary estimand. The total effect captures the overall effect of treatment on outcome, has a clear causal interpretation, and directly informs policy. Path-specific estimates should be presented as supplementary and accompanied by clear assumptions and sensitivity analysis. Report path-specific effects only when sequential ignorability is plausible, assumptions are explicitly discussed, and sensitivity analyses assess robustness to violations.

Avoid conditioning on post-treatment variables without explicit justification. When feasible, use experimental variation in mediators—for example, randomising reward reminders—to isolate specific channels with design-based credibility. Path-specific effects should never be reported as headline results in observational marketing panels. They belong in sensitivity analysis, with clear caveats, after you have established credible total effects.

## 10.6 Inference

Inference for dynamic treatment effects builds on the clustering and bootstrap machinery developed in Chapter 16, but introduces three complications specific to the dynamic setting: serial dependence in impulse responses creates correlation across lag-specific estimates, common shocks induce cross-unit correlation within periods, and event-study designs require testing effects at many lags simultaneously. We focus only on complications introduced by lagged structure and event-time profiles. Foundational material on cluster-robust variance, bootstrap methods, and multiple testing appears in Chapter 16 and Chapters 4–5. None of these inference tools repair violations of the identification assumptions in Section 10.3. They quantify uncertainty conditional on those assumptions holding.

Table 10.9 summarises the main inference approaches for quick reference.

**Table 10.9** Inference Methods for Dynamic Effects

Method	Assumptions	When to Use	Software
Cluster-robust (unit)	Arbitrary within-unit correlation	Default for panels	<code>fixest</code> , <code>plm</code>
Two-way cluster	Within-unit + within-period correlation	Common shocks present	<code>fixest</code> , <code>reghdfe</code>
HAC (Newey–West)	Correlation decays with lag	Moderate serial dependence	<code>sandwich</code> , <code>plm</code>
Wild cluster bootstrap	Small number of clusters	$G < 30$ clusters	<code>fwildclusterboot</code>
Romano–Wolf	Multiple testing correction	Many lags tested	<code>rwolf</code>

## Clustering for Dynamic Panels

The default approach clusters standard errors by unit, allowing arbitrary within-unit correlation of errors over time. The independent sampling unit is the cluster that you treat as independent, typically the unit  $i$  or a higher-level cluster  $c$ . The effective sample size is therefore the number of independent units,  $N$  or  $G$ , not the total number of unit–time observations. Two-way clustering extends this to account for common shocks across units within periods—macroeconomic fluctuations, regulatory changes, or competitive disruptions that affect all firms simultaneously. The cost is larger standard errors, but this correctly reflects uncertainty from cross-sectional dependence. When the number of independent clusters is small ( $G < 30$ ), the wild cluster bootstrap provides finite-sample corrections as detailed in Chapter 16.

HAC (heteroskedasticity and autocorrelation consistent) estimators offer an alternative when serial dependence is moderate and decays with lag distance. The Newey–West estimator downweights lags beyond a bandwidth, typically set by the rule  $\lfloor T^{1/4} \rfloor$  or chosen via data-driven methods. HAC standard errors are often smaller than cluster-robust ones because they impose structure on the decay pattern, but they may underestimate uncertainty if the decay assumption fails. In marketing panels with many cross-sectional units

and moderate  $T$ , cluster-robust SEs are typically more appropriate than HAC. HAC is more natural when a single long time series is analysed.

## Multiple Testing Across Lags

Event-study designs estimate effects at many lags, and reporting pointwise confidence intervals for each  $\theta_k$  inflates the family-wise error rate. If we test  $\bar{K} + 1$  hypotheses at level  $\alpha$ , the probability of at least one false rejection can approach  $1 - (1 - \alpha)^{\bar{K}+1}$ , which exceeds 0.5 when testing more than a dozen lags at the 5% level.

Three corrections address this problem. The Bonferroni correction rejects the null  $H_{0,k} : \theta_k = 0$  only if  $p_k < \alpha/(\bar{K} + 1)$ . It is conservative because it ignores correlation among test statistics. The Holm stepdown procedure orders p-values and applies sequentially adjusted thresholds, gaining power over Bonferroni while still controlling the family-wise error rate. The Romano–Wolf stepdown procedure [Romano and Wolf, 2005] bootstraps the joint distribution of test statistics, accounting for the correlation structure among lag-specific estimates. This is the preferred method for event studies because it exploits the dependence among  $\hat{\theta}_k$  to provide tighter bounds.

## Pre-Trend Testing

**Proposition 10.5 (Pre-Trend Test Statistic)** *The joint null of no pre-trends is  $H_0 : \theta_{-K_{\text{pre}}} = \dots = \theta_{-1} = 0$ . The Wald test statistic is*

$$W = \hat{\boldsymbol{\theta}}'_{\text{pre}} \hat{\mathbf{V}}_{\text{pre}}^{-1} \hat{\boldsymbol{\theta}}_{\text{pre}},$$

where  $\hat{\boldsymbol{\theta}}_{\text{pre}} = (\hat{\theta}_{-K_{\text{pre}}}, \dots, \hat{\theta}_{-1})'$  and  $\hat{\mathbf{V}}_{\text{pre}}$  is the corresponding submatrix of the cluster-robust covariance. Under  $H_0$ ,

$$W \xrightarrow{d} \chi^2_{K_{\text{pre}}} \quad \text{as } G \rightarrow \infty.$$

Rejection indicates either anticipation effects or parallel trends failure.

The pre-trend test is central to event-study credibility. However, failure to reject does not validate parallel trends—the test has limited power against small violations. Roth [2022] shows that pre-testing can bias subsequent treatment effect estimates toward zero, a phenomenon known as "pre-test bias." Report the test statistic and p-value, but interpret non-rejection cautiously. See Chapter 5, Section 5.8, for complementary discussion of pre-trend power, pre-test bias, and graphical diagnostics.

## Long-Run Multiplier Uncertainty

The long-run multiplier (LRM) aggregates dynamic effects into a single summary statistic. Uncertainty quantification requires accounting for the covariance among lag-specific estimates.

**Proposition 10.6 (Variance of the LRM)** *Let  $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_{\bar{L}})'$  denote the estimated impulse responses from a distributed-lag model with covariance matrix  $\hat{\mathbf{V}}_\beta$ . Define the long-run effect as*

$$\hat{L}_\beta = \sum_{s=0}^{\bar{L}} \hat{\beta}_s,$$

with variance

$$\text{Var}(\hat{L}_\beta) = \mathbf{1}' \hat{\mathbf{V}}_\beta \mathbf{1} = \sum_{s=0}^{\bar{L}} \text{Var}(\hat{\beta}_s) + 2 \sum_{s < s'} \text{Cov}(\hat{\beta}_s, \hat{\beta}_{s'}).$$

The corresponding long-run multiplier is  $\widehat{\text{LRM}}_\beta = \hat{L}_\beta / \hat{\beta}_0$ , with variance obtained by the delta method.

For event-time effects  $\theta_k$  interpreted as period-specific flow impacts (as in Chapters 4 and 5), the long-run effect over a horizon  $\bar{K}$  is  $\hat{L}_\theta = \sum_{k=0}^{\bar{K}} \hat{\theta}_k$  with variance  $\mathbf{1}' \hat{\mathbf{V}}_\theta \mathbf{1}$ . The associated long-run multiplier is  $\widehat{\text{LRM}}_\theta = \hat{L}_\theta / \hat{\theta}_0$ . For stock outcomes, the long-run effect at horizon  $\bar{K}$  is  $\hat{\theta}_{\bar{K}}$ , with variance  $\text{Var}(\hat{\theta}_{\bar{K}})$ .

For geometric ad-stock models with long-run effect  $L(\beta_0, \beta_1, \delta) = \beta_0 + \beta_1 / (1 - \delta)$ , the delta method gives

$$\text{Var}(\hat{L}) \approx \nabla L(\hat{\eta})' \hat{\mathbf{V}}_\eta \nabla L(\hat{\eta}),$$

where  $\nabla L = (1, (1 - \delta)^{-1}, \beta_1(1 - \delta)^{-2})'$  and  $\hat{\eta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\delta})'$ . The variance of the corresponding multiplier  $\widehat{\text{LRM}} = \hat{L} / \hat{\beta}_0$  follows from an additional delta-method step.

Note the distinction in notation: we use  $\theta_k$  for event-time coefficients indexed by time relative to treatment adoption ( $k \in \{-K, \dots, \bar{K}\}$ ), and  $\beta_s$  for impulse response coefficients from distributed-lag models indexed by lag ( $s \in \{0, \dots, \bar{L}\}$ ). For flow outcomes,  $\theta_k$  represents the period-specific effect at lag  $k$ , and the LRM is  $\sum_k \theta_k$ . For stock outcomes,  $\theta_{\bar{K}}$  is the long-run level effect. The impulse responses  $\beta_s$  are always period-specific.

### Reporting Standards for Dynamic Inference

Standard errors should be cluster-robust at the unit level as a default. Use two-way clustering when common shocks are plausible (e.g., macro shocks, regulatory changes) and wild cluster bootstrap when clusters number fewer than thirty.

Report joint tests for both pre-trends ( $H_0 : \theta_k = 0$  for  $k < 0$ ) and post-treatment effects ( $H_0 : \theta_k = 0$  for  $k \geq 0$ ). For individual lag estimates, report unadjusted p-values but note the number of tests. Consider Romano–Wolf corrections when formal family-wise error rate control is required.

For summary statistics, report confidence intervals for long-run effects and, when you use them, long-run multipliers, using the delta method or bootstrap. If the half-life is estimated, report its confidence interval as well. Discuss the overall pattern of effects rather than fixating on which individual lags cross significance thresholds. When presenting full dynamic profiles, consider sup- $t$  or Romano–Wolf-based joint confidence bands (see Chapter 16) so that statements about the entire path (“no effect at any lag” or “effects are non-zero for all post-treatment lags”) have a clear inferential basis.

## 10.7 Diagnostics

Credible dynamic treatment effect analysis requires diagnostics that go beyond the general design checks covered in Chapter 17. We assume the general diagnostic tools from Chapters 17 and 5 (Section 5.8). Here we focus only on issues that arise because treatment effects are dynamic: event-time support, lag/window choices, and functional-form sensitivity in distributed lags. For identification assumptions underlying these diagnostics, see Section 10.3.

Table 10.10 summarises the key diagnostics.

**Table 10.10** Diagnostics for Dynamic Effects

Diagnostic	What It Checks	Action if Failed
Support per event time	Adequate treated/control observations at each lag	Bin sparse lags or restrict window
Binning sensitivity	Stability across bin thresholds	Report multiple thresholds
Window sensitivity	Stability across event-time windows	Use restricted window with better support
Lag length sensitivity	Cumulative effect stability as $\bar{L}$ increases	Effects persist beyond window
Functional form	Agreement across specifications	Report multiple specifications

### Support per Event Time

Event-study estimates can become unreliable when few observations contribute to particular event times. Support quantifies the number of treated and control observations at each lag, following the cohort notation established in Section 10.3.

**Definition 10.10 (Event-Time Support)** For each event time  $k$ , define treated support as  $N_k^{\text{tr}} = \sum_{g \in \mathcal{G}_k} N_g$ , the count of treated observations at event time  $k$ . Define control support as  $N_k^{\text{co}}$ , the count of not-yet-treated or never-treated observations serving as comparisons. Define effective support as the harmonic mean

$$N_k^{\text{eff}} = \frac{2N_k^{\text{tr}}N_k^{\text{co}}}{N_k^{\text{tr}} + N_k^{\text{co}}}.$$

Event times with  $N_k^{\text{eff}} < N_{\min}$  should be binned or excluded.  $N_{\min}$  should be chosen ex ante and reflect the unit of analysis and clustering (e.g., at least 10 treated *clusters* rather than 10 treated observations). A common threshold is  $N_{\min} = 10$  clusters.

Plotting the number of treated and control units at each event time  $k$ , separately for leads ( $k < 0$ ) and lags ( $k > 0$ ), reveals where support becomes thin. Lags with support below the threshold should be flagged and either binned into endpoint categories or excluded from the analysis.

## Binning and Window Sensitivity

Binning pools multiple lags into a single estimate, improving stability at the cost of coarser resolution. The standard practice bins distant leads (e.g.,  $k \leq -5$ ) into a single pre-period category and distant lags (e.g.,  $k \geq 10$ ) into a single post-period category. The choice of bin thresholds should not drive conclusions.

Sensitivity analysis compares results across bin thresholds—for example,  $k \geq 8$  versus  $k \geq 10$  versus  $k \geq 12$  for the post-period bin. If estimates remain stable across thresholds, conclusions are robust to binning choices. If estimates vary widely, the analyst should report results for multiple thresholds and discuss which is most credible given the support structure.

Window sensitivity assesses how estimates change when the event-time window is restricted. Estimating with the full window (e.g.,  $k \in [-10, 15]$ ) and then re-estimating with a restricted window (e.g.,  $k \in [-5, 10]$ ) reveals whether distant lags are driven by sparse support. If estimates are similar, the full window is credible; if they differ substantially, the restricted window with better support is more reliable and should be preferred for primary conclusions.

## Pre-Trends and Placebo Tests

Pre-trend diagnostics assess whether event-study leads are close to zero, providing evidence on parallel trends. Individual diagnostics examine each lead  $\theta_k$  for  $k < 0$ , while the joint F-test (Proposition 10.5) evaluates whether all leads are jointly zero. Interpret these together with power considerations and design knowledge. For implementation details and power/interpretation issues, see Chapter 17 and Chapter 5. We summarise only how these diagnostics interact with dynamic designs here. Rejection of the joint statistic in a dynamic context indicates either anticipation effects or differential pre-trends that vary with time-to-treatment.

Placebo tests assign pseudo-treatment dates to never-treated units and estimate pseudo event-time effects. Under correct specification, these pseudo effects should be zero. Large pseudo effects indicate that the design captures spurious variation—seasonality, common trends, or other confounders unrelated to treatment—and conclusions from the actual analysis are unreliable. For inference considerations when implementing placebo checks, see Section 10.6. For the full placebo testing framework, see Chapter 17.

## Lag Length and Functional Form Robustness

For distributed-lag models, the choice of maximum lag  $\bar{L}$  affects the estimated cumulative effect. Define the cumulative effect of the impulse response through  $\bar{L}$  as

$$\hat{L}_\beta(\bar{L}) = \sum_{s=0}^{\bar{L}} \hat{\beta}_s,$$

and, analogously, for flow-type event-time effects,

$$\widehat{L}_\theta(\bar{K}) = \sum_{k=0}^{\bar{K}} \widehat{\theta}_k$$

(see Section 10.2). Estimating for multiple lag lengths (e.g.,  $\bar{L} \in \{4, 6, 8, 10\}$ ) and plotting  $\widehat{L}_\beta(\bar{L})$  against  $\bar{L}$  reveals whether the cumulative effect stabilises. If it does, the chosen  $\bar{L}$  is adequate. If  $\widehat{L}_\beta(\bar{L})$  increases without bound as  $\bar{L}$  grows, effects persist beyond the estimation window and the reported cumulative effect understates the true long-run impact.

Functional form robustness compares geometric decay, Almon polynomial, and flexible lag specifications. Plotting the impulse response profile for each specification reveals whether conclusions depend on modelling assumptions. If profiles agree, conclusions are robust. If they disagree, results for multiple specifications should be reported with discussion of the economic reasoning favouring each. Model selection criteria such as AIC or BIC, or cross-validation, can provide data-driven guidance but should not override substantive considerations.

#### Diagnostic Workflow for Dynamic Effects

The diagnostic workflow proceeds in five stages. First, plot support per event time and flag sparse lags with  $N_k^{\text{eff}} < N_{\min}$ . Bin or exclude these lags before proceeding. Second, document binning choices and test sensitivity to bin thresholds and window restrictions. Report multiple specifications if results are sensitive. Third, examine pre-trend evidence by plotting lead coefficients with confidence intervals and reporting the joint F-test. If the diagnostic reveals concerns, diagnose the source before interpreting post-treatment effects. Fourth, run placebo tests on never-treated units to verify the design does not capture spurious variation. Fifth, assess robustness by varying lag length (plotting  $\widehat{L}_\beta(\bar{L})$  versus  $\bar{L}$ ), comparing functional forms, and reporting results for multiple specifications when conclusions differ across models.

These steps should be layered on top of the general DiD and event-study workflows in Sections 4.10 and 5.11, with lag length, binning, and functional-form checks treated as dynamic-specific extensions rather than standalone diagnostics. As with all diagnostics in this book, these checks can reveal tensions between the data and your assumptions but cannot turn a weak or violated design into a credible one.

## 10.8 Marketing Applications

Dynamic treatment effects are ubiquitous in marketing, where interventions generate time-varying responses through awareness-building, habit formation, stockpiling, and decay. This section demonstrates dynamic methods through three marketing applications that showcase how dynamic analysis proceeds in practice. The numerical values are illustrative, designed to show the workflow rather than report actual empirical findings.

Table 10.11 summarises the three applications.

**Table 10.11** Marketing Applications of Dynamic Methods

Application	Method	Data	Key Finding	Policy Implication
TV advertising	Geometric ad-stock	Weekly DMA sales	Carryover with decay	Pulsing vs continuous
Promotions	Event study	Weekly store sales	Spike then dip	Promotion spacing
Loyalty programme	Dynamic DiD	Quarterly revenue	Gradual ramp-up	ROI forecasting

### Television Advertising with Carryover

Television advertising provides a canonical setting for dynamic methods. The distributed-lag framework has a rich history in advertising response modelling, and meta-analytic evidence shows that advertising elasticities average 0.10 in the short run and 0.22 in the long run, with substantial heterogeneity across contexts [Sethuraman et al., 2011].

Consider a CPG brand that launches a national TV campaign in 50 DMAs over four weeks, with varying gross rating points (GRPs) across markets. Let  $D_{it}$  denote GRPs for market  $i$  in week  $t$ . Weekly sales are observed for 52 weeks before and 26 weeks after the campaign. The goal is to estimate the dynamic sales response and quantify the half-life and long-run multiplier.

The geometric ad-stock model takes the form

$$\log(\text{Sales}_{it}) = \beta_0 D_{it} + \beta_1 A_{it} + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $A_{it} = \delta A_{i,t-1} + D_{i,t-1}$  is ad-stock with decay rate  $\delta$ . Estimation proceeds via NLS grid search over  $\delta \in [0.5, 0.95]$ .

For causal interpretation we assume that variation in GRPs across DMAs is exogenous or instrumented, for example via budget rules, historical buying patterns, or quasi-experimental shifts in available inventory that are unrelated to contemporaneous demand shocks after controlling for  $\alpha_i$  and  $\lambda_t$ . Without such design or instrumental-variables structure, the ad-stock model describes predictive dynamics rather than causal responses.

Suppose the estimates yield  $\hat{\beta}_0 = 0.03$  (3% immediate lift per 100 GRPs),  $\hat{\beta}_1 = 0.02$  (2% lift per 100 units of ad-stock), and  $\hat{\delta} = 0.8$  (20% weekly decay). The implied half-life is  $\log 2 / (-\log 0.8) \approx 3.1$  weeks. The long-run effect of a one-week, 100-GRP pulse is

$$\hat{L}_\beta = \left( \hat{\beta}_0 + \frac{\hat{\beta}_1}{1 - \hat{\delta}} \right) \times 100 = 13\%,$$

meaning the campaign generates roughly 13% cumulative incremental sales over all weeks. The corresponding long-run multiplier is

$$\text{LRM} = \frac{\hat{\beta}_0 + \hat{\beta}_1 / (1 - \hat{\delta})}{\hat{\beta}_0} \approx 4.3,$$

which expresses the long-run effect in units of the immediate lift. These dynamic metrics (half-life, long-run effect, and multiplier) follow the definitions in Section 10.2 and the uncertainty methods in Section 10.6. Diagnostics should examine whether residual dependence is material for inference (for example, by comparing cluster-robust vs HAC-type standard errors) and whether the geometric decay restriction materially changes the estimated profile relative to flexible alternatives. Diagnostics should also verify that geometric decay fits better than Almon polynomial alternatives (via BIC) and that standard errors are cluster-robust at the DMA level. For lag-length and functional-form checks, see Section 10.7.

The policy implication is that a 3-week half-life favours pulsed schedules that maintain awareness through periodic bursts rather than continuous low-intensity spending. The LRM informs budget allocation across markets based on cumulative responsiveness.

### Meta-Analytic Evidence on Advertising Elasticity

Sethuraman et al. [2011] conduct a meta-analysis of 751 short-term and 402 long-term advertising elasticities from 56 studies spanning 1960–2008. The average short-run elasticity is 0.12, while long-run elasticity averages 0.22, yielding a long-run/short-run ratio of approximately 2.2 that reflects carryover through awareness and habit formation.

A critical methodological finding concerns endogeneity correction. Studies that fail to account for endogeneity report systematically lower elasticities than those that do. This negative bias is consistent with Villas-Boas and Winer [1999]: managers typically increase advertising when facing demand shortfalls, creating a negative correlation between advertising and the unobserved demand shock. Ignoring this correlation attenuates estimated effects toward zero. Endogeneity concerns apply to dynamic models as well: if managers increase GRPs in response to negative shocks, even a well-specified ad-stock structure will not solve selection-on-shocks without appropriate instruments or design-based variation.

Durables show higher elasticities than non-durables, television outperforms print, and new brands respond more than established brands. Weekly data yield higher elasticities than monthly due to aggregation bias. The key implication is that practitioners should expect modest average effects—a 12% elasticity means doubling ad spending increases sales by 12%—but both endogeneity correction and substantial heterogeneity make context-specific, causally credible estimation essential.

### Promotions with Post-Promotion Dips

Multi-period promotions induce stockpiling, creating post-promotion dips that reduce net cumulative effect. Consider a retailer that offers a 20% discount every four weeks for six months (six cycles) in 30 stores. Weekly sales are observed for 12 weeks before and 12 weeks after each promotion. The goal is to estimate the dynamic sales response, quantify the post-promotion dip, and compute the net cumulative effect.

The event-study specification is

$$\log(\text{Sales}_{it}) = \sum_{k \in \{-3, -2, 0, 1, 2, 3\}} \theta_k \mathbf{1}\{t - G_{im} = k\} + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $G_{im}$  denotes the week of the  $m$ -th promotion cycle for store  $i$ , with  $k = -1$  as reference. This formulation allows multiple events per unit, treating each promotion cycle as a separate treatment adoption.

For these estimates to have a causal interpretation, promotion timing across stores and cycles must be unrelated to short-run demand shocks after conditioning on store and time effects (and any covariates), so that promotion weeks are not systematically scheduled in response to unusually weak or strong recent sales.

Suppose the event-study estimates show  $\hat{\theta}_0 = 0.35$  (42% sales lift in promotion week), followed by  $\hat{\theta}_1 = -0.15$  (14% dip one week after) and  $\hat{\theta}_2 = -0.10$  (10% dip two weeks after), with return to baseline by week three. If pre-trends are near zero, the net cumulative effect over horizon  $H = 2$  is  $\hat{L}_\theta(2) = \sum_{k=0}^2 \hat{\theta}_k =$

$0.35 - 0.15 - 0.10 = 0.10$ , a 10% net lift (using  $\theta_k$  as per-period flow effects). Diagnostics would check that support is adequate across all six cycles, report the joint lead statistic (e.g.,  $p = 0.42$ ) as limited diagnostic evidence, confirm that results are stable across window choices, and use wild cluster bootstrap with 1,000 replications for inference. For repeated-event diagnostics and support checks across cycles, see Section 5.8 and Section 10.7.

The policy implication is that the post-promotion dip reduces the net effect to less than one-third of the immediate lift. Spacing promotions further apart allows customer inventories to deplete and increases the net cumulative effect per promotion.

## Loyalty Programmes with Ramp-Up

Loyalty programmes generate effects that build gradually as customers enrol, learn benefits, and develop habits. Consider a retailer that launches a loyalty programme in 50 stores over six quarters with staggered adoption (10 stores per cohort). Quarterly revenue is observed for eight quarters before and eight quarters after programme launch in each store. The goal is to estimate the dynamic revenue response, quantify the ramp-up period, and assess persistence.

The dynamic DiD specification uses heterogeneity-robust aggregation:

$$\hat{\theta}_k = \frac{\sum_{g \in \mathcal{G}_k} N_g \widehat{\text{ATT}}(g, g+k)}{\sum_{g \in \mathcal{G}_k} N_g},$$

where  $\text{ATT}(g, t)$  compares cohort  $g$  to never-treated stores at time  $t$ , and  $\mathcal{G}_k$  denotes the set of cohorts for which event time  $k$  is observed. This is the cohort-size-weighted aggregation described in Proposition 10.3. As in the general staggered-adoption setting, identification relies on parallel trends and no anticipation for the untreated potential outcomes of treated and comparison stores (Section 10.3), along with adequate overlap in adoption times and event-time support across cohorts.

Suppose the estimates show a gradual ramp-up:  $\hat{\theta}_0 = 0.02$  (2% lift in first quarter),  $\hat{\theta}_1 = 0.05$  (5% in second quarter),  $\hat{\theta}_2 = 0.08$  (8% in third quarter), stabilising at approximately 10% lift thereafter. If pre-trends are near zero, this pattern suggests a three-quarter ramp-up period with persistent long-run effects. Diagnostics would verify that support is adequate for  $k \in [-3, 8]$ , report the joint lead statistic (e.g.,  $p = 0.68$ ) as limited diagnostic evidence, check whether uniform and size-weighted cohort aggregations produce similar estimates, and use cluster-robust standard errors at the store level. For heterogeneity and cohort-weighting diagnostics, see Chapter 4 and Chapter 5.

The policy implication is that the three-quarter ramp-up informs ROI forecasting: early-quarter revenue lifts understate the programme's steady-state value. Cumulative revenue lift compared to operating costs determines programme viability, but this comparison requires waiting until effects stabilise.

### Matching Method to Application

The choice of dynamic method depends on the treatment structure and the phenomenon of interest. For continuous treatments with carryover, such as advertising, distributed-lag models with geometric or flexible ad-stock specifications estimate half-life and long-run effects (and their multipliers) that inform media scheduling decisions. For discrete treatments where stockpiling or anticipation creates intertemporal substitution, such as promotions, event-study designs visualise the spike-dip pattern and compute net cumulative effects that guide promotion timing. For staggered binary adoptions where effects build gradually, such as loyalty programmes, dynamic difference-in-differences with heterogeneity-robust aggregation estimates ramp-up periods and steady-state effects that inform ROI forecasting. When the goal is to trace feedback among marketing investments, consumer response, and firm value, vector autoregressions with impulse response functions can map how shocks to advertising or customer satisfaction propagate through sales and user-generated content to stock returns. Identification then requires timing restrictions or external instruments [Srinivasan et al., 2009, Tirunillai and Tellis, 2012]. In all these applications, credible dynamic estimates still depend on a strong design or valid instruments. Advanced dynamic models cannot by themselves resolve endogeneity in marketing investments.

## 10.9 Workflow Checklist

This section provides a compact, reproducible protocol for conducting dynamic treatment effect analyses in marketing panels. The workflow integrates design, estimation, diagnostics, inference, and reporting. This checklist synthesises the DiD workflow in Section 4.10, the event-study workflow in Section 5.11, and the dynamic-specific diagnostics and inference in Sections 10.7 and 10.6. This checklist quantifies uncertainty and stress-tests dynamic specifications conditional on the identification assumptions in Section 10.3. It does not, by itself, create identification.

For estimation methods, see Section 10.4. For identification, see Section 10.3.

Box 10.2: Dynamic Effects Workflow Summary

1. Define the estimand. Event-time effects, impulse responses, long-run effect and/or long-run multiplier, or half-life. Pre-specify the window based on domain knowledge.
2. Map support. Plot treated and control units for each lag. Bin sparse tails.
3. Select the estimator. Event study (discrete), distributed-lag (continuous), or dynamic DiD (staggered).
4. Run pre-trend diagnostics. Estimate leads. Interpret large or patterned leads as tension with no-anticipation or with the design.
5. Report the dynamic profile. Event-time plot with uniform confidence bands. Report LRM and half-life with bootstrap or delta-method standard errors.
6. Inference. Cluster-robust standard errors, joint statistics, multiple testing adjustments.
7. Diagnostics. Follow Section 10.7. Run placebo checks and sensitivity checks.

### Step 1: Define the Dynamic Estimand

Clarify the substantive question and define the target estimand:

- Event-time effects  $\theta_k$  (average effect at each lag relative to treatment adoption).
- Impulse responses  $\beta_s$  (marginal effect of treatment intensity at each lag).
- Long-run effect (cumulative effect over a pre-specified horizon, for example  $\widehat{L}_\beta(\bar{L}) = \sum_{s=0}^{\bar{L}} \hat{\beta}_s$  or  $\widehat{L}_\theta(\bar{K}) = \sum_{k=0}^{\bar{K}} \hat{\theta}_k$ ).
- Long-run multiplier (long-run effect expressed in units of the impact effect, for example  $\widehat{\text{LRM}}_\beta = \widehat{L}_\beta(\bar{L})/\hat{\beta}_0$  or  $\widehat{\text{LRM}}_\theta = \widehat{L}_\theta(\bar{K})/\hat{\theta}_0$ ).
- Half-life (time required for effects to decay to half).

Pre-specify the event-time window based on substantive knowledge about expected effect duration. Document the estimand, window, and rationale.

## Step 2: Map Event-Time Support

For each lag  $k$ , compute the number of treated and control observations:

1. Plot support per event time.
2. Flag lags where effective support is sparse (e.g., fewer than 10 treated and 10 control *clusters* contributing). See Definition 10.10.
3. Decide whether to bin distant lags or restrict the window.
4. Document binning choices and rationale.
5. Conduct sensitivity analyses varying bin thresholds.

## Step 3: Select the Estimator

Choose based on data structure and identification assumptions:

- **Event studies:** Discrete treatment with common or staggered adoption.
- **Distributed-lag models:** Continuous treatment intensity with expected carryover.
- **Dynamic DiD:** Staggered adoption with heterogeneity-robust estimates.
- **Continuous dose-response:** Treatment intensity with high-dimensional confounders.

Justify choice based on data structure and estimand. See Section 10.4.

## Step 4: Specify Leads, Lags, and Bins

- **Lead window:** Leads  $k < 0$  for pre-trend diagnostics. Include at least 3–5 leads.
- **Lag window:** Lags  $k \geq 0$  for post-treatment effects.
- **Reference category:** Typically  $k = -1$  for event studies.
- **Maximum lag  $\bar{L}$ :** For distributed-lag models, based on substantive knowledge or cross-validation.

Document the window, reference category, and rationale.

## Step 5: Run Anticipation and Pre-Trend Checks

1. Estimate the specification including leads ( $k < 0$ ).
2. Plot lead coefficients with confidence intervals.
3. Compute the joint lead statistic for the null that all lead coefficients are zero.
4. If the joint lead statistic does not reject, treat this as limited diagnostic evidence and interpret it alongside plots and design knowledge.

5. If the joint lead statistic rejects, diagnose differential trends, anticipation, or confounders. See Section 10.5 for interpretation of significant leads and design responses.
6. Consider covariate adjustment, factor models, or alternative controls.

Report the joint lead statistic and any adjustments. See Sections 10.5 and 10.7 for interpretation of significant leads and design responses.

### Step 6: Choose Inference Procedure

- **Cluster-robust SE:** Cluster by the independent sampling unit (often unit  $i$  or cluster  $c$ ). Report the number of independent units ( $N$  or  $G$ ).
- **Two-way clustering:** By unit and time when common shocks are plausible.
- **HAC (Newey–West):** When serial dependence decays with lag.
- **Wild cluster bootstrap:** When  $G < 30$  independent clusters.

Document choice and rationale. Report number of clusters and bootstrap replications. See Section 10.6 and Chapter 16 for implementation and trade-offs.

### Step 7: Report Dynamic Paths and Summaries

#### For event studies:

- Report  $\hat{\theta}_k$  for all lags with confidence intervals.
- Plot event-time profile marking reference category.
- For flow outcomes, report the long-run *effect* over horizon  $\bar{K}$  as  $\widehat{L}_\theta(\bar{K}) = \sum_{k=0}^{\bar{K}} \hat{\theta}_k$ , with variance computed from the joint covariance of  $(\hat{\theta}_0, \dots, \hat{\theta}_{\bar{K}})'$ . Optionally also report the long-run *multiplier*  $\widehat{\text{LRM}}_\theta = \widehat{L}_\theta(\bar{K})/\hat{\theta}_0$ . For stock outcomes (levels), the long-run effect at horizon  $\bar{K}$  is  $\hat{\theta}_{\bar{K}}$  with its standard error.

#### For distributed-lag models:

- Report impulse response  $\hat{\beta}_s$  with confidence intervals.
- Report the long-run effect  $\widehat{L}_\beta(\bar{L}) = \sum_{s=0}^{\bar{L}} \hat{\beta}_s$  and, if useful, the long-run multiplier  $\widehat{\text{LRM}}_\beta = \widehat{L}_\beta(\bar{L})/\hat{\beta}_0$ .
- Report half-life with confidence interval.

Discuss substantive interpretation of dynamic profile, LRM, and half-life.

### Step 8: Conduct Sensitivity Analyses

- **Window:** Compare full vs restricted window.

- **Binning:** Compare different bin thresholds.
- **Lag length:** Compare  $\bar{L} = 4, 6, 8, 10$ .
- **Functional form:** Compare geometric, polynomial, and flexible lags.
- **Control group:** Compare never-treated vs not-yet-treated.

Report results for multiple specifications. Discuss most plausible based on diagnostics.

### Step 9: Document Assumptions and Threats

State assumptions: parallel trends, no anticipation, overlap. Provide evidence (pre-trend diagnostics, support plots, balance checks). If you trim or restrict the sample to address overlap, state the trimmed estimand explicitly (for example, an overlap population).

Discuss threats:

- Pre-trends (differential trends unrelated to treatment).
- Anticipation (forward-looking behaviour).
- Spillovers (contamination of control units).
- Structural breaks (regime shifts at intervention).
- Sparse support (unstable estimates at distant lags).

Provide replication materials: data, scripts, software versions.

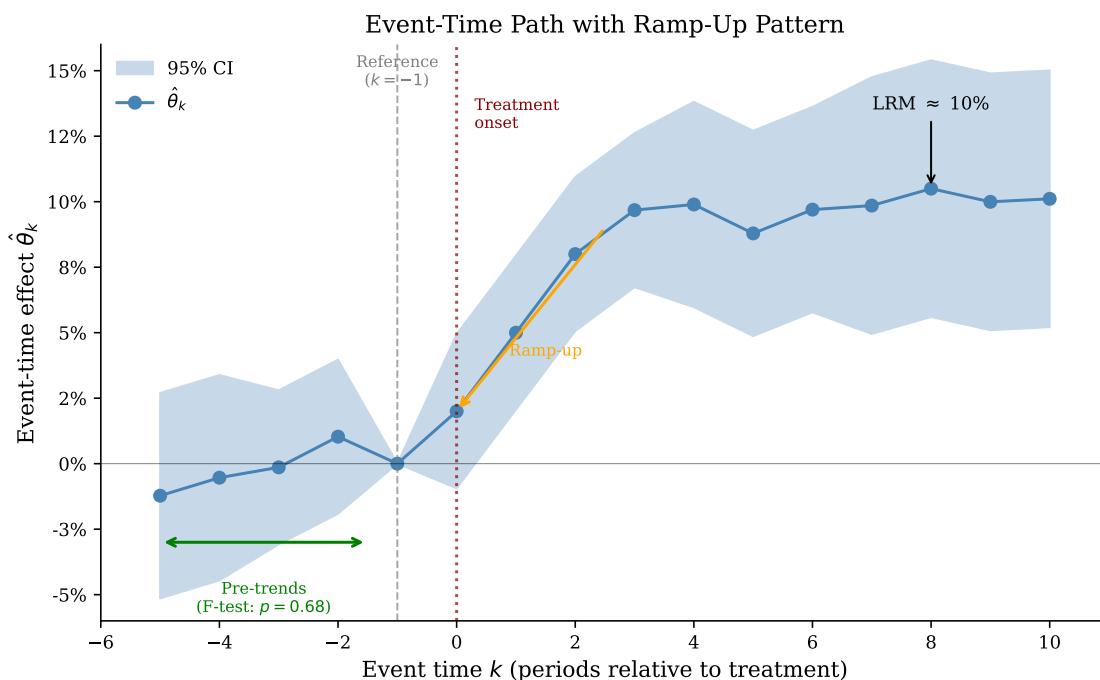
#### Box 10.3: Dynamic Effects Checklist

1. Define estimand. Specify  $\theta_k, \beta_s$ , long-run effect/multiplier, or half-life. Pre-specify the window.
2. Map support. Plot treated and control per event time. Flag sparse lags. Bin or restrict.
3. Select estimator. Event study, distributed-lag, or dynamic DiD. Justify the choice.
4. Specify leads and lags. Define windows. Specify reference category. Document the rationale.
5. Pre-trend diagnostics. Estimate leads. Plot with confidence intervals. Report the joint lead statistic. Diagnose the source if it rejects.
6. Choose inference. Cluster-robust, two-way, HAC, or wild bootstrap. Document the choice and report  $N$  or  $G$ .
7. Report paths. Plot event-time or impulse response. Report LRM and half-life with standard errors.
8. Sensitivity. Vary window, binning, lag length, functional form, control group.
9. Document. State assumptions. Discuss threats. Provide replication materials.

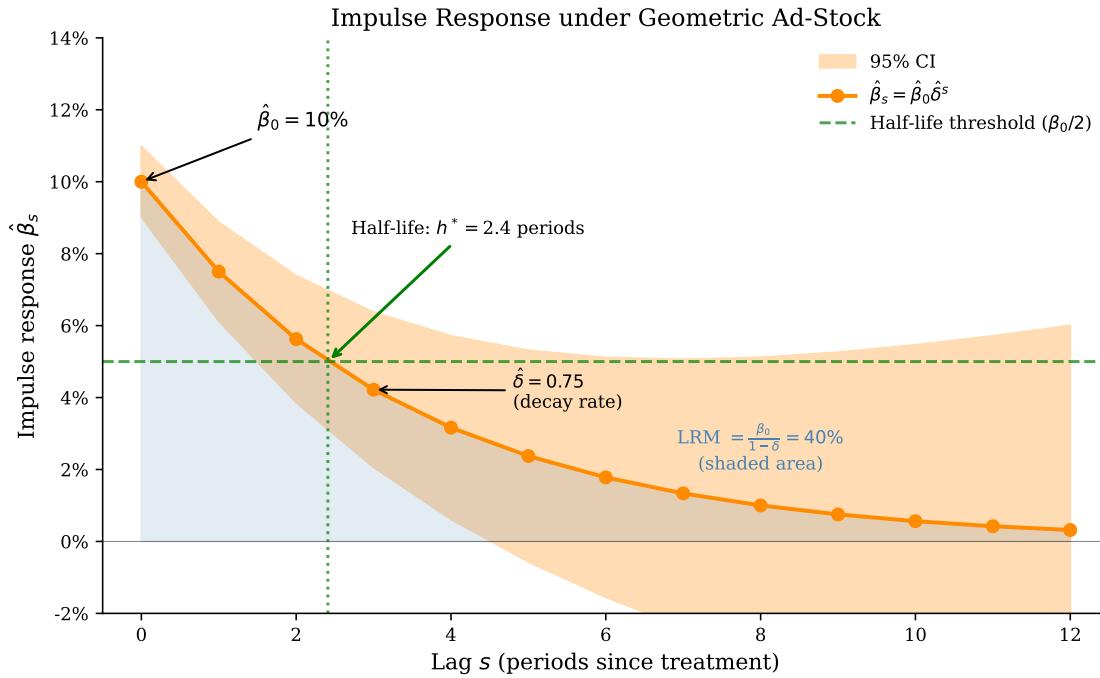
### Summary Tables and Figures

**Table 10.12** Estimand–Estimator Mapping for Dynamic Treatment Effects

Estimand	Data Structure	Estimator	Reference
Event-time effects $\theta_k$	Binary treatment with staggered adoption	Heterogeneity-robust event study	Chapter 4, 5
Impulse responses $\beta_s$	Continuous intensity with panel variation	Geometric ad-stock, Almon, flexible lags	This chapter
Long-run effect / multiplier	Binary or continuous treatment with sufficient lags	Long-run effect: $\sum_{k=0}^{\bar{K}} \theta_k$ (flows) or $\theta_{\bar{K}}$ (stocks), and $\sum_s \beta_s$ (impulse). Multiplier: long-run effect divided by the impact effect ( $\theta_0$ or $\beta_0$ ), delta method	This chapter
Half-life	Decaying effects with geometric structure	$\log 2 / (-\log \delta)$	This chapter
Dose-response $\mu(d, k)$	Continuous treatment with high-dimensional confounders	DML for panels	Chapter 12, 14



**Fig. 10.1** Event-Time Path with Ramp-Up Pattern. The figure shows illustrative event-time effects  $\hat{\theta}_k$  with 95% confidence intervals. Pre-treatment leads ( $k < -1$ ) are near zero, which is consistent with the identification assumptions under common diagnostics but not conclusive. Post-treatment effects show a gradual ramp-up, stabilising at approximately 10% after three periods. The vertical dashed line marks the reference category ( $k = -1$ ).



**Fig. 10.2** Impulse Response under Geometric Ad-Stock. The figure shows the impulse response  $\hat{\beta}_s = \hat{\beta}_0 \hat{\delta}^s$  with geometric decay ( $\delta = 0.75$ ). The contemporaneous effect is  $\beta_0 = 10\%$ . The half-life (time to decay to  $\beta_0/2$ ) is approximately 2.4 periods. The long-run effect (shaded area) is  $\beta_0/(1-\delta) = 40\%$ . The corresponding long-run multiplier is  $LRM = 1/(1-\delta) = 4$ , which expresses the long-run effect in units of the contemporaneous effect.



# Chapter 11

## Interference and Spillovers

We study violations of the no-interference component of SUTVA in marketing panels, while maintaining well-defined versions of treatment for each unit and time. Marketing interventions often change outcomes beyond the targeted units, so a unit's outcome can depend on others' treatments. We distinguish between network, geographic, and competitive spillovers, and show how each propagates through different market structures. We formalise interference using exposure mappings, writing potential outcomes as  $Y_{it}(d, h_i(D_{-i,t}))$ , where  $h_i(\cdot)$  summarises others' treatments.

Under partial interference and a credible assignment mechanism (for example, cluster randomisation or saturation designs), direct and spillover estimands can be identified and estimated. Partial-interference designs allow spillovers within pre-specified clusters of units (for example, DMAs, social circles, or product markets) but rule them out across clusters. Together with a design that generates variation in within-cluster treatment intensity, this can identify direct and spillover contrasts defined relative to the chosen exposure mapping.

We address the challenges of competition and saturation in marketing panels, showing how strategic interactions change both the relevant estimand (the policy-specific joint assignment of treatments) and the identification problem (competitors' treatments are typically endogenous). Throughout, we emphasise identification strategies and diagnostics for detecting and quantifying interference. For event-study and DiD mechanics, see Chapters 5 and 4.

## 11.1 The Challenge of Interference in Marketing

Marketing interventions often change outcomes beyond the targeted units. A unit's outcome can then depend on others' treatments. These interconnections create interference, one of the most pervasive violations of standard causal inference assumptions in marketing settings. When one customer receives a promotion, others may hear about it through word-of-mouth. When one firm advertises, competitors respond. When one store in a chain runs a sale, nearby stores experience traffic changes.

### Why Interference Matters

Three examples illustrate the practical stakes.

Word-of-mouth spillovers arise when a firm targets 10% of customers with a referral incentive. Treated customers tell untreated friends about the product, generating sales among the "control" group. Ignoring this spillover underestimates the total treatment effect and contaminates the control group.

Competitive response arises when a retailer launches a price promotion. Competitors observe and match the price cut. The treated firm's sales lift is attenuated by competitive response, and competitors' sales become endogenous to the focal firm's treatment.

Geographic spillovers arise when a media campaign runs in 30 DMAs but not in 20 control DMAs. Consumers in control DMAs see the campaign when travelling or on social media. The control group is contaminated, biasing treatment effect estimates toward zero.

In each case,  $Y_{it}$  depends on others' treatments. Standard methods that assume no interference produce biased estimates because they treat the control group as unaffected by the treated group and ignore equilibrium responses.

### When SUTVA Fails

SUTVA (Assumption 3) combines two requirements: no interference across units and no hidden versions of treatment. Under SUTVA and a binary treatment, we write potential outcomes as  $Y_{it}(0)$  and  $Y_{it}(1)$ , indexed only by the unit's own treatment. The examples above show why the no-interference component of SUTVA is often implausible in marketing panels: customers talk to each other, firms compete, and geographic markets overlap. Once  $Y_{it}$  depends on others' treatments, population averages such as ATE and ATT become policy-indexed: they are defined relative to a specified intervention that changes the joint assignment of treatments across units. Naive DiD or event-study designs that treat the control group as unaffected therefore conflate direct and spillover effects.

## Potential Outcomes Under Interference

When SUTVA fails, we can write potential outcomes as  $Y_{it}(d_{it}, D_{-i,t})$ , indexed by both the unit's own treatment and the treatments of all other units. This extends the panel potential-outcomes framework of Section 2.1. Here we emphasise the cross-sectional dimension at a given time  $t$ . The chapter then restricts this dependence through an exposure mapping  $h_i(D_{-i,t})$  so that  $Y_{it}(d, h)$  is tractable. Later sections layer dynamic dependence on the treatment history on top of this cross-sectional structure.

**Definition 11.1 (Potential Outcomes Under Interference)** For each  $\mathbf{d} \in \{0, 1\}^N$ , let  $Y_{it}(\mathbf{d})$  denote the potential outcome for unit  $i$  at time  $t$  under cross-sectional assignment  $\mathbf{d}$ . The observed outcome satisfies

$$Y_{it} = Y_{it}(\mathbf{D}_t),$$

where  $\mathbf{D}_t = (D_{1t}, \dots, D_{Nt})'$  is the realised treatment vector at time  $t$ . For each unit  $i$  and time  $t$ , the set of potential outcomes comprises  $2^N$  possible values when treatment is binary, as noted in Section 2.1. This combinatorial explosion motivates dimensionality reduction through exposure mappings (Section 11.3), which summarise others' treatments through lower-dimensional exposure variables [Hudgens and Halloran, 2008].

When dynamic dependence matters, we use the path notation from the notation guide and write  $Y_{it}(\underline{d}_i^t, h_i(D_{-i,t}))$ . In this section we suppress path dependence to focus on cross-sectional interference at time  $t$ .

This definition formalises the idea that, under interference, causal questions concern both direct effects of a unit's own treatment and spillover effects of others' treatments. The remainder of the chapter develops tools to classify spillover mechanisms, construct exposure mappings, and estimate direct and spillover effects under workable assumptions. In the rest of the chapter we focus on estimands that separate the effect of a unit's own treatment from the effect of others' treatments as summarised by exposure mappings. Under partial interference, we study contrasts of the form  $\mathbb{E}[Y_{it}(1, e) - Y_{it}(0, e)]$  and  $\mathbb{E}[Y_{it}(d, e') - Y_{it}(d, e)]$ , defined over the analysis population of unit-period cells for which the required exposure levels have support.

## Chapter Overview

Table 11.1 provides the roadmap for this chapter.

For comparison with standard DiD methods that assume no interference, see Chapter 4. For dynamic effects under SUTVA, see Chapter 10.

**Table 11.1** Chapter Roadmap: Interference and Spillovers

Section	Topic
11.2	Taxonomy of spillover mechanisms in marketing
11.3	Exposure mappings and potential outcomes
11.4	Partial interference framework
11.5	Estimation of direct and spillover effects
11.6	Competitive reactions and saturation effects
11.7	Diagnostic procedures
11.8	Marketing applications
11.9	Conclusions and workflow

## 11.2 Types of Spillovers in Marketing Panels

Interference manifests through distinct mechanisms in marketing data. We focus on three spillover mechanisms that recur in marketing panels: geographic, network, and competitive spillovers. For the general framework of exposure mappings that enable estimation under interference, see Section 11.3. For partial interference assumptions that enable tractable identification, see Section 11.4.

### Overview of Spillover Types

Table 11.2 summarises the three primary spillover mechanisms.

**Table 11.2** Comparison of Spillover Types in Marketing

Type	Exposure Measure	Decay Function	Marketing Example
Geographic	$\sum_j D_{jt} \cdot k(d_{ij})$	Inverse distance or exponential	Retail promotions, local advertising
Network	$\sum_{j \in \mathcal{N}_i} D_{jt}$ or fraction	None (binary connection)	Word-of-mouth, referral programmes
Competitive	$\sum_{j \neq i} D_{jt} \cdot \rho_{ij}$	Market overlap or similarity	Price competition, ad wars

### Interference Structure

**Definition 11.2 (Interference Structure)** An interference structure specifies which units can affect which other units. Let units be indexed by  $i = 1, \dots, N$ . An interference structure is a collection  $\{\mathcal{S}_i\}_{i=1}^N$ , where  $\mathcal{S}_i \subseteq \{1, \dots, N\} \setminus \{i\}$  denotes the set of units whose treatment can affect unit  $i$ 's outcome. For simplicity we treat the interference structure as fixed over the analysis window. This is a substantive restriction: if treatment changes the network or the spatial weights, then exposure becomes endogenous and many identification arguments in this chapter no longer apply without additional structure. Three common structures are:

- (i) Geographic interference, where  $\mathcal{S}_i = \{j : d_{ij} \leq \bar{d}\}$ ,  $d_{ij}$  is distance, and  $\bar{d}$  is a threshold.
- (ii) Network interference, where  $\mathcal{S}_i = \mathcal{N}_i$  and  $\mathcal{N}_i = \{j : A_{ij} = 1\}$  is the set of network neighbours.
- (iii) Competitive interference, where  $\mathcal{S}_i = \{j : \rho_{ij} > 0\}$  and  $\rho_{ij}$  captures product similarity or market overlap.

The interference structure  $\{\mathcal{S}_i\}$  is the structural skeleton that exposure mappings (Section 11.3) will later turn into scalar or low-dimensional exposures. In the geographic case, for example, the exposure mapping will be the distance-weighted treated-neighbour count over  $\mathcal{S}_i$ . Partial interference (Section 11.4) will impose that units can be partitioned into clusters such that  $\mathcal{S}_i$  is contained within unit  $i$ 's cluster for all  $i$ . Spillovers are allowed within clusters but ruled out across clusters.

## Geographic Spillovers

Geographic spillovers occur when treatment in one spatial unit affects outcomes in nearby units.

**Setting.** A retailer operates 200 stores across 50 metropolitan areas. The retailer launches a promotional campaign in 20 randomly selected stores. Customers near treated stores travel to take advantage of the promotion. Customers near untreated stores may also travel to treated stores.

**Mechanism.** Cross-shopping creates a negative spillover on untreated stores (sales decline) and a positive spillover on treated stores (additional sales from customers travelling from untreated areas).

**Formal specification.** Let  $d_{ij}$  denote distance between stores  $i$  and  $j$ . Define exposure as  $E_{it}^{\text{geo}} = \sum_{j \neq i} D_{jt} k(d_{ij})$ , where  $D_{jt} \in \{0, 1\}$  denotes treatment status (or  $D_{jt} \in \mathbb{R}$  for intensity, if applicable). We then write potential outcomes as  $Y_{it}(d, e)$  with  $e = E_{it}^{\text{geo}}$ . The potential outcome depends on own treatment and distance-weighted neighbour treatment:

$$Y_{it} = Y_{it}(D_{it}, E_{it}^{\text{geo}}),$$

where  $k(d_{ij})$  is a decay function: inverse distance  $k(d) = 1/d$ , inverse squared  $k(d) = 1/d^2$ , or exponential  $k(d) = \exp(-\lambda d)$ . As in Definition 11.1,  $Y_{it}(D_{it}, \sum_{j \neq i} D_{jt} k(d_{ij}))$  is shorthand for the potential outcome evaluated at the realised treatment paths whose period- $t$  cross-section has components  $D_{it}$  and  $\{D_{jt}\}_{j \neq i}$ . The choice of decay function  $k(d)$  and the radius defining “nearby” stores should be justified by data and domain knowledge. In practice, analysts compare alternative specifications in diagnostics (Section 11.7). In observational settings, the functional form of  $k(\cdot)$  is typically not point identified from the data alone. Analysts therefore treat  $k$  and the cluster radius as design choices, report diagnostics across plausible specifications, and interpret estimated spillover effects as conditional on those choices rather than nonparametrically identified objects.

**Distance decay.** Spillovers often decay with distance, but the relevant scale is setting-specific (urban density, category, travel costs). In applications you should justify a distance cut-off using domain knowledge and pre-treatment mobility or catchment data. This decay motivates the partial interference framework (Section 11.4), which partitions space into clusters and assumes spillovers are contained within clusters.

**Data requirements.** Geographic information systems (GIS) data on store locations, customer addresses, and travel patterns inform the specification of decay functions and cluster boundaries.

## Network Spillovers

Network spillovers propagate through social or professional connections between individuals.

**Setting.** A social media platform launches a new feature for a randomly selected subset of users. Treated users discuss the feature with friends, generating awareness and adoption among untreated users.

**Mechanism.** Word-of-mouth creates a positive spillover. Direct platform interactions (messages, content sharing) create additional spillover pathways.

**Formal specification.** Let  $\mathcal{N}_i$  denote user  $i$ 's network neighbours. Let  $E_{it}^{\text{net}} = \sum_{j \in \mathcal{N}_i} D_{jt}$  or  $E_{it}^{\text{net}} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} D_{jt}$ , depending on whether the estimand targets total or average peer exposure, and write  $Y_{it}(D_{it}, E_{it}^{\text{net}})$ . The potential outcome depends on own treatment and neighbour treatment:

$$Y_{it} = Y_{it}(D_{it}, E_{it}^{\text{net}}).$$

Alternatively, you can use the fraction of treated neighbours to normalise by degree. Counts of treated neighbours identify effects per additional treated neighbour (holding network size fixed), whereas fractions identify effects of increasing the share of treated neighbours (holding network size fixed). These correspond to different policy questions and should be pre-specified before estimation. As in Definition 11.1,  $Y_{it}(D_{it}, E_{it}^{\text{net}})$  is shorthand for the potential outcome evaluated at the realised treatment paths whose period- $t$  cross-section has components  $D_{it}$  and  $\{D_{jt}\}_{j \neq i}$ .

**Identification challenges.** Network spillovers raise several identification concerns. Selection on network position arises when high-degree users are more likely to be treated and have different baseline outcomes. Homophily creates confounding because users choose friends with similar characteristics. Mitigations include pre-stratifying the randomisation by degree or centrality, and using designs that induce exogenous variation in peers' treatment (for example, saturation designs within predefined communities). In observational settings, homophily and reflection typically require instruments or experimental variation. Fixed effects alone rarely suffice.

## Competitive Spillovers

Competitive spillovers arise from strategic interactions between firms.

**Setting.** Firm A launches an advertising campaign. If effective, Firm A gains market share at the expense of Firms B and C. Competitors respond with their own advertising, promotions, or price adjustments.

**Mechanism.** Competitive reactions create negative spillovers on Firm A (campaign effect dampened) and positive spillovers on competitors (outcomes improve from reactions).

**Formal specification.** Let  $D_{it}$  denote firm  $i$ 's treatment intensity at time  $t$ . Define  $E_{it}^{\text{comp}} = \sum_{j \neq i} D_{jt} \rho_{ij}$  and write  $Y_{it}(D_{it}, E_{it}^{\text{comp}})$ . The potential outcome depends on own treatment and competitor treatment:

$$Y_{it} = Y_{it}(D_{it}, E_{it}^{\text{comp}}),$$

where  $\rho_{ij}$  is market share, geographic proximity, or product similarity. As in Definition 11.1,  $Y_{it}(D_{it}, E_{it}^{\text{comp}})$  is shorthand for the potential outcome evaluated at the realised treatment paths whose period- $t$  cross-section has components  $D_{it}$  and  $\{D_{jt}\}_{j \neq i}$ .

**Strategic interactions.** Competitive spillovers can operate in either direction. Under strategic substitutes, competitors' actions dampen the focal firm's effect, creating negative spillovers. Under strategic complements, competitors' actions reinforce the focal firm's effect, creating positive spillovers. Throughout this chapter we treat the observed competitive responses as part of the spillover mechanism within the sampled market. We do not attempt to solve the full dynamic game or characterise counterfactual equilibria. When such questions matter, structural IO or dynamic oligopoly models are more appropriate than the reduced-form designs we emphasise here.

**Identification.** Panel data on firm actions and outcomes allow you to model reaction functions and measure co-movement. Causal interpretation of competitive spillovers typically requires instruments, policy shocks, or experimental variation that shifts one firm's treatment without directly shifting rivals' demand. When such instruments are unavailable, dynamic game or structural models may be required rather than reduced-form IV.

**Interaction with dynamics.** Competitive spillovers complicate staggered adoption designs (Chapter 4). If early adopters trigger competitive reactions, treatment effects may decline over time due to competition, not treatment decay. Distinguishing dynamic effects (Chapter 10) from competitive spillovers requires explicit modelling or sensitivity analysis.

**Box 11.1: Spillover Magnitudes in Practice**

The following stylised scenarios illustrate how spillovers can bias treatment effect estimates. These numbers are hypothetical and are used only to illustrate orders of magnitude. They are not evidence and should not be treated as benchmarks.

**Geographic spillovers (retail).** Treated stores experience a 15% sales increase, while untreated stores within 5 km experience a 3% sales decline due to cross-shopping. The net effect is 12%, compared to the 15% naive estimate that ignores spillovers. Some field experiments report economically meaningful geographic spillovers that can materially attenuate naive estimates. The magnitude is context-specific.

Retail spillovers operate over space. On platforms, they propagate through networks.

**Network spillovers (social media).** Users with one treated friend are 20% more likely to adopt a feature, while users with three or more treated friends are 40% more likely. This non-linear pattern suggests saturation effects as exposure increases.

In product markets, interference arises from strategic reactions rather than travel or communication.

**Competitive spillovers (retail pricing).** A 10% price cut by the focal store induces a 5% cut by competitors within 1 km. The direct effect might be a 12% sales increase, but the net effect is only 8% after accounting for competitive response.

**Implication.** Ignoring spillovers biases treatment effect estimates. The direction of bias depends on spillover sign (positive or negative) and mechanism. These magnitudes make clear that even “modest” spillovers can change business conclusions and should be treated as first-order design considerations, not second-order nuisances.

The next sections translate these interference structures into exposure mappings and partial-interference assumptions that make it possible to estimate direct and spillover effects in marketing panels.

### 11.3 Exposure Mappings and Potential Outcomes Under Interference

Exposure mappings formalise how treatment of one unit affects outcomes of others. Building on the panel potential-outcomes framework in Section 2.1, we now summarise the cross-sectional treatment vector via exposure mappings so that interference becomes tractable. Under SUTVA, potential outcomes depend only on own treatment. Here we allow them to depend on the treatment vector via a summary statistic. For the taxonomy of spillover types, see Section 11.2. For partial interference assumptions, see Section 11.4.

#### Exposure Mapping Framework

**Definition 11.3 (Exposure Mapping)** An exposure mapping is a function  $h_i(\cdot)$  that maps others' treatments to an exposure variable. In a binary treatment setting, write  $E_{it} = h_i(D_{-i,t})$ , where  $D_{-i,t} = (D_{1t}, \dots, D_{i-1,t}, D_{i+1,t}, \dots, D_{Nt})$ . The exposure mapping takes the form  $h_i : \{0,1\}^{N-1} \rightarrow \mathcal{E}$ , where  $\mathcal{E} \subseteq \mathbb{R}^d$  with  $d \ll N$ . Different cross-sectional assignments that generate the same exposure  $E_{it}$  are treated as causally equivalent for unit  $i$ .

In panels we write  $E_{it} = h_i(D_{-i,t})$  to emphasise that exposure is indexed by both unit  $i$  and time  $t$ .

**Assumption 44 (Exposure Mapping Sufficiency)** The exposure mapping  $h_i$  is sufficient for potential outcomes if:

$$Y_{it}(\mathbf{d}) = Y_{it}(\mathbf{d}') \quad \text{whenever } d_i = d'_i \text{ and } h_i(\mathbf{d}_{-i}) = h_i(\mathbf{d}'_{-i}).$$

This is a substantive restriction: it defines which spillover patterns are treated as observationally and causally equivalent for unit  $i$ . Under this assumption, potential outcomes depend on the treatment vector only through own treatment  $d_i$  and exposure  $E_i = h_i(\mathbf{d}_{-i})$ :

$$Y_{it}(\mathbf{d}) = Y_{it}(d_i, E_i).$$

This reduces the indexing from the full  $\{0,1\}^N$  assignment to own treatment and a low-dimensional exposure. In panels we write  $E_{it} = h_i(D_{-i,t})$  for the realised exposure and treat  $E_{it} \in \mathcal{E}$  as the relevant summary of others' treatments. This sufficiency assumption is strong. It rules out cases where the pattern of who is treated matters beyond what is summarised by the exposure, for example when one high-degree neighbour has a different impact from several low-degree neighbours with the same total exposure. In practice,  $h_i$  encodes a modelling choice about which features of others' treatments you believe are relevant for  $Y_{it}$ . Diagnostic checks in Section 11.7 should probe the sensitivity of conclusions to this choice.

#### Parametric Exposure Mappings

In the examples below we work with scalar exposure variables for simplicity. Vector-valued exposures can be handled by stacking multiple scalar mappings.

Table 11.3 summarises common parametric exposure mappings.

**Table 11.3** Common Exposure Mappings

Type	Formula	Interpretation
Geographic	$E_{it}^{\text{geo}} = \sum_{j \neq i} D_{jt} \cdot k(d_{ij})$	Distance-weighted treatment of neighbours
Network (count)	$E_{it}^{\text{net}} = \sum_{j \in \mathcal{N}_i} D_{jt}$	Number of treated network neighbours
Network (fraction)	$E_{it}^{\text{net}} = \frac{1}{ \mathcal{N}_i } \sum_{j \in \mathcal{N}_i} D_{jt}$	Fraction of treated neighbours
Competitive	$E_{it}^{\text{comp}} = \sum_{j \neq i} \rho_{ij} D_{jt}$	Similarity-weighted competitor treatment

**Distance decay functions.** For geographic exposure, common kernels include inverse distance  $k(d) = d^{-1}$ , inverse squared  $k(d) = d^{-2}$ , and exponential  $k(d) = \exp(-\lambda d)$ .

**Network normalisation.** Count exposure captures absolute spillovers. Fraction exposure captures relative spillovers (normalised by degree).

**Competitive weights.** Common choices for competitive weights include market share  $s_j$ , product similarity  $\rho_{ij}$ , or geographic proximity.

## Regularity Conditions

**Assumption 45 (Exposure Regularity Conditions)** The exposure mapping  $h_i$  satisfies:

- (i) There exists  $M < \infty$  such that  $\sup_{\mathbf{d}_{-i}} \|h_i(\mathbf{d}_{-i})\| \leq M$ , where  $\|\cdot\|$  is a fixed norm on  $\mathbb{R}^d$ .
- (ii) Exposure depends on a bounded number of neighbours (for example, within a radius or within a network neighbourhood).
- (iii) In designs that construct exposure using both treatments and covariates, you may strengthen boundedness by requiring that  $\|h_i(D_{-i,t}) - h_j(D_{-j,t})\| \leq L \|X_{it} - X_{jt}\|$  for some  $L < \infty$  and a fixed norm on covariates.

## Treatment Effects Under Interference

With exposure mappings, potential outcomes are  $Y_{it}(d, e)$ , where  $d \in \{0, 1\}$  is own treatment and  $e$  is exposure. The observed outcome is  $Y_{it} = Y_{it}(D_{it}, E_{it})$ .

Unless otherwise stated, expectations  $\mathbb{E}[\cdot]$  in this chapter are taken over the empirical distribution of unit-period cells  $(i, t)$  in the analysis sample. To avoid collisions with the book's staggered-adoption notation,

we write the direct, spillover, and total effects using roman labels. The resulting objects  $\text{DE}(e)$ ,  $\text{SE}(d, e, e')$ , and  $\text{TE}(e, e')$  are average effects across units and times at the specified exposure levels.

### Treatment Effects Under Interference

**Direct effect** (own treatment, holding exposure constant). For exposure level  $e$  with support in the analysis population, define the average direct effect as:

$$\text{DE}(e) = \mathbb{E}[Y_{it}(1, e) - Y_{it}(0, e)],$$

where the expectation averages over the target population of unit-period cells.

**Spillover effect** (exposure change, holding own treatment constant):

$$\text{SE}(d, e, e') = \mathbb{E}[Y_{it}(d, e') - Y_{it}(d, e)].$$

**Total effect** (own treatment and exposure change):

$$\text{TE}(e, e') = \mathbb{E}[Y_{it}(1, e') - Y_{it}(0, e)].$$

**Decomposition:**

$$\text{TE}(e, e') = \underbrace{\text{DE}(e)}_{\text{direct}} + \underbrace{\mathbb{E}[Y_{it}(1, e') - Y_{it}(1, e)]}_{\text{spillover for treated}}.$$

This decomposition uses the direct effect evaluated at baseline exposure  $e$ . If  $\text{DE}(e')$  differs from  $\text{DE}(e)$ , then the “own-treatment” contribution to the total effect depends on which exposure level we take as reference. In Section 11.4 we adopt the alternative convention that evaluates the direct effect at the *target* intensity  $e'$  (or  $p'$ ). Both decompositions are algebraically valid. What differs is which component you interpret as “own treatment” versus “exposure” when both change.

**Interpretation.** The direct effect may vary with exposure  $e$  if there are interactions between own treatment and spillovers. The spillover effect may differ for treated ( $d = 1$ ) and untreated ( $d = 0$ ) units.

### Identification Strategies

Identifying direct and spillover effects requires variation in both own treatment and exposure.

**Individual randomisation limitation.** If treatment is i.i.d. across units in a large dense network, exposure  $E_{it}$  concentrates around its expectation and has little effective variation. Direct effects are well identified, but spillover effects are typically weakly identified and estimated imprecisely unless the design is very large or the network is sparse.

**Table 11.4** Identification Strategies for Spillover Effects

Strategy	Mechanism	Identifies
Individual randomisation	Random $D_{it}$ , exposure variation driven by network or geography	Identifies direct effects. Spillovers may be weakly identified unless networks are sparse and the design is large
Cluster randomisation	All units in cluster treated/untreated	Total effect. Saturation designs separate direct and spillover effects
Saturation design	Varying treatment intensity across clusters	Direct + spillover effects
Observational with exogenous exposure	Control for unit characteristics and exploit exposure variation	Can identify spillover effects under strong assumptions

**Cluster randomisation.** Randomise treatment at cluster level. Within treated clusters, all units are treated (high exposure). Within control clusters, no units are treated (zero exposure). Comparison across clusters identifies total effect.

**Saturation designs.** Assign varying treatment intensities to different clusters (e.g., 25%, 50%, 75% treated). Variation in exposure across clusters identifies spillover effects. See Section 11.4 for the formal framework.

**Observational designs.** Treatment is assigned based on unit characteristics and exposure varies due to geographic or network structure. Analysts control for unit characteristics (fixed effects, matching). Residual exposure variation identifies spillover effects under conditional exogeneity.

Observational designs rely on conditional exogeneity assumptions such as  $\{Y_{it}(d, e) : d \in \{0, 1\}, e \in \mathcal{E}\} \perp (D_{it}, E_{it}) | X_{it}^{\text{pre}}, \alpha_i, \lambda_t$ , where  $E_{it}$  is constructed from neighbours' treatments and  $X_{it}^{\text{pre}}$  denotes pre-treatment covariates. In panel terms, this is a restriction on the assignment mechanism  $\Pr(\mathbf{D} | \{Y_{it}(d_i^t)\}_{i,t}, \mathbf{X})$  once you have fixed the exposure mapping and the no-anticipation restrictions that justify the shorthand  $Y_{it}(d, e)$ . It requires that the joint process  $(D_{it}, E_{it})$  is as good as random conditional on pre-treatment covariates and fixed effects. Because exposure is itself a function of neighbours' choices, homophily and reflection make this assumption demanding. In practice, analysts combine rich covariates, unit and cluster fixed effects, and sensitivity analyses rather than claiming full non-experimental identification.

In practice, credible identification of spillover effects typically relies on partial interference (clusters within which interference is allowed but across which it is ruled out) and on experimental saturation designs. The next section formalises partial interference and describes estimation strategies for direct and spillover effects under these assumptions.

## 11.4 Partial Interference and Cluster Designs

Partial interference provides a tractable middle ground between SUTVA, which rules out any spillovers, and full interference, which allows arbitrary spillovers across all units. Partial interference assumes that the population can be partitioned into clusters, with interference occurring within clusters but not between clusters.

This assumption is plausible in many marketing settings where spillovers are geographically or socially localised. For exposure mappings, see Section 11.3. For estimation, see Section 11.5. For diagnostics, see Chapter 17. For inference, see Chapter 16.

### Formal Framework

**Assumption 46 (Partial Interference)** Assume the units can be partitioned into clusters  $\{\mathcal{C}_c\}_{c=1}^G$  such that, for any time  $t$ , unit  $i \in \mathcal{C}_c$  satisfies  $Y_{it}(d_{it}, D_{-i,t}) = Y_{it}(d_{it}, D_{\mathcal{C}_c \setminus \{i\}, t})$ . That is, unit  $i$ 's potential outcomes depend on others' treatments only through treatments within its cluster. Treat clusters as independent sampling units for inference.

To make estimation feasible, we combine partial interference with a within-cluster exposure mapping, for example  $E_{it} = p_{ct}$ , and assume exposure sufficiency so that  $Y_{it}(\cdot)$  depends on within-cluster assignments only through  $(d_{it}, p_{ct})$ .

**Definition 11.4 (Cluster Treatment Intensity)** Let  $Z_c \in \{p^{(1)}, \dots, p^{(K)}\}$  denote the assigned cluster intensity, and define realised intensity as  $p_{ct} = N_c^{-1} \sum_{i \in \mathcal{C}_c} D_{it}$ , where  $N_c$  is the number of units in cluster  $c$ . Under perfect compliance,  $p_{ct} = Z_c$  (up to rounding). Under partial interference with exposure mapping sufficiency (Assumption 44) applied at the cluster level,  $p_{ct}$  is sufficient for within-cluster exposure, so we can write potential outcomes as  $Y_{it}(D_{it}, p_{ct})$  without loss of generality given our chosen exposure mapping.

### Estimands Under Partial Interference

State explicitly whether the target estimand is unit-weighted (averaging over unit-period cells) or cluster-weighted (averaging over clusters). Unless otherwise stated, we use unit weighting and average over unit-period cells  $(i, t)$  within the analysis window. For brevity we suppress unit and time subscripts and write  $Y(d, p)$  for  $Y_{it}(d, p)$ .

Table 11.5 summarises the estimands.

**Definition 11.5 (Estimands Under Partial Interference)** Under Assumption 46:

- (i) The direct effect changes own treatment while holding intensity fixed. For intensity  $p$ ,  $DE(p) = \mathbb{E}[Y(1, p) - Y(0, p)]$ .

**Table 11.5** Estimands Under Partial Interference

Estimand	Formula	Interpretation
Direct effect at $p$	$\text{DE}(p) = \mathbb{E}[Y(1, p) - Y(0, p)]$	Own treatment effect, holding intensity fixed
Spillover effect	$\text{SE}(p, p') = \mathbb{E}[Y(0, p') - Y(0, p)]$	Intensity change effect on untreated
Total effect	$\text{TE}(p, p') = \mathbb{E}[Y(1, p') - Y(0, p)]$	Combined own treatment and intensity change

- (ii) The spillover effect changes cluster intensity for untreated units. For intensities  $p$  and  $p'$ ,  $\text{SE}(p, p') = \mathbb{E}[Y(0, p') - Y(0, p)]$ . We focus on spillovers onto untreated units to avoid mixing direct and spillover channels. The spillover effect on treated units,  $\mathbb{E}[Y(1, p') - Y(1, p)]$ , can be defined analogously if needed.
- (iii) The total effect changes both own treatment and cluster intensity. For intensities  $p$  and  $p'$ ,  $\text{TE}(p, p') = \mathbb{E}[Y(1, p') - Y(0, p)]$ .

$$\text{TE}(p, p') = \text{DE}(p') + \text{SE}(p, p').$$

This decomposition evaluates the direct effect at the new intensity  $p'$ . If  $\text{DE}(p) \neq \text{DE}(p')$ , alternative decompositions are possible (for example, using  $\text{DE}(p)$  as the baseline). Compare the decomposition in the gray box in Section 11.3. Our convention attributes changes in own treatment at the *target* intensity to the direct effect and the change in cluster intensity from  $p$  to  $p'$  to the spillover term.

## Identification and Estimation

**Assumption 47 (Cluster Randomisation)** Treatment is randomised at two levels:

- (i) Clusters are randomly assigned to treatment intensities  $Z_c \in \{p^{(1)}, \dots, p^{(K)}\}$  with probabilities  $\{\pi_k\}_{k=1}^K$ .
- (ii) Within each cluster assigned to intensity  $p^{(k)}$ , a fraction  $p^{(k)}$  of units are randomly selected for treatment.

**Theorem 11.1 (Identification Under Cluster Randomisation)** *Under Assumptions 46 and 47, the estimands are identified by conditioning on the assigned cluster intensity  $Z_c$  and individual treatment assignment, averaging over the two-stage randomisation distribution. For assigned intensities  $p^{(k)}$  and  $p^{(k')}$ :*

- (i) *Direct effect:*  $\text{DE}(p^{(k)}) = \mathbb{E}[Y_{it} | D_{it} = 1, Z_c = p^{(k)}] - \mathbb{E}[Y_{it} | D_{it} = 0, Z_c = p^{(k)}]$ .
- (ii) *Spillover effect:*  $\text{SE}(p^{(k)}, p^{(k')}) = \mathbb{E}[Y_{it} | D_{it} = 0, Z_c = p^{(k')}] - \mathbb{E}[Y_{it} | D_{it} = 0, Z_c = p^{(k)}]$ .
- (iii) *Total effect:*  $\text{TE}(p^{(k)}, p^{(k')}) = \mathbb{E}[Y_{it} | D_{it} = 1, Z_c = p^{(k')}] - \mathbb{E}[Y_{it} | D_{it} = 0, Z_c = p^{(k)}]$ .

Here expectations are taken over the randomisation distribution induced by Assumption 47, averaging across clusters assigned to the specified intensities, units within those clusters, and time periods in the analysis window.

**Proposition 11.1 (Horvitz-Thompson Estimator)** *Under two-stage randomisation, the unbiased estimator is the difference in appropriately weighted means across the four design cells (cluster intensity  $\times$*

*individual assignment), with weights given by the known assignment probabilities. The estimator compares unit-level outcomes weighted by inverse probabilities of cluster assignment and within-cluster individual assignment. Standard finite-population inference (Chapter 16) requires treating clusters as primary sampling units to account for within-cluster correlation.*

### Worked Example: Retail Promotion Experiment

**Setting.** A retailer operates 100 stores across 20 metropolitan areas (clusters). The retailer randomises promotion intensity across metros: 5 metros at 0% (control), 5 at 25%, 5 at 50%, 5 at 75%.

**First stage.** Assign metros to intensities:  $Z_c \in \{0, 0.25, 0.50, 0.75\}$  with  $\pi_k = 0.25$  each.

**Second stage.** Within each metro assigned to intensity  $p$ , randomly select  $p \times N_c$  stores for promotion.

**Identification.** The direct effect at  $p = 0.50$  is identified by comparing treated versus untreated stores within 50% metros. The spillover effect is identified by comparing untreated stores in 75% versus 25% metros. The total effect is identified by comparing treated stores in 75% metros versus untreated stores in 25% metros.

While identification follows directly from the design, valid uncertainty quantification must treat clusters as the primary sampling units, so the effective sample size is  $G$ . Chapter 16 discusses cluster-robust and randomisation-based variance estimators for these designs.

### When Partial Interference Fails

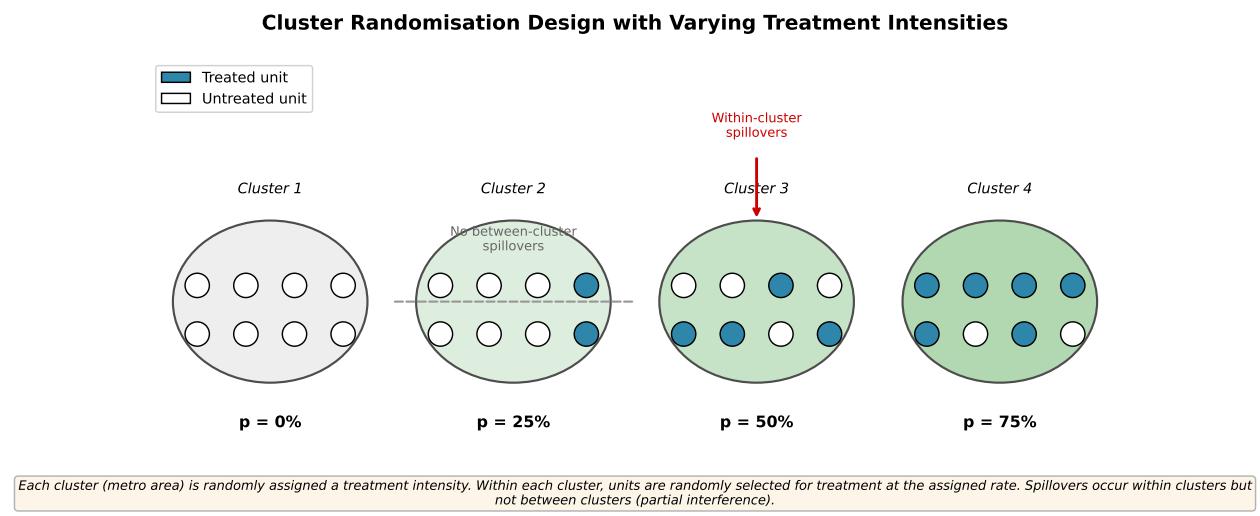
The partial interference assumption may fail in several scenarios:

**Between-cluster spillovers.** If units in different clusters interact (e.g., customers shop across metro areas, social networks span geographic boundaries), the assumption  $Y_{it}(d_{it}, D_{-i,t}) = Y_{it}(d_{it}, D_{C_c \setminus \{i\},t})$  is violated. Treat this as a diagnostic: check whether outcomes in nominally untreated boundary clusters move with neighbouring clusters' assigned intensity *in pre-treatment periods* and in placebo windows. Interpret co-movement as consistent with contamination or shared shocks. In typical marketing settings, such violations bias estimated spillover effects toward zero, because exposure in nominally “control” clusters is higher than assumed. However, the sign and magnitude of bias ultimately depend on how misclassified clusters differ in baseline outcomes and in true exposure.

**Misspecified clusters.** If the true interference structure differs from the assumed partition, estimates are biased. Example: if spillovers occur within 5 km but clusters are defined as 20 km metros, untreated units may be affected by treated units in other clusters.

**Endogenous cluster size.** If cluster sizes  $N_c$  respond to treatment, the exposure mapping  $p_{ct} = N_c^{-1} \sum_{i \in C_c} D_{it}$  is confounded. Example: promotions attract customers to a metro area, changing the denominator.

**Remedies.** Vary cluster definitions and buffer widths, and report how estimates change, as part of the robustness protocol in Section 11.7. These robustness checks should vary both the cluster partition and the exposure mapping (for example, alternative definitions of  $p_{ct}$  or distance-based exposure sets) so that conclusions do not hinge on a single, possibly misspecified interference structure. Use buffer zones (exclude units near cluster boundaries) to reduce between-cluster contamination.



**Fig. 11.1** Cluster Randomisation Design with Varying Treatment Intensities. Each cluster (metro area) is randomly assigned a treatment intensity  $p \in \{0\%, 25\%, 50\%, 75\%\}$ . Within each cluster, units are randomly selected for treatment at the assigned rate. Filled circles denote treated units; hollow circles denote untreated units. Under partial interference, spillovers occur within clusters but not between clusters.

### Practical Guidance: Cluster Construction

Cluster selection depends on the spillover mechanism under study. Geographic clusters such as metros, DMAs, or regions suit settings where spillovers decay with distance. Network clusters based on communities or cliques suit social spillover settings. Institutional boundaries—stores within a chain, schools within a district, firms within an industry—provide natural partitions when spillovers operate through organisational channels.

Cluster size involves a trade-off. Clusters must be large enough to provide within-cluster variation in treatment status, enabling estimation of direct effects. They must be small enough that the assumption of no between-cluster spillovers is credible. Similar sizes across clusters avoid heterogeneity in statistical power and simplify inference.

Intensity selection shapes what estimands are identified. Including both  $p = 0$  (pure control) and  $p = 1$  (saturation) enables estimation of the full spillover curve. Multiple intermediate intensities (e.g., 25%, 50%, 75%) allow researchers to trace how effects vary with cluster treatment density. Equal allocation across intensities maximises power for pairwise contrasts. Sparse or unbalanced choices of intensities (e.g., many clusters at  $p = 0.5$  but few at  $p = 0$  or  $p = 1$ ) will leave parts of the spillover curve poorly identified; plan the intensity grid and allocation ex ante to match the specific contrasts you care about.

Diagnostics should accompany any partial interference analysis. Treat checks for between-cluster spillovers as diagnostics rather than tests. Examine whether outcomes in cluster  $c$  correlate with treatment intensity in adjacent clusters, particularly in pre-treatment periods. Check covariate balance across clusters assigned to different intensities. Conduct placebo tests as described in Chapter 17 to assess design validity.

## 11.5 Estimation Strategies for Direct and Spillover Effects

Estimating direct and spillover effects requires methods that exploit variation in own treatment and exposure. We start with designs that generate credible exposure variation (two-stage/saturation designs and spillover-group DiD), then discuss regression adjustments that require stronger exogeneity assumptions, and finally mention specialised spatial-panel models as modelling extensions rather than primary identification strategies.

For the partial interference framework, see Section 11.4. For regression and DiD specifications, see Chapters 5 and 4. For inference, see Chapter 16. For design diagnostics, see Chapter 17.

### Overview of Estimation Approaches

Table 11.6 summarises the five estimation approaches.

**Table 11.6** Estimation Methods for Spillover Effects

Method	Data Requirements	Identifies
Regression with exposure	Panel, exposure mapping, exogenous exposure	Direct + spillover (under additivity and exogeneity)
DiD with spillover groups	Panel, treated/spillover/control partition	Total, spillover, and direct (by difference)
Two-stage design	Cluster + individual randomisation	Direct, spillover, total (by comparison)
Spatial Hausman–Taylor	Panel, spatial weights matrix, instruments	Time-invariant coefficients with spatial dependence
Spatial Vertical Regression	Location-based treatment, distance rings	Effect surface over distance rings $r$

### Regression with Exposure Controls

Under exposure sufficiency (Assumption 44) and a linear-additive structural restriction  $\mathbb{E}[Y_{it}(d, e) | \alpha_i, \lambda_t] = \alpha_i + \lambda_t + \tau d + \delta e$ , and under an assignment mechanism that renders  $(D_{it}, E_{it})$  strictly exogenous conditional on unit and time fixed effects, the fixed-effects regression

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \delta E_{it} + \varepsilon_{it},$$

where  $E_{it}$  is the exposure mapping from Definition 11.3, can be used to estimate average marginal effects. Strict exogeneity conditional on  $(\alpha_i, \lambda_t)$  requires that  $\mathbb{E}[\varepsilon_{it} | D_{i1}, \dots, D_{iT}, E_{i1}, \dots, E_{iT}, \alpha_i, \lambda_t] = 0$ . Since exposure is a function of others' treatments, this is a design and behaviour assumption: it rules out feedback

from future outcomes to future own treatment and to future exposure through network or market reconfiguration. In panels with such feedback (for example, stores increase promotions following poor sales, or customers move toward promoted stores over time), strict exogeneity is unlikely to hold. In those settings, the exposure regression remains a useful descriptive device but does not deliver causal direct or spillover effects without stronger design-based or instrumental-variables structure.

Under these conditions and assuming non-degenerate variation in both  $D_{it}$  and  $E_{it}$  after removing unit and time fixed effects (no perfect multicollinearity) and bounded moments so that standard fixed-effects asymptotics apply, the within-group estimators  $\hat{\tau}$  and  $\hat{\delta}$  are consistent:  $\hat{\tau} \xrightarrow{P} \tau$ ,  $\hat{\delta} \xrightarrow{P} \delta$ .

**Proposition 11.2 (Asymptotic Normality)** *If clusters  $c = 1, \dots, G$  are independent by design (partial interference, Assumption 46) and  $G$  is large, cluster-robust inference clustered at  $c$  is a defensible default. If exposure links clusters (for example, through geographic or network ties that span cluster boundaries), cluster-robust variance estimators will underestimate uncertainty. In that case, consider alternative clustering schemes or robust variance estimators that account for cross-cluster dependence.*

Under cluster independence and regularity conditions on within-cluster dependence and moments, with  $G$  clusters:

$$\sqrt{G} \begin{pmatrix} \hat{\tau} - \tau \\ \hat{\delta} - \delta \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

where  $\mathbf{V}$  is the asymptotic variance. Use cluster-robust standard errors (Chapter 16).

Here  $G$  is the number of independent clusters (for example, markets or DMAs) as defined in the notation guide. The limit distribution is taken as  $G \rightarrow \infty$  with cluster sizes allowed to grow or remain bounded. Cluster-robust variance estimators in Chapter 16 implement this asymptotic framework in practice.

**Interpretation.** Under the maintained linear-additive model,  $\tau$  captures the average change in  $Y$  from switching  $D$  holding  $E$  fixed, and  $\delta$  captures the average change in  $Y$  from a one-unit increase in  $E$  holding  $D$  fixed, both averaged over the estimation sample. Without the linear-additive structural restriction and strict exogeneity, they are descriptive partial correlations rather than causal quantities.

**Interaction term.** To allow direct effect to vary with exposure:

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \delta E_{it} + \theta(D_{it} \times E_{it}) + \varepsilon_{it}.$$

Interpret  $\theta$  as an interaction in the linear approximation: the marginal effect of  $D$  varies linearly with  $E$  over the support of the data. Consider centring  $E$  so that  $\tau$  is the direct effect at mean exposure. With the interaction term included,  $\tau$  is the direct effect when exposure is zero and  $\delta$  is the spillover effect when own treatment is zero. For exposure levels far from zero, the relevant marginal effects combine  $\tau$ ,  $\delta$ , and  $\theta$  evaluated at the observed  $(D_{it}, E_{it})$ . Interpret these coefficients as local derivatives over the support of the data rather than global effects.

**Identification threat.** Exposure must be exogenous conditional on fixed effects—a strong requirement that is often violated in observational settings where units with high exposure differ systematically from

those with low exposure. In such cases, the spillover coefficient  $\delta$  conflates true spillovers with selection. Design-based variation (e.g., saturation designs) or credible instruments are needed for causal interpretation.

### Difference-in-Differences with Spillover Groups

When treatment is staggered or clustered, partition untreated units into spillover units (exposed) and pure control units (not exposed).

**Three-group partition.** Table 11.7 defines the groups.

**Table 11.7** DiD Spillover Groups

Group	Definition	Indicator	Identifies (vs Pure Control)
Treated	Received treatment	$S_{it}^{\text{treat}}$	Total effect TE
Spillover	Untreated but exposed	$S_{it}^{\text{spillover}}$	Spillover effect SE
Pure control	Untreated and unexposed	Reference	Baseline

**Regression specification.** Let  $S_{it}^{\text{treat}}$  be an indicator that unit  $i$  is treated at time  $t$ , and let  $S_{it}^{\text{spillover}}$  be an indicator that unit  $i$  is untreated but exposed according to the chosen spillover rule (for example, within  $\bar{d}$  of a treated store).

$$Y_{it} = \alpha_i + \lambda_t + \beta^{\text{TE}} S_{it}^{\text{treat}} + \beta^{\text{SE}} S_{it}^{\text{spillover}} + \varepsilon_{it}.$$

If spillovers onto treated and untreated units are equal conditional on exposure (a substantive restriction), then the direct effect can be recovered as the difference  $\text{DE} = \beta^{\text{TE}} - \beta^{\text{SE}}$ . Otherwise the three-group regression does not identify a unique direct effect.

Identification requires that, in the absence of treatment and exposure, treated, spillover, and pure-control groups would have followed parallel trends in outcomes (conditional on fixed effects and controls). Violations will typically bias  $\hat{\beta}^{\text{SE}}$  toward zero if spillover units are drawn from higher-growth areas, or in unpredictable directions when selection into spillover status is driven by unobservables.

**Group definition.** For geographic spillovers, define spillover units as those within distance  $\bar{d}$  of treated units and pure controls as those beyond  $\bar{d}$ . For network spillovers, define spillover units as those with at least one treated neighbour and pure controls as those with no treated neighbours. Robustness checks should vary  $\bar{d}$  or the neighbour threshold.

## Two-Stage Designs

Two-stage designs randomise at cluster and unit levels (see Section 11.4).

**Four groups.** Two-stage designs create four groups: (1) treated units in treated clusters (high exposure), (2) untreated units in treated clusters (high exposure), (3) treated units in control clusters (if any, zero exposure), and (4) untreated units in control clusters (zero exposure).

**Identification.** The direct effect is identified by comparing groups 1 versus 2 (within treated clusters). The spillover effect is identified by comparing groups 2 versus 4 (untreated units across cluster types). The total effect is identified by comparing groups 1 versus 4 (treated in treated clusters versus untreated in control clusters).

These group contrasts correspond directly to the estimands in Definition 11.5: group 1 vs 2 recovers the direct effect  $DE(p)$  at the treated-cluster intensity, group 2 vs 4 recovers the spillover effect  $SE(p, 0)$  for untreated units, and group 1 vs 4 recovers the total effect  $TE(0, p)$  when moving from zero to high intensity.

**Implementation.** Balance at both stages required. If the number of clusters is small, stratify randomisation on pre-treatment cluster covariates and use randomisation inference or small- $G$  corrections for cluster-robust inference.

## Spatial Hausman–Taylor

This is a spatial-econometrics modelling approach for correlated effects and spatial dependence. It is not, by itself, an interference identification strategy. When spillovers follow a spatial process and time-invariant regressors are central, spatial variants of Hausman–Taylor combine quasi-demeaning with instrumental variables in spatial error or spatial autoregressive panels [Baltagi et al., 2012].

**Idea.** Use within-unit deviations of exogenous time-varying regressors as instruments for endogenous time-varying regressors. Use unit means as instruments for potentially endogenous time-invariant regressors. Model spatial dependence through weights matrix  $W$ .

**Identification caveat.** Spatial Hausman–Taylor does not on its own solve identification: credible instruments or design-based variation remain essential. In the spirit of this book’s design-first approach, treat spatial Hausman–Taylor as a modelling device layered on top of a credible identification strategy, not as a substitute for it. The method addresses spatial dependence in errors and allows estimation of time-invariant coefficients, but if the underlying instruments are weak or invalid, the estimates will be biased.

**Diagnostics.** Report first-stage fits, alternative  $W$  specifications, sensitivity to instrument partitions.

**Use case.** Coefficients on time-invariant covariates with correlated unit effects and spatial dependence.

## Spatial Vertical Regression (SVR)

For treatments at specific locations (new store, infrastructure), Grossi et al. [2025] introduce SVR, a Bayesian extension of Synthetic Control modelling spatial decay.

**Definition 11.6 (SVR Estimator)** For treated areas grouped into distance bands  $r$  from the intervention site, SVR estimates

$$\hat{\tau}(r, t) = Y_{rt} - \sum_{j \in \mathcal{C}} \hat{w}_j(r) Y_{jt},$$

where  $Y_{rt}$  is the average outcome in distance band  $r$  at time  $t$ , and weights  $\hat{w}_j(r)$  are posterior means from a Gaussian Process:

$$w_j(\cdot) \sim \mathcal{GP}(0, K_\theta).$$

The kernel  $K_\theta$  enforces smooth spatial weights.

The estimand  $\hat{\tau}(r, t)$  is a contrast of observed outcomes between treated distance ring  $r$  and a synthetic combination of control rings, interpreted through the synthetic control identification logic (Chapter 6). Map it to potential outcomes only after stating the specific intervention and what “distance ring treated” means under the design.

**Use case.** “Impact zone” studies: cannibalisation radius of flagship store, effective reach of billboard campaign. Recovers the full profile  $\hat{\tau}(r, t)$  to identify where the effect fades to zero across distance bands.

**Identification caveat.** SVR relies on the synthetic control logic: control units must provide a valid counterfactual for treated distance rings. If treatment assignment is correlated with unobserved spatial heterogeneity, the estimated effect profile  $\hat{\tau}(r, t)$  will be biased. Design-based variation or careful donor selection remains essential.

### Method Selection Guidance

Different interference problems call for different tools. You should match methods to the structure of spillovers in your data rather than defaulting to a single favourite model.

**Regression with exposure** suits settings like a retail panel with distance-weighted exposure to promoted stores. This approach requires a continuous exposure measure, exposure that is plausibly exogenous conditional on fixed effects (a strong assumption often violated in observational data), and direct and spillover effects that are approximately additive.

**DiD with spillover groups** applies when a new store opening affects nearby stores (spillover) versus distant stores (control). This method requires a clear partition into treated, spillover, and pure control groups, staggered or clustered treatment adoption, and parallel trends that are plausible for all three groups.

**Two-stage designs** work for geo-experiments with varying treatment intensity across DMAs. These designs require randomisation at both cluster and unit levels, a credible partial interference assumption, and sufficient clusters for cluster-level inference.

**Spatial Hausman–Taylor** applies when estimating brand equity effects on sales with spatial correlation. This method requires time-invariant coefficients, a known or estimable spatial weights matrix, and instruments for time-invariant regressors.

**SVR** suits impact-zone studies such as measuring the cannibalisation radius of a flagship store opening. This approach requires treatment at a specific geographic location, interest in how treatment effects decay with distance, and sufficient distance rings with control units.

In every case, state the estimand and the assignment mechanism for both  $D_{it}$  and the exposure variable before interpreting coefficients causally.

## 11.6 Competition and Saturation Effects

Competitive reactions and saturation effects represent two distinct but related forms of interference in marketing. Both phenomena violate SUTVA and require careful modelling. For spillover types, see Section 11.2. For estimation strategies, see Section 11.5. For dynamic effects, see Chapter 10.

### Overview: Two Sources of Declining Effects

Table 11.8 compares the two mechanisms.

**Table 11.8** Competition vs Saturation Effects

Feature	Competitive Reactions	Saturation Effects
Mechanism	Rivals respond to focal firm's treatment	Treatment effect declines with aggregate treatment proportion
Mathematical signature	$\partial Y_{it}(d, e) / \partial D_{jt} \neq 0$ for rivals $j$ with $\rho_{ji} > 0$	$DE(p') < DE(p)$ for $p' > p$
Diagnostic	Competitors' actions co-move with focal treatment, conditional on pre-trends	Later cohorts exhibit smaller effects
Bias if ignored	Overestimate treatment effect	Misattribute declining effects to unit heterogeneity
Implication	Counterfactual effects depend on rivals' behaviour	Diminishing marginal returns to expanding coverage

Here  $DE(p)$  is the direct-effect function from Definition 11.5, written as a function of the aggregate treatment proportion  $p$ .

Three distinct mechanisms can generate declining treatment effects over time or across cohorts. Dynamic decay concerns the event-time profile  $\theta_k$  (Chapter 10). Saturation concerns how effects vary with cluster intensity  $p$  (Section 11.4). Competitive reactions concern how rivals' treatments respond over time. Distinguishing these mechanisms is essential for correct interpretation and optimal policy.

### Competitive Reactions

Building on the competitive spillover framework introduced in Section 11.2, we now formalise the estimation of reaction functions. Competitive reactions create dynamic spillovers that evolve over time. Consider a market with  $N$  firms. Firm  $i$  increases advertising expenditure at time  $t$ . If the advertising is effective, firm  $i$  gains market share. In many categories competitors may respond with a lag (for example, by adjusting advertising budgets in the next period).

This competitive response dampens firm  $i$ 's market share gain. The observed treatment effect at time  $t+1$  is smaller than at time  $t$ , not because the advertising wears off but because competition intensifies.

We now move to a reduced-form description of firms' dynamic best responses.

**Definition 11.7 (Reduced-Form Competitive Reaction Function)** In a market with  $N$  firms, the competitive reaction function for firm  $j$  in response to firm  $i$ 's treatment is:

$$D_{jt} = R_j(D_{i,t-1}, \mathbf{X}_{jt}, \mathbf{D}_{-ij,t-1}) + \varepsilon_{jt},$$

where  $D_{i,t-1}$  is firm  $i$ 's lagged treatment,  $\mathbf{X}_{jt}$  are firm  $j$ 's characteristics, and  $\mathbf{D}_{-ij,t-1}$  are other firms' lagged treatments. The marginal reaction coefficient is:

$$\rho_{ji} = \frac{\partial R_j}{\partial D_{i,t-1}}.$$

Interpret  $\rho_{ji}$  as a reduced-form sensitivity of  $D_{jt}$  to lagged  $D_{i,t-1}$ , conditional on included controls. Causal interpretation requires an instrument or experimental variation in  $D_{i,t-1}$ . The sign of  $\rho_{ji}$  is positive for strategic complements and negative for strategic substitutes.

**Proposition 11.3 (Total Effect Accounting for Competition)** Let  $Y_{it}(D_{it}, \mathbf{D}_{-i,t})$  denote firm  $i$ 's outcome under its own treatment  $D_{it}$  and competitors' treatments  $\mathbf{D}_{-i,t}$ . The total effect of firm  $i$  increasing treatment by  $\Delta D_i$ , accounting for competitive reactions, can be approximated by a chain-rule decomposition. In a simple two-period setting where firm  $i$  changes treatment at  $t - 1$  and competitors respond at  $t$ :

$$\frac{\partial Y_{it}}{\partial D_{i,t-1}} = \underbrace{\frac{\partial Y_{it}}{\partial D_{i,t-1}}|_{direct}}_{direct persistence} + \underbrace{\sum_{j \neq i} \frac{\partial Y_{it}}{\partial D_{jt}} \cdot \rho_{ji}}_{spillover via reactions}.$$

The direct term captures the persistent effect of firm  $i$ 's lagged treatment on its own current outcome, holding competitors' current actions fixed. The spillover term captures how competitors' reactions (measured by  $\rho_{ji}$ ) affect firm  $i$ 's outcome through the contemporaneous spillover channel  $\partial Y_{it}/\partial D_{jt}$ .

If competitors' treatments harm firm  $i$  ( $\partial Y_{it}/\partial D_{jt} < 0$ ) and competitors respond positively ( $\rho_{ji} > 0$ ), the spillover component is negative, reducing the total effect below the direct persistence effect.

Identifying competitive reactions requires observing multiple firms and their actions over time. VAR-type models provide a descriptive framework for co-movement. In small  $N$  or short  $T$ , they are best treated as exploratory diagnostics unless supported by a clear instrument or design. We estimate a system of equations where each firm's advertising depends on lagged advertising of all firms:

$$D_{it} = \sum_{j=1}^N \sum_{\ell=1}^L \rho_{ij\ell} D_{j,t-\ell} + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $\rho_{ij\ell}$  captures the effect of firm  $j$ 's advertising at lag  $\ell$  on firm  $i$ 's advertising. The coefficients  $\rho_{ij\ell}$  for  $j \neq i$  measure competitive reactions. Impulse response functions trace out the dynamic path of competitive reactions following a shock to one firm's advertising.

These coefficients capture predictive responses rather than pure causal effects. Common shocks to advertising demand or unobserved market conditions can induce apparent reactions even when firms do not respond strategically.

Reduced-form approaches detect competitive reactions without estimating a full structural model. We test whether treatment of one firm predicts changes in competitors' actions. For example, in a difference-in-differences framework, we compare competitors of treated firms to competitors of control firms. If competitors of treated firms increase advertising more than competitors of control firms, this indicates competitive reactions.

## Saturation Effects

Saturation effects occur when treatment effects decline as the proportion of treated units increases. Consider a loyalty programme rolled out to stores in a retail chain. The first stores to adopt see large gains in customer retention because the programme is novel and provides a competitive advantage.

As more stores adopt, the novelty wears off and the competitive advantage diminishes. The last stores to adopt see small gains because most customers are already enrolled through other stores.

**Definition 11.8 (Saturation Function)** The saturation function describes how the direct effect varies with the aggregate treatment proportion  $p$ :

$$\text{DE}(p) = \mathbb{E}[Y_{it}(1, p) - Y_{it}(0, p)].$$

This  $\text{DE}(p)$  is the direct-effect function from Definition 11.5, now emphasised as a function of the aggregate treatment proportion  $p$ . In partial-interference designs,  $p$  typically corresponds to the cluster treatment intensity  $p_{ct}$  in Definition 11.4, averaged over clusters at a given intensity. Saturation is present if effects are smaller at higher treatment intensities, that is  $\text{DE}(p') < \text{DE}(p)$  for  $p' > p$  on the support of the design. A parametric saturation model is:

$$\text{DE}(p) = \text{DE}_0(1 - \kappa p)^+,$$

where  $\text{DE}_0$  is the effect at zero saturation,  $\kappa > 0$  is the saturation rate, and  $(x)^+ = \max(0, x)$ .

**Proposition 11.4 (Saturation Diagnostic)** *In a staggered adoption design with cohorts indexed by adoption time  $G_i$ , let  $p_t$  denote the aggregate treatment proportion at time  $t$  (which increases as more cohorts adopt). Treat the following regression as a diagnostic for patterns consistent with saturation in a monotone-adoption setting:*

$$\hat{\theta}_k(G_i) = \beta_0 + \beta_1 p_{G_i+k} + u_{G_i},$$

where  $\hat{\theta}_k(G_i)$  is the cohort-specific estimate for cohort  $G_i$  at event time  $k$ , and  $p_{G_i+k}$  is the aggregate treatment proportion at calendar time  $G_i+k$ . A negative slope ( $\beta_1 < 0$ ) is consistent with saturation, but it can also arise from cohort-specific confounding, differential dynamics, or policy targeting. Interpret together with design structure and additional diagnostics rather than as definitive causal evidence.

Distinguishing saturation from dynamic treatment effects is challenging. Both generate declining treatment effects over time. Dynamic effects occur because the treatment wears off for individual units. Saturation effects occur because the market-level treatment intensity increases. We can distinguish them by examining heterogeneity. If effects decline more for units in markets with high treatment intensity, this suggests saturation. If effects decline uniformly across markets, this suggests dynamic decay. See Chapter 10 for dynamic decay patterns in the absence of saturation, and compare those to the intensity-correlated patterns described here.

## Implications for Causal Inference

Competitive reactions and saturation effects have important implications for causal inference in marketing panels. First, they bias standard estimators that assume SUTVA. Ignoring competitive reactions will often overestimate treatment effects when competition is zero-sum, because it fails to account for the dampening effect of rivals' responses. In markets where competitors' actions generate positive externalities, the direction of bias can be more complex. Ignoring saturation effects leads to incorrect inferences about treatment effect heterogeneity (attributing declining effects to unit characteristics rather than market-level saturation).

Second, they complicate the interpretation of treatment effects in staggered adoption designs. A declining treatment effect over cohorts may reflect saturation, competitive reactions, or dynamic decay. Distinguishing these mechanisms requires additional data (on competitors' actions, market-level treatment intensity) or structural modelling (reaction functions, saturation curves).

Third, they affect optimal treatment assignment. These mechanisms matter for policy choice. Saturation implies diminishing marginal returns to expanding coverage. Competitive reactions imply that counterfactual effects depend on rivals' behaviour. Quantifying optimal policies typically requires additional structure beyond this chapter. Chapter 12 discusses policy-learning methods that can incorporate saturation and competitive reactions when optimising treatment assignment.

**Practical Guidance: When to Worry About Each Mechanism**

Competitive reactions warrant concern when firms observe each other's actions (advertising is visible, prices are public), market share is roughly fixed creating zero-sum competition, competitors have capacity to respond quickly, and treatment effects decline even when market saturation is low.

Saturation effects warrant concern when treatment creates positive externalities that are exhaustible (novelty, network effects), later adopters show smaller effects than early adopters, effects decline more in high-penetration markets, and customer overlap across treated units is high.

Distinguishing the two mechanisms requires examining correlations. Treat checks for co-movement as diagnostics: if declining effects correlate with competitor actions (conditional on pre-trends and common shocks), competition is a plausible cause. If declining effects correlate with aggregate treatment proportion, saturation is a plausible cause. If effects decline uniformly across markets regardless of competitor actions or treatment proportion, dynamic decay (Chapter 10) is the likely explanation. These correlations are diagnostic rather than definitive. They should be interpreted alongside experimental variation or quasi-experimental shifts in competition or saturation wherever possible.

## 11.7 Diagnostics for Detecting Interference

Detecting interference requires diagnostics that assess whether patterns in outcomes and exposure are consistent with no interference under a chosen exposure mapping. These checks ask whether outcomes and exposures co-move in ways that are inconsistent with SUTVA under that mapping. They are not omnibus tests for all possible forms of interference. They are design-focused diagnostics that stress-test specific interference structures. These checks complement the general design-diagnostic framework in Chapter 17, focusing on exposure mappings and cross-unit dependence. We present three approaches: balance diagnostics on exposure, placebo diagnostics on untreated units, and sensitivity analyses. For exposure mappings, see Section 11.3. For partial interference, see Section 11.4. For general design diagnostics, see Chapter 17.

### Overview of Diagnostic Approaches

Table 11.9 summarises the three diagnostic approaches.

**Table 11.9** Diagnostics for Interference

Diagnostic	Null Hypothesis	Data Required	Limitation
Exposure balance	$\mathbb{E}[E_{it}   D_{it} = 1] = \mathbb{E}[E_{it}   D_{it} = 0]$	Exposure computed using pre-treatment network/weights	Confounded by network structure
Placebo on untreated	$\mathbb{E}[Y_{it}   D_{it} = 0, E_{it}] = \mathbb{E}[Y_{it}   D_{it} = 0]$	Outcomes for untreated, exposure variation	Requires exposure variation among controls
Sensitivity analysis	Estimates stable across specifications	Multiple exposure mappings	No formal decision rule

### Balance Diagnostics on Exposure

If treatment assignment is random and independent across units, exposure  $E_{it}$  from Definition 11.3 should be balanced across treated and untreated units. Imbalance in exposure suggests that treatment assignment is correlated with network or geographic structure, which may confound spillover estimates.

**Proposition 11.5 (Exposure Balance Diagnostic)** *Under complete randomisation, the expected difference in exposure between treated and untreated units is zero, but finite-sample imbalance is common in networks and spatial settings. Prefer a randomisation/permutation check that reassigns treatment according to the actual design and recomputes the exposure imbalance statistic.*

*For a standard two-sample comparison, a simple diagnostic statistic is:*

$$T_{balance} = \frac{\bar{E}_1 - \bar{E}_0}{\sqrt{\hat{\sigma}_E^2(1/n_1 + 1/n_0)}},$$

where  $\bar{E}_1$  and  $\bar{E}_0$  are mean exposures for treated and untreated units,  $n_1$  and  $n_0$  are the sample sizes, and  $\hat{\sigma}_E^2$  is the pooled variance. This is the usual two-sample t-statistic for a completely randomised design with equal treatment probabilities. In cluster-randomised or unequal-probability designs, the same idea applies but variances and weights must reflect the assignment scheme. See Chapter 17 for design-based balance diagnostics.

In smaller experiments, a randomisation or permutation test based on re-assigning treatment according to the original assignment mechanism (individual-, cluster-, or two-stage randomisation) within the actual network or geographic structure is preferable.

Even under perfectly random assignment, network or geographic topology can induce apparent imbalance in  $E_{it}$  if treated units happen to occupy higher-degree or more-central positions. Under a finite-population randomisation view, such imbalances are real but may reflect unlucky assignment draws rather than structural design flaws, so they should trigger further diagnostics rather than an automatic rejection of the design. A large imbalance is a signal that treatment correlates with the interference structure. It is a cue to stratify or reweight on exposure or topology.

Balance diagnostics should be conducted on pre-treatment exposure (computed using the pre-treatment network or geographic structure, and evaluating balance of exposure induced by the experimental assignment mechanism). Post-treatment exposure may be endogenous if treatment affects network formation or geographic sorting.

For example, if treated users befriend each other, post-treatment network exposure will be higher for treated users, but this does not indicate a violation of random assignment.

## Placebo Diagnostics on Untreated Units

If spillovers are present, untreated units with high exposure should have different outcomes than untreated units with low exposure. You can assess this by comparing outcomes of untreated units across exposure levels.

**Proposition 11.6 (Placebo Diagnostic on Untreated Units)** *Under the null of no spillovers (SUTVA holds), exposure should not predict outcomes among untreated units:*

$$H_0 : \mathbb{E}[Y_{it} | D_{it} = 0, E_{it} = e] = \mathbb{E}[Y_{it} | D_{it} = 0] \text{ for all } e.$$

Regressing  $Y_{it}$  on  $E_{it}$  among untreated units:

$$Y_{it} = \alpha + \delta E_{it} + \varepsilon_{it}, \quad \text{for } \{(i, t) : D_{it} = 0\},$$

and assessing whether  $\delta$  is close to zero provides a placebo diagnostic. A clear association is consistent with spillovers under the maintained exposure mapping. It can also arise from correlated shocks, sorting on network position, or exposure mismeasurement. A null finding is weak evidence, particularly when exposure has limited

*variation. In randomised designs, you can also compute a randomisation-inference p-value under the sharp null that exposure has no effect on any untreated unit-period cell under the chosen mapping. In observational settings, an association may also reflect unobserved heterogeneity correlated with exposure even when SUTVA holds, so interpret it as a warning sign rather than definitive proof of interference.*

In observational settings,  $E_{it}$  may be correlated with unobserved determinants of  $Y_{it}$  even under SUTVA (e.g., high-degree users differ systematically). Including pre-treatment covariates and fixed effects in this regression can reduce confounding, but it does not address reflection or unmeasured network heterogeneity.

This diagnostic requires variation in exposure among untreated units. If all untreated units have the same exposure (for example, in a cluster-randomised trial where control clusters have zero exposure), the diagnostic is not informative. Partial treatment within clusters (where some units are treated and others are untreated) generates variation in exposure among untreated units.

## Sensitivity Analyses

Spillover estimates depend on the specification of the exposure mapping. Sensitivity analyses assess how estimates change when we vary the exposure mapping.

For geographic spillovers, we vary the distance decay function (inverse distance, inverse squared distance, exponential decay) and the distance threshold (spillovers within 5 km, 10 km, 20 km). For network spillovers, we vary the definition of neighbours (direct connections, connections within two steps, connections within three steps).

If estimates are stable across specifications, this suggests that the results are robust to the choice of exposure mapping. If estimates vary widely, this indicates sensitivity to functional form. These variations probe the robustness of Assumption 44 to alternative plausible exposure mappings. If results change dramatically when  $h_i(\cdot)$  is modified within a reasonable class, conclusions about spillovers are fragile. In such cases, it is more accurate to say that the data are consistent with several competing exposure mappings rather than that a single mapping has been identified. Substantive judgement about which exposure mechanisms are plausible should then drive interpretation. In such cases, we should report results for multiple specifications and discuss the economic reasoning for each.

Another sensitivity analysis examines the impact of excluding units with extreme exposure. Units with very high or very low exposure may be outliers that drive the results. Trimming changes the target estimand to an “overlap” subpopulation defined by exposure. Report this explicitly and treat trimming as sensitivity analysis, not as a default correction.

**Common sensitivity specifications.** Table 11.10 lists common variations.

**Table 11.10** Sensitivity Specifications for Spillover Analysis

Spillover Type	Parameter	Common Variations
Geographic	Distance decay	Inverse, inverse squared, exponential
Geographic	Distance threshold	5 km, 10 km, 20 km
Network	Neighbour definition	Direct (1-hop), 2-hop, 3-hop
Network	Normalisation	Count, fraction, weighted by strength
Both	Outlier exclusion	Trim 5%, 10% of exposure distribution

### Diagnostic Workflow for Interference

In practice you will rarely rely on a single diagnostic. Treat the following steps as an iterative workflow that you can loop through as you refine your exposure mapping and design.

The first step is a balance check. Compute pre-treatment exposure for all units and assess whether exposure is balanced across treatment groups using the diagnostic in Proposition 11.5. If exposure imbalance is large, consider re-randomisation or stratification at the design stage (if feasible) or report reweighting as an estimand-changing adjustment. Avoid conditioning on post-treatment exposure.

The second step is a placebo diagnostic. Among untreated units, regress outcomes on exposure. If the coefficient  $\delta$  is distinguishable from zero in magnitude, the pattern is consistent with spillovers and you should estimate spillover effects using the methods in Section 11.5. If  $\delta$  is close to zero, the result is consistent with SUTVA, though not conclusive.

The third step is sensitivity analysis. Re-estimate spillover effects with alternative distance decay functions or network neighbour definitions. Re-estimate after trimming units with extreme exposures (top and bottom 5–10 per cent). Report the range of estimates across specifications and discuss the economic reasoning for preferring one specification over others.

The fourth step is reporting. Document the exposure mapping used and the alternatives considered. Report balance diagnostic and placebo diagnostic results in tables or text. If SUTVA appears violated, discuss the implications for interpreting treatment effect estimates and the adjustments made to account for spillovers. Use the estimation strategies in Section 11.5 (exposure regression, DiD with spillover groups, saturation designs) to quantify direct and spillover effects once diagnostics indicate interference.

## 11.8 Marketing Applications

We illustrate the methods developed in this chapter through three hypothetical marketing applications: geographic spillovers in retail promotions, network spillovers in social media, and competitive spillovers in advertising. Each example applies the exposure mappings from Section 11.3 and the estimation strategies from Sections 11.4 and 11.5 in concrete marketing settings. The numerical examples are illustrative and demonstrate how spillover analysis would proceed in practice. These examples illustrate workflows. Causal interpretation requires the identification assumptions stated in earlier sections. In particular, you need an assignment mechanism for  $(D_{it}, E_{it})$ , support for the exposure levels being compared, and an exposure mapping that is substantively defensible. Table 11.11 summarises the three applications.

**Table 11.11** Summary of Marketing Applications

Application	Spillover Type	Exposure Mapping	Sign	Effect on Total
Retail promotions	Geographic	Inverse distance	Negative	Reduces total effect
Social media feature	Network	Treated friends count	Positive	Amplifies
Advertising	Competitive	Competitor actions	Negative	Reduces total effect

### Application 1: Geographic Spillovers in Retail Promotions

Suppose a grocery retailer operates 150 stores across 30 metropolitan areas. The retailer launches a promotional campaign for a new product in 50 stores, randomly selected within each metro area. The outcome is weekly sales of the new product. The goal is to estimate the direct effect of the promotion and the spillover effect on nearby untreated stores.

If treatment is randomised within metro and the spillover exposure is computed from this randomisation, then the resulting variation in  $E_{it}$  can be design-induced. Interpreting the exposure regression causally still requires that (i) the exposure mapping is sufficient (Assumption 44), (ii) outcomes do not feed back into future treatment or exposure through mobility or store behaviour over the analysis window (strict exogeneity), and (iii) interference across metros is negligible (partial interference, Assumption 46) or explicitly modelled.

We define exposure using a bounded kernel. For each store  $i$ , exposure is:

$$E_{it} = \sum_{j \neq i} D_{jt} k(d_{ij}),$$

where  $k(d_{ij})$  is a bounded distance kernel, for example  $k(d) = (d + \epsilon)^{-1}$  with  $\epsilon > 0$  to avoid singularities, or an exponential kernel  $k(d) = \exp(-\theta d)$ . Treat this as the exposure mapping  $h_i(D_{-i,t}) = E_{it}$  from Definition 11.3.

We estimate a regression with store and week fixed effects:

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \delta E_{it} + \varepsilon_{it}.$$

Suppose the direct effect estimate is  $\hat{\tau} = 120$  units, indicating that treated stores sell 120 more units per week. The spillover effect estimate is  $\hat{\delta} = -8$  units per unit of exposure, indicating negative spillovers.

Untreated stores near treated stores experience sales declines as customers travel to treated stores to take advantage of the promotion.

Under the linear specification  $Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \delta E_{it} + \varepsilon_{it}$ , a rough estimate of the difference in expected outcomes between treated and untreated stores at their average exposures can be computed. Suppose the average exposure for treated stores is  $\bar{E}_{\text{treat}} = 0.8$  and for untreated stores is  $\bar{E}_{\text{untreat}} = 0.4$ . A simple linear-model back-of-the-envelope calculation is:

$$\hat{\tau} + \hat{\delta}(\bar{E}_{\text{treat}} - \bar{E}_{\text{untreat}}) = 120 + (-8)(0.8 - 0.4) = 116.8.$$

This back-of-the-envelope “total effect per treated store” incorporates both the direct lift and the average net spillover via differences in exposure between treated and untreated stores. Alternative definitions (for example, total effect at the metro level) would weight spillovers differently. In this hypothetical example, the spillover effect reduces the total effect by about 3 units, a modest but non-negligible adjustment. A more precise total-effect estimand at the metro level would aggregate both direct and spillover effects across all stores within each metro, weighting by store size or revenue as appropriate.

## Application 2: Network Spillovers in Social Media

Suppose a social media platform launches a new feature (video stories) for 10,000 users, randomly selected from a network of 100,000 users. The outcome is weekly engagement (time spent on the platform). The goal is to estimate the direct effect of the feature and the spillover effect on friends of treated users.

If treatment is assigned at random, then conditional on the realised network, exposure variation is induced by design. In finite samples, treatment can still be imbalanced by degree or centrality. It is often prudent to stratify randomisation on degree or to report degree-binned balance checks.

We compute network exposure as the number of treated friends. For each user  $i$ , exposure is:

$$E_{it} = \sum_{j \in \mathcal{N}_i} D_{jt},$$

where  $\mathcal{N}_i$  is the set of user  $i$ 's friends. We estimate a regression with user and week fixed effects:

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \delta E_{it} + \varepsilon_{it}.$$

Suppose the direct effect estimate is  $\hat{\tau} = 15$  minutes per week, indicating that treated users spend 15 more minutes on the platform. The spillover effect estimate is  $\hat{\delta} = 2$  minutes.

Interpret  $\delta$  as a linear approximation over the support of exposures observed among untreated users. If exposures are concentrated (for example, most untreated users have 0–2 treated friends), use exposure bins or nonparametric smoothing rather than a single linear slope to capture potential non-linearity.

Untreated users with treated friends spend more time on the platform, likely due to increased content (video stories from friends) and social influence.

The spillover effect may be non-linear. Estimating a specification with exposure bins (0, 1, 2–3, 4+ treated friends) might reveal that the spillover effect is largest for users with one treated friend and declines for users with more treated friends, suggesting saturation in spillover effects. This is an example of an exposure mapping that groups treated-friends counts into bins, as discussed in Section 11.3. Each bin defines a distinct exposure level  $e$  for which we can estimate  $DE(e)$  and  $SE(d, e, e')$ .

### Application 3: Competitive Spillovers in Advertising

Suppose a market has 5 firms competing in the same product category. Firm 1 increases advertising expenditure by 50 per cent for 12 weeks. The outcome is weekly market share. The goal is to estimate the direct effect on Firm 1 and the spillover effect on competitors.

Treat the VAR as descriptive unless you have an instrument or shock that shifts one firm's advertising. With few firms, reduce dimensionality (for example, by constructing a competitor-index exposure variable) and focus on design-based shifts in advertising rather than relying on VAR coefficients for causal interpretation.

We estimate a panel VAR to capture competitive reactions. Each firm's advertising depends on lagged advertising of all firms:

$$D_{it} = \sum_{j=1}^5 \rho_{ij} D_{j,t-1} + \alpha_i + \lambda_t + \varepsilon_{it}.$$

Suppose the estimated reaction coefficients indicate that competitors increase advertising in response to Firm 1's increase. If the average competitive reaction is around 0.2, competitors increase advertising by about 20 per cent of Firm 1's increase.

In the absence of randomised advertising or strong instruments, both the VAR and the regression below should be interpreted as reduced-form descriptions of how advertising and market share co-move, not as fully identified causal models.

We estimate the effect of advertising on market share using a regression with firm and week fixed effects:

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \delta \sum_{j \neq i} D_{jt} + \varepsilon_{it}.$$

Suppose the direct effect estimate is positive (own advertising increases market share) and the spillover effect estimate is negative (competitors' advertising reduces own market share).

These  $\rho_{ij}$  correspond to the reaction coefficients in Definition 11.7. Combining them with the regression of market share on own and rivals' advertising implements the total-effect decomposition in Proposition 11.3. The total effect of Firm 1's advertising increase, accounting for competitive reactions, combines the direct

effect and the spillover effect weighted by competitors' responses. If competitive reactions are substantial, they can reduce the total effect by a meaningful fraction, potentially 20–40 per cent of the direct effect. When credible instruments or experiments are available, they should be integrated into both stages so that the total-effect decomposition reflects causal reactions and causal outcome responses rather than purely predictive relationships.

### Key Takeaways from Applications

Spillover sign varies by context. Geographic spillovers from promotions are often negative because customers travel to treated stores, reducing sales at untreated competitors. Network spillovers from new features are typically positive in engagement metrics because content and social influence spread through connections. Competitive spillovers from advertising are often negative because rivals respond with their own campaigns.

Magnitude matters for decisions. Small spillover adjustments may not change managerial recommendations, but large adjustments can substantially affect ROI calculations and optimal budget allocation. Large negative spillovers can materially change ROI calculations and therefore the decision, relative to an analysis that assumes no interference. Non-linear effects such as saturation should inform the optimal treatment intensity—treating more units may yield diminishing returns if spillover effects saturate.

Estimation requires methods appropriate to the spillover mechanism. Geographic spillovers call for regression with exposure controls (Section 11.5). Network spillovers may require exposure bins to capture non-linearity (Section 11.3). Competitive spillovers require panel VAR or similar dynamic models to capture reaction functions (Section 11.6).

In all three settings, you should first run the diagnostic workflow in Section 11.7 to check balance and placebo patterns under your chosen exposure mapping before committing to a particular estimation strategy.

## 11.9 Conclusion

Interference is pervasive in marketing. Geographic spillovers arise from customer mobility and overlapping media markets. Network spillovers propagate through social connections and word-of-mouth. Competitive spillovers emerge from strategic interactions between firms. Ignoring spillovers can bias estimators that assume no interference, because treated and control units need not represent the intended counterfactual states. The direction and magnitude depend on the spillover structure and the estimand. In this chapter we have replaced the single SUTVA-based treatment effect with a set of design- and exposure-mapping-dependent objects: direct effects of own treatment, spillover effects of others' treatments as summarised by exposure mappings, and policy-specific total effects that combine both.

## Chapter Roadmap

Table 11.12 summarises the chapter sections for easy reference.

**Table 11.12** Chapter 11 Roadmap

Topic	Section	Key Content
SUTVA and violations	Section 11.1	Why SUTVA fails, interference types
Spillover taxonomy	Section 11.2	Geographic, network, competitive
Exposure mappings	Section 11.3	Formalising spillover mechanisms
Partial interference	Section 11.4	Cluster designs, estimands
Estimation	Section 11.5	Five estimation approaches
Competition & saturation	Section 11.6	Dynamic and market-level effects
Diagnostics	Section 11.7	Balance, placebo, sensitivity
Applications	Section 11.8	Retail, social media, advertising

This chapter has developed a framework for causal inference under interference. Exposure mappings [Aronow and Samii, 2017] give us a way to collapse the full treatment vector into low-dimensional exposure variables. We have adapted this framework to panel marketing settings, tying exposure to geographic, network, and competitive structures and aligning it with the partial-interference and dynamic designs developed in earlier chapters. Partial interference provides a tractable structure by partitioning units into clusters. Estimation strategies exploit variation in own treatment and exposure to identify direct and spillover effects. Diagnostics detect interference and assess sensitivity to modelling choices. Throughout, credible identification of both direct and spillover effects still rests on the same design principles as in earlier chapters. You need to understand the assignment mechanism for both own treatment and exposure, ensure overlap and support for the exposure levels you want to study, and justify parallel-trends or exogeneity assumptions for the joint process ( $D_{it}, E_{it}$ ).

**Box 11.2: Interference and Spillovers Workflow**

Building on the general design workflow in Chapter 17 and the event-study and dynamic workflows in Chapters 5 and 10, the interference-specific workflow is as follows.

Begin by defining the exposure mapping (Section 11.3). Specify whether spillovers operate through geographic proximity, network connections, or competitive dynamics, and formalise the mechanism using an appropriate functional form such as inverse distance or fraction of treated neighbours.

Next, choose the research design (Section 11.4). Cluster-randomised trials, two-stage designs, and observational studies with plausibly exogenous exposure variation (for example, instruments or natural experiments that move neighbours' treatments without directly affecting outcomes) can identify spillover contrasts under their respective assumptions, provided the design delivers support for the exposure levels being compared and the exposure mapping is pre-specified and defensible. If invoking partial interference, justify why between-cluster spillovers are negligible.

Then estimate direct and spillover effects (Section 11.5). Regression with exposure controls, DiD with spillover groups, or two-stage estimation approaches can separate own-treatment effects from exposure-mediated effects. Report estimated direct, spillover, and total effects for the exposure levels that matter for your decision.

Conduct diagnostics to stress-test the analysis (Section 11.7). Balance diagnostics assess whether exposure is balanced across treatment groups. Placebo diagnostics check whether exposure predicts outcomes among untreated units. Sensitivity analyses vary the exposure mapping to assess robustness. If competitive reactions or saturation effects are plausible (Section 11.6), diagnose them explicitly. Competitive reactions show up as changes in competitor behaviour after treatment. Saturation shows up when effects decline as the proportion of treated units grows.

Finally, apply appropriate inference methods (Chapter 16). Cluster at the level at which treatment is randomised and interference is assumed to be contained (typically  $c = 1, \dots, G$  clusters under partial interference). Randomisation inference is exact only under the sharp null and when you condition on the realised assignment mechanism. Report the total effect alongside direct and spillover components, and follow the full diagnostic protocol in Chapter 17.

Three lessons emerge for practitioners. First, always question the SUTVA assumption. Consider whether spillovers are plausible given the market structure, the treatment, and the outcome. If spillovers are likely, model them explicitly using exposure mappings.

Second, design studies to generate variation in exposure. Cluster randomisation, two-stage designs, and observational designs with exogenous exposure variation enable identification of spillover effects.

Third, report both direct and spillover effects. The total effect (combining direct and spillover components) is the policy-relevant estimand for evaluating market-level interventions. Which total effect is relevant depends on the exposure mapping and the decision context (per unit, per cluster, or per market). Specify this clearly when reporting results.

Future research should address several open questions. How can we estimate spillover effects when the network or geographic structure is unknown or measured with error? How can we distinguish spillovers from

correlated shocks (for example, common demand shocks that affect connected units)? The factor-model and synthetic-control methods in Chapters 8 and 6 provide tools for separating common shocks from unit-specific variation. An open problem is how to combine these low-rank structures with exposure mappings so that we can distinguish genuine spillovers from correlated shocks in high-dimensional marketing panels.

Large networks raise computational constraints because exposure computation can be expensive. Practical work often relies on sparse representations, neighbourhood sampling, or coarse exposure summaries. How can we incorporate spillovers into machine learning methods for heterogeneous treatment effects? These questions are central to advancing causal inference in interconnected markets.

## **Part VI**

# **Machine Learning and High-Dimensional Methods**



## Chapter 12

# Machine Learning for Nuisance and Heterogeneity

Machine learning enters this book as a tool for estimating nuisance functions in panels while preserving the design-based identification strategies from earlier chapters. We use flexible algorithms to learn objects such as propensity scores and outcome regressions, but we do not relax the underlying assumptions that justify DiD, event-study, or factor-based designs. Under the same identification assumptions as earlier chapters and under cluster-level regularity conditions that justify inference, this chapter shows how to combine flexible nuisance estimation with orthogonal scores to estimate average effects such as ATT, staggered-adoption effects  $\tau(g, t)$  and event-time responses  $\theta_k$ , and dose-response functions  $\mu(d)$ .

We formalise Neyman-orthogonal scores and cross-fitting schemes that split along the independent sampling units (typically units  $i$ , or higher-level clusters  $c$  when interference or shared shocks make units dependent), with explicit caveats for cross-sectional dependence and spillovers. Building on this foundation, we construct double/debiased machine-learning (DML) estimators [Chernozhukov et al., 2018] for average and dynamic effects, and for dose-response functions after explicitly defining the unit-level dose and outcome summaries (for example,  $D_i$  and  $Y_i(d)$ ) tied to a stated calendar window and estimand. We then turn to heterogeneous treatment effects and policy learning [Wager and Athey, 2018, Athey and Wager, 2021], emphasising that their standard identification arguments rely on unit-level unconfoundedness and overlap unless they are embedded inside a panel design such as staggered adoption with credible conditional parallel trends. Throughout, the contract is simple: we keep the design and change only the nuisance estimation.

The methods here build on the design foundations from our discussions of DiD (Chapter 4), event studies (Chapter 5), and factor models (Chapters 8 and 7). They also complement our treatment of continuous treatments (Chapter 14), high-dimensional controls (Chapter 13), and diagnostics (Chapters 17 and 16).

## 12.1 Motivation and Setup

We start from a causal estimand. For a binary treatment  $D_{it} \in \{0, 1\}$ , the average effect on treated unit-time cells is

$$\text{ATT} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1].$$

In staggered adoption and other dynamic designs, we also target cohort-time effects  $\tau(g, t)$ , event-time effects  $\theta_k$ , and dose-response functions  $\mu(d)$ , all defined in earlier chapters.

Modern marketing panels confront you with high-dimensional covariates, complex nonlinear relationships, and treatment effects that vary across units and over time. A retailer analysing a promotional campaign may observe hundreds of store characteristics and dozens of time-varying market conditions. Traditional parametric specifications impose strong functional-form restrictions that can miss important interactions and nonlinearities. Machine learning offers flexible, data-driven algorithms that adapt to rich covariate spaces without prespecifying a narrow functional form.

The difficulty is to keep causal interpretations credible when you use ML methods. Causal identification in panels rests on the assumptions laid out in Chapter 2, and DML does not weaken those requirements. Off-the-shelf prediction algorithms are tuned for out-of-sample prediction accuracy, not for recovering causal parameters. If you simply regress outcomes on treatment and high-dimensional controls using a black-box ML method, the resulting treatment coefficient will generally suffer from regularisation and model-selection bias, and from failing to respect the panel's design and dependence structure.

Double/debiased machine learning (DML) addresses this tension by combining flexible ML estimation of nuisance functions with score-based estimators that target well-defined causal estimands. In the simplest unconfoundedness setup, the nuisance functions are the outcome regressions  $m_d(X_{it}, \alpha_i, \lambda_t) = \mathbb{E}[Y_{it} \mid D_{it} = d, X_{it}, \alpha_i, \lambda_t]$  and the propensity score  $e(X_{it}, \alpha_i, \lambda_t) = \mathbb{P}(D_{it} = 1 \mid X_{it}, \alpha_i, \lambda_t)$ . When we use covariates in these nuisance models, they must be fixed before treatment, or in staggered adoption, before adoption time  $G_i$ .

The two key ingredients are Neyman-orthogonal scores and sample splitting. Orthogonalisation constructs score functions whose first-order behaviour is insensitive to small errors in the estimated nuisance functions when scores are evaluated out-of-fold and regularity conditions hold at the cluster level. Sample splitting and cross-fitting train nuisance models on one subset of the data and evaluate them on disjoint folds, breaking the feedback loop between fitting and evaluation. In panels, we treat units (or higher-level groups) as the independent sampling clusters. Let  $G$  denote the number of such clusters in the asymptotics, and impose weak-dependence conditions that justify a cluster-level central limit theorem. Section 12.3 makes this precise.

DML therefore does not create identification where none exists. It builds on a credible panel design—difference-in-differences, event studies, or factor- and synthetic-control-based designs from earlier chapters—with the same unconfoundedness, parallel-trends, or factor-structure assumptions as those designs require. This allows you to condition on richer covariate sets and to model heterogeneous and dynamic responses, while retaining design-based interpretations of estimands like ATT and event-time effects  $\theta_k$ . As in Chapter 2 and Chapter 13, identification still hinges on conditioning on the right covariates. ML can help with function

approximation, but it does not protect you from bad controls or collider bias. In particular, covariates used in nuisance models must be fixed before treatment, or in staggered adoption, before adoption time  $G_i$ .

## Relation to Factor and Hybrid Methods

Factor models (Chapter 8) replace parallel-trends assumptions with a factor-stability restriction. DML can complement such designs by flexibly modelling residual outcome components or auxiliary nuisance functions after you specify the factor structure and run the diagnostics that support it. Hybrid methods such as augmented synthetic control and synthetic difference-in-differences combine weighting with regression adjustment. DML extends these hybrids by replacing parametric regressions with ML-based nuisance estimators, making it easier to incorporate rich covariate information while keeping orthogonalisation and cluster-respecting inference.

## Marketing Motivation

Consider again a retailer rolling out a loyalty programme in 200 stores over six quarters under staggered adoption, with around 50 pre-treatment covariates per store. The goal is to estimate the dynamic cohort-time profile  $\tau(g, t)$  and its event-time analogue  $\theta_k$  from Chapters 4 and 5, and to assess whether effects differ across store types.

A naïve two-way fixed-effects regression offers limited flexibility to capture rich interactions between store characteristics and time-varying market conditions. Synthetic control methods are less natural when many units are treated and adoption is staggered (see Chapter 7 for hybrid approaches). DML offers a middle ground. It uses ML to estimate high-dimensional nuisance functions, constructs Neyman-orthogonal scores that are insensitive to small nuisance-estimation errors, and aggregates cohort-time effects into event-time profiles consistent with the estimands introduced earlier in the book. The result can support a transparent design-based interpretation when the underlying design assumptions are credible and the nuisance training respects treatment timing.

## Chapter Roadmap

We begin by formalising Neyman-orthogonal scores in panel settings (Section 12.2) and cross-fitting under dependence across units and over time (Section 12.3). We then develop DML estimators for average effects, staggered-adoption designs, and dose-response functions (Sections 12.4 and 12.9), before turning to heterogeneous treatment effects (Section 12.5) and policy learning (Section 12.6). The chapter concludes with a detailed discussion of identification assumptions in these ML-enhanced designs (Section 12.7), tun-

ing and overlap management (Section 12.8), diagnostics (Section 12.10), inference and small-sample issues (Section 12.11), marketing applications (Section 12.12), and a workflow checklist (Section 12.13).

## 12.2 Orthogonalisation and Neyman Orthogonality

Orthogonalisation is the conceptual foundation of double machine learning. It gives score functions whose dependence on nuisance estimates is weak enough that small errors in those estimates do not distort the target treatment effect. This is what lets us combine flexible ML methods for nuisance functions with valid inference for parameters such as ATT.

We formalise this idea using Neyman orthogonality, following Chernozhukov et al. [2018]. For clarity we work first in cross-sectional notation and then explain how the objects map to panel settings. Let  $Z_{it} = (Y_{it}, D_{it}, X_{it})$  denote a generic unit-period observation, with outcome  $Y_{it}$ , treatment indicator  $D_{it} \in \{0, 1\}$ , and covariates  $X_{it}$ . In cross-sectional settings  $\{Z_{it}\}$  are i.i.d. In panels, we treat units (or higher-level groups) as the independent sampling clusters and allow dependence within cluster, such as serial correlation within unit. When we write expectations and  $L^2$  norms for functions of  $(Y_{it}, D_{it}, X_{it})$ , we mean the norm induced by the empirical distribution of unit-time cells within a cluster, averaged across clusters. A single observation corresponds to a unit-period triple  $(Y_{it}, D_{it}, X_{it})$ . Section 12.3 discusses the dependence structure that panels introduce and how it affects asymptotics.

**Definition 12.1 (Neyman Orthogonality)** Let  $\psi(Z_{it}, \beta, \eta)$  be a score function for a target parameter  $\beta \in \mathbb{R}$  with nuisance parameter  $\eta \in \mathcal{H}$ , where  $\mathcal{H}$  is a function space. The score is Neyman-orthogonal at  $(\beta_0, \eta_0)$  if:

- (i) the moment condition holds,  $\mathbb{E}[\psi(Z_{it}, \beta_0, \eta_0)] = 0$ .
- (ii) the Gateaux derivative of the expected score with respect to  $\eta$  vanishes at  $\eta_0$ . For any direction  $h$  in the tangent space,

$$\frac{d}{dt} \mathbb{E}[\psi(Z_{it}, \beta_0, \eta_0 + th)] \Big|_{t=0} = 0.$$

Orthogonality means that small perturbations in the nuisance functions do not affect the expected score to first order. When we combine this property with suitable convergence rates for nuisance estimators, the resulting treatment-effect estimator remains  $\sqrt{G}$ -consistent and asymptotically normal even if the nuisance functions themselves converge more slowly than  $G^{-1/2}$ . Throughout this chapter,  $G$  denotes the number of independent clusters used for inference, which may coincide with  $N$  when units are independent clusters. The effective sample size for inference is therefore  $G$ , not the number of unit-time cells  $NT$ .

We now state the rate conditions in cross-sectional notation and then restate how they translate when units are the independent clusters in panels. Throughout this subsection,  $\beta$  denotes a generic scalar target parameter, such as ATT or a dynamic analogue defined in Chapters 4 and 5, and  $\beta_0$  refers to its true value.

**Assumption 48 (Nuisance Rate Conditions)** Let  $\hat{\eta} = (\hat{m}_0, \hat{e})$  denote the estimated nuisance functions, where  $m_0(x) = \mathbb{E}[Y_{it} \mid D_{it} = 0, X_{it} = x]$  is the outcome regression under control and  $e(x) = \mathbb{P}(D_{it} = 1 \mid X_{it} = x)$  is the propensity score. The estimators satisfy:

- (i) (product rate)  $\|\hat{m}_0 - m_0\|_{L^2} \cdot \|\hat{e} - e\|_{L^2} = o_P(G^{-1/2})$ .
- (ii) (individual rates)  $\|\hat{m}_0 - m_0\|_{L^2} = o_P(G^{-1/4})$  and  $\|\hat{e} - e\|_{L^2} = o_P(G^{-1/4})$ .

(iii) (complexity condition) the function classes containing  $m_0$  and  $e$  are regular enough to permit the required uniform convergence.

The product rate ensures that the bias induced by nuisance-estimation error is  $o_P(G^{-1/2})$ , negligible relative to sampling variation. These conditions can hold for specific learners under additional structure, such as approximate sparsity for lasso or regularity conditions for certain forest and boosting procedures. See Chernozhukov et al. [2018] for representative sufficient conditions.

In panel applications we take  $G$  to be the number of independent clusters used for inference, so these rates are formulated in terms of the effective number of independent observations rather than the raw cell count  $NT$ . When both the number of clusters and the time dimension grow, valid inference requires additional conditions on serial and cross-sectional dependence. We return to these in Section 12.3 and Chapter 16.

Given these conditions, the orthogonality property implies that the impact of nuisance-estimation error on the target estimator is of second order: if  $\hat{\eta}$  converges to  $\eta_0$  at rate  $r_G$ , then the induced bias in the estimator of  $\beta$  is of order  $r_G^2$  rather than  $r_G$ .

We now construct a score that both enjoys Neyman orthogonality and is doubly robust for the ATT, using the familiar outcome-regression and propensity-score components from earlier chapters.

**Definition 12.2 (Doubly Robust Score for ATT)** Let  $\text{ATT} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1]$  denote the Average Treatment Effect on the Treated, as defined in Chapter 2. For this estimand, let  $p = \mathbb{P}(D = 1)$  denote the marginal treatment probability. Following Sant'Anna and Zhao [2020], the doubly robust score is

$$\psi^{\text{ATT}}(Z_{it}, \beta, \eta) = \frac{D_{it}}{p} (Y_{it} - m_0(X_{it}) - \beta) - \frac{e(X_{it})(1 - D_{it})}{p(1 - e(X_{it}))} (Y_{it} - m_0(X_{it})),$$

where  $\eta = (m_0, e)$  collects the nuisance functions. The overlap condition from Chapter 2 requires  $0 < e(X_{it}) < 1$  on the covariate support of the treated group. For stable finite-sample behaviour,  $e(X_{it})$  should be bounded away from 0 and 1 on that support. These conditions ensure that the weights in  $\psi^{\text{ATT}}$  are well-defined and prevent extreme reweighting. Under the unconfoundedness condition from Chapter 2,

$$m_0(x) = \mathbb{E}[Y_{it} \mid D_{it} = 0, X_{it} = x] = \mathbb{E}[Y_{it}(0) \mid X_{it} = x],$$

so the outcome regression identifies the counterfactual mean for treated units.

**Proposition 12.1 (Double Robustness)** Let  $\beta_0$  denote the true ATT. Under the unconfoundedness and overlap conditions stated in Chapter 2, the moment condition  $\mathbb{E}[\psi^{\text{ATT}}(Z_{it}, \beta_0, \eta)] = 0$  holds, and hence the estimator  $\hat{\beta}$  that solves the sample analogue is consistent for  $\beta_0$ , if either of the following holds:

- (i) the outcome regression is correctly specified:  $m_0(x) = \mathbb{E}[Y_{it}(0) \mid X_{it} = x]$ .
- (ii) the propensity score is correctly specified:  $e(x) = \mathbb{P}(D_{it} = 1 \mid X_{it} = x)$ .

When both nuisance functions are correctly specified, this score attains the semiparametric efficiency bound for the ATT in the corresponding model. In panel applications, we typically apply this score either to cross-sectional aggregates (for example, post-treatment averages) or to cohort-time or event-time cells, with aggregation schemes that match the  $\tau(g, t)$  and  $\theta_k$  estimands from Chapters 4 and 5.

**Proposition 12.2 (Neyman Orthogonality of DR Score)** *The doubly robust score  $\psi^{\text{ATT}}$  is Neyman-orthogonal with respect to  $\eta = (m_0, e)$ . For any perturbations  $h_m$  and  $h_e$ ,*

$$\frac{d}{dt} \mathbb{E}[\psi^{\text{ATT}}(Z_{it}, \beta_0, \eta_0 + t(h_m, 0))] \Big|_{t=0} = 0, \quad \frac{d}{dt} \mathbb{E}[\psi^{\text{ATT}}(Z_{it}, \beta_0, \eta_0 + t(0, h_e))] \Big|_{t=0} = 0.$$

Consequently, estimation errors in  $\hat{m}_0$  and  $\hat{e}$  contribute only second-order bias to  $\hat{\beta}$  under Assumption 48.

Proof sketch. The result follows by direct differentiation. Write the expected score as a functional of  $(m_0, e)$  and note that, at  $(m_0, e)$  equal to the true nuisance functions, the residuals  $Y_{it} - m_0(X_{it})$  have zero conditional mean given  $(D_{it}, X_{it})$ , while the weighting terms ensure  $\mathbb{E}[D_{it}/p - e(X_{it})(1 - D_{it})/(p(1 - e(X_{it}))) | X_{it}] = 0$ . Differentiating with respect to either  $m_0$  or  $e$  in any direction  $h$  and evaluating at the truth yields zero because the residualised treatment term has the required moment-orthogonality property. See Chernozhukov et al. [2018] and Sant'Anna and Zhao [2020] for full details.

The two components of the score play complementary roles familiar from earlier chapters. The first term uses the outcome regression to impute counterfactual outcomes for treated units. The second term reweights control observations so that their covariate distribution matches that of the treated group. When the outcome regression is slightly misspecified, the weighting term corrects the discrepancy. When the propensity score is slightly misspecified, the outcome regression still delivers the correct mean for treated units. The orthogonality property formalises this complementarity: perturbing one nuisance function does not change the expected score at the true parameter values because the other nuisance function absorbs the perturbation. Orthogonality protects the estimand from small, regularisation-induced errors in the nuisance functions, but it does not fix gross misspecification, violations of overlap, or design failures. You address those with the diagnostics and tuning choices in Sections 12.8 and 12.10.

### 12.3 Cross-Fitting under Panel Dependence

Cross-fitting (sample splitting) underpins DML by decoupling nuisance estimation from score evaluation. We train nuisance models on one subset of the data and evaluate the scores on a disjoint subset, so that the estimation error in the nuisance functions is approximately independent of the sampling noise in the scores. This prevents the bias that would arise from using the same observations both to fit and to evaluate the nuisance components, while still allowing flexible ML methods inside the training folds.

The key requirement is that the nuisance estimates used to construct the score for an observation  $Z_{it}$  are trained on data that exclude  $Z_{it}$  itself. In cross-sectional settings, random partitioning of observations into folds suffices. Panel data introduce two complications: observations for a given unit are dependent over time, and observations across units may be dependent within a period because of common shocks or spillovers. Cross-fitting must respect this dependence structure.

**Definition 12.3 (K-Fold Cross-Fitting for Panels)** A  $K$ -fold cross-fitting structure partitions the set of independent clusters  $\{1, \dots, G\}$  into  $K$  disjoint folds  $\mathcal{G}_1, \dots, \mathcal{G}_K$  with  $\bigcup_{k=1}^K \mathcal{G}_k = \{1, \dots, G\}$ . When units are the independent clusters,  $G = N$  and each cluster contains a single unit.

For each fold  $k$ :

- (i) (training set)  $\mathcal{G}_{-k} = \{1, \dots, G\} \setminus \mathcal{G}_k$  contains all clusters not in fold  $k$ .
- (ii) (nuisance estimation)  $\hat{\eta}^{(-k)}$  is estimated using observations  $\{Z_{it} : i \in \mathcal{C}_c, c \in \mathcal{G}_{-k}, t = 1, \dots, T\}$ .
- (iii) (score evaluation) for any cluster  $c \in \mathcal{G}_k$ , for all units  $i \in \mathcal{C}_c$  and all periods  $t$ , the score  $\psi(Z_{it}, \beta, \hat{\eta}^{(-k)})$  uses out-of-fold nuisance estimates.

The cross-fitted estimator  $\hat{\beta}$  solves

$$\frac{1}{NT} \sum_{k=1}^K \sum_{c \in \mathcal{G}_k} \sum_{i \in \mathcal{C}_c} \sum_{t=1}^T \psi(Z_{it}, \hat{\beta}, \hat{\eta}^{(-k)}) = 0,$$

averaging scores over all  $NT$  unit-period observations. For inference, however, we work with the equivalent cluster-level moment condition formed by summing scores within cluster and averaging over  $G$  independent clusters.

The fold construction must align with the clustering scheme used for inference. When you cluster by unit in variance estimation, cross-fitting must also hold out entire units rather than individual unit-time cells, so that training and evaluation are independent across clusters. The  $1/(NT)$  normalisation is conventional for presenting the estimating equation. For variance estimation we aggregate scores within cluster and normalise by the number of independent clusters, consistent with Chapter 16. In the default asymptotic regime with  $G \rightarrow \infty$  and  $T$  fixed (or bounded), the effective sample size for inference is  $G$ , not  $NT$ .

The cluster-based folds in Definition 12.3 ensure that all observations within a cluster are either in the training set or in the evaluation set, not both. This accommodates serial correlation within units and allows for common time shocks that affect all units symmetrically.

**Assumption 49 (Panel Cross-Fitting)** For panel data with units  $i = 1, \dots, N$  and periods  $t = 1, \dots, T$ :

- (i) (cluster-level folds) partition the independent clusters  $c = 1, \dots, G$  into  $K$  folds. For a cluster  $c$  in fold  $k$ , all observations  $(Y_{it}, D_{it}, X_{it})$  with  $i \in \mathcal{C}_c$  are held out for score evaluation, and nuisances are estimated using clusters in  $\mathcal{G}_{-k}$ . Unit-level folds are the special case with  $G = N$ .
- (ii) (weak dependence across folds) dependence between clusters in different folds is weak enough that a cluster-level central limit theorem applies to the cluster-level influence contributions. See Chapter 16 for the relevant CLT and dependence assumptions. Common time shocks are allowed and are typically absorbed by time effects  $\lambda_t$ , but they must be accommodated by the chosen clustering scheme when they induce strong within-period correlation. Strong spatial or network dependence across folds requires more conservative block constructions (see below and Chapter 11). If highly connected units are split across folds and treated as independent clusters, both DML bias corrections and cluster-robust standard errors will be misleading because the effective number of independent pieces of information is overstated.
- (iii) (no post-treatment training data for nuisance models meant to represent untreated relationships) exclude post-treatment observations from treated units whenever a nuisance model is intended to represent an untreated or pre-treatment data-generating relationship. The same timing restriction applies to feature construction for propensity or adoption models and to any nuisance trained on outcomes.

The weak-dependence requirement in (ii) permits common time shocks yet rules out strong idiosyncratic dependence between units assigned to different folds. If such dependence is present—for example, with tight spatial competition or network spillovers—simple random partitioning may fail, and folds must be constructed to keep strongly connected units in the same block.

Condition (iii) is the familiar no-leakage or *clean control* restriction. Continuing the loyalty-programme example from Section 12.1, suppose store  $i$  adopts the programme in quarter  $G_i = 3$ . The outcome regression  $m_0(x) = \mathbb{E}[Y | D = 0, X = x]$  is meant to capture counterfactual sales for treated stores had they not adopted. If we train  $\hat{m}_0$  using store  $i$ 's post-adoption sales  $(Y_{i3}, Y_{i4}, \dots)$ , the nuisance estimator learns a function that already embeds treatment effects and therefore cannot represent the untreated path. The remedy is to use only pre-treatment observations for the treated cohort under study, and to use observations from comparison units in periods where they are untreated (never-treated, or not-yet-treated relative to the cohort-time cell). This clean-control principle applies equally to dynamic ATT and event-time estimands. Any outcome used to train a nuisance model meant to represent the untreated process must correspond to a period in which the unit is untreated under the relevant potential-outcome path.

When the independent clusters are units, unit-level folds are the default for panels with many units and moderate time spans. Two alternatives sometimes prove useful.

*Time-based folds* partition periods rather than units. This approach suits settings with few units but many periods, such as aggregate time series with a single treated unit. In synthetic-control applications (Chapters 6 and 7), time-based splitting is natural because identification already hinges on pre-treatment fit and post-treatment extrapolation, so training on pre-periods and evaluating on post-periods respects the SC/SDID design. The limitation is that in staggered-adoption panels, naive time-based folds can mix pre- and post-treatment observations across cohorts in ways that complicate the clean-control restriction, so they must be constructed with care around adoption dates.

*Block folds* partition both units and time into rectangular blocks. This approach handles settings where dependence is localised in both dimensions, such as spatial panels with regional shocks or network interference. Block folds are more conservative than unit or time folds alone, reducing effective sample size but providing stronger protection against bias when dependence across units and over time is pronounced.

The choice among splitting strategies depends on the dominance of within-unit serial correlation versus cross-sectional dependence. When within-unit serial correlation dominates and cross-sectional dependence is modest, unit-level folds combined with cluster-robust variance estimation at the unit level typically suffice. When cross-sectional correlation within periods dominates, carefully designed time-based folds may be preferable. When both forms of dependence are material, block folds provide the most robust protection at some cost in statistical efficiency. As a rule of thumb, design folds to keep highly dependent units and periods in the same block used for both training and score evaluation, and then cluster standard errors at that block level.

## 12.4 Double/Debiased ML Estimators in Panels

This section develops double/debiased machine learning estimators for panel data, providing formal convergence results for ATT and extending to staggered adoption designs common in marketing. We work under the nuisance-rate and cross-fitting conditions from Sections 12.2 and 12.3, together with the usual overlap condition.

**Assumption 50 (Overlap)** The propensity score  $e(x) = \mathbb{P}(D_{it} = 1 | X_{it} = x)$  is bounded away from zero and one. There exists  $\epsilon > 0$  such that  $\epsilon < e(x) < 1 - \epsilon$  for all  $x$  in the support of  $X_{it}$ , as in the overlap assumption from Chapter 2.

Overlap ensures that for any covariate profile observed among treated units, there are comparable control units to support counterfactual construction. When overlap fails, weights become extreme and estimates unstable. The diagnostics and trimming strategies from Chapters 17 and 13 become essential.

For panel data we treat units as the independent sampling clusters, so the number of independent clusters is  $G = N$ . Let

$$\Psi_i = \sum_{t=1}^T \psi^{\text{ATT}}(Z_{it}, \beta_0, \eta_0)$$

denote the unit-level influence contribution, aggregating scores over time for unit  $i$ .

We work in the default regime with  $N \rightarrow \infty$  and  $T$  fixed (or bounded), treating units  $i$  as the independent sampling clusters. When both  $N$  and  $T$  grow, additional conditions on serial and cross-sectional dependence are required. See Sections 12.3 and 12.2 and Chapter 16.

**Theorem 12.1 (DML Asymptotic Theory)** *Here  $\beta_0$  denotes a panel ATT estimand, for example the average of  $Y_{it}(1) - Y_{it}(0)$  over unit-period cells in a specified post-treatment window (as defined in Chapter 2). Under Assumptions 48, 49, and 50, and assuming the unit-level influence contributions  $\Psi_i$  have finite second moment and satisfy a cluster-level central limit theorem under the dependence conditions in Chapter 16, the cross-fitted DML estimator  $\hat{\beta}$  for ATT satisfies:*

(i) (consistency)  $\hat{\beta} \xrightarrow{P} \beta_0$  as  $G \rightarrow \infty$ .

(ii) (asymptotic normality)

$$\sqrt{G}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, V),$$

where  $V = \mathbb{E}[\Psi_i^2]$  is the variance of the unit-level influence function. With  $\Psi_i$  defined as the within-unit sum of orthogonal scores, the estimator admits the linear expansion

$$\sqrt{G}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{G}} \sum_{i=1}^G \Psi_i + o_p(1).$$

Consequently  $V = \mathbb{E}[\Psi_i^2]$  is the asymptotic variance of the influence-function representation.

(iii) (nuisance-induced remainder) under the product-rate condition in Assumption 48, the nuisance-induced remainder term in the influence-function expansion is

$$O_p(\|\hat{m}_0 - m_0\|_{L^2} \|\hat{e} - e\|_{L^2}) = o_p(G^{-1/2}).$$

The  $\sqrt{G}$  rate reflects that inference is over  $G$  independent units (clusters), not over all  $NT$  unit-period observations. Within-unit dependence across time periods affects  $V$  but not the convergence rate.

#### Warning: Rate-Robustness vs Model-Robustness

Practitioners often conflate two different notions of robustness. The first is *rate robustness*: standard DML estimators remain  $\sqrt{G}$ -consistent and asymptotically normal even when nuisance estimators converge slowly, for example at order  $G^{-1/4}$ , provided both nuisance models are consistent. This is the central promise of the orthogonal-score plus cross-fitting construction.

The second is *model double robustness*, familiar from augmented inverse probability weighting, where the estimator is consistent if either the propensity score or the outcome regression is correctly specified, even if the other is structurally misspecified [Bang and Robins, 2005]. Generic regression-based DML estimators of the Robinson type typically deliver the first property but not the second unless they are explicitly constructed from efficient influence functions that retain the AIPW structure. When the treatment model is badly misspecified, the residualised treatment term  $D_{it} - \hat{e}(X_{it})$  no longer delivers the intended orthogonality in the estimating equation, and no amount of accuracy in the outcome regression can fully repair the resulting bias.

Recent work develops augmented DML procedures that adaptively downweight unreliable treatment models and move closer to AIPW-style behaviour when the estimated propensity score appears misaligned with the data. In continuous-treatment settings this perspective connects directly to efficient-influence-function constructions [Kennedy et al., 2017]. In binary-treatment settings, the substantive point is simpler: if you want model double robustness, use an estimator that is explicitly built from an efficient influence function.

For applied work the message is simple. Do not rely on the word “double” in an estimator’s name to guarantee protection against arbitrary misspecification. Diagnose the treatment model  $\hat{e}(X_{it})$  directly, for example by checking whether residuals  $D_{it} - \hat{e}(X_{it})$  are approximately orthogonal to covariates (see the balance and residual diagnostics in Chapter 17). Where feasible, compare Robinson-style DML estimates with estimators that have genuine model double robustness. When estimates diverge, the treatment model is usually the place to look first.

Inference uses the influence-function representation. We estimate  $V$  by the empirical second moment of cluster-level score sums and apply the clustering guidance from Chapter 16. We cluster by unit by default, and use more conservative schemes when shared shocks induce additional dependence.

**Proposition 12.3 (Cluster-Robust Variance Estimator)** *Under panel dependence with clustering by unit, a consistent estimator of the asymptotic variance  $V$  is*

$$\hat{V} = \frac{1}{G} \sum_{i=1}^G \left( \sum_{t=1}^T \psi^{\text{ATT}}(Z_{it}, \hat{\beta}, \hat{\eta}^{(-k(i))}) \right)^2,$$

where  $k(i)$  denotes the fold containing unit  $i$ . Then

$$\frac{\sqrt{G}(\hat{\beta} - \beta_0)}{\sqrt{\hat{V}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

## Direct Debiased ML and Bregman-Riesz Regression

Double machine learning is often introduced as a generic recipe in which we choose flexible learners for the regression function and the propensity score, plug them into an orthogonal score, and cross fit. Recent work by Kato [2025b] shows that there is more structure to exploit. Many familiar estimators, including inverse probability weighting, doubly robust scores, covariate balancing and targeted maximum likelihood, can be viewed through a single object, the Riesz representer  $a(D_{it}, X_{it})$ . In the ATT example this is the function  $a(D_{it}, X_{it})$  that re-weights residuals in Definition 12.2. Estimating  $a$  well is as central as estimating  $m_0$ , because  $a$  determines which regions of covariate space receive weight when we construct the score.

Bregman-Riesz regression treats the representer  $a$  as the solution to an optimisation problem built from a Bregman divergence. Squared-loss Bregman-Riesz regression reproduces the Riesz regression used in automatic debiased machine learning, which coincides with least-squares density-ratio estimation in simple settings. Alternative convex generators such as Kullback–Leibler losses recover entropy-balancing style estimators, where the dual problem matches covariate moments exactly. From this perspective inverse probability weighting, entropy balancing, stable balancing weights and related procedures are different ways of targeting the same orthogonal score with different choices of loss and model class for the representer  $a$ .

For applied work this unified view has two practical implications. First, it encourages you to think explicitly about the weighting scheme implicit in your DML implementation rather than treating the choice of loss and basis for  $a$  as a black box. Second, it suggests that you can borrow regularisation and diagnostics from the density-ratio literature when fitting Riesz regressions, for example by inspecting whether estimated weights concentrate on a small number of observations. In marketing panels where extreme weights correspond to a handful of unusual stores or customers, combining DML with stable Bregman-Riesz objectives delivers more stable estimates while preserving the orthogonality and efficiency properties emphasised in this chapter.

## Staggered Adoption and DML-DiD

Marketing interventions rarely occur simultaneously across all units. Loyalty programmes roll out region by region, advertising campaigns launch in waves, and pricing experiments stagger across stores. This staggered adoption creates multiple treatment cohorts, each defined by its adoption time  $G_i$ . The DML framework extends naturally to this setting by estimating cohort-time-specific effects and aggregating them.

We apply DML to cohort-time-specific DiD scores of the type used in Chapter 4: for each  $(g, t)$  we construct an orthogonal score for  $\tau(g, t)$  based on treated units with  $G_i = g$  and an appropriate comparison group  $\mathcal{C}$ ,

with nuisance functions (propensity scores and outcome regressions) estimated using cross-fitting within the union of cohort  $g$  and  $\mathcal{C}$ .

**Theorem 12.2 (DML-DiD with Staggered Adoption)** *Consider staggered adoption with cohorts  $g \in \mathcal{G}$  and comparison group  $\mathcal{C}$  (never-treated or not-yet-treated units), and define cohort-time effects*

$$\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g], \quad t \geq g,$$

where  $Y_{it}(g)$  and  $Y_{it}(\infty)$  denote potential outcomes as in Chapters 4 and 5. Suppose the conditional parallel-trends assumption from Chapter 4 holds, along with Assumptions 48–49 and overlap within each cohort-comparison cell. Let  $N_g$  denote the number of units in cohort  $g$ . The number of cohorts is assumed to be fixed.

- (i) For each  $(g, t)$  with  $t \geq g$ , the DML estimator  $\hat{\tau}(g, t)$  based on cohort-specific nuisance functions is  $\sqrt{N_g}$ -consistent and asymptotically normal under an asymptotic regime where  $\min_{g \in \mathcal{G}} N_g \rightarrow \infty$  and  $|\mathcal{G}|$  is fixed.
- (ii) Event-time aggregators of the form

$$\hat{\theta}_k = \sum_{g \in \mathcal{G}} \omega_g \hat{\tau}(g, g + k), \quad \omega_g \propto N_g,$$

are  $\sqrt{N}$ -consistent and asymptotically normal,

$$\sqrt{N}(\hat{\theta}_k - \theta_k) \xrightarrow{d} \mathcal{N}(0, V_k),$$

where  $N = \sum_{g \in \mathcal{G}} N_g$  and  $V_k$  accounts for dependence across cohorts that share comparison units.

See Chapter 16 for explicit cluster-robust variance formulas in this multi-cohort setting. As in Chapter 4, identification hinges on conditional parallel trends within each cohort-comparison cell. DML improves the flexibility of nuisance estimation but does not relax these design assumptions.

## DML in Panel Settings: The Role of Unobserved Heterogeneity

Applying DML to panel data requires care in handling unobserved unit heterogeneity  $\alpha_i$ . A naive approach might time-demean the outcome and covariates (as in linear fixed effects) and then apply DML. However, time-demeaning removes  $\alpha_i$  only under strict additivity and linearity. When the true relationship between covariates and outcomes is nonlinear, the within-transformation fails.

The Correlated Random Effects (CRE) approach, also known as the Mundlak device, provides a practical modelling choice. This adapts the Mundlak idea from Chapter 10 to the DML nuisance context: we treat  $(X_{it}, \bar{X}_i)$  as the feature set for flexible outcome models, rather than imposing a linear fixed-effects structure. For every time-varying covariate  $X_{it}$ , we augment the feature set with its unit-level time average  $\bar{X}_i = T^{-1} \sum_t X_{it}$ . The ML learner then estimates  $Y_{it} = f(X_{it}, \bar{X}_i) + \varepsilon_{it}$ , where the function  $f$  can be arbitrarily

nonlinear. Conditioning on  $\bar{X}_i$  can reduce sensitivity to time-invariant confounding correlated with covariates, but it does not by itself justify unconfoundedness or parallel trends.

This approach performs better in simulations with nonlinear confounding [Fuhr and Papies, 2024] and is a common robustness device in applied panel settings where unobserved heterogeneity is a concern.

## Semi-Supervised DML with Unlabelled Covariates

The Riesz regression viewpoint also clarifies how to exploit unlabelled covariates when treatments and outcomes are scarce. In many marketing databases we have detailed profiles for large numbers of users or stores but observe assignment and revenue outcomes only for a managed test sample. Under additional regularity and design conditions, using auxiliary observations of  $X$  alone can reduce the asymptotic variance of average treatment effect estimators relative to procedures that ignore them. This variance reduction requires an explicit model for the observation or labelling mechanism and a same-population assumption linking the  $X$ -only sample to the target population. Otherwise the target estimand changes or is not identified.

The key idea is to treat unlabelled covariates as additional information about the distribution of  $X$  that enters the estimation of the Riesz representer. In the one-sample or censoring design, a single dataset contains both fully observed units and units for which  $D$  and  $Y$  are missing. In the two-sample or case-control design, we observe one dataset with complete triples  $(X, D, Y)$  and a second dataset with  $X$  only. In both cases the efficient influence function still has the usual doubly robust form, but the optimal weighting function  $a$  (the Riesz representer) now depends on the joint law of  $X$ ,  $D$ , and the observation process. Identification of the same ATE or ATT estimands as in Chapter 2 requires that the observation process (labelled vs unlabelled units) is conditionally independent of potential outcomes, given covariates and treatment, as formalised in Chapter 16. Generalised Bregman-Riesz regression uses both labelled and unlabelled covariates when minimising its loss, so the estimated weights exploit all available information about covariate frequencies.

In standard applications you can treat the labelled sample as the relevant universe and apply standard panel DML. The semi-supervised perspective becomes valuable when you truly have a much larger pool of passive covariate data, for example platform level browsing histories or impression logs, but can only measure outcomes and treatment on a subset because of logging costs or privacy constraints. In those cases semi-supervised DML provides a principled way to translate abundant information about  $X$  into tighter confidence intervals for treatment effects rather than simply using the extra data for prediction.

*Positive-unlabelled average treatment effect (PUATE).* A related but distinct scenario arises when outcomes are observed widely but treatment status is only known for a subset of units. This is the positive-unlabelled ATE (PUATE) problem: we observe a confirmed treatment group and a large unknown group that mixes treated and untreated units. The resulting estimators combine observation models, propensity scores, and outcome regressions in a way that resembles doubly robust scores but with a modified structure, because the propensity for appearing in the labelled treatment group plays a central role. These designs require additional assumptions about the labelling mechanism and the share of treated units in the unlabelled pool. Without credible bounds or models for the labelling mechanism and the treated share in the unlabelled pool,

PUATE designs are under-identified, and any effect estimate is driven by untestable assumptions rather than panel variation.

This is a different identification problem from the panel designs emphasised in this chapter. Treat PUATE methods as specialised tools that require additional, design-specific assumptions about labelling.

## 12.5 Heterogeneous Treatment Effects

Heterogeneous treatment effects (HTE) describe how causal impacts vary across units with different pre-treatment characteristics. In marketing applications this heterogeneity often drives strategy: a loyalty programme might lift sales substantially in urban flagship stores but have modest effects in small rural outlets. The goal is to move beyond a single average effect and recover how treatment response varies across the kinds of units defined in Chapter 2 and Chapter 10.

We formalise heterogeneity through the Conditional Average Treatment Effect (CATE), specialising the general potential-outcomes framework from Chapter 2. In this chapter we treat CATE identification primarily under unit-level unconfoundedness and overlap, unless we explicitly embed the heterogeneity analysis inside a design such as staggered adoption with credible conditional parallel trends. In panel settings we usually define CATEs at the unit level, conditioning on time-invariant or pre-treatment features.

**Definition 12.4 (CATE)** Let  $X_i$  collect pre-treatment characteristics for unit  $i$ , such as store demographics and summaries of pre-intervention sales paths. Fix a scalar post-treatment estimand  $Y_i(1) - Y_i(0)$  consistent with the aggregation schemes used for ATT and dynamic ATT in Chapters 4 and 10 (for example, an average over a specified post-treatment horizon, or an average of event-time effects over a window). The Conditional Average Treatment Effect (CATE) for units with  $X_i = x$  is

$$\delta(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x],$$

so the CATE function  $\delta : \mathcal{X} \rightarrow \mathbb{R}$  maps pre-treatment covariate profiles to unit-level treatment effects.

In panels,  $X_i$  typically comprises time-invariant characteristics (store geography, format, target segment) and pre-treatment summaries of time-varying covariates (such as average and variance of baseline sales, baseline category mix). Time-varying covariates observed after treatment should not enter  $X_i$ . The unconfoundedness condition at the unit level— $D_i \perp (Y_i(1), Y_i(0)) \mid X_i$ —relies on  $X_i$  being fixed before treatment and on overlap in  $e(x) = \mathbb{P}(D_i = 1 \mid X_i = x)$ . In staggered adoption, you must define  $D_i$  (ever-treated, cohort membership, or a window-specific indicator) to match the estimand. Including variables affected by treatment in  $X_i$  (for example, post-programme category mix) violates this condition and undermines the CATE interpretation.

Causal forests estimate  $\delta(x)$  by adapting random forests to the causal setting [Wager and Athey, 2018]. The central device is honesty: the data used to grow the trees and choose splits are separated from the data used to estimate treatment effects in the terminal leaves. In panel applications we treat each unit as the basic observational unit for the forest. By aggregating to one observation per unit, we recover a cross-sectional setting in which the unit-level data can plausibly be treated as independent clusters. Each unit contributes one observation  $(X_i, D_i, Y_i)$ , with serial dependence over  $t$  already absorbed into the construction of  $D_i$  and  $Y_i$ . Here  $D_i$  is a unit-level treatment indicator summarising the panel treatment path  $D_{it}$ , and  $Y_i(\cdot)$  aggregates the corresponding potential outcome path as defined in Chapter 10.

**Assumption 51 (Honest Splitting)** The causal-forest estimator is built from an honest sample split at the unit level:

- (i) (sample splitting) the set of units is partitioned into a tree-building sample  $\mathcal{I}_{\text{tree}}$  and an estimation sample  $\mathcal{I}_{\text{est}}$  with  $\mathcal{I}_{\text{tree}} \cap \mathcal{I}_{\text{est}} = \emptyset$ .
- (ii) (honest trees) tree structure (splits) is determined using  $\mathcal{I}_{\text{tree}}$ , and leaf-level treatment effects use only units in  $\mathcal{I}_{\text{est}}$ .
- (iii) (regularity) each leaf contains at least  $n$  treated and  $n$  control units, with  $n \rightarrow \infty$  as the number of independent units  $G$  grows.

Honesty plays a role analogous to cross-fitting in DML: tree-building and effect estimation use disjoint samples, so that the leaf-level CATE estimates behave as if they were evaluated using out-of-sample predictions.

**Theorem 12.3 (Consistency of Causal-Forest CATE)** *Suppose the unconfoundedness and overlap conditions from Chapter 2 hold at the unit level, so that  $D_i$  is independent of  $(Y_i(1), Y_i(0))$  conditional on  $X_i$ , and  $0 < e(x) < 1$  for all  $x$  in the support of  $X$ . Under Assumption 51 and standard regularity conditions on the forest (tree depth, subsampling rate, and smoothness of  $\delta(\cdot)$ ), the causal-forest estimator  $\hat{\delta}(x)$  satisfies:*

- (i) (pointwise consistency)  $\hat{\delta}(x) \xrightarrow{P} \delta(x)$  for each interior point  $x$  of  $\mathcal{X}$ .
- (ii) (asymptotic normality) for interior points  $x$ ,

$$\frac{\hat{\delta}(x) - \delta(x)}{\hat{\sigma}(x)} \xrightarrow{d} \mathcal{N}(0, 1),$$

where  $\hat{\sigma}(x)$  denotes a forest-based standard-error estimate under the regularity conditions in [Wager and Athey, 2018]. In practice, treat these standard errors as approximate and validate stability across folds and tuning.

The convergence rate depends on the dimension of  $X$  and the smoothness of  $\delta(\cdot)$ . In favourable, low-dimensional cases with smooth effects, rates can approach  $N^{-1/2}$ . In high-dimensional settings they are typically slower.

For strategic decisions a continuous CATE surface is often less useful than a small number of interpretable segments. Marketing teams want to know which store or customer types respond best, rather than inspecting a high-dimensional function. We therefore aggregate CATEs into grouped effects.

**Definition 12.5 (Grouped Average Treatment Effect)** Let  $\{S_1, \dots, S_J\}$  be a partition of the covariate space. The grouped average treatment effect (GATE) for subgroup  $j$  is

$$\delta_j = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i \in S_j].$$

When the CATE function  $\delta(x)$  is well defined, this equals  $\mathbb{E}[\delta(X_i) \mid X_i \in S_j]$ . A plug-in estimator averages the estimated CATEs within each subgroup,

$$\hat{\delta}_j = \frac{1}{|\{i : X_i \in S_j\}|} \sum_{i:X_i \in S_j} \hat{\delta}(X_i).$$

To preserve identification, the partition  $\{S_j\}$  should depend only on pre-treatment features  $X_i$ , not on post-treatment outcomes or estimated treatment effects. Otherwise, selection into groups becomes outcome-driven, and  $\delta_j$  no longer has a causal interpretation.

### 12.5.1 Panel Clustering Estimators (PaCE)

While causal forests estimate a smooth CATE surface, many marketing decisions are implemented through discrete segments. A retailer may want to label stores as “high-response”, “medium-response”, or “low-response” and design separate policies for each. Levi et al. [2024] introduce a Panel Clustering Estimator (PaCE) that directly partitions units into clusters with distinct treatment effects, using methods from earlier chapters as building blocks.

To keep the cluster definitions exogenous, the splitting or partitioning step must be learned on a separate sample or with cross-fitting at the unit level. Otherwise the resulting “high-response” clusters are mechanically selected on estimated outcomes.

PaCE combines recursive partitioning with panel-appropriate effect estimation. In the first stage, a regression tree greedily and exclusively splits units on pre-treatment covariates to reduce within-cluster variation in treatment effects. The splitting criterion is tailored to panels: it is defined in terms of estimated unit-level effects or panel-based loss functions computed out-of-fold, so that the criterion does not reuse the same outcomes used for within-leaf effect estimation. In the second stage, the average treatment effect within each leaf is estimated using the panel estimators developed earlier in the book, such as matrix-completion or synthetic-control methods from Chapters 6–8, rather than simple post-minus-pre means. This two-stage approach respects the missing-counterfactual structure of panels. Because the second stage uses the same DiD, SC, or matrix-completion estimators as earlier chapters, all identification assumptions and diagnostics from those methods apply leaf by leaf.

The method is particularly attractive when the objective is to recover a small number of interpretable, panel-consistent segments for targeted interventions. A causal forest might suggest that treatment effects vary smoothly with income, age, and baseline sales. PaCE distils that variation into a few operational clusters that marketing teams can use for rollout and budget allocation.

Estimating heterogeneous effects in panels introduces additional complications beyond the cross-sectional setting. When units are observed repeatedly, the effective information for  $\delta(x)$  depends on how much each additional period contributes. If treatment effects are approximately constant over time, pooling periods to form a scalar summary  $Y_i(1) - Y_i(0)$  reduces variance. If effects evolve dynamically, pooling may obscure important event-time patterns. In that case it is preferable to estimate dynamic profiles first (as in Chapter 10) and then study heterogeneity in those profiles.

The correlated-random-effects (CRE) approach from Section 12.4 extends naturally to HTE and should be treated as the default when unobserved heterogeneity is a concern. For time-varying covariates  $X_{it}$  we augment the feature set with unit-level averages  $\bar{X}_i = T^{-1} \sum_t X_{it}$ , so that the HTE learner conditions on both  $X_{it}$  and  $\bar{X}_i$ . Causal forests, PaCE, and related methods can all incorporate CRE-style features.

This helps ensure that estimated heterogeneity reflects genuine differences in treatment response rather than confounding from omitted unit-level factors that are correlated with both covariates and treatment. When you first estimate dynamic treatment profiles and then study heterogeneity in those profiles, use sample splitting across units where feasible so that the same outcomes are not used both to estimate effects and to learn segmentation rules. Otherwise, apparent heterogeneity may partly reflect overfitting rather than genuine differences in response.

## 12.6 Policy Learning

Estimating heterogeneous treatment effects is only a means to an end. The ultimate goal is to use those estimates to make better decisions: which customers should receive a promotion, which stores should adopt a loyalty programme, which users should see an advertisement. Policy learning formalises this decision problem and provides methods to learn targeting rules from data.

Consider a retailer deciding which customers to target with a discount coupon. The coupon costs \$5 to send and administer. If a customer's expected incremental purchase from receiving the coupon exceeds \$5 over a specified post-treatment horizon, targeting that customer is profitable. Policy learning translates the CATE estimates from Section 12.5 into such targeting rules.

**Definition 12.6 (Policy and Value Function)** A treatment policy is a mapping  $\pi : \mathcal{X} \rightarrow \{0, 1\}$  that assigns treatment based on pre-treatment covariates  $X_i$ . Let  $Y_i(1)$  and  $Y_i(0)$  denote the unit-level potential outcomes for a fixed post-treatment estimand constructed from the panel outcome paths, such as a long-run ATT or an event-time average defined in Chapter 10. Concretely,  $Y_i(d)$  aggregates the path  $\{Y_{it}(d)\}_t$  over a specified post-treatment window using the same averaging or event-time schemes as in Chapters 4 and 10. The value of policy  $\pi$  is the expected outcome under that policy,

$$V(\pi) = \mathbb{E}[Y_i(\pi(X_i))].$$

Expanding in terms of potential outcomes and defining the unit-level conditional effect function

$$\delta(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x],$$

this can be written as:

$$V(\pi) = \mathbb{E}[\delta(X_i) \pi(X_i)] + \mathbb{E}[Y_i(0)].$$

The welfare gain relative to treating no one is

$$W(\pi) = V(\pi) - V(0) = \mathbb{E}[\delta(X_i) \pi(X_i)].$$

If treatment has a known per-unit cost  $\kappa$  expressed in the same units as  $Y_i$  (for example, expected margin), the net welfare is

$$W_\kappa(\pi) = \mathbb{E}[(\delta(X_i) - \kappa) \pi(X_i)].$$

To interpret  $V(\pi)$  as written, we need the same design discipline as elsewhere in the book. Identification of  $\delta(x)$  and  $V(\pi)$  typically relies on unit-level unconfoundedness and overlap:  $D_i \perp (Y_i(1), Y_i(0)) \mid X_i$  and  $0 < e(x) < 1$  on the covariate support of the treated, with  $X_i$  fixed before treatment. We also need no interference at the unit level (SUTVA). When spillovers are material, you must redefine the policy value using exposure mappings (Chapter 11). Otherwise  $V(\pi)$  is not the relevant causal object.

**Proposition 12.4 (Optimal Targeting Rule)** *Under the assumptions from Chapter 2 that identify  $\delta(x)$  (unconfoundedness and overlap at the unit level), and among policies that treat units independently and face a*

constant per-unit cost  $\kappa$  with no budget or capacity constraints, the welfare-maximising policy is the threshold rule

$$\pi^*(x) = \mathbf{1}\{\delta(x) > \kappa\}.$$

A plug-in policy based on estimated CATEs takes the form

$$\hat{\pi}(x) = \mathbf{1}\{\hat{\delta}(x) > \kappa\}.$$

The rule has a simple interpretation: treat unit  $i$  if and only if the expected benefit  $\delta(X_i)$  exceeds the cost  $\kappa$ . In the coupon example, send the coupon to customers whose expected incremental spend over the evaluation horizon exceeds \$5.

Regret measures how much welfare the estimated policy sacrifices relative to the oracle policy that knows  $\delta(\cdot)$ , that is,  $W_\kappa(\pi^*) - W_\kappa(\hat{\pi})$ . For a fixed policy class and under unconfoundedness, regret typically converges to zero when  $\hat{\delta}(x)$  converges uniformly to  $\delta(x)$  and the distribution of  $\delta(X_i)$  does not place too much mass near the threshold  $\kappa$  (a margin condition). Mere pointwise consistency of  $\hat{\delta}(x)$  is not enough if many units have effects close to  $\kappa$ . Small errors can then flip many treatment decisions. Formal regret bounds (see the policy-learning literature cited in Section 12.5) typically require uniform rates for  $\hat{\delta}(x)$  over the policy class and a margin condition that limits the mass of units with  $\delta(X_i)$  close to  $\kappa$ . To avoid optimistic regret assessments driven by overfitting, you should estimate  $\hat{\delta}(x)$  and learn  $\hat{\pi}$  on one set of units and evaluate  $\hat{W}_\kappa(\hat{\pi})$  on a disjoint evaluation set, with units treated as independent clusters.

Before deploying a learned policy, we need to estimate its value. When unconfoundedness and overlap hold at the unit level, doubly robust estimators combine outcome regressions and propensity scores to evaluate a given policy.

**Proposition 12.5 (Doubly Robust Policy Value Estimator)** *Let  $m_1(x) = \mathbb{E}[Y_i | D_i = 1, X_i = x]$  and  $m_0(x) = \mathbb{E}[Y_i | D_i = 0, X_i = x]$  denote the outcome regressions under treatment and control, and let  $e(x) = \mathbb{P}(D_i = 1 | X_i = x)$  be the propensity score, as in Chapter 2. For a fixed policy  $\pi$  and  $G$  units (stores, customers, or markets) treated as the independent sampling clusters, a doubly robust estimator of  $V(\pi)$  is*

$$\hat{V}(\pi) = \frac{1}{G} \sum_{i=1}^G \left[ \frac{\pi(X_i) D_i (Y_i - \hat{m}_1(X_i))}{\hat{e}(X_i)} - \frac{(1 - \pi(X_i))(1 - D_i)(Y_i - \hat{m}_0(X_i))}{1 - \hat{e}(X_i)} + \pi(X_i) \hat{m}_1(X_i) + (1 - \pi(X_i)) \hat{m}_0(X_i) \right].$$

The welfare gain  $W(\pi) = V(\pi) - V(0)$  is estimated by  $\hat{W}(\pi) = \hat{V}(\pi) - \hat{V}(0)$ , where  $\hat{V}(0)$  sets  $\pi \equiv 0$ . Under the unconfoundedness and overlap conditions from Chapter 2, together with the nuisance-rate and cross-fitting assumptions from Sections 12.2 and 12.3,  $\hat{V}(\pi)$  is consistent and asymptotically normal when  $\pi$  is fixed. When  $\pi$  is learned from the data, valid evaluation typically requires an additional holdout sample or nested cross-fitting so that the same outcomes are not used both to choose  $\hat{\pi}$  and to evaluate  $\hat{V}(\hat{\pi})$ . As in Theorem 12.1, cluster-robust variance estimation must use the same unit-level clustering that defines the influence contributions.

The structure of  $\hat{V}(\pi)$  mirrors the ATT score from Section 12.2. The first two terms reweight residuals to correct for imbalance between the policy  $\pi$  and the observed assignment mechanism  $D_i$ , while the last two

terms use the outcome regressions for prediction. If either (a) both outcome regressions  $m_0, m_1$  are correctly specified or (b) the propensity score  $e$  is correctly specified, then  $\hat{V}(\pi)$  is consistent for  $V(\pi)$ . When combined with cross-fitting of the nuisance functions at the unit level, this score is Neyman-orthogonal with respect to  $(m_0, m_1, e)$ , so small regularisation errors in the fitted nuisances affect  $\hat{V}(\pi)$  only through second-order bias.

When  $\pi$  is learned from the same data (for example,  $\pi = \hat{\pi}$  based on  $\hat{\delta}$ ), additional care is needed. Overfitting in the policy-learning step can invalidate naive inference for  $\hat{V}(\hat{\pi})$ , even when  $\hat{V}(\pi)$  behaves well for fixed  $\pi$ . For learned policies  $\hat{\pi}$ , a standard approach is two-way sample splitting: partition units into folds, use one subset to estimate  $(\hat{m}_0, \hat{m}_1, \hat{e})$  and learn  $\hat{\pi}$ , and use a disjoint subset to compute  $\hat{V}(\hat{\pi})$  with out-of-fold nuisance predictions, clustering by unit for inference.

*Practical considerations.* In marketing applications the cost  $\kappa$  often includes more than direct expenses. A coupon's cost includes printing, distribution, and the margin erosion from customers who would have purchased anyway. The threshold rule  $\pi^*(x) = \mathbf{1}\{\delta(x) > \kappa\}$  treats these costs as known and constant. When costs vary across units or are uncertain, more general formulations, such as constrained or robust policy learning, replace the simple threshold with rules that respect budget or capacity limits. Examples include targeting only the top  $q$  per cent of customers by  $\hat{\delta}(x)$  under a budget constraint, or sequential allocation rules when each store can handle only a fixed flow of new loyalty members per period. When budget or capacity constraints induce interactions across units (for example, a finite call-centre or limited in-store staff), the simple value function  $V(\pi)$  no longer captures the full welfare problem. In those cases, you must either model the constraint explicitly or interpret  $V(\pi)$  as an approximation valid only in marginal, small-scale rollout scenarios. When the number of independent clusters is small, treat regret and value estimates as descriptive and prioritise re-randomisation or fresh experiments for validation. Before treating  $\hat{\pi}$  as a deployable targeting strategy, revisit the threats in Chapters 15 and 11. Policy learning amplifies any residual bias from unobserved confounding, bad controls, or interference, because it concentrates treatment where estimated effects are largest.

## 12.7 Assumptions

Double machine learning rests on two layers of assumptions. The first layer comprises identification assumptions that justify a causal interpretation of the target estimands, such as ATT or CATE. These originate from the research design and from the broader panel framework developed in earlier chapters. The second layer comprises regularity conditions that ensure valid estimation and inference when we estimate nuisance functions with flexible ML methods. This section consolidates the key assumptions used in this chapter, clarifies their roles, and links them to diagnostics.

### Identification Assumptions

The following assumptions are substantive causal requirements. They must be defended using research design, institutional knowledge, and diagnostics from the threats and diagnostics chapters. DML does not weaken these assumptions. It provides a flexible estimation framework that is compatible with them.

**Assumption 52 (Unconfoundedness)** Treatment assignment is independent of potential outcomes conditional on observed covariates, as in Chapter 2:

$$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i.$$

Here  $D_i$  is a unit-level treatment indicator summarising the treatment path  $\{D_{it}\}_t$ , and  $Y_i(1), Y_i(0)$  denote unit-level aggregates of the potential-outcome paths  $Y_{it}(1), Y_{it}(0)$ —for example, a long-run ATT or an event-time average defined in Chapter 10. The construction of  $D_i$  and  $Y_i(d)$  depends on the estimand and design (for example, post-period averages under DiD or event-time averages under staggered adoption) as in Chapters 4 and 10. Unconfoundedness must hold for that particular aggregation. For example, in staggered adoption you might define  $D_i = \mathbf{1}\{G_i < \infty\}$  and  $Y_i(d)$  as the average of  $Y_{it}(d)$  over a fixed post-adoption window. Other choices define different estimands and require different unconfoundedness statements.

In panel settings with correlated unit effects, a common refinement conditions also on unit-level heterogeneity:

$$D_{it} \perp\!\!\!\perp (Y_{it}(1), Y_{it}(0)) \mid X_{it}, \alpha_i,$$

where CRE features such as  $\bar{X}_i$  are used as a robustness device to reduce sensitivity to time-invariant confounding (Section 12.4). This does not by itself justify unconfoundedness when time-varying unobservables remain.

Unconfoundedness is the central selection-on-observables assumption. It asserts that, after conditioning on observed covariates (and CRE-style proxies for unit effects), treatment assignment is as good as random. Violations arise when unobserved factors drive both treatment and outcomes. In marketing, stores that adopt a loyalty programme may differ from non-adopters in management quality, local competitive dynamics, or algorithmic treatment rules that are not fully captured in  $X_{it}$  or  $\bar{X}_i$ . CRE reduces bias from time-invariant

heterogeneity but cannot eliminate confounding from time-varying unobservables. You should treat CRE as a robustness tool, not as evidence that selection-on-observables holds.

Overlap (Assumption 50) and limited interference (Assumption 54 below) supplement unconfoundedness for designs identified by selection on observables. In DML-DiD and factor-based settings, these assumptions operate alongside the conditional parallel-trends and factor-structure conditions introduced in Chapters 4, 5, and 8. Those design-specific assumptions are not repeated here.

## Regularity Conditions

The second layer of assumptions concerns the estimation procedure. These are technical conditions that ensure DML estimators achieve  $\sqrt{G}$ -consistent, asymptotically normal behaviour when nuisance functions are estimated flexibly.

Neyman orthogonality (Definition 12.1) requires that the expected score be first-order insensitive to perturbations in the nuisance functions at the truth. The doubly robust scores used in this chapter satisfy this property by construction. Without orthogonality—for example, under naïve regression adjustment without a properly constructed bias-correction term—regularisation bias from ML methods would feed directly into the treatment-effect estimator.

The nuisance-rate conditions (Assumption 48) require that nuisance estimators converge fast enough that their product error is  $o(G^{-1/2})$ , where  $G$  is the number of independent clusters used for inference. In some structured settings, modern ML methods such as random forests, gradient boosting, and neural networks can achieve the necessary  $G^{-1/4}$ -type rates, but this is not automatic and depends on smoothness or sparsity and on careful regularisation. In practice we cannot verify these rates directly. Instead we use diagnostics such as out-of-fold prediction error, calibration checks for propensity scores, and stability of DML estimates across reasonable learner choices and hyperparameters (see Section 12.8).

Panel cross-fitting (Assumption 49) requires that sample splitting respects the dependence structure: folds partition units, not individual observations, and the “clean control” restriction excludes post-treatment observations from treated units when estimating counterfactual outcome models. These design choices ensure that nuisance-estimation error behaves approximately independently across folds, supporting the orthogonality arguments developed earlier in the chapter. Combined with unit-level cluster-robust variance estimators from Chapter 16, these choices deliver the asymptotic guarantees in Theorem 12.1.

## Stability and Interference

Two further assumptions are often implicit but deserve explicit discussion because they connect DML to the broader threats-to-validity story.

**Assumption 53 (Stability)** The conditional expectation of untreated outcomes is stable over the sample period:

$$\mathbb{E}[Y_{it}(0) \mid X_{it}, \alpha_i]$$

is invariant across the pre- and post-treatment calendar periods used for identification, in the sense that the same conditional-mean function applies. This is a functional-form stability requirement for the nuisance model, not a causal identifying assumption by itself. Stability should be assessed on the same time windows and covariate sets that feed the DML nuisance models. Stability in a broader sample does not guarantee stability on the specific subsample used for identification.

Stability is a form of temporal external validity for the nuisance functions. It fails when regime changes, seasonality shifts, platform policy updates, or competitor entries alter the relationship between covariates and outcomes. In marketing, a pandemic or major algorithm change on a platform may invalidate outcome regressions estimated on pre-crisis data. Diagnostics include stability checks across subperiods, comparison of nuisance model fit in early versus late periods, and the broader stress tests for non-stationarity in Chapters 15 and 17.

**Assumption 54 (Limited Interference)** Treatment effects satisfy a limited-interference condition. Either Rubin’s Stable Unit Treatment Value Assumption (SUTVA) holds for the unit definition used in the analysis, or you must redefine the estimand using an exposure mapping and derive a corresponding score under that interference model (Chapter 11).

Limited interference is often violated in marketing panels. Loyalty programmes in one store may cannibalise sales from nearby stores. Targeted online ads may spill over through word-of-mouth or competition. In such cases, DML estimators that assume SUTVA will generally conflate direct and spillover effects. Orthogonal scores and cross-fitting mitigate regularisation bias but do not repair misspecification from unmodelled interference. In high-interference environments, the correct object is rarely a simple unit-level ATT. DML must be combined with the exposure mappings and estimands from Chapter 11, or the resulting estimates should be interpreted as mixtures of direct and spillover effects. Diagnostics for spillovers—such as spatial or network placebo regressions and comparisons of estimates under alternative exposure mappings—belong to the toolbox developed in Chapter 11 and Chapter 15.

## Summary

The assumptions relevant for DML in panels fall into two broad categories.

Identification assumptions—unconfoundedness, overlap, stability, and limited interference, together with design-specific conditions such as conditional parallel trends and factor-structure assumptions from earlier chapters—are substantive causal requirements. They determine whether the target estimands (ATT, dynamic effects, CATEs, policy values) have a credible causal interpretation in the first place. These assumptions must be defended using institutional knowledge, balance diagnostics, placebo and pre-trend tests, and sensitivity analyses, as laid out in the validity and diagnostics chapters.

Regularity conditions—orthogonal scores, nuisance-rate conditions, dependence-respecting cross-fitting, and cluster-robust variance estimation—are technical requirements that ensure DML estimators behave well

once identification is secured. They allow you to use flexible ML methods for nuisance estimation without sacrificing  $\sqrt{G}$ -rate convergence or valid asymptotic inference, provided the nuisance models achieve adequate predictive performance and are trained with careful sample splitting.

DML therefore changes how you estimate and aggregate effects, not what you need to assume for those effects to be causal. Any design that is not credible without ML remains not credible with it. The substantive assumptions come from design and data. ML enters to make high-dimensional, nonlinear conditioning feasible while preserving the inferential guarantees of the underlying panel designs.

## 12.8 Tuning and Implementation

Implementing DML in marketing panels forces you to make concrete choices about overlap, hyperparameter tuning, learner selection, shape constraints, and leakage avoidance. These choices sit at the intersection of three aims: preserving the identification assumptions from earlier chapters, achieving good predictive performance for nuisance functions, and keeping computation tractable. In this section we focus on how to tune learners and design validation schemes so that prediction supports, rather than undermines, causal identification.

### Overlap and Trimming

In panel DML we typically work with unit-level propensity scores, such as the probability of ever receiving treatment or joining a particular rollout cohort, conditional on pre-treatment features  $X_i$  that summarise the history  $\{X_{it} : t < G_i\}$ . Assumption 50 requires these propensities to be bounded away from zero and one on the population where we define the estimand. As in the design and diagnostics chapters, you diagnose overlap by plotting the distribution of estimated propensities  $\hat{e}(X_i)$  for treated and comparison units and by reporting simple summaries, such as the proportion of units whose propensities lie in regions where both groups are represented.

Units with extreme propensities near zero or one are natural candidates for trimming. Removing a small fraction of the most extreme cases often improves precision and reduces the influence of a handful of poorly matched units. Trimming, however, changes the target estimand. If  $\mathcal{X}_{\text{overlap}}$  denotes the subset of covariate space that remains after trimming, then an average treatment effect on the treated becomes

$$\text{ATT}_{\text{overlap}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1, X_i \in \mathcal{X}_{\text{overlap}}],$$

for the unit-level aggregated outcome used in the analysis. You should report explicitly that your estimates target  $\text{ATT}_{\text{overlap}}$  rather than the original ATT. Trimming sets should depend only on pre-treatment features  $X_i$ , not on post-treatment outcomes or estimated effects, to preserve the interpretation of  $\text{ATT}_{\text{overlap}}$  as a causal effect for a covariate-defined subpopulation.

In staggered-adoption designs, overlap is cohort-specific. For each cohort  $g$ , you care about overlap between that cohort and its comparison group over the pre-periods that identify  $\tau(g, t)$  or  $\theta_k$ . When you trim, report how many treated and comparison units remain in each cohort, how the support of  $X_i$  has changed, and interpret estimates explicitly as effects on the trimmed population. Overlap tables that only report global counts can be misleading. Always present cohort-by-comparison overlap and trimming statistics, and report the implied trimmed population by cohort since cohort-specific trimming changes the aggregation weights for  $\theta_k$ .

## Hyperparameter Tuning

Hyperparameters control the complexity of the ML learners that estimate nuisance functions. For regularised GLMs, the penalty parameter  $\lambda$  governs the trade-off between fit and sparsity. For random forests, key levers include the number of trees, maximum depth, minimum leaf size, and the fraction of covariates considered at each split. For gradient boosting, learning rate, number of iterations, tree depth, and subsampling fractions play similar roles.

Tuning has to respect the panel cross-fitting structure from Section 12.3 and the no-leakage conditions in Assumption 49. Within each training fold, you partition the training units into sub-folds, train learners across a hyperparameter grid, and evaluate out-of-sample fit on held-out sub-folds using prediction error metrics appropriate for each nuisance function, such as mean squared error for outcome regressions and log-likelihood or Brier score for propensity scores. Sub-folds should respect the same unit-level clustering used for inference: split on units, not on individual unit-period observations, so that validation error reflects performance on independent clusters. Random search or Bayesian optimisation can be more efficient than exhaustive grid search when the hyperparameter space is large, but the essential point is that all tuning occurs within the same fold structure used for cross-fitting.

When treatment timing matters, you must restrict tuning for nuisance models of untreated potential outcomes to pre-treatment observations for treated units and to all periods for never-treated units. Validation losses computed on post-treatment periods for treated units can favour models that leak information about treatment effects into counterfactual predictions. In practice, this means that hyperparameter search and selection for outcome regressions are based only on pre-treatment windows and control units.

For propensity scores, tuning should use only information available at the time of treatment assignment. In staggered-adoption designs, this typically means pre-treatment covariate histories and baseline summaries, excluding any variables that are measured after adoption or that blend pre- and post-treatment information.

## Stability Checks

Hyperparameter choice should not drive your substantive conclusions. After selecting a preferred configuration on predictive grounds, re-estimate treatment effects under a small set of alternative, reasonable tuning settings, such as varying  $\lambda$  in the lasso over an order of magnitude or changing the number of trees and minimum leaf size in a forest. Plot or tabulate how key estimands such as ATT, event-time effects  $\theta_k$ , or conditional effects  $\delta(x)$  move with these choices.

If estimates remain stable across tuning variations that all deliver good out-of-sample fit, that supports robustness. If estimates swing widely, your results are fragile to modelling choices. In that case you should report multiple specifications, explain which you view as most credible, and treat conclusions cautiously. When design-based estimators are available (for example, two-way fixed-effects DiD, Sun–Abraham, or SDID), treat them as benchmarks after aligning estimands and comparison groups. If estimates diverge, first check whether the methods target the same causal object, then investigate leakage and overlap.

## Prediction versus Identification

Breiman's essay on the two statistical cultures [Breiman, 2001], discussed in Chapter 1, highlights the gap between pure prediction and model-based explanation. In causal panel work this gap has a specific incarnation. Cross-validation rewards models that predict observed outcomes accurately, but accurate prediction does not guarantee credible counterfactuals. An outcome regression that predicts treated outcomes well by implicitly using post-treatment information, or by overfitting to idiosyncrasies of a few treated units, can still generate biased counterfactual paths.

For DML, tuning is therefore constrained by the design. Nuisance models for untreated potential outcomes should be trained only on pre-treatment data for treated units and on control units, using folds that respect the panel structure. Features that are themselves affected by treatment, or that blend pre- and post-treatment information, must be excluded from the covariate sets used in these nuisance models. Design-based estimators from earlier chapters—event studies, differences-in-differences, synthetic control—provide useful benchmarks. In this book's design-first perspective (Chapters 1 and 2), prediction quality is subordinate to identification: nuisance models are acceptable only if they respect the design's timing and covariate restrictions. When identification is strong, large and systematic discrepancies between ML-based counterfactuals and design-based estimates are a warning sign that prediction has won out over identification. A common failure mode is to feed lagged outcomes that include treated periods into the control regression, effectively letting the model 'peek' at post-treatment sales paths when constructing counterfactuals. This violates the no-leakage principle and typically biases effects toward zero.

## Learner Selection

DML treats nuisance functions as prediction problems, so almost any supervised learner can be used in principle. In practice, different learners are better suited to different nuisances. Regularised GLMs are stable and interpretable for propensity scores, especially when treatment is rare and you want to avoid spuriously extreme propensities in regions with thin support. Boosting and forests are often more effective for outcome regressions, where nonlinearities and interactions may be important. Neural networks offer flexibility but typically require large samples, careful tuning, and strong regularisation to avoid overfitting.

For marketing panels, a sensible default is to combine a relatively simple, well-regularised learner for the propensity score, such as logistic lasso or elastic net, with a more flexible learner for the outcome regression, such as gradient boosting or forests. Sensitivity analyses that swap in alternative learners—for example, comparing lasso versus forests for the outcome regression—help assess robustness to modelling choices. The key diagnostic remains the behaviour of the DML estimand and its confidence intervals, not marginal differences in nuisance predictive accuracy. A learner that reduces mean-squared error for  $Y$  but worsens balance or inflates extreme weights can degrade identification even as predictive metrics improve.

## Shape Constraints

Shape constraints encode prior knowledge about how outcomes or propensities should behave. Propensity scores must lie in  $(0, 1)$ , which you can enforce by using logistic links or by truncating extreme predictions to stay away from zero and one, in line with the overlap region. Outcome regressions may be constrained to be monotone over a narrowly justified local range, if domain knowledge supports it. In many marketing settings monotonicity is contestable due to substitution, saturation, and targeting. Conditional effects  $\delta(x)$  can be constrained in sign only when prior information genuinely rules out negative effects. Many marketing interventions plausibly have adverse effects for some units, so non-negativity constraints should be used sparingly and explicitly justified.

Imposing constraints reduces flexibility but can improve credibility when the constraints are well grounded. In practice, constrained boosting, isotonic regression, and projection methods onto constrained function classes provide ways to incorporate such structure into nuisance estimation or into post-processing of CATEs. Truncation choices should be reported alongside overlap diagnostics, and any monotonicity or sign constraints should be justified with substantive arguments (see Chapter 15). Otherwise, they risk baking optimistic priors into the estimator.

## Leakage Avoidance

Leakage arises when nuisance models are trained on data that already reflect treatment effects and are then used to impute counterfactuals. In panels this typically happens when outcome regressions for the control path use post-treatment observations from treated units, or when features include post-treatment variables that themselves carry treatment effects—for example, lagged outcomes that span treated periods or cumulative spend measures.

The cross-fitting protocol in Section 12.3 and Assumption 49 encode a no-leakage principle: folds are defined at the unit level, and nuisance models for untreated potential outcomes are trained using only pre-treatment periods for treated units and all periods for never-treated units. In implementation, you should verify that the training datasets for outcome regressions and propensity scores obey this restriction and that tuning procedures do not silently reintroduce post-treatment data. A simple implementation check is to tabulate, for each treated unit, the maximum calendar period included in the training data for the control regression. This maximum should be strictly less than the adoption period  $G_i$ . Comparing DML estimates based on correctly restricted nuisances with naïve estimates that ignore leakage is a useful diagnostic. Large discrepancies indicate that leakage was materially biasing counterfactuals.

## Block-Gap Cross-Validation for Time-Series ML

Standard cross-validation that randomly shuffles observations across time can badly underestimate prediction error in time-series and panel settings, because short-run dependence between nearby periods makes future observations look easier to predict than they really are. Block-gap, or *hv-block*, cross-validation addresses this by holding out contiguous blocks in time for validation and leaving a gap of several periods between those blocks and the data used for training. The gap weakens short-run serial correlation between the last training observations and the first validation observations, so that validation errors better reflect true out-of-sample performance.

In panels, block-gap ideas apply along both the time and unit dimensions. You can hold out future time windows for a subset of units while training on earlier periods, or hold out spatial or network neighbourhoods of units while training on more distant units. Gap-based cross-validation is designed to reduce optimistic validation under serial dependence. See the time-series cross-validation discussion in Chapter 16 (add citation). In this book we treat block-gap cross-validation as a sensible default when serial dependence is pronounced, provided the blocks and gaps respect treatment timing to avoid leakage.

When tuning nuisance learners for DML, tune within the same unit-level folds used for cross-fitting, block by unit or time, and, for treated units, ensure that both training and validation windows lie strictly in pre-treatment periods when you are fitting models for untreated potential outcomes. Leaving at least a modest temporal gap between training and validation windows whenever serial dependence is pronounced turns overly optimistic cross-validation scores into realistic estimates of how nuisance models will perform when you roll them forward on truly unseen data. Block-gap schemes that ignore treatment timing can inadvertently leak post-treatment information into validation folds. Always align gaps and validation windows with treatment adoption dates.

## 12.9 Dose-Response Extensions

Many marketing treatments are continuous rather than binary. Advertising spend, discount depth, price levels, and promotion duration all vary in intensity. The central question is not simply whether to advertise, but how much to spend and at what level diminishing returns set in. Building on Chapter 14, we show how DML delivers orthogonal, doubly robust estimators of dose-response functions in this setting.

Chapter 14 develops the potential-outcomes framework for continuous treatments, including identification via the generalised propensity score (GPS) and a range of estimation strategies. Here we focus on the contribution of DML: constructing orthogonal scores that combine flexible ML estimation of nuisance functions with doubly robust estimation of the average dose-response function and its marginal effects.

**Definition 12.7 (Dose-Response and Marginal Effect)** For a realised unit-level dose  $D_i \in \mathcal{D} \subseteq \mathbb{R}$  and scalar dose level  $d \in \mathcal{D}$ , the average dose-response function (ADRF) is

$$\mu(d) = \mathbb{E}[Y_i(d)],$$

where  $Y_i(d)$  denotes the potential outcome for unit  $i$  under dose  $d$ , constructed over a fixed post-treatment horizon as in Chapter 14 from the underlying path  $\{Y_{it}(d)\}_t$ . The aggregation window and weighting scheme for constructing  $Y_i(d)$  from  $\{Y_{it}(d)\}_t$  should match the dose-response estimands in Chapter 14 (for example, average outcome over a fixed post-treatment horizon), and must be stated explicitly in applications so that  $\mu(d)$  and  $\tau(d)$  are interpreted on the same scale as the ATT-type estimands earlier in the book. The marginal treatment effect at dose  $d$  is

$$\tau(d) = \frac{\partial \mu(d)}{\partial d},$$

which quantifies the incremental effect of a small increase in treatment intensity. In marketing terms,  $\tau(d)$  answers questions such as: “What is the marginal return to an additional dollar of advertising when current spend is  $d$ ?”

We adopt the continuous-treatment unconfoundedness and overlap conditions from Chapter 14, specialised to the notation used for DML.

**Assumption 55 (Continuous-Treatment Unconfoundedness and Overlap)** For all  $d \in \mathcal{D}$ , potential outcomes satisfy

$$Y_i(d) \perp\!\!\!\perp D_i | X_i,$$

where  $X_i$  collects pre-treatment covariates that may summarise the history  $\{X_{it}\}_t$  and include correlated-random-effects proxies for unit heterogeneity. The generalised propensity score is

$$r(d | X_i) = f_{D|X}(d | X_i),$$

which we assume is bounded away from zero on a trimmed region of doses and covariates where you report  $\mu(d)$  and  $\tau(d)$ . Treat extreme doses with very small  $r(d | X_i)$  as extrapolation rather than design-based inference. Chapter 14 shows that the GPS has a balancing property analogous to the binary propensity

score. As in Chapter 14, this assumption is understood to hold for all  $d$  in the region where you report  $\mu(d)$  and  $\tau(d)$ , and conditional on pre-treatment covariates only. In panels,  $X_i$  should therefore comprise features fixed before or contemporaneously with dose assignment, together with CRE-style proxies for  $\alpha_i$ , and must not include variables affected by  $D_i$  or by earlier doses.

Under these assumptions,  $\mu(d)$  is identified and can be estimated using a variety of methods. The DML approach uses orthogonal scores that combine outcome regression and GPS weighting.

## Doubly Robust Estimation

**Definition 12.8 (Doubly Robust Orthogonal Score for ADRF)** Let the outcome regression at dose  $d$  be

$$m(d, x) = \mathbb{E}[Y_i | D_i, X_i] \text{ evaluated at } (d, x).$$

The GPS  $r(d | x)$  is the conditional density of  $D_i$  given  $X_i$ . For a target dose  $d$ , a doubly robust, orthogonal estimator of  $\mu(d)$  takes the form

$$\hat{\mu}(d) = \frac{1}{G} \sum_{i=1}^G \left[ \hat{m}(d, X_i) + \frac{K_h(D_i - d)}{\hat{r}(D_i | X_i)} \{Y_i - \hat{m}(D_i, X_i)\} \right],$$

where  $K_h(\cdot)$  is a kernel with bandwidth  $h$  that localises around dose  $d$ , and  $\hat{r}$  and  $\hat{m}$  are estimated GPS and outcome-regression functions obtained from flexible ML learners. Here  $G$  denotes the number of independent units (clusters), not the total number of unit-time cells.

Define the orthogonal score for  $\mu(d)$  by

$$\psi_d(Z_i, m, r) = m(d, X_i) + \frac{K_h(D_i - d)}{r(D_i | X_i)} \{Y_i - m(D_i, X_i)\}.$$

The estimator is the sample average of this score evaluated at  $(\hat{m}, \hat{r})$ . The term in brackets is the influence-function-based orthogonal score for  $\mu(d)$  in the continuous-treatment model. It combines an imputed outcome  $\hat{m}(d, X_i)$  with a bias-correction term that reweights residuals  $Y_i - \hat{m}(D_i, X_i)$  by the inverse GPS and localises to doses near  $d$  via the kernel. Under Assumption 55, orthogonality ensures that small errors in  $\hat{m}$  and  $\hat{r}$  have only second-order effects on  $\hat{\mu}(d)$  when nuisance estimators are trained on separate folds and converge at suitable rates.

This estimator is locally doubly robust in the sense that, for fixed  $d$  in the interior of  $\mathcal{D}$  and under the kernel and bandwidth regularity conditions, the leading bias from nuisance estimation vanishes if either  $m$  or  $r$  is consistently estimated. Under standard kernel conditions with  $h \rightarrow 0$  and  $Gh \rightarrow \infty$ ,  $\hat{\mu}(d)$  admits an asymptotic normal approximation under  $\sqrt{Gh}$  scaling, with a bias-variance trade-off governed by  $h$ .

Marginal effects  $\tau(d)$  can then be obtained by differentiating a smooth estimate of  $\mu(d)$ , either numerically or using local-linear variants of this doubly robust score as in Kennedy et al. [2017]. Finite-difference derivatives amplify estimation noise. As in Chapter 14, it is often preferable to estimate  $\tau(d)$  using local-linear or

spline-based methods that smooth  $\hat{\mu}(d)$  in  $d$  rather than naive finite differences, especially when  $D_i$  is noisy or heavily discretised.

## Panel Considerations

In panels,  $Y_i(d)$  typically summarises an outcome path  $\{Y_{it}(d)\}_t$  over a window tied to the treatment dose  $D_i$ . For example,  $D_i$  might measure average advertising intensity over a quarter and  $Y_i(d)$  total conversions over the following quarter. The continuous-treatment unconfoundedness assumption must then hold conditional on both observed covariates and proxies for unit-level heterogeneity.

The correlated-random-effects (CRE) approach from Section 12.4 extends naturally. You augment  $X_i$  with unit-level averages of key covariates and, where appropriate, average or baseline treatment intensity, such as  $\bar{X}_i$  and  $\bar{D}_i$ , when estimating both  $m(d, x)$  and  $r(d | x)$ . This helps account for time-invariant heterogeneity correlated with long-run treatment intensity. It does not, however, address time-varying unobservables correlated with both  $D_i$  and  $Y_i(d)$ —for example, campaign-specific shocks, competitor promotions, or algorithmic allocation rules that respond to unobserved demand signals.

Cross-fitting proceeds as in the binary-treatment case. You partition units into folds, estimate  $\hat{m}$  and  $\hat{r}$  on out-of-fold units, and evaluate the orthogonal scores on held-out units. When treatments evolve over time, you define the dose and outcome windows so that the nuisance functions are always trained on pre-treatment or contemporaneous information consistent with the identification strategy in Chapter 14, and you avoid using post-outcome periods in nuisance training for the same unit.

## Marketing Applications

Dose-response DML addresses central marketing questions about intensity and saturation. Consider a retailer varying promotional discount depth across stores. Let  $D_i$  be the discount percentage offered in a given promotion window and  $Y_i$  the corresponding units sold or revenue over that window. The ADRF  $\mu(d)$  traces expected sales as a function of discount depth, while the marginal effect  $\tau(d) = \partial\mu(d)/\partial d$  reveals whether deeper discounts yield proportionally more or fewer sales at different price points. Profit maximisation requires knowing where marginal revenue from a deeper discount equals the associated margin cost.

DML permits  $\mu(d)$  and  $\tau(d)$  to be estimated flexibly, without imposing a particular parametric shape on the sales-discount relationship, while maintaining orthogonality and double robustness under Assumption 55. At the same time, identification in this setting is demanding. Discount depth is typically chosen in response to expected demand, competitive actions, and inventory constraints. High-dimensional controls and CRE features can reduce bias, but they cannot by themselves guarantee unconfoundedness. In most observational pricing and advertising problems, purely selection-on-observables assumptions for continuous treatments are weak. Treat DML-based ADRF estimates as exploratory unless they are anchored by experimental or instrumental-variables variation as in Chapters 14 and 16.

## 12.10 Diagnostics and Robustness

Credible DML analysis requires the same level of diagnostic discipline as any other causal panel method. This section outlines a diagnostic workflow that integrates DML with the tools from Chapter 17 and the method chapters on DiD, event studies, and synthetic control. Section 12.8 focused on implementation choices such as overlap diagnostics and learner sensitivity. Here we concentrate on how to use those tools to assess the credibility of causal conclusions.

### Nuisance Fit

The first question is whether ML-estimated nuisance functions predict outcomes and treatment assignment well enough to be useful without clearly overfitting. For outcome regressions, compute out-of-fold measures such as  $R^2$  on held-out units or time blocks and examine plots of predicted versus observed outcomes. These must respect the same unit-level folds used for cross-fitting: predictions for a given unit should always be evaluated on data that were excluded from the training set for that unit’s fold. For propensity scores, examine out-of-fold metrics such as the area under the ROC curve (AUC) and visual plots of predicted propensities against observed treatment indicators.

High apparent fit is not automatically good news. Extremely high  $R^2$  or AUC can signal overfitting or leakage, especially when models use rich feature sets or post-treatment information. A simple check is to verify that, for each treated unit, nuisance training and tuning windows for control models end strictly before adoption time  $G_i$ . Modest but non-trivial fit that improves clearly on simple benchmarks (for example, a two-way fixed-effects regression) often indicates that the nuisance models are capturing systematic structure without simply memorising noise. See Chapter 2 for the canonical TWFE baseline. Interpret diagnostics relative to these benchmarks and to the underlying design, rather than against rigid numerical thresholds.

### Balance Improvement

DML should improve balance on observed covariates, either through propensity-score weighting or through residualisation. To check this, compute standardised mean differences (SMDs) for key covariates before and after adjustment, following the conventions in Chapter 17. For a covariate  $X$ , the SMD is

$$\text{SMD} = \frac{\bar{X}_{\text{treat}} - \bar{X}_{\text{control}}}{\sqrt{(s_{\text{treat}}^2 + s_{\text{control}}^2)/2}},$$

where  $\bar{X}$  and  $s^2$  denote means and variances in the relevant group. When using weights implied by DML (for example, from propensity scores or Riesz regressions), compute weighted versions of  $\bar{X}$  and  $s^2$  that match the weighting scheme used in the treatment-effect estimator, and treat each unit as the basic observational unit.

Compute SMDs at baseline or over pre-treatment periods, both for raw comparisons and after applying DML-derived weights or residuals. Balance improves if most absolute SMDs shrink relative to their raw values and move toward ranges often treated as acceptable (for example, below conventional 0.1–0.25 thresholds, depending on context). Treat these thresholds as rough heuristics, not pass or fail criteria. What matters is whether remaining imbalances are plausibly large enough to confound the estimand. For staggered-adoption designs, examine balance within cohorts and between cohorts and their comparison groups over the pre-periods that identify their effects. Persistent imbalance after adjustment suggests either weak overlap or misspecified nuisance models. It should trigger design or model revisions rather than blind trust in DML.

## Overlap

Overlap diagnostics for DML follow the same principles as in Chapters 17 and 13. Plot estimated propensities for treated and comparison units, quantify the extent of common support, and report effective sample sizes and covariate support after any trimming. For staggered designs, compute these diagnostics by cohort and comparison group.

In the DML context the emphasis is on understanding whether the units that dominate the orthogonal score lie in regions of good overlap and whether extreme propensities drive a disproportionate share of the influence function. If important parts of the score are supported only by a handful of highly treated or highly control-like units, the resulting estimates will be fragile, regardless of how sophisticated the nuisance learners are. Influence-function diagnostics from Chapter 17 help quantify how much individual units drive the result. For example, compute the share of  $\sum_i \Psi_i^2$  contributed by the top 1–5 per cent of units, and report how  $\hat{\beta}$  changes when those units are removed.

## CATE Stability

When estimating heterogeneous treatment effects, stability across folds and tuning choices is as important as point estimates. For each cross-fitting fold  $f$ , obtain CATE estimates  $\hat{\delta}^{(f)}(x)$  based on fold-specific nuisance functions, or GATEs  $\hat{\delta}_j^{(f)}$  for pre-specified subgroups. Compare these across folds by computing correlations or simple differences.

High agreement across folds—for example, similar rankings of subgroups by effect size and reasonably high correlations between fold-specific estimates—indicates that heterogeneity patterns are driven by signal rather than noise. Low agreement suggests that sample composition and tuning choices materially change who appears to benefit most. In that case you should report the range of estimates across folds for key subgroups and treat any segmentation decisions as tentative. Crucially, each fold-specific estimate must recompute the full DML pipeline (nuisance fitting, cross-fitting, and CATE estimation) within that fold. Re-using common nuisance fits defeats the purpose of the stability check.

## Placebo Diagnostics

Placebo diagnostics extend the pre-trend logic from Chapters 4 and 17 to the DML setting. Apply the full DML procedure—including cross-fitting, learner tuning, and the same leakage restrictions as in the main analysis—to pre-treatment periods, treating a pseudo-intervention date as if it were the true treatment date. Estimate pseudo treatment effects or event-time profiles where no real effect should exist.

If the design is credible and nuisance models capture only legitimate confounding structure, placebo effects should be small and centred near zero, with confidence intervals that mostly cover zero. Occasional non-zero placebo estimates are expected in finite samples, but systematic patterns—such as sustained positive or negative pseudo effects or placebo magnitudes comparable to post-treatment effects—signal problems with parallel trends, unobserved confounding, or leakage. Plot placebo and post-treatment effects on the same scale, using the same unit clustering and confidence-band construction as in the main results. This way differences reflect genuine post-treatment signal rather than changes in variance estimators.

## Learner Sensitivity

Sensitivity to the choice of learner complements hyperparameter stability checks from Section 12.8. Re-estimate key estimands using alternative nuisance learners—for example, lasso, forests, and boosting—each tuned with design-respecting cross-validation. Compare point estimates and confidence intervals across these implementations.

If all reasonable learners deliver estimates that agree within sampling uncertainty and lead to the same substantive conclusions, the analysis is relatively robust to modelling choices. If estimates diverge meaningfully across learners, report the full range explicitly and investigate why. Differences can arise because some learners capture important nonlinearities or interactions that others miss, but they can also reflect overfitting or leakage specific to a given learner. Use standard-error comparisons only as a descriptive summary.

Ensemble estimates that average across learners (for example, weighting by inverse variance) can summarise results when learners broadly agree, but they do not fix design problems or mis-specification shared across learners. Interpretation should still be grounded in design diagnostics.

## Integration with Design-Based Diagnostics

DML should be integrated with, not substituted for, the diagnostic playbooks from earlier chapters. Run the usual pre-trend checks on leads to assess parallel trends, plot event-time profiles with confidence bands, and implement placebo diagnostics on never-treated units or on periods and locations where no intervention occurred, following Chapter 17.

Compare DML estimates of ATT, dynamic effects  $\theta_k$ , or CATE summaries  $\delta(x)$  with estimates from simpler designs: two-way fixed-effects DiD (where appropriate), event-study estimators that handle heterogeneity, and

synthetic-control or matrix-completion approaches from Chapters 6–8. Agreement across these methods, after accounting for their differing estimands and variance, increases confidence that the design, rather than the functional form, is driving the result. Make the estimand alignment explicit (population, weighting, time window). Otherwise disagreement can be an artefact of targeting different causal objects. When methods disagree, the first question should always be which identification assumptions are most plausible in the setting at hand, not which method delivers the largest or most ‘exciting’ effect. Disagreement requires explanation. In some cases DML may legitimately adjust for rich covariate structure that simpler estimators ignore. In others it may be overfitting idiosyncrasies or extrapolating counterfactuals in regions of poor support.

## Triangulation

Triangulation formalises this comparison across estimators. Estimate a focal effect—such as the long-run ATT for a loyalty programme—using a baseline DiD estimator, a synthetic-control or matrix-completion approach, and a DML estimator with high-dimensional covariates. Examine not only the point estimates and confidence intervals, but also the identification assumptions each method relies on. TWFE DiD may be vulnerable to treatment heterogeneity and negative weights. Synthetic control may target a single treated unit or a subset of cohorts. DML may rely heavily on overlap and nuisance-model quality.

If, after accounting for these differences, all methods tell a broadly consistent story, the substantive conclusion is likely robust. If DML departs substantially from the others, ask whether the difference is plausibly due to better covariate adjustment, perhaps correcting an obvious imbalance documented in Chapter 17. Also ask whether it reflects extrapolation in regions of weak support or sensitivity to particular learners. Transparent reporting of multiple estimators, along with an explicit ranking of which identification assumptions you are willing to defend and why you privilege a particular estimator, allows readers to see whether your conclusions are design-driven or method-driven.

## 12.11 Inference

Inference for DML estimators in panels must respect three features: clustering induced by dependence within units and over time, randomness introduced by sample splitting and cross-fitting, and multiplicity when we estimate many effects (subgroups, event times, or dynamic profiles). This section builds on the variance estimator in Proposition 12.3 and connects DML to the panel-inference framework developed in Chapter 16.

### Influence Functions and Clustering

The doubly robust score  $\psi^{\text{ATT}}(Z_{it}, \beta, \eta)$  from Definition 12.2 underlies inference for the ATT estimator. At the cluster level, the influence function contribution of unit  $i$  is

$$\Psi_i = \sum_{t=1}^T \psi^{\text{ATT}}(Z_{it}, \beta_0, \eta_0),$$

so that  $\sqrt{G}(\hat{\beta} - \beta_0)$  has asymptotic variance  $V = \mathbb{E}[\Psi_i^2]$ , with  $G$  denoting the number of independent units. We work under an asymptotic regime with  $G \rightarrow \infty$  and  $T$  fixed or bounded, so that units are the asymptotically independent clusters. When both  $G$  and  $T$  grow, additional dependence conditions from Chapter 16 are required. Proposition 12.3 estimates  $V$  with the empirical second moment of these unit-level influence contributions:

$$\hat{V} = \frac{1}{G} \sum_{i=1}^G \left( \sum_{t=1}^T \psi^{\text{ATT}}(Z_{it}, \hat{\beta}, \hat{\eta}^{(-k(i))}) \right)^2.$$

The inner sum aggregates influence contributions over time for each unit, respecting serial dependence. The outer sum aggregates across units. Cluster-robust standard errors and confidence intervals in Chapter 16 can be interpreted as working with this same cluster-level influence-function representation.

In settings where both unit and time effects may induce dependence, two-way clustering generalises this idea by constructing influence contributions aggregated by unit, by time, and by their combination, and then applying the inclusion–exclusion formulas described in Chapter 16. Two-way clustering is appropriate only when both unit and time dimensions contain enough clusters for asymptotics to be plausible. With very short panels, time clustering adds noise without improving coverage (see Chapter 16). DML enters only through the construction of  $\psi^{\text{ATT}}$ . The clustering logic is identical to that for design-based estimators. Cross-fitting does not change the influence-function form. It only affects how  $\hat{\eta}^{(-k(i))}$  is obtained within each cluster, so the same cluster-robust logic applies under the regularity conditions that justify the influence-function expansion.

## Small-Sample Adjustments

When the number of clusters is modest, asymptotic approximations can be fragile even with cluster-robust variance estimators. Chapter 16 discusses two main remedies that apply equally to DML.

First, degrees-of-freedom corrections adjust cluster-robust variance estimates to account for the finite number of clusters. A common heuristic multiplies  $\hat{V}$  by  $G/(G - p)$ , with  $p$  the number of low-dimensional target parameters (often  $p = 1$ ). In high-dimensional nuisance settings  $p$  refers to the parameters of interest, not to the dimension of the ML models. This does not change the asymptotic properties of the estimator but can slightly widen confidence intervals when  $G$  is small.

Second, wild cluster bootstrap methods approximate the sampling distribution of  $\hat{\beta}$  without relying solely on large- $G$  theory. In a DML context, the natural bootstrap object is the unit-level influence function  $\Psi_i$ . Wild bootstrap procedures reweight these cluster-level contributions with random multipliers (for example, Rademacher signs), recompute the estimator across many bootstrap draws, and form empirical critical values for confidence intervals and hypothesis tests. Bootstrap multipliers must be applied at the unit level (or at the chosen cluster level for two-way clustering), never at the individual unit-time cell level, to preserve the dependence structure assumed in the asymptotics. Chapter 16 provides detailed algorithms and guidance on when wild bootstrap is preferable to analytical variance formulas.

## Sample-Splitting Randomness

Random sample splitting introduces an additional layer of variability: different partitions into folds yield different nuisance estimates and, in finite samples, different treatment-effect estimates. The asymptotic theory in Chernozhukov et al. [2018] shows that, under our regularity conditions, a single K-fold cross-fitted estimator is already  $\sqrt{G}$ -consistent and asymptotically normal. In practice, however, analysts often reduce Monte Carlo noise from any particular partition by repeating the procedure over several random splits.

Suppose we run the full DML pipeline for  $R$  independent random fold assignments, obtaining estimates  $\hat{\beta}_1, \dots, \hat{\beta}_R$  and associated cluster-robust variance estimates  $\widehat{\text{Var}}(\hat{\beta}_r)$ . A natural aggregate estimator is

$$\hat{\beta}_{\text{avg}} = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_r.$$

A corresponding variance estimator that reflects both within-split uncertainty and between-split variation is

$$\widehat{\text{Var}}(\hat{\beta}_{\text{avg}}) = \frac{1}{R^2} \sum_{r=1}^R \widehat{\text{Var}}(\hat{\beta}_r) + \frac{1}{R(R-1)} \sum_{r=1}^R (\hat{\beta}_r - \hat{\beta}_{\text{avg}})^2.$$

The first term averages the cluster-robust variance estimates from each split. The second term captures how much the point estimates move across splits. Because all  $\hat{\beta}_r$  are computed from the same underlying data, interpret the between-split term as a measure of algorithmic instability, not as an independent sampling

variance component. This decomposition is a pragmatic stabilisation device. It should not be over-interpreted as a principled variance decomposition.

An alternative, used in much of the DML literature, is to work with K-fold cross-fitted scores that are already averaged over folds. In that case the influence-function representation and cluster-robust variance in Proposition 12.3 already account for the fold structure, and repeated random splitting becomes optional rather than mandatory.

## Multiplicity Adjustments

When we estimate many effects—CATEs across subgroups, event-time effects, or dynamic profiles across multiple horizons—standard pointwise inference can dramatically underestimate uncertainty about the overall pattern. If we test ten subgroup effects at the 5 per cent level, we expect roughly half a false rejection even when all true effects are zero.

Chapter 16 surveys multiple-testing and joint-confidence-band methods for panel settings, including Bonferroni (add citation) and Holm corrections for family-wise error control [Holm, 1979], Benjamini–Hochberg false-discovery-rate procedures [Benjamini and Hochberg, 1995], Romano–Wolf stepdown methods [Romano and Wolf, 2005], and cluster-robust joint bands built from influence-function covariance matrices or bootstrap draws. In the DML context, these methods are applied to vectors of estimators whose covariance structure is determined by their joint influence functions. Concretely, the cluster-robust covariance matrix is estimated as the empirical covariance of the vector-valued influence contributions across units, with the same clustering scheme as in scalar DML. For example, a vector of GATEs or event-time effects has an associated cluster-robust covariance matrix that can be used in F-tests, chi-squared tests, or stepdown procedures.

Practical guidance is to report unadjusted estimates and p-values for individual subgroups or event times alongside at least one joint test of heterogeneity (for example, an F-test of equality of subgroup effects or a joint test that all pre-treatment event-time coefficients are zero). Use the joint tests as primary evidence for or against heterogeneity. If they reject, treat detailed subgroup patterns as exploratory and interpret individual adjusted p-values or joint bands rather than relying on a single unadjusted comparison. If subgroups are chosen after inspecting the data (for example, via trees), treat subgroup p-values as descriptive unless you use an honest sample split or selective-inference machinery.

## Connection to Panel Inference Frameworks

DML inference extends rather than replaces the panel-inference frameworks developed earlier in the book. Once we have an influence-function representation for our estimator—whether for an average effect, a cohort-time ATT, a dynamic profile, or a GATE—the machinery from Chapter 16 applies directly: we aggregate influence contributions at the relevant cluster level, choose an appropriate clustering scheme (unit, time, or two-way), and, where needed, apply wild bootstrap or joint-band procedures for multiple testing.

For staggered adoption (Chapter 4), DML estimators of the cohort–time effects  $\tau(g, t)$  and their event-time aggregations use the same cohort weights and inference structures as their design-based counterparts. The difference is that parametric or low-dimensional trend adjustments are replaced by ML-estimated nuisance functions. For heterogeneous effects and dose–response functions, the same influence-function and clustering principles carry over to vectors of estimators, with multiple-testing and joint-band tools from Chapter 16 providing the appropriate adjustments.

In short, DML affects how we estimate nuisance functions and construct orthogonal scores, but once the influence-function representation is in hand, inference is governed entirely by the same clustering and multiple-testing principles as for design-based estimators.

## 12.12 Marketing Applications

DML methods are particularly valuable in marketing problems where treatment effects are plausibly heterogeneous, confounding is complex and high-dimensional, and flexible functional forms are needed. This section sketches methodological blueprints for five recurring problems, linking back to the estimands and designs developed in earlier chapters. Chapter 18 revisits these and other applications in full detail.

### Loyalty Programme Heterogeneity

Consider a retailer rolling out a loyalty programme across stores under staggered adoption. Outcomes are store-period sales  $Y_{it}$ , and the design from Chapters 4 and 5 delivers cohort-time effects  $\tau(g, t)$  and event-time profiles  $\theta_k$ . Here we summarise each store’s dynamic effect into a scalar post-treatment estimand, for example the average ATT over four post-adoption quarters, constructed by aggregating the cohort-time or event-time effects from Chapters 4 and 5 over the chosen horizon, so that  $Y_i(1) - Y_i(0)$  is defined on the same scale as those estimands. We then study heterogeneity in that scalar across store characteristics.

The target is the CATE  $\delta(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$ , where  $X_i$  collects pre-programme characteristics such as demographics, competitive density, and baseline sales patterns, and  $Y_i(1) - Y_i(0)$  is a long-run ATT summary as defined in Chapter 10. The underlying average effects inherit the staggered-adoption identification assumptions from Chapter 4. Interpreting second-stage heterogeneity in a constructed unit-level summary additionally requires that the summarisation and segmentation steps do not reintroduce post-treatment selection. Use honest splitting across units when learning segments or forests.

A DML implementation partitions stores into folds, estimates nuisances on  $K - 1$  folds using flexible learners such as gradient-boosted trees, and evaluates doubly robust scores on the held-out fold. CATEs can then be estimated with causal forests or grouped into GATEs over interpretable segments such as urban versus rural or high- versus low-income catchments, using the machinery from Section 12.5.

Key diagnostics mirror the staggered-DiD chapters. Pseudo-treatment effects in pre-programme quarters should be small, overlap diagnostics for store-level propensities should show common support across cohorts and comparison groups, and CATE patterns should be stable across folds and learners. Substantively, loyalty programme adoption is often highly selected: better-managed or faster-growing stores may adopt earlier. Correlated-random-effects controls help proxy for unit-level heterogeneity, but time-varying unobservables remain a threat. The DML framework does not fix violations of conditional parallel trends; it simply makes high-dimensional conditioning feasible once a staggered-DiD design is judged credible.

### Advertising Dose–Response

Digital advertisers routinely vary impression intensity across markets and devices and want to recover dose–response functions. Let  $D_i$  denote impression intensity for market  $i$  over a defined campaign window, and

let  $Y_i$  be conversions over an aligned outcome window. As in Chapter 14, the estimand is a continuous-treatment response  $\mu(d) = \mathbb{E}[Y_i(d)]$  or its derivative  $\tau(d) = \partial\mu(d)/\partial d$  at relevant impression levels, defined over a specified post-campaign horizon. The continuous-treatment unconfoundedness and overlap conditions of Assumption 55 must hold conditional on pre-campaign features  $X_i$ , which is a demanding requirement in observational ad-spend data.

Nuisance functions follow the continuous-treatment DML setup: a conditional mean function  $m(d, x)$  and a generalised propensity score  $r(d | x)$ , the conditional density of treatment intensity given pre-campaign features  $X_i$  such as baseline demand, competition, and platform-side signals. Cross-fitting and orthogonal scores from Definition 12.8 deliver estimates of  $\mu(d)$  on a grid, with marginal effects obtained by numerical differentiation or local-slope estimation [Kennedy et al., 2017]. Even with orthogonal scores, these estimates remain design-based only if the continuous-treatment unconfoundedness condition is credible or if you embed the nuisance machinery inside an IV or quasi-experimental design.

Diagnostics focus on whether the continuous-treatment assumptions from Chapter 14 are remotely plausible: generalised overlap across device types and markets, sensible pre-campaign placebo checks, and robustness to learner choice in both outcome and density models. Economically, advertising intensity is highly endogenous: spend rises where returns are expected to be high. High-dimensional controls can reduce bias but rarely eliminate it. When credible cost-shifter or delivery instruments exist, Chapter 16.10 provides IV-based dose-response designs that should be used alongside or in place of pure selection-on-observables DML.

## Competition-Conditioned Price Sensitivity

Retailers often suspect that price elasticity varies with local competition. Let  $C_i$  count nearby competitors for store  $i$ . The object of interest is how a pre-specified price effect—such as the short-run ATT of a discount on sales over a promotion window (for example, the ATT on revenue over the promotion window as in Chapter 14 or Chapter 10)—varies with  $C_i$ . Formally, we target the CATE function

$$\delta(u) = \mathbb{E}[Y_i(1) - Y_i(0) | C_i = u],$$

where  $u$  indexes values of the pre-treatment covariate  $C_i$ .

Nuisance functions include an outcome regression for sales as a function of price, competition, and other covariates, and a propensity model for price changes or promotions given those covariates. A DML strategy partitions stores into folds, estimates nuisances with flexible learners such as elastic net or boosting, constructs orthogonal scores for the promotion effect, and then uses causal forests or grouped analyses to estimate how  $\delta(\cdot)$  evolves with competitor count.

Here identification challenges are even more acute. Price changes are strategic responses to anticipated demand and competitive pressure, so unconfoundedness is a strong assumption. Correlated-random-effects controls help remove time-invariant store effects, but time-varying demand shocks and strategic rival responses remain. Whenever possible, combine DML with credible instruments, for example cost shifters, or quasi-experimental variation such as competitor entry or policy changes, using the designs in Chapter 16 and

Chapter 18. DML should be treated as a flexible nuisance engine within IV or quasi-experimental designs, not as a standalone fix for strategic pricing endogeneity.

## Multi-Touch Attribution

Multi-touch attribution aims to disentangle the contributions of multiple channels—email, display, search, social—to conversion. Let  $D_{ij}$  indicate exposure of user  $i$  to channel  $j$  over a specified pre-conversion window, and let  $Y_i$  be a binary conversion indicator or value measure over a post-exposure horizon. This is a windowed cross-sectional summary rather than a full panel:  $j$  indexes channels and the exposure window is fixed *ex ante*. Ideally we would like to know how outcomes would change under interventions that toggle specific channels while holding broader campaign design fixed.

One approach is to define, for each channel  $j$ , an average marginal effect conditional on the observed exposure structure,

$$\tau_j = \mathbb{E}[Y_i(D_{ij} = 1, D_{i,-j}) - Y_i(D_{ij} = 0, D_{i,-j})],$$

where the expectation is taken over the joint distribution of  $(X_i, D_{i,-j})$  in the observed campaign environment. This notation is shorthand for potential outcomes under joint channel-exposure vectors, which are rarely identified without experiments or very strong selection-on-observables assumptions. Standard SUTVA is typically violated here, because channel exposures interact and may affect outcomes jointly. Chapter 11 provides exposure-mapping frameworks more appropriate for such dense multi-channel settings. Treat channel-level effects as exploratory unless supported by randomised variation or a clearly defended assignment mechanism (Chapter 2).

A DML implementation partitions users into folds, trains multi-channel outcome models and channel-wise propensity models on  $K - 1$  folds, and constructs channel-specific orthogonal scores on the held-out fold. Aggregating these scores yields channel-level effect estimates that can be compared to industry attribution heuristics. Diagnostics must be severe: check positivity for each channel conditional on key covariates and other exposures, inspect overlap in exposure patterns, and run pre-period placebos where possible.

Even under strong selection-on-observables assumptions and positivity across channel-exposure patterns, identification in attribution is fragile. Exposures across channels are often highly collinear, and unobserved intent drives both exposure and conversion. In many cases only randomised experiments or strong natural experiments can deliver credible channel effects. DML provides a structured way to combine high-dimensional data with causal scores, but it cannot manufacture exogenous variation where none exists.

## Customer Lifetime Value Targeting

Subscription and repeat-purchase businesses often wish to target acquisition discounts or retention offers based on their impact on customer lifetime value (CLV). Let  $CLV_i(1)$  and  $CLV_i(0)$  denote long-run value measures constructed from the customer's panel of transactions under treatment and control, following the

CLV definitions in Chapter 18. The target is a CATE  $\delta(x) = \mathbb{E}[\text{CLV}_i(1) - \text{CLV}_i(0) \mid X_i = x]$ , and the policy-learning problem from Section 12.6 is to identify customers with  $\delta(x) > \kappa$ , where  $\kappa$  is the cost of the discount.

Nuisance functions are CLV regressions under treatment and control,  $m_1(x)$  and  $m_0(x)$ , and a treatment propensity  $e(x)$  capturing how discounts were assigned historically. Because CLV is a constructed outcome that may involve censoring and extrapolation, outcome models often use survival or hazard components in addition to standard regressions. Any misspecification in the CLV construction propagates directly into  $\delta(x)$  and the learned policy, so you should validate CLV models separately before layering DML on top. Cross-fitted DML scores feed into CATE estimators, for example causal forests, and then into the policy-learning framework to produce targeting rules.

Diagnostics include CLV-prediction accuracy on held-out customers, overlap in propensities across covariate space, and doubly robust estimates of policy value as in Proposition 12.5. Comparing the learned policy's estimated value to simple baselines—treating everyone, treating no one, or treating by simple rules such as income or tenure thresholds—clarifies how much incremental business value DML targeting appears to generate.

Identification in CLV targeting faces the usual observational challenges. Discount recipients may differ systematically from non-recipients in unobserved purchase propensity or platform engagement. Historical A/B tests and pilots provide valuable ground truth for validating DML-based targeting rules before large-scale deployment. In their absence, CLV targeting should be presented as hypothesis-generating rather than as definitive evidence. Before operational rollout, always validate DML-based CLV targeting rules in fresh A/B tests or staggered pilots. Observational CLV targeting should not be deployed at scale without such experimental confirmation.

## Summary

These blueprints illustrate how DML can be embedded in realistic marketing designs rather than used in isolation. CATE estimation  $\delta(x)$  informs segmentation and targeting decisions, dose-response curves  $\mu(d)$  and  $\tau(d)$  guide budget allocation, effect modification clarifies competitive mechanisms, attribution frameworks attempt to decompose multi-channel journeys via  $\tau_j$ , and CLV targeting links heterogeneous effects to long-run business value.

Across all these problems, DML changes only the estimation layer. It lets you fit rich, nonlinear nuisance models while preserving orthogonal-score and clustering structures, but it does nothing to fix weak designs. If unconfoundedness, overlap, or interference assumptions are not credible, the resulting estimates remain untrustworthy, regardless of how impressive the ML machinery looks.

## 12.13 Workflow Checklist

This section sets out a compact, reproducible protocol for conducting DML analyses in marketing panels. The workflow integrates estimand definition, nuisance estimation, cross-fitting, diagnostics, inference, and reporting, and is intended to be read alongside the more detailed discussions in earlier sections of this chapter and in Chapters 2, 4, 14, and 16.

### Step 1: Define the Estimand

Begin by specifying the causal quantity of interest in the potential-outcomes framework from Chapter 2. Clarify whether you are targeting an average effect—ATE, ATT, or cohort-time effects  $\tau(g, t)$  and event-time effects  $\theta_k$  from Chapter 4—a continuous-treatment response  $\mu(d)$  from Chapter 14, a CATE  $\delta(x)$  or a grouped effect, or a policy value as in Section 12.6. Write the estimand explicitly in notation and explain its business meaning. For staggered adoption, document the aggregation strategy clearly: which  $\tau(g, t)$  enter, how they are weighted, and which event-time summary (if any) you will report.

### Step 2: Select an Orthogonal Score

Identify the nuisance functions required for your estimand, drawing on Sections 12.2 and 12.4. For binary treatments these typically include outcome regressions  $m_0(x)$  and  $m_1(x)$  and a propensity score  $e(x)$ . For continuous treatments they include a conditional mean  $m(d, x)$  and a conditional density or Riesz representer. Select the corresponding doubly robust, Neyman-orthogonal score  $\psi(\cdot)$  from the definitions in this chapter (for example, the DR ATT score, the continuous-treatment score, or the policy-value score).

Rather than deriving new scores from scratch, use ones that are known to satisfy Neyman orthogonality and double robustness in the relevant semiparametric model. This ensures first-order insensitivity to nuisance-estimation error and aligns your implementation with the asymptotic results in Section 12.11.

### Step 3: Design Dependence-Respecting Cross-Fitting

Design the sample-splitting scheme to respect the dependence structure of your panel, as described in Section 12.3. Partition the data into  $K$  folds, often  $K = 5$ . Use larger  $K$  only when the number of independent clusters is comfortably large.

When many units are observed over a modest number of periods, use unit-level folds: each unit belongs entirely to a single fold, and nuisance models are trained on other units. When there are few units but a long time series, time-based folds that hold out contiguous time blocks for validation may be more appropriate. When both cross-sectional and temporal dependence are important, block folds in the unit-time plane provide

more conservative protection. In all cases, enforce the clean-control restriction from Assumption 49: nuisance models intended to represent untreated paths must not be trained on post-treatment observations from treated units.

#### Step 4: Choose Learners and Tune

Select ML learners to estimate the identified nuisance functions, taking into account both the structure of the data and the role of each nuisance. Regularised GLMs such as lasso or elastic net are natural for high-dimensional but approximately linear propensities. Gradient boosting and forests are often preferred for outcome regressions where nonlinearities and interactions matter. When dynamic outcomes are central, consider learners that incorporate time trends or splines explicitly.

Tune hyperparameters using cross-validation restricted to the training folds defined in Step 3. Within each training set, create internal validation splits, fit learners across a hyperparameter grid or via random/Bayesian search, and select configurations that perform well on appropriate out-of-sample metrics (for example, mean squared error for outcomes, log-likelihood or Brier score for propensities). Avoid tuning on held-out folds or post-treatment data for treated units to prevent leakage.

#### Step 5: Assess Overlap and Trim if Needed

Before relying on DML estimates, examine overlap between treated and comparison units in the covariate space that underpins your estimand. Plot estimated propensities (or generalised propensity scores) for treated and comparison groups, and compute simple overlap summaries as in Chapter 17. In staggered-adoption designs, do this by cohort and comparison group over the pre-treatment periods that identify  $\tau(g, t)$ .

If treatment probabilities are extremely close to zero or one in parts of the covariate space, consider trimming those regions, recognising that this changes the estimand to one defined on the overlap region. Trimming a small fraction of the most extreme units can improve precision and reduce the influence of a few poorly matched observations, but the fraction should be guided by overlap diagnostics rather than by a fixed percentage. Always report how many units are excluded, how the covariate support changes, and how you interpret the resulting estimand.

#### Step 6: Estimate Effects with Clustered Inference

With nuisances and cross-fitting in place, estimate the target parameter using the selected orthogonal score. For each fold  $k$ , fit the nuisance models on the  $K - 1$  training folds, evaluate them on the held-out fold, compute the orthogonal scores  $\hat{\psi}_{it}$  for that fold, and solve the corresponding estimating equation or build the relevant CATE/dose-response model. Denote the fold-specific estimate (scalar or vector) by  $\hat{\beta}_k$ .

Aggregate across folds to obtain the final estimator  $\hat{\beta}$ , either by averaging fold-specific scalars or, for vector parameters, by stacking fold-level scores and solving the global moment conditions. Use the influence-function representation to compute standard errors, clustering at the unit level by default. Cluster at a higher level (market or geo blocks) when interference or shared shocks imply fewer independent pieces of information, and use two-way clustering only when the time dimension contains enough clusters to justify it. See Proposition 12.3 and Chapter 16. Report point estimates, confidence intervals, and p-values in a way that is transparent about the clustering scheme and effective number of clusters.

## Step 7: Run Diagnostics

Diagnostics for DML mirror those for design-based estimators, with additional attention to nuisance performance. Run placebo exercises in pre-treatment periods by applying the full DML pipeline to pseudo-intervention dates and checking that pseudo effects are small and centred near zero. Assess nuisance fit using out-of-fold metrics (for example,  $R^2$  and AUC) and inspect whether DML-based weighting or residualisation improves balance on key covariates, especially at baseline and within cohorts.

Check the stability of estimates across folds, learners, and tuning choices, particularly for heterogeneous effects and policy rules. Compare DML estimates with simpler benchmarks—such as event-study estimators that handle heterogeneity and synthetic-control designs—paying attention to whether differences can be explained by richer covariate adjustment or appear driven by extrapolation in regions of weak support. Chapter 17 and Section 12.10 provide detailed templates for these checks.

## Step 8: Report Sensitivity and Policy Implications

Finally, present results in a way that makes model and design dependence clear. Report treatment-effect estimates across a small number of reasonable learner classes and tuning grids, explaining which specifications you regard as most credible and why. For heterogeneous effects, summarise CATEs into grouped effects or segments that are meaningful for decision-making, and link them explicitly to policy-learning results when targeting rules are proposed.

Discuss how sensitive the main conclusions are to alternative specifications, trimming choices, and clustering schemes. Where possible, benchmark learned policies against simple baselines (treat-all, treat-none, or naive rules) using doubly robust policy-value estimators. Provide enough detail—on estimands, nuisance learners, cross-fitting design, diagnostics, and code or pseudo-code—that other analysts can reproduce and stress test the analysis.

**Table 12.1** Mapping from Estimands to Nuisance Components, Scores, and Aggregation

Estimand	Nuisance Components	Orthogonal Score	Aggregation
ATE (high-dimensional)	$m_1(x), m_0(x), e(x)$	Doubly robust score	Average of unit-level scores
Cohort-time effects $\tau(g, t)$	Cohort-specific nuisances	Cohort-time doubly robust scores	Weight $\tau(g, t)$ to event-time effects $\theta_k$
Dose–Response	$m(d, x), f_{D X}(d \mid x)$ or Riesz representer	Continuous-treatment orthogonal score	Recover $\mu(d)$ and its derivative $\tau(d)$
CATE $\delta(x)$	$m_1(x), m_0(x), e(x)$	Doubly robust scores by unit	Causal forests or grouped effects
Policy Value	$m_1(x), m_0(x), e(x)$ , policy $\pi$	Policy-value score $\psi_\pi$	Average under policy $\pi$



# Chapter 13

## High-Dimensional Controls and Regularisation

Modern marketing panels routinely contain more potential controls than traditional econometric recipes can handle. Store- and customer-level characteristics, platform and market signals, competitor actions, and rich lag structures quickly push the covariate dimension  $p$  into the hundreds or thousands, rivalling or exceeding the effective sample size. This chapter summarises how to use *regularisation* to estimate nuisance components—conditional means and treatment models—when  $p$  is large, while keeping identification anchored in the designs from earlier chapters.

We start with the substantive step: restricting the *candidate* covariate library to design-admissible, pre-treatment variables, defined relative to the estimand at hand. We then use lasso-family penalties—lasso, elastic net, and grouped penalties—to manage dimensionality within that library. We show how to use these tools inside (i) DiD and event-study estimators under (conditional) parallel trends and (ii) double machine learning estimators under selection-on-observables with overlap.

We also discuss post-selection inference under clustering, emphasising where asymptotic cluster-robust approximations are fragile (few clusters, volatile selection) and when bootstrap-based alternatives are preferable. Throughout, identification continues to rest on the assumptions stated in earlier chapters (for example, unconfoundedness with overlap in selection-on-observables settings, or conditional parallel trends in DiD). Regularisation only changes how we estimate the associated nuisance components. Section 13.7 states the key conditions explicitly.

### 13.1 Motivation and Setup

Modern marketing panels confront practitioners with an abundance of potential confounders. Your causal target is an estimand such as ATT or an event-time effect  $\theta_k$ , defined in terms of potential outcomes. Controls enter only through the identification assumptions you are willing to maintain, for example unconfoundedness given pre-treatment covariates, or conditional parallel trends in DiD designs. A retailer estimating a promotion's effect on sales may observe hundreds of store characteristics, market conditions, competitor actions, and platform signals. The covariate dimension  $p$  can easily rival or exceed the effective sample size, and a naive regression that includes all covariates produces unstable estimates, multicollinearity, and overfitting. That undermines precision and makes any causal interpretation harder to defend.

Regularisation offers a practical way to manage dimensionality in this regime, but it is a prediction tool and does not, by itself, justify a causal interpretation. By shrinking or nullifying the coefficients of many covariates, lasso, elastic net, and grouped penalties retain only the most predictive variables. Lasso (an  $\ell_1$  penalty) encourages sparsity, performing variable selection by setting many coefficients exactly to zero. Elastic net (a mix of  $\ell_1$  and  $\ell_2$  penalties) balances sparsity with stability, allowing correlated covariates to enter together. Group lasso penalises entire groups of covariates, enabling structured selection—for example, keeping or dropping all lags of a variable as a block. These methods can approximate a rich *candidate* conditioning set without manual selection, provided that the candidate library has already been restricted to design-admissible, pre-treatment covariates.

The challenge is that regularisation is optimised for prediction, not for causal identification. As Breiman [2001] emphasised, prediction and explanation pursue different goals. Prediction seeks to minimise forecast error, while identification aims to isolate a treatment effect by blocking backdoor paths without conditioning on colliders or mediators. A naive application of lasso may drop important confounders because they do not predict outcomes strongly, or may include variables that introduce bias. Regularisation must be disciplined by the study design to serve causal inference rather than undermine it.

The potential-outcomes framework in Chapter 2 clarifies what regularisation can and cannot do. It does not relax the identification assumptions—unconfoundedness, overlap, and SUTVA. Instead, it provides a data-driven way to choose among many pre-treatment covariates when theory does not uniquely specify which ones matter. The goal is to approximate a conditioning set that is defensible under your maintained causal story—typically, a set of pre-treatment variables that you believe captures the main drivers of treatment assignment and outcomes—without making the model so complex that it becomes unstable.

High-dimensional theory often frames this problem through approximate sparsity. This is a modelling convenience, not a verifiable property. In practice, you lean on it when the selected sets and the resulting treatment-effect estimates are stable across folds and penalty choices, and when the effective number of independent clusters  $G$  is comfortably larger than the implied complexity of the nuisance models.

The design-based perspective emphasised by Angrist and Pischke [2010] clarifies the source of identifying variation. Difference-in-differences (Chapter 4) and event studies (Chapter 5) identify effects through (conditional) parallel trends in untreated potential outcomes, not through predictive fit. Regularisation complements these designs by enabling richer covariate adjustment—for example, augmenting a DiD with lasso-selected controls—while keeping the identifying variation in the cohort–time contrasts.

The discipline is to avoid selecting controls that conflict with identification (post-treatment variables, mediators, or colliders), and to report transparently which covariates are candidates, which are excluded, and how estimates vary with penalty choice. In practice, we enforce the design by restricting the candidate library to covariates measured strictly before the relevant treatment onset for the estimand at hand (for staggered adoption, cohort-specific pre-periods). When we say that admissible controls are not descendants of  $D_{it}$  in a causal graph, we mean the operational rule you can actually implement: if business process knowledge, timing, or measurement makes it plausible that  $D_{it}$  can change the variable within the analysis window, it does not belong in the candidate library.

Positioning regularisation relative to factor models (Chapter 8) and double machine learning (Chapter 12) highlights trade-offs. Factor models capture shared shocks with heterogeneous loadings. Regularisation can then select additional unit-specific controls after partialling out the factor structure, improving precision. Double machine learning methods as in Chernozhukov et al. [2018] use orthogonalisation and cross-fitting to deliver robust causal estimates even when nuisance functions are estimated by flexible ML, with lasso and related regularisers providing high-dimensional building blocks. In all these cases, regularisation is a tool inside a design-based identification strategy, not a substitute for that strategy. None of these procedures can rescue a design that conditions on descendants of treatment or lacks credible untreated counterfactuals. They only help approximate a defensible conditioning set once a valid design has been specified.

A concrete example illustrates the stakes. Consider a digital advertiser estimating the effect of display advertising on conversions using a panel of 500 DMAs observed over 52 weeks. The advertiser observes roughly 200 covariates per DMA-week: demographics, competitor ad spend, seasonality dummies, holiday and weather indicators, search trends, social media engagement, website metrics, and lagged outcomes. Here  $p \approx 200$  is large relative to the number of independent clusters  $G = 500$  and is comparable to  $T = 52$ . Under clustering by DMA, the relevant sample size for high-dimensional regularity conditions and penalty calibration is the number of independent DMAs,  $G$ , rather than the total number of DMA-week cells  $NT$ , because shocks are serially dependent within DMAs.

A naive regression including all 200 covariates produces unstable estimates and poor out-of-sample performance. Regularisation via lasso or elastic net selects a smaller set of covariates—say, 20–30 variables that theory and the data agree are important—yielding a more stable nuisance model. Double-selection procedures in the spirit of Belloni et al. [2014] select covariates from both the outcome and treatment models, helping ensure that variables that predict treatment assignment (and thus may be confounders under an unconfoundedness story, or that stabilise trends under a conditional-parallel-trends story) are retained even if they are weak outcome predictors. Conditional on the maintained design assumptions and a defensible pre-treatment candidate library, this can produce estimates that are more precise and less sensitive to ad hoc control choice than a high-variance, fully saturated regression.

The remainder of this chapter formalises high-dimensional panel models (Section 13.2), presents regularisation methods and dependence-aware tuning (Section 13.3), develops double selection and post-selection inference (Sections 13.4–13.5), and shows how to integrate regularisation with DiD and event-study designs (Section 13.6). Later sections cover identification assumptions, tuning, diagnostics, and a practical workflow checklist that ties these tools back to the book’s core designs.

## 13.2 High-Dimensional Controls in Panels

High-dimensional panel models extend standard panel regression to accommodate many controls while respecting the dependence and fixed-effects structures that underlie identification. This section formalises the baseline model, discusses how to partial out fixed effects and latent factors, and clarifies how dependence affects selection and inference.

The baseline specification extends the standard two-way fixed-effects model to a high-dimensional setting:

$$Y_{it} = \tau D_{it} + X'_{it}\beta + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $D_{it}$  is the treatment variable (often binary in this chapter),  $\tau$  the treatment coefficient in the TWFE regression,  $X_{it} \in \mathbb{R}^p$  the control vector, and  $\alpha_i, \lambda_t$  unit and time fixed effects. The causal interpretation of  $\tau$  depends on the estimand and the identification assumptions stated in the relevant design section.

The model is high-dimensional when  $p$  is large relative to the effective number of independent observations, which is typically the number of independent clusters  $G$  under clustered dependence. Including valid pre-treatment confounders in  $X_{it}$  can reduce omitted-variable bias when identification relies on selection-on-observables or conditional parallel trends. Conditioning on post-treatment variables, colliders, or proxies for treatment-induced processes can instead introduce bias. Regularisation manages dimensionality within a candidate library that has already been restricted by design.

### Fixed Effects and the Within-Transformation

As established in Chapter 2, unit fixed effects  $\alpha_i$  absorb time-invariant heterogeneity (store location, brand equity, baseline loyalty), while time fixed effects  $\lambda_t$  capture common shocks (seasonality, macroeconomic conditions). In high-dimensional settings, fixed effects reduce the effective covariate dimension: time-invariant controls are collinear with  $\alpha_i$  and drop out, and time-varying controls enter only after the within-transformation.

A common approach is to apply the two-way within transformation that removes  $\alpha_i$  and  $\lambda_t$ , then run the regularised regression on the transformed variables  $(\tilde{Y}_{it}, \tilde{D}_{it}, \tilde{X}_{it})$ . Concretely, we run lasso on the within-transformed variables rather than on the raw levels, so the penalty operates on variation orthogonal to  $(\alpha_i, \lambda_t)$ . This partialling-out step is an estimation device for handling nuisance variation conditional on an identification strategy. It does not alter the underlying unconfoundedness or design assumptions.

### Factor-Augmented Models

Factor-augmented models (Chapter 8) extend this logic to heterogeneous exposure to common shocks. Writing

$$Y_{it} = \tau D_{it} + X'_{it}\beta + \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

where  $f_{tr}$  are common factors and  $\lambda_{ir}$  are unit-specific loadings, allows rich time-varying unobservables to enter through a low-rank structure, with  $R \ll \min(N, T)$  as in Chapter 8. This equation is written in levels. After the within transformation,  $\alpha_i$  and  $\lambda_t$  drop out, and we work with within-transformed residuals.

In practice, we estimate the factor components under the maintained factor-model assumptions, residualise outcomes (and, in some designs, treatments and controls) with respect to the estimated factor structure, and then apply regularisation to the residualised variables. This combines factor flexibility for unobserved shocks with covariate parsimony for observed controls.

## Dependence Structures in Panels

Panel data exhibit dependence across both units and time. This matters for regularisation in two ways. First, penalty choice must account for dependence to avoid selecting too many or too few controls. Second, post-selection inference must adjust for dependence to avoid understating uncertainty.

**Clustering by unit.** The most common assumption is that errors are independent across units but arbitrarily correlated within units over time. This is plausible when units are distinct entities (stores, DMAs, customers) that do not strongly interact, and when within-unit correlation arises from persistent unobserved characteristics or serially dependent shocks. Cluster-robust variance estimators aggregate within-unit residuals to account for this serial dependence, providing valid standard errors without specifying the exact form of the time-series process.

Penalty choice under unit clustering should respect this structure. In practice, we either use cluster-level theoretical penalties that scale with  $\sqrt{\log(p)/G}$ , where  $G$  is the number of independent clusters, or blocked cross-validation that leaves whole units out of fold, so that validation error reflects out-of-cluster prediction rather than reusing serially correlated observations. Blocked cross-validation, described in Section 13.8, partitions units into folds and ensures that all observations for a given unit are either in the training set or in the validation set, never split across folds. The effective sample size for tuning is the number of independent clusters  $G$ , not the total number of cell observations  $NT$ .

Clustering by time is appropriate when dependence is primarily cross-sectional within periods (for example, due to common shocks or spatial spillovers) and when the number of periods  $T$  is large relative to  $N$ . Errors are assumed to be independent across periods but arbitrarily correlated across units within periods. Cluster-robust variance estimators then aggregate within-period residuals, and penalty choice uses time-blocked cross-validation that partitions periods into folds.

**Two-way clustering.** When both within-unit serial correlation and within-period cross-sectional dependence are material, two-way clustering aggregates residuals along both dimensions. Two-way cluster-robust variance estimators, described in Chapter 16, combine unit- and time-cluster covariance matrices using inclusion–exclusion formulas. They are often more conservative (yielding wider intervals) but more robust to complex dependence structures. Cross-validation under two-way clustering is more challenging, as folds must block both units and time, reducing the effective validation sample and increasing computational cost. Two-

way blocked cross-validation that leaves out entire unit-time rectangles is often too data-hungry in marketing panels.

In that case, tuning under unit blocking is a pragmatic compromise, but you should treat unstable selection and large time-cluster correlations as a warning that the chosen penalty may be too small. For inference, you can still rely on two-way CRVE, as discussed in Chapter 16.

## Implications for Selection and Inference

Regularisation and post-selection procedures must respect these dependence structures. Naive lasso that treats all  $(i, t)$  observations as independent under-penalises in the presence of strong within-cluster correlation: the effective noise level is mis-estimated, so the chosen penalty is too small, leading to overly complex nuisance models, which can inflate variance and destabilise post-selection or orthogonal-score estimators. Naive standard errors that ignore clustering underestimate uncertainty, producing intervals that are too narrow and p-values that are too small.

The remedy is twofold. First, use dependence-adjusted penalties—either via theoretical penalty formulas calibrated to the cluster dimension or via blocked cross-validation that respects the clustering scheme—to select covariates at an appropriate complexity level. Second, conduct inference with cluster-robust or two-way cluster-robust variance estimators and, when clusters are few, use wild cluster bootstrap procedures from Chapter 16.

These are variance and uncertainty problems. They do not repair omitted-variable bias if key pre-treatment confounders are missing from  $X_{it}$  or are aggressively shrunk towards zero. Identification still rests on the design assumptions from earlier chapters: conditional parallel trends for DiD, unconfoundedness for selection-on-observables designs, and factor-structure assumptions for synthetic control and matrix methods.

### 13.3 Regularisation Methods

This section presents regularisation methods adapted for panel data, including lasso, elastic net, and group lasso, and discusses how to choose the penalty parameter when observations are clustered.

Lasso (Least Absolute Shrinkage and Selection Operator) applies an  $\ell_1$  penalty to regression coefficients, encouraging a sparse model by shrinking many coefficients exactly to zero.

**Definition 13.1 (Lasso Estimator)** For the panel model  $Y_{it} = \tau D_{it} + X'_{it}\beta + \alpha_i + \lambda_t + \varepsilon_{it}$  with within-transformed variables  $\tilde{Y}_{it}, \tilde{D}_{it}, \tilde{X}_{it}$ , the lasso estimator solves

$$(\hat{\tau}^{\text{lasso}}, \hat{\beta}^{\text{lasso}}) = \arg \min_{\tau, \beta} \left\{ \frac{1}{NT} \sum_{i,t} (\tilde{Y}_{it} - \tau \tilde{D}_{it} - \tilde{X}'_{it}\beta)^2 + \eta \sum_{j=1}^p |\beta_j| \right\},$$

where  $\eta > 0$  is the regularisation parameter. The  $1/(NT)$  term is a scaling convention. Under clustered dependence, penalty choices and rates should be interpreted in terms of the number of independent clusters  $G$ , not the cell count  $NT$  (see Section 13.8). The penalty is applied only to the control coefficients  $\beta$ , not to the target parameter  $\tau$ , so that the treatment effect is not directly shrunk towards zero.

While lasso offers interpretability, it can be unstable when covariates are highly correlated, arbitrarily selecting one from a group. For causal targets, this instability matters when it changes the implied adjustment set enough to move  $\hat{\tau}$  materially across equally defensible penalties or folds.

**Definition 13.2 (Elastic Net Estimator)** The elastic net estimator combines  $\ell_1$  and  $\ell_2$  penalties:

$$(\hat{\tau}^{\text{enet}}, \hat{\beta}^{\text{enet}}) = \arg \min_{\tau, \beta} \left\{ \frac{1}{NT} \sum_{i,t} (\tilde{Y}_{it} - \tau \tilde{D}_{it} - \tilde{X}'_{it}\beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\},$$

where  $\lambda_1 > 0$  controls sparsity ( $\ell_1$  penalty) and  $\lambda_2 > 0$  controls stability ( $\ell_2$  penalty). The mixing weight  $\rho = \lambda_1/(\lambda_1 + \lambda_2) \in [0, 1]$  interpolates between ridge ( $\rho = 0$ ) and lasso ( $\rho = 1$ ).

**Definition 13.3 (Group Lasso Estimator)** For controls partitioned into  $M$  groups  $\mathcal{G}_1, \dots, \mathcal{G}_M$  with  $\bigcup_{m=1}^M \mathcal{G}_m = \{1, \dots, p\}$ , the group lasso estimator solves

$$(\hat{\tau}^{\text{group}}, \hat{\beta}^{\text{group}}) = \arg \min_{\tau, \beta} \left\{ \frac{1}{NT} \sum_{i,t} (\tilde{Y}_{it} - \tau \tilde{D}_{it} - \tilde{X}'_{it}\beta)^2 + \eta \sum_{m=1}^M \sqrt{|\mathcal{G}_m|} \|\beta_{\mathcal{G}_m}\|_2 \right\},$$

where  $\beta_{\mathcal{G}_m} = (\beta_j)_{j \in \mathcal{G}_m}$  is the coefficient subvector for group  $m$ , and  $\sqrt{|\mathcal{G}_m|}$  scales the penalty by group size. Group lasso sets entire groups to zero ( $\beta_{\mathcal{G}_m} = \mathbf{0}$ ) or retains all coefficients in a group.

Hierarchical group lasso imposes structure on nested groups, encouraging coarser groups to be selected before finer groups. For example, a hierarchical structure might require that main effects be included before interactions, or that lower-order lags be included before higher-order lags. The details depend on the specific hierarchical penalty you choose. In this chapter we use hierarchical group lasso as intuition for how grouping

can encode structure. If you need a particular constraint (for example, “interactions can enter only if corresponding main effects enter”), you should state the objective or constraint explicitly and treat it as a design choice.

### 13.3.1 Theoretical Foundations

The usefulness of regularisation for causal inference rests on formal properties of the design matrix and on sparsity of the true model, conditional on the identification assumptions discussed earlier in the chapter.

**Assumption 56 (Restricted Eigenvalue Condition (iid-style benchmark))** The design matrix  $\tilde{\mathbf{X}} \in \mathbb{R}^{NT \times p}$  of within-transformed controls satisfies a restricted eigenvalue (RE) condition with parameters  $(\kappa, c_0)$ : for all vectors  $\delta \in \mathbb{R}^p$  with  $\|\delta_{S^c}\|_1 \leq c_0 \|\delta_S\|_1$ , where  $S$  is the support of the true coefficient vector  $\beta_0$ ,

$$\frac{1}{NT} \|\tilde{\mathbf{X}}\delta\|_2^2 \geq \kappa \|\delta_S\|_2^2.$$

The constant  $\kappa > 0$  ensures that the design is not too collinear in directions relevant to the sparse solution. Intuitively, relevant controls cannot be almost perfectly replicated by linear combinations of irrelevant ones after fixed effects. This is the standard iid-style RE condition. In clustered panels, analogous conditions require additional assumptions on within-cluster dependence and cluster sizes, so we treat this statement as a benchmark regularity condition.

**Assumption 57 (Approximate Sparsity)** The true coefficient vector  $\beta_0$  is approximately  $s$ -sparse: most of the signal is concentrated on an effective support of size  $s$ , with remaining coefficients small enough that their aggregate contribution is negligible. Formally, there exists a set  $S \subseteq \{1, \dots, p\}$  with  $|S| = s$  such that  $\|\beta_{0,S^c}\|_1 \leq c \|\beta_{0,S}\|_1$  for a small constant  $c$ , and the effective sparsity level satisfies

$$s^2 \frac{\log p}{n_{\text{eff}}} \rightarrow 0 \quad \text{as } n_{\text{eff}} \rightarrow \infty,$$

where  $n_{\text{eff}}$  denotes the effective number of independent observations. In iid settings  $n_{\text{eff}} = NT$ ; in clustered panels with  $G$  independent clusters (for example, units),  $n_{\text{eff}} = G$ , giving the condition  $s^2 \log p/G \rightarrow 0$ . This approximate-sparsity formulation aligns with the DML literature and with the earlier discussion of nuisance-function sparsity: what matters is that the bulk of the predictive signal comes from a modest number of covariates, not that all other coefficients are exactly zero.

**Theorem 13.1 (Lasso Estimation Error—iid Benchmark)** *Under Assumptions 56 and 57 with iid observations ( $n_{\text{eff}} = NT$ ), and for penalties  $\eta \asymp \sigma \sqrt{\log p/(NT)}$ , the lasso estimator satisfies*

- (i)  $\ell_2$  error:  $\|\hat{\beta}^{\text{lasso}} - \beta_0\|_2 \leq C_1 \eta \sqrt{s}/\kappa$ .
- (ii)  $\ell_1$  error:  $\|\hat{\beta}^{\text{lasso}} - \beta_0\|_1 \leq C_2 \eta s/\kappa$ .
- (iii) Prediction error:  $\frac{1}{NT} \|\tilde{\mathbf{X}}(\hat{\beta}^{\text{lasso}} - \beta_0)\|_2^2 \leq C_3 \eta^2 s/\kappa$ .

for constants  $C_1, C_2, C_3$  that do not depend on  $N, T, p$  [see Bickel et al., 2009, Bühlmann and van de Geer, 2011, for detailed proofs]. The familiar rate  $\|\hat{\beta}^{\text{lasso}} - \beta_0\|_2 = O_P(\sqrt{s \log p / (NT)})$  follows.

In clustered panels, these iid results do not apply directly. When observations within a unit are serially correlated, the effective sample size for high-dimensional asymptotics is the number of independent clusters  $G$ , not the cell count  $NT$ . Heuristically, replacing  $NT$  by  $G$  in the penalty formula  $\eta \asymp \sigma \sqrt{\log p / G}$  and in the rate bounds yields analogous convergence guarantees when clusters are independent and of comparable size. In practice, we treat these clustered-panel rates as guiding heuristics for penalty scaling and rely on blocked cross-validation and cluster-robust inference to handle dependence. The formal theory for general clustered high-dimensional panels remains an active research area.

**Theorem 13.2 (Support Recovery (Strong Conditions, Rarely Plausible in Marketing Panels))**  
*If, in addition to Assumptions 56 and 57, the irrepresentability condition*

$$\|\tilde{\mathbf{X}}'_{S^c} \tilde{\mathbf{X}}_S (\tilde{\mathbf{X}}'_S \tilde{\mathbf{X}}_S)^{-1}\|_\infty < 1 - \delta$$

*holds for some  $\delta > 0$ , then lasso recovers the true support with high probability:*

$$P(\text{supp}(\hat{\beta}^{\text{lasso}}) = \text{supp}(\beta_0)) \rightarrow 1 \quad \text{as } n_{\text{eff}} \rightarrow \infty.$$

*This oracle-style property is a useful theoretical benchmark but is typically stronger than what is required for valid treatment-effect estimation in practice. Many causal applications rely on good prediction of the nuisance functions rather than exact recovery of all non-zero coefficients. In marketing panels with many correlated controls—for example, overlapping lag structures, regional indicators, and category dummies—the irrepresentability condition is rarely plausible. We do not use support recovery as a diagnostic criterion in this book. Instead, we use lasso and related methods for prediction of nuisance functions and combine them with double-selection or DML procedures that are robust to imperfect variable selection.*

## Penalty Choice under Dependence

Penalty choice under dependence requires adapting theoretical penalty formulas or cross-validation to account for clustering or serial correlation. For independent data, theoretical penalties for lasso often take the form

$$\eta \approx \Phi^{-1}(1 - a/(2p)) \cdot \frac{\sigma}{\sqrt{n_{\text{eff}}}},$$

where  $\Phi^{-1}$  is the inverse normal CDF,  $a$  a nominal level, and  $n_{\text{eff}}$  the effective sample size [see Bickel et al., 2009, Bühlmann and van de Geer, 2011, for derivations]. In clustered panels,  $n_{\text{eff}}$  should be taken as the number of independent clusters  $G$ , not the raw cell count  $NT$ . In practice,  $\sigma$  is unknown and must be estimated using a cluster-robust method consistent with the chosen dependence structure—for example, by computing residual variance from cluster-aggregated residuals rather than cell-level residuals. Theory-based

formulas should be complemented by blocked cross-validation because the true dependence strength and scale can vary across applications.

## Blocked Cross-Validation

Blocked cross-validation partitions units (or periods) into folds, ensuring that all observations from each unit (or period) are either in the training set or the validation set. For unit-blocked cross-validation with  $K$  folds, partition the  $N$  units into  $K$  groups, train the lasso on  $K - 1$  groups (using all time periods for those units), predict outcomes for the held-out group, and compute the validation error. Repeat for all folds and select the penalty  $\eta$  that minimises the average validation error.

Unit-blocking respects within-unit dependence by preventing observations from the same unit from appearing in both the training and validation sets. Time-blocking respects within-period dependence by partitioning periods rather than units. Two-dimensional blocking respects both forms of dependence but is computationally expensive and reduces the effective validation set size.

A critical implementation detail: all preprocessing—including standardisation, imputation, within transformation parameters, factor estimation, and any demeaning or residualisation steps—must be performed inside each training fold only, with the resulting parameters then applied to the corresponding validation fold. Computing preprocessing statistics on the full sample before splitting leaks information from validation to training, breaking the cross-fitting logic used in DML and inflating apparent predictive performance.

## Practical Penalty Grids

Practical penalty grids for cross-validation use a sequence of penalties  $\eta_1 > \eta_2 > \dots > \eta_L$  spaced logarithmically (for example,  $\eta_\ell = \eta_{\max} \cdot r^{\ell-1}$ , where  $\eta_{\max}$  is the smallest penalty that shrinks all coefficients to zero and  $r \in (0, 1)$  is a decay rate, typically  $r = 0.9$  or  $r = 0.95$ ).

For each  $\eta_\ell$ , compute the lasso solution and the validation error. Plot the validation error against  $\eta$  (the penalty path), identify the penalty that minimises validation error ( $\eta_{\min}$ ), and report estimates for  $\eta_{\min}$  and for a slightly larger penalty that produces a sparser model within one standard error of the minimum ( $\eta_{1se}$ ). The  $\eta_{1se}$  rule trades a small increase in validation error for a simpler, more interpretable model, aligning with the principle that parsimony is valuable for causal inference when predictive performance is comparable.

## Standardisation and Preprocessing

Standardising controls before regularisation is essential. Because the penalties are scale-dependent, covariates with different scales can be penalised unevenly. We standardise the within-transformed controls  $\tilde{X}_{it}$  by dividing each covariate by its standard deviation, ensuring all are on the same scale. The resulting coefficients

are transformed back to their original scale for interpretation. When using cross-validation, standardisation must be computed within each training fold and then applied to the corresponding validation fold to prevent data leakage.

## Interactions and Lags

Handling interactions and lags requires care to respect hierarchical structure and to avoid constructing an unmanageably large design matrix. Interactions between treatment and key covariates (for example,  $D_{it} \times X_{j,it}$ ) capture effect modification and should be included when heterogeneity is of substantive interest. Group lasso can select interactions jointly with main effects, ensuring that an interaction enters only if its main effects are included.

Lags of covariates and treatment (for example,  $X_{j,i,t-1}, X_{j,i,t-2}, \dots, X_{j,i,t-L}, D_{i,t-1}, \dots$ ) capture dynamic relationships and distributed effects. Group or hierarchical lasso can select all lags of a covariate jointly, or enforce that lower-order lags are included before higher-order ones. Chapter 10 provides comprehensive coverage of distributed lags and dynamic effects in panels. In practice, theory and prior evidence should guide which families of interactions and lags are entered into the candidate library before regularisation is applied; blindly generating all possible combinations can overwhelm the data even with strong penalties.

A non-negotiable design constraint applies to all interaction and lag construction: the candidate library must contain only pre-treatment covariates and, when the estimand permits, pre-treatment values of the treatment variable. Post-treatment variables must never enter the candidate set, regardless of how they are generated. No algorithm should be allowed to create or select post-treatment covariates, mediators, or colliders. This constraint is enforced by construction—by restricting the input data to pre-treatment periods or pre-treatment covariate values—not by hoping the lasso will exclude problematic variables. See Section 13.1 and Chapter 4 for the identification logic underlying this requirement.

### 13.4 Variable Selection for Causal Targets

Variable selection for causal inference differs from selection for prediction because the target is a causal estimand (for example ATT or an event-time effect  $\theta_k$ ), identified under a specific design. In selection-on-observables settings, identification rests on unconfoundedness with overlap. In DiD and event studies, identification rests on (conditional) parallel trends in untreated potential outcomes. Regularisation helps you estimate the nuisance components and choose among many candidate controls within those designs. It does not create identification.

This section presents two key approaches: double selection, which builds a control set from both outcome and treatment models, and orthogonalised machine learning, which formalises the same principle through Neyman-orthogonal scores.

Double selection addresses a critical problem. A naive lasso regression of an outcome on a treatment and controls may omit important confounders if those variables strongly predict treatment but only weakly predict the outcome. Excluding such variables reintroduces omitted-variable bias even when the lasso is excellent at forecasting outcomes.

Double selection corrects this by selecting controls from both the outcome model and the treatment model and including the union of selected controls in the final regression [Belloni et al., 2014].

**Definition 13.4 (Double Selection)** For the within-transformed model, the double-selection procedure for estimating  $\tau$  proceeds in three steps:

1. **Outcome model selection:** run a lasso of  $\tilde{Y}_{it}$  on  $\tilde{X}_{it}$  (excluding treatment),

$$\hat{\beta}^Y = \arg \min_{\beta} \left\{ \frac{1}{NT} \sum_{i,t} (\tilde{Y}_{it} - \tilde{X}'_{it}\beta)^2 + \eta^Y \|\beta\|_1 \right\},$$

and let  $S^Y = \{j : \hat{\beta}_j^Y \neq 0\}$  denote the selected controls.

2. **Treatment model selection:** run a lasso of  $\tilde{D}_{it}$  on  $\tilde{X}_{it}$ ,

$$\hat{\gamma}^D = \arg \min_{\gamma} \left\{ \frac{1}{NT} \sum_{i,t} (\tilde{D}_{it} - \tilde{X}'_{it}\gamma)^2 + \eta^D \|\gamma\|_1 \right\},$$

and let  $S^D = \{j : \hat{\gamma}_j^D \neq 0\}$  denote the selected controls.

3. **Post-selection regression:** run OLS of  $\tilde{Y}_{it}$  on  $\tilde{D}_{it}$  and the union of selected controls,

$$(\hat{\tau}^{\text{DS}}, \hat{\beta}_S^{\text{DS}}) = \arg \min_{\tau, \beta_S} \sum_{i,t} (\tilde{Y}_{it} - \tau \tilde{D}_{it} - \tilde{X}'_{S,it}\beta_S)^2,$$

where  $S = S^Y \cup S^D$  and  $\tilde{X}_{S,it}$  contains only the selected controls.

The union  $S^Y \cup S^D$  is designed to capture variables that predict the outcome (improving precision), variables that predict the treatment (reducing confounding), and variables that predict both. Under approximate sparsity and suitable regularity conditions, this construction yields valid inference for  $\tau$  under the maintained

identification assumptions for the design (for example, unconfoundedness conditional on the candidate pre-treatment library, plus overlap), even though the individual lasso models are tuned for prediction.

A critical identification constraint applies: the candidate covariate library  $X_{it}$  must be restricted to pre-treatment variables that are not descendants of  $D_{it}$  in the causal graph from Chapter 2. The algorithm chooses weights and selection within this library, but it never decides which variables are causally admissible. Including post-treatment variables, mediators, or colliders in the candidate set can introduce bias that no amount of regularisation or double selection can repair.

**Theorem 13.3 (Double-Selection Asymptotic Normality)** *Let  $S_0^Y = \text{supp}(\beta_0^Y)$  and  $S_0^D = \text{supp}(\gamma_0^D)$  denote the true supports for the outcome and treatment models. Suppose:*

- (i) *the RE and approximate-sparsity conditions from Assumptions 56–57 hold for both models, with effective sample size given by the number of independent clusters  $G$ .*
- (ii) *penalties satisfy  $\eta^Y, \eta^D \asymp \sqrt{\log p/G}$ , chosen via blocked cross-validation or theory-based formulas.*
- (iii) *the lasso estimators approximate the true nuisance functions well in the sense that prediction errors for the outcome and treatment models shrink at the usual high-dimensional rates.*
- (iv) *overlap holds: there exist  $0 < \underline{e} < \bar{e} < 1$  such that  $\underline{e} \leq \mathbb{P}(D_{it} = 1 \mid X_{it}, \alpha_i, \lambda_t) \leq \bar{e}$  for all  $(i, t)$  in the estimation sample.<sup>1</sup>*

*Under these conditions, the post-double-selection estimator is asymptotically normal [Belloni et al., 2014],*

$$\sqrt{G}(\hat{\tau}^{\text{DS}} - \tau_0) \xrightarrow{d} \mathcal{N}(0, V^{\text{DS}}),$$

*where  $V^{\text{DS}}$  is the asymptotic variance accounting for clustering. The  $\sqrt{G}$ -normal limit is derived under specific regularity conditions on cluster sizes, dependence, and sparsity that are only approximately met in finite marketing panels; in practice, we rely on cluster-robust standard errors and diagnostics to assess reliability. The union  $S^Y \cup S^D$  approximates the set of confounders—variables that jointly affect  $Y$  and  $D$ —well enough that the remaining bias in  $\hat{\tau}^{\text{DS}}$  is asymptotically negligible.*

In practice we rarely expect lasso to recover the exact support of the true model. What matters is that the selected controls deliver good prediction of the nuisance functions; exact support recovery is a stringent benchmark rather than a requirement for valid causal inference.

## Orthogonalised Machine Learning

Orthogonalised machine learning provides a formal framework for the logic underlying double selection. As detailed in Chapter 12, DML uses ML or regularised regressions to estimate nuisance functions—outcome regressions and propensity scores—and then combines them into Neyman-orthogonal scores. Orthogonality

---

<sup>1</sup> The overlap condition is stated in terms of the full covariate vector and fixed effects. Because  $S$  is data-dependent (it depends on the lasso outcome), formal proofs work with overlap on the full conditioning set and then show that selection preserves it under approximate sparsity. See Belloni et al. [2014] for details.

makes the resulting estimator first-order insensitive to small errors in the nuisances, which is precisely what we need when nuisances are estimated in high dimensions. Double selection can be seen as constructing an approximate orthogonal score via partialling-out, while general DML uses influence-function-based scores that are exactly Neyman-orthogonal.

Using lasso or elastic net to estimate nuisance functions makes DML computationally feasible and stable in panels with rich  $X_{it}$ . To keep the discussion concrete, we illustrate an orthogonal-score construction for the cell-level ATT as defined in Chapter 2. In panels with dependence, you should treat the independent sampling unit as the cluster (often the unit  $i$ ) and aggregate score contributions within clusters when forming standard errors.

A typical orthogonalised lasso+DML procedure proceeds as follows. Partition clusters into  $K$  folds using blocked cross-fitting to respect dependence. For staggered adoption and DiD-style designs, folds must also respect cohorts and pre/post timing so that no fold mixes treated and control observations in ways that violate the identification window. See Section 13.6 for details. For each fold  $k$ :

Train outcome and propensity-score lasso models on the  $K - 1$  training folds, selecting penalties via blocked cross-validation. Evaluate these models on fold  $k$  to produce predictions  $\hat{\mu}_{it}^{(-k)}(0, X_{it})$  (the predicted outcome under control) and  $\hat{e}_{it}^{(-k)}(X_{it}, \alpha_i, \lambda_t)$  (the predicted propensity score), where the superscript  $(-k)$  indicates that the cluster containing unit  $i$  is held out from training.

Construct the doubly robust ATT score for observations in fold  $k$ , using the same orthogonal score as in Section 12.2, for example

$$\psi_{it,k}^{\text{ATT}}(a) = \frac{D_{it}}{\pi} (Y_{it} - \hat{\mu}_{it}^{(-k)}(0, X_{it}) - a) - \frac{\hat{e}_{it}^{(-k)}(X_{it}, \alpha_i, \lambda_t)(1 - D_{it})}{\pi(1 - \hat{e}_{it}^{(-k)}(X_{it}, \alpha_i, \lambda_t))} (Y_{it} - \hat{\mu}_{it}^{(-k)}(0, X_{it})),$$

where  $\pi = \mathbb{P}(D_{it} = 1)$  is the marginal treatment probability. In practice,  $\pi$  is estimated by the sample mean of  $D_{it}$ , and inference treats clusters as the independent sampling unit by aggregating score contributions within clusters and applying a  $\sqrt{G}$ -based cluster-robust variance estimator. Aggregate the scores across folds and solve the sample moment equation for  $a$  to obtain  $\widehat{\text{ATT}}$ . This score matches the doubly robust ATT score derived in Section 12.2 and in Sant'Anna and Zhao [2020].

This procedure combines the robustness of DML (orthogonality + cross-fitting) with the parsimony of regularisation (lasso selects a small set of controls in each nuisance model). Conceptually, it extends the double-selection principle: both outcome and treatment models matter, but their influence on the final estimator enters only through an orthogonal score.

## Integration with Staggered DiD

Integration with staggered DiD (Chapter 4) requires aligning selection and nuisance estimation with the cohort-time structure. For cohort  $g$  and period  $t \geq g$ , estimate outcome and treatment lasso models using only pre-treatment periods  $s < g$  for cohort- $g$  units, and using not-yet-treated or never-treated units as comparisons in period  $t$ . Use blocked cross-validation confined to these pre-periods to choose penalties.

Select controls for a given cohort  $g$  and period  $t$  based on their predictive power in the pre-treatment sample, ensuring that post-treatment information does not leak into nuisance estimates. Under no circumstances should post-treatment periods for cohort  $g$  enter the nuisance training sample for  $\tau(g, t)$ , including via cross-validation and hyperparameter tuning. Selection and tuning should be carried out using only the cohort-specific pre-treatment window used to support conditional parallel trends. Construct doubly robust scores for  $\tau(g, t)$  under conditional parallel trends given pre-treatment covariates, then aggregate across cohorts to obtain event-time effects  $\theta_k$  as defined in Chapter 5. This preserves the parallel-trends logic while enabling flexible, high-dimensional covariate adjustment.

## Balancing Weights and Riesz Regression

Balancing weights such as entropy balancing and stable balancing weights were introduced earlier in the book as tools for constructing comparison groups that match treated units on observed covariates. Recent work in debiased ML shows that many of these weighting schemes can be viewed as special cases of Riesz regression [see, for example, Kato, 2025a,b].

In average-treatment-effect problems, the Riesz representer for the target functional can often be written as a density ratio between treated and control covariate distributions. Squared-loss Riesz regression coincides with least-squares density-ratio estimation. Kullback–Leibler-type losses lead, via duality, to the entropy-balancing and covariate-balancing propensity-score problems studied by Hainmueller [2012], Zubizarreta [2015], Zhao [2019]. Chapter 12 discusses these connections in the context of DML.

For applied work the message is that there is no sharp divide between “regression-based” and “balancing-based” estimators. When you fit Riesz regressions inside DML, you are implicitly choosing a weighting scheme through the loss, basis functions, and regularisation used for the weight model. Conversely, when you solve an entropy-balancing problem, you can view the resulting weights as a particular Riesz-regression estimate that targets an orthogonal score. This unifying perspective makes it easier to compare results across methods and to design diagnostics that focus on estimated weights—their dispersion, balance, and influence—rather than on whether they came from a lasso, an entropy-balancing routine, or a named balancing method.

The equivalence is at the level of optimisation problems for a chosen function class and loss. It does not imply that different balancing and regression procedures are interchangeable in finite samples.

All these weighting and regression schemes still rely on the identification assumptions from earlier chapters: unconfoundedness conditional on covariates and adequate overlap. Weights reorganise estimation—they can improve efficiency and reduce sensitivity to functional-form choices—but they cannot create identification where none exists. If key confounders are omitted from  $X_{it}$ , no weighting scheme will recover the true ATT.

### 13.5 Post-Selection and Debiased Inference

Inference after selection is not straightforward. Because the same data are used both to select controls and to estimate coefficients, treating the selected model as if it were pre-specified typically understates uncertainty. This section presents methods that correct for this, including debiased (or desparsified) lasso, and discusses robust inference for panel data.

A naive approach runs lasso to select controls and then runs OLS on the selected model, treating it as fixed. Standard OLS variance formulas then ignore both the shrinkage bias and the extra variability from the selection step, producing confidence intervals that are too narrow and p-values that are too small. The selection process itself introduces uncertainty, and inference must account for it.

Debiased (or desparsified) lasso constructs bias-corrected estimators that are asymptotically normal under approximate sparsity and regularity conditions. The key idea is to adjust the lasso estimate by a term that counteracts the shrinkage induced by the  $\ell_1$  penalty.

**Definition 13.5 (Debiased Lasso)** Consider the within-transformed model  $\tilde{Y}_{it} = \tau \tilde{D}_{it} + \tilde{X}'_{it}\beta + \tilde{\varepsilon}_{it}$ , and let  $(\hat{\tau}^{\text{lasso}}, \hat{\beta}^{\text{lasso}})$  denote the lasso solution. The debiased (or desparsified) lasso estimator for  $\tau$  corrects for regularisation bias using the residualisation residuals  $\hat{r}_{it}$  (defined below):

$$\hat{\tau}^{\text{deb}} = \hat{\tau}^{\text{lasso}} + \frac{1}{\frac{1}{NT} \sum_{i,t} \hat{r}_{it}^2} \cdot \frac{1}{NT} \sum_{i,t} \hat{r}_{it} (\tilde{Y}_{it} - \hat{\tau}^{\text{lasso}} \tilde{D}_{it} - \tilde{X}'_{it} \hat{\beta}^{\text{lasso}}).$$

The second term corrects for the regularisation bias introduced by the lasso penalty. In more general multivariate settings, debiased lasso can be written using an estimated precision matrix. We do not construct that object explicitly here. See Zhang and Zhang [2014], Javanmard and Montanari [2014], van de Geer et al. [2014] for the formal constructions.

**Definition 13.6 (Residualisation of Treatment)** The residuals  $\hat{r}_{it}$  used in the debiasing correction are constructed by residualising the within-transformed treatment on within-transformed controls using lasso:

$$\hat{\gamma} = \arg \min_{\gamma} \left\{ \frac{1}{NT} \sum_{i,t} (\tilde{D}_{it} - \tilde{X}'_{it}\gamma)^2 + \eta_{\text{node}} \|\gamma\|_1 \right\},$$

and defining  $\hat{r}_{it} = \tilde{D}_{it} - \tilde{X}'_{it}\hat{\gamma}$ . These residuals represent the variation in treatment that is orthogonal to the high-dimensional controls. The corresponding sparsity condition—that the regression of treatment on controls is itself approximately sparse—connects to the treatment-model sparsity in the double-selection framework and ensures that  $\hat{r}_{it}$  is a good approximation to the population residual.

Under approximate sparsity—both in the outcome model and in this residualisation regression—the debiased estimator enjoys asymptotic normality.

**Theorem 13.4 (Debiased Lasso Inference—iid Benchmark)** Suppose the RE and approximate-sparsity conditions from Assumptions 56–57 hold with iid observations, the relevant residualisation regression is approximately sparse, and  $s^2 \log p/(NT) \rightarrow 0$  as  $NT \rightarrow \infty$ . Then the debiased lasso estimator satisfies

$$\sqrt{NT}(\hat{\tau}^{\text{deb}} - \tau_0) \xrightarrow{d} \mathcal{N}(0, V^{\text{deb}}),$$

where  $V^{\text{deb}}$  is the asymptotic variance determined by the long-run variance of  $\tilde{\varepsilon}_{it}\hat{r}_{it}$  [see Zhang and Zhang, 2014, Javanmard and Montanari, 2014, van de Geer et al., 2014, for detailed proofs].

Here we treat the within-transformed cell-level observations as iid for expositional purposes. In genuine panel settings with serial dependence, this benchmark result motivates but does not fully characterise the clustered-panel case discussed below.

Do not use iid scaling such as  $\sqrt{NT}$  or iid penalty formulas in clustered panels. Use  $G$ -based scaling and blocked validation as discussed in Section 13.8.

In clustered panels, these iid results do not apply directly. When observations within a unit are serially correlated, the effective sample size for high-dimensional asymptotics is the number of independent clusters  $G$ , not the cell count  $NT$ . Heuristically, replacing  $NT$  by  $G$  in the rate condition  $s^2 \log p/G \rightarrow 0$  and in the CLT yields a  $\sqrt{G}$ -normal approximation, but formal proofs require additional regularity conditions on within-cluster dependence and cluster sizes. In practice, we treat these clustered-panel rates as guiding heuristics and rely on cluster-robust variance estimators and wild cluster bootstrap to handle dependence. The formal theory for debiased inference in general clustered high-dimensional panels remains an active research area.

## Cluster-Robust Inference

Cluster-robust inference for debiased and post-double-selection estimators accounts for within-cluster dependence by aggregating score contributions within clusters. Let clusters be indexed by  $c = 1, \dots, G$ , with cluster membership sets  $\mathcal{C}_c$  as in the notation guide. Define the cluster-level influence contribution as

$$\Psi_c = \sum_{i \in \mathcal{C}_c} \sum_{t=1}^T \hat{r}_{it} \hat{\varepsilon}_{it} / \left( \frac{1}{NT} \sum_{i,t} \hat{r}_{it}^2 \right),$$

where  $\hat{\varepsilon}_{it} = \hat{Y}_{it} - \hat{\tau}^{\text{deb}} \hat{D}_{it} - \hat{X}'_{it} \hat{\beta}^{\text{lasso}}$  are residuals and  $\hat{r}_{it}$  are residualisation residuals. The cluster-robust variance estimator is then

$$\hat{V}_{\text{cluster}}^{\text{deb}} = \frac{1}{G-1} \sum_{c=1}^G (\Psi_c - \bar{\Psi})^2,$$

where  $\bar{\Psi} = \frac{1}{G} \sum_{c=1}^G \Psi_c$ . The cluster-robust standard error for  $\hat{\tau}^{\text{deb}}$  is then

$$\hat{\sigma}_{\text{cluster}}^{\text{deb}} = \sqrt{\hat{V}_{\text{cluster}}^{\text{deb}}/G},$$

so that  $\sqrt{G}(\hat{\tau}^{\text{deb}} - \tau_0)$  is approximated by  $\mathcal{N}(0, \hat{V}_{\text{cluster}}^{\text{deb}})$ . This expression mirrors the cluster-robust influence-function formulas from Chapter 16, where  $\Psi_c$  plays the same role as the cluster-level influence contribution in CRVE. For two-way clustering (by unit and time), sum influence contributions along both dimensions

and subtract the overlap, as described in Chapter 16. For post-double-selection OLS, standard cluster-robust variance estimators applied to the selected model provide inference under the same cluster structure.

## Wild Cluster Bootstrap

When the number of clusters  $G$  is small, asymptotic approximations—even with clustered variances—can be fragile. Wild cluster bootstrap methods provide a finite-sample alternative. The general idea is to treat the estimated cluster-level influence contributions as pseudo-observations, reweight them with random multipliers (for example, Rademacher weights), recompute the estimator under each bootstrap draw, and use the empirical distribution of these bootstrap estimates to form confidence intervals.

In post-selection settings, a conservative but robust approach is to re-run the full selection plus estimation procedure (lasso, double selection, or debiased lasso) within each bootstrap replication. This accounts for selection uncertainty at the cost of increased computation. Re-running selection and debiasing inside each bootstrap replication can be very expensive in large panels; practitioners should consider reducing the candidate set  $X_{it}$  using theory and design constraints before applying fully iterated bootstrap procedures. Chapter 16 discusses practical variants (Rademacher, Mammen, Webb weights) and when wild cluster bootstrap is preferable to purely analytical cluster-robust intervals. This is a pragmatic approach when  $G$  is small, but it is computationally heavy and still relies on the maintained dependence structure for the chosen bootstrap scheme (Chapter 16).

## Small-Sample Cautions

Debiased lasso requires that the number of truly non-zero coefficients  $s$  is small relative to the effective sample size  $G$ ; formally,  $s^2 \log p/G \rightarrow 0$ . In marketing terms, approximate sparsity means that only a relatively small subset of the many candidate controls (for example, specific lags of prices, promotions, and search) materially affect the outcome, even if many coefficients are non-zero in practice. When  $G$  is small (for example, fewer than 30 clusters) or sparsity is only approximate and  $s$  is large, the debiasing correction can become unstable or biased. In such cases, double selection combined with cluster-robust standard errors, possibly supplemented by wild cluster bootstrap, is often more reliable.

Reporting estimates and confidence intervals from multiple methods—post-double selection, debiased lasso, and, where feasible, bootstrap-based intervals—helps gauge robustness. Agreement across methods increases confidence that results are not an artefact of a particular post-selection adjustment. Divergence signals that asymptotic approximations may be weak and that design or sparsity assumptions deserve closer scrutiny.

## High-Dimensional Granger Causality

Forecasting problems in high-dimensional panels often conceal a sharper predictive question. You may wish to know whether a particular signal has incremental predictive content once you condition on a very rich information set. In time-series language, this is the question of whether a process  $x_t$  Granger-causes an outcome  $y_t$  at horizon  $h$ . Granger non-causality is a predictive concept: rejection means the driver adds forecasting power beyond controls, not necessarily that it has a structural causal effect in the potential-outcomes sense of Chapter 2. The methods below are most naturally applied to aggregate or market-level time series; for full panels with a unit index  $i$ , one can either aggregate to a single series or run separate time-series models by unit and then aggregate results.

Transfer entropy provides a closely related language for directed predictive dependence. For two processes  $(x_t, y_t)$ , the  $h$ -step transfer entropy from  $x$  to  $y$  is the conditional mutual information

$$\text{TE}_{x \rightarrow y}(h) = I(\mathbf{x}_t; y_{t+h} | \mathcal{F}_t),$$

where  $\mathbf{x}_t$  and  $\mathbf{y}_t$  denote chosen lag windows and  $\mathcal{F}_t$  denotes the remaining conditioning information set. When  $\text{TE}_{x \rightarrow y}(h) = 0$ , the past of  $x$  adds no incremental predictive information about  $y_{t+h}$  once we condition on the chosen lags and  $\mathcal{F}_t$ . For Gaussian variables, transfer entropy and Granger causality are equivalent measures of this incremental predictability [Barnett et al., 2009].

In nonlinear or heavy-tailed settings, transfer-entropy estimators often rely on discretisation or other regularisation. Symbolic transfer entropy applies the same idea after mapping a series into a symbolic representation (for example, ordinal patterns). This can stabilise estimation, but it changes the target object and introduces tuning choices that should be treated as part of the analysis specification [Zaremba and Aste, 2014]. Time variation adds another complication. Local or time-varying Granger measures emphasise that directed predictive dependence can shift across regimes, which is useful for diagnosing instability even when it does not justify a structural causal claim [Stramaglia et al., 2021].

In high-dimensional settings this becomes a block-testing problem: testing whether a set of lag coefficients on the candidate driver is jointly zero once you include many other lags and covariates. Standard Wald and F tests fail after lasso selection because they ignore shrinkage and serial correlation. Naively treating the selected model as fixed produces p-values that are too small and a false sense of discovery.

Babii et al. [2024] show how to recover asymptotically valid Granger tests in this setting by combining debiasing with long-run variance estimation, under standard time-series regularity conditions (stationarity and weak dependence) and appropriately tuned HAC estimators. First, estimate a high-dimensional regression of  $y_{t+h}$  on lagged outcomes, lagged values of the candidate driver  $x_t$ , and other controls using lasso or sparse group lasso. Then construct a debiased version of the lasso coefficients for the block of interest and estimate their long-run covariance with a HAC estimator that respects time dependence. A Wald statistic based on the bias-corrected coefficients and the HAC covariance matrix converges to a chi-squared distribution under the null that  $x_t$  does not Granger-cause  $y_t$ .

In marketing panels the interpretation is predictive rather than structural. Suppose you run a weekly panel regression of brand sales on its own lags, lagged brand sentiment from Twitter, and a large set of

controls capturing prices, promotions, search activity, and macro shocks. A bias-corrected Wald test on the block of sentiment lags answers whether sentiment adds incremental forecasting power once you control for everything else. A non-rejection supports the view that sentiment is only a proxy for other drivers already in the model. A strong rejection—combined with stable coefficients and good diagnostics—supports the claim that sentiment helps forecast sales beyond what is explained by observed fundamentals, without itself proving a structural causal effect.

Even with debiasing and HAC, these tests answer a model-dependent predictive question: given the chosen lag structure and controls, does  $x_t$  add forecasting power for  $y_{t+h}$ ? They do not, by themselves, establish structural causality. Model mis-specification—omitted lags, incorrect lag length, misspecified dynamics—remains a serious risk. The test says “incremental predictive power given this model,” not “true non-causality” in any structural sense. In high-dimensional marketing settings, many signals (search, social, display) are highly collinear and respond to common shocks. Even with debiasing and HAC, Granger tests can attribute “incremental” predictive power to whichever signal survives selection, so interpret rejections as model- and design-specific, not as definitive causal evidence for one channel. Lag length choice and multiple testing are central practical risks. When you try many candidate drivers, many horizons  $h$ , or many lag lengths, some rejections will occur by chance. Treat Granger findings as exploratory unless you have a pre-specified lag structure and adjust for multiplicity using the tools from Chapter 16.

## 13.6 Panels with DiD/Event-Study and Many Controls

Integrating regularisation with DiD and event-study designs requires that we respect the logic of those methods. The treatment variation that identifies effects comes from staggered adoption and cohort-time contrasts, not from arbitrary shifts in high-dimensional controls. Regularisation must therefore avoid colliders, use only pre-treatment controls, and align selection with the cohort-time structure.

Sparse augmentation of DiD uses lasso to select a small set of controls to include in the regression. The goal is to justify a *Maintained* conditional parallel-trends assumption given a pre-treatment covariate library. Regularisation only helps estimate the adjustment. It does not make conditional parallel trends hold.

The augmented DiD specification with high-dimensional controls is

$$Y_{it} = \tau D_{it} + X'_{it}\beta + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $D_{it}$  is a binary treatment indicator,  $X_{it}$  are time-varying controls, and  $\beta$  are their coefficients. In high dimensions we apply lasso or related penalties to  $\beta$  after within-transforming the data, leaving  $\tau$  and the fixed effects unpenalised.

**Assumption 58 (Conditional Parallel Trends)** For units in cohort  $g$  and comparison units  $\mathcal{C}$ , there exists a (possibly unknown) low-dimensional index set  $S_0 \subseteq \{1, \dots, p\}$  with  $|S_0| = s \ll p$  such that conditional parallel trends hold approximately given  $X_{S_0}$ , as in the conditional parallel-trends assumption of Chapter 4. Formally, for pre-treatment periods  $t' < t < g$ ,

$$\mathbb{E}[Y_{it}(0) - Y_{it'}(0) | G_i = g, X_{S_0, it}] \approx \mathbb{E}[Y_{jt}(0) - Y_{jt'}(0) | j \in \mathcal{C}, X_{S_0, jt}].$$

Here  $X_{S_0, it}$  can be interpreted as baseline covariates or time-varying covariates measured strictly before treatment for the estimand at hand. The controls in  $S_0$  are sufficient *under the maintained causal model* to render untreated outcome changes comparable across treated and comparison units. Lasso and double-selection procedures attempt to approximate  $S_0$  under the same approximate-sparsity conditions as in Section 13.7.

Combined with the sparsity and RE conditions from Section 13.3, this assumption allows us to replace the unknown high-dimensional conditioning set with a sparse approximation, without invalidating the DiD logic.

**Proposition 13.1 (Sparse-Augmented DiD Estimator)** *Under Assumption 58, the high-dimensional RE and sparsity conditions, and overlap conditions appropriate to staggered DiD, the sparse-augmented DiD estimator based on double selection satisfies*

$$\sqrt{G}(\hat{\tau}^{\text{sparse-DiD}} - \tau_0) \xrightarrow{d} \mathcal{N}(0, V^{\text{DiD}}),$$

where  $G$  is the number of independent clusters (for example, units) and  $V^{\text{DiD}}$  is the asymptotic variance that incorporates clustering. The estimator combines within-transformation (removing  $\alpha_i$  and  $\lambda_t$ ), lasso selection of controls (achieving parsimony), and post-selection OLS with cluster-robust inference.

In finite marketing panels with modest  $G$ , we rely on cluster-robust standard errors or wild cluster bootstrap intervals rather than treating the asymptotic CLT as exact. The  $\sqrt{G}$ -normal limit is derived under specific

*regularity conditions on cluster sizes, dependence, and sparsity that are only approximately met in practice. See Belloni et al. [2014] for related high-dimensional post-selection theory.*

The practical challenge is selecting which controls to include when  $p$  is large. A straightforward approach uses lasso on the within-transformed regression,

$$(\hat{\tau}^{\text{lasso}}, \hat{\beta}^{\text{lasso}}) = \arg \min_{\tau, \beta} \left\{ \sum_{i,t} (\tilde{Y}_{it} - \tau \tilde{D}_{it} - \tilde{X}'_{it} \beta)^2 + \eta \sum_{j=1}^p |\beta_j| \right\},$$

with  $\tilde{Y}_{it}$ ,  $\tilde{D}_{it}$ ,  $\tilde{X}_{it}$  denoting within-transformed variables and  $\eta$  chosen via unit-blocked cross-validation. To reduce sensitivity to any one lasso fit and to better capture confounders that primarily predict treatment, double selection—selecting controls from both the outcome and treatment models and then running OLS on their union—provides a more robust baseline, as in Section 13.4.

## Avoiding Colliders and Post-Treatment Variables

Double selection must respect identification. Including colliders or post-treatment variables in the candidate control set can introduce bias even if lasso is used. A collider is a variable affected by both treatment and outcome; conditioning on it induces spurious associations. A post-treatment variable is measured after treatment begins and may lie on the causal path from treatment to outcome.

The remedy is to pre-specify a set of pre-treatment controls that are plausibly exogenous—variables measured before treatment adoption that are not themselves affected by treatment or outcomes—and to exclude post-treatment or clearly endogenous variables from the lasso. The candidate library for  $X_{it}$  is fixed by design to pre-treatment, plausibly exogenous covariates. Regularisation chooses which of these to emphasise; it never introduces post-treatment or clearly endogenous variables. This pre-specification is a design decision that the algorithm is not allowed to override. No amount of cross-validation or lasso tuning can rescue a bad candidate set.

Lagged outcomes require the same discipline. Lags measured strictly before treatment onset for the relevant cohort can be admissible proxies for baseline dynamics. Lags measured after treatment begins for treated cohorts are post-treatment variables and must be excluded from the candidate library.

For staggered adoption, pre-treatment controls for cohort  $g$  are those measured in periods  $t < g$ . Outcome and treatment lasso models for cohort  $g$  should be trained only on pre-treatment data for cohort- $g$  units and on not-yet-treated or never-treated units as comparisons. Hyperparameter tuning (penalty selection) must also use only those pre-period windows, not combined pre+post data.

Transparent reporting of which controls are in the candidate set (pre-treatment, time-varying but plausibly exogenous) and which are excluded (post-treatment, endogenous, potential colliders) builds confidence that regularisation respects the DiD identification strategy.

## Aligning with Cohort–Time Structure

Aligning regularisation with the cohort–time structure of staggered DiD (Chapters 4 and 5) requires estimating cohort–time effects  $\tau(g, t)$  separately and allowing selection to vary by cohort where appropriate.

A typical procedure is as follows. For each cohort  $g$ , define the treated group ( $G_i = g$ ) and a comparison group of not-yet-treated or never-treated units. For each post-treatment period  $t \geq g$ , use pre-treatment data ( $s < g$ ) from cohort- $g$  units and comparison units to run double selection, identifying a sparse set of controls that stabilise pre-trends. Both selection and penalty tuning use only pre-period data appropriate for each cohort and estimand; no post-treatment observations for cohort  $g$  enter folds used to tune nuisances for  $\tau(g, t)$ . Then estimate  $\tau(g, t)$  by OLS of  $\tilde{Y}_{it}$  on  $\tilde{D}_{it}$  and the union of selected controls, with cluster-robust standard errors.

Overlap in this DiD context requires that, for each relevant cohort–time cell  $(g, t)$ , there are sufficient not-yet-treated or never-treated comparison observations and non-degenerate support of pre-treatment covariates in both treated and comparison groups. When overlap is weak, any conditional parallel-trends argument becomes fragile regardless of how controls are selected.

Aggregate  $\hat{\tau}(g, t)$  across cohorts using cohort weights to obtain event-time effects. For example, define weights  $w_g \propto \mathbb{P}(G_i = g | G_i < \infty)$  in the target population, normalised so that  $\sum_g w_g = 1$ , and set

$$\theta_k = \sum_g w_g \hat{\tau}(g, g + k),$$

where  $k = t - g$  is event time. This preserves the cohort–time identification structure while using regularisation to manage high-dimensional controls.

## Event-Time with Rich Covariates

Regularisation also stabilises event-time profiles when many controls are available. Event-study specifications estimate treatment effects at each lag  $k$  relative to adoption,

$$Y_{it} = \sum_{k \neq -1} \delta_k \mathbf{1}\{t - G_i = k\} + X'_{it} \beta + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $\delta_k$  are event-time regression coefficients and  $k = -1$  is the omitted reference period. This regression form is the classic TWFE event study. With heterogeneous effects, it does not in general identify the heterogeneity-robust event-time estimand  $\theta_k$  defined in earlier chapters. In this book, we treat heterogeneity-robust  $\theta_k$  as an aggregation of cohort–time estimates  $\hat{\tau}(g, t)$ .

When  $p$  is large, including all controls can produce noisy and unstable estimates of the event-time coefficients. Applying lasso to the control coefficients only,

$$(\hat{\delta}^{\text{lasso}}, \hat{\beta}^{\text{lasso}}) = \arg \min_{\delta, \beta} \left\{ \sum_{i,t} \left( \tilde{Y}_{it} - \sum_{k \neq -1} \delta_k \mathbf{1}\{t - G_i = k\} - \tilde{X}'_{it} \beta \right)^2 + \eta \sum_{j=1}^p |\beta_j| \right\},$$

produces sparser models (fewer controls) and smoother event-time profiles, while leaving the event-time dummy coefficients unpenalised. This is a deliberate design choice: event-time dummies are never penalised. Only the control coefficients  $\beta$  are subject to regularisation.

## Support-by-Event-Time Reporting

Support-by-event-time reporting assesses whether the set of selected controls is stable across lags, providing evidence about *selection stability*. For each lag  $k$ , report which controls are selected by lasso or double selection when estimating event-time effects. If the selected controls change dramatically across lags, it suggests that the outcome–covariate relationship is evolving over time, which may signal violations of the stability conditions needed for regularised adjustment.

Conditional parallel trends should still be assessed via pre-treatment leads, placebo checks, and other falsification diagnostics from Chapter 17. Control stability is not, by itself, evidence for parallel trends.

## Alignment with Dynamic Effects

Aligning regularisation with Chapter 10 (dynamic effects) and Chapter 5 (event studies) ensures that selection serves identification rather than undermining it. Dynamic effects estimate distributed lags of treatment—how effects evolve over time and how past treatments affect current outcomes. Group and hierarchical lasso can select treatment lags as blocks or in nested fashion, enforcing natural temporal hierarchies. Event studies estimate heterogeneity-robust event-time effects by comparing treated cohorts to appropriate controls at each lag; regularisation selects controls that stabilise these comparisons without biasing them.

The unifying theme is that regularisation provides flexibility in covariate adjustment while remaining subordinate to the design-based logic that delivers identification. High-dimensional tools should sharpen DiD and event-study designs, not replace them.

## 13.7 Assumptions

Causal identification using regularisation rests on assumptions about sparsity, overlap, dependence, stability, and design alignment. This section articulates these assumptions, discusses their implications, and connects them to the broader panel-data frameworks developed in earlier chapters.

Throughout, our notation for potential outcomes follows the book-wide conventions. When we write  $Y_{it}(0)$  and  $Y_{it}(1)$ , we are using the binary-treatment special case of the dynamic primitive  $Y_{it}(\underline{d}_i^t)$  from Chapter 2. In settings with dynamics, adoption timing, or carryover, the relevant causal model is still the path-dependent object, and regularised adjustment must be aligned with the identification restrictions that justify the short-hand (for example, no anticipation for staggered adoption).

**Assumption 59 (Approximate Sparsity After Partialling Fixed Effects or Factors)** After removing unit and time fixed effects (or latent factors), the high-dimensional nuisance objects needed by regularised estimators are approximately sparse.

In particular, we require approximate sparsity of the conditional mean of outcomes given controls (after partialling-out), and of the conditional mean of treatment given controls (after partialling-out). If we use bias-correction constructions such as debiased lasso, additional sparsity conditions may be needed for the corresponding residualisation or Riesz-type objects. In all cases, the effective complexity  $s$  must be small relative to the effective sample size  $n_{\text{eff}}$ . In iid panels  $n_{\text{eff}} = NT$ . In clustered panels  $n_{\text{eff}}$  is the number of independent clusters  $G$ .

Approximate sparsity is the key regularity condition that makes lasso-style methods useful in high dimensions. It asserts that, conditional on fixed effects or factors, only a modest subset of covariates is needed to approximate the nuisance functions. If the true model is dense, shrinking many coefficients to zero induces bias; if it is (approximately) sparse, regularisation can recover a good approximation to the required nuisance components.

Diagnostics cannot prove sparsity, but stability of selected sets across folds and robustness of estimates to modest changes in penalty strength can flag gross violations. Instability can also arise from collinearity among acceptable substitutes. Treat it as a warning flag, not a falsification.

**Assumption 60 (Overlap and Support for Treatment Given Controls)** The propensity score—the conditional probability of treatment given the admissible pre-treatment covariate library and fixed effects—is bounded away from zero and one:

$$0 < \underline{e} \leq \mathbb{P}(D_{it} = 1 | X_{it}, \alpha_i, \lambda_t) \leq \bar{e} < 1.$$

In other words, for units with similar covariate values and fixed effects, both treated and untreated observations occur with non-negligible probability.

Formally, overlap is a property of the data-generating process conditional on the full admissible covariate library and fixed effects. Approximate sparsity is an estimation device that can help approximate these

conditioning sets in finite samples. The underlying requirement remains that treated and control units are comparable given the true confounders.

Violations arise when treatment is nearly deterministic given high-dimensional  $X$  (for example, strict eligibility rules) or when selection fails to capture key predictors of treatment. Diagnostics include plotting estimated propensity scores for treated and control units, computing overlap statistics, and assessing covariate balance before and after adjustment. These diagnostics can reveal severe violations (for example, extreme propensities or wildly unstable selected sets) but they cannot prove that the assumptions hold.

Weak overlap calls for trimming extreme propensities and being explicit that the target estimand is restricted to the overlap region rather than the full-sample ATE or ATT (see Section 3.6 for the estimand discipline).

**Assumption 61 (Dependence-Aware Sampling and Penalty Choice)** Cross-validation, penalty choice, and—when applicable—cross-fitting respect the panel’s dependence structure. Folds partition units or periods (not individual observations), and penalties are scaled for the effective sample size  $n_{\text{eff}}$  (for example, the number of clusters  $G$  under unit clustering) rather than the raw cell count  $NT$ .

Dependence-aware sampling ensures that cross-validation provides honest estimates of out-of-sample performance and that penalties do not under-penalise (selecting too many controls and overfitting) or over-penalise (omitting important confounders). Violations occur when folds are defined by random observation splits that ignore within-unit or within-period dependence, or when penalty formulas are calibrated to  $NT$  instead of  $G$ .

Diagnostics include verifying fold construction, comparing penalties from blocked cross-validation to theory-based values, and examining the stability of selected controls across folds. In DML implementations, any cross-validation used to tune nuisance models is performed within each cross-fitting training fold to prevent information leakage from validation to held-out folds. When factor residualisation is used inside cross-fitting, factor estimation must also be performed using only the training fold.

**Assumption 62 (Stability and No Leakage from Post-Treatment Information)** The relationship between outcomes and controls is sufficiently stable across training and validation folds—and across pre-treatment subperiods—that controls selected on training data remain useful on validation data. For treated cohorts, control selection uses only pre-treatment observations; post-treatment outcomes and covariates that may be affected by treatment are excluded from the selection stage.

Stability supports the idea that a single sparse set of controls can be used across pre- and post-treatment periods for counterfactual segments. No leakage ensures that selection does not exploit post-treatment patterns, which would introduce collider or mediator bias. Violations occur when the outcome-control relationship changes sharply over time (for example, due to regime shifts or composition changes), or when controls are selected using post-treatment data for treated units. Diagnostics include comparing selected controls across early and late pre-treatment windows, checking the sensitivity of estimates to the pre-period length used for selection, and running placebo interventions with pseudo treatment dates. If selected controls differ dramatically across pre-treatment windows or across folds, practitioners should interpret that as evidence against stability and be cautious about leaning on regularised estimates.

**Assumption 63 (Correct Design Alignment with DiD/Event Studies and Factor Models)** When regularisation is combined with difference-in-differences, event studies, or factor models, the selection procedure respects the underlying design. For DiD and event studies, controls are pre-treatment, plausibly exogenous, and aligned with cohort–time identification (as in Section 13.6). For factor models (Chapter 8), regularisation is applied after residualising with respect to estimated factors (and, when appropriate, residualising treatments and controls), so that selection does not reintroduce factor-like collinearity or undermine rank and coverage conditions.

Design alignment ensures that regularisation serves the identification strategy rather than competing with it. Including post-treatment variables, colliders, or factor proxies that break the low-rank structure can undo the benefits of careful design. Transparent reporting of which controls are candidates, how they are pre-processed (demeaning, factor residualisation), and which are ultimately selected is central to design-based transparency.

## Regularisation and Fundamental Assumptions

Regularisation does not relax the fundamental assumptions of causal inference—unconfoundedness, SUTVA, consistency, and overlap—as formalised in Chapter 2 and by Pearl [2009]. It is a tool for implementing those assumptions in high-dimensional settings by selecting or shrinking covariates, but it cannot redeem a design that lacks credible identifying variation.

In marketing panels, interference and spillovers are often plausible. If SUTVA fails, regularised adjustment targets a different estimand. See Chapter 11.

This framing is consistent with modern panel surveys such as Arkhangelsky and Imbens [2024]. Heterogeneity-robust identification (avoiding biased aggregation across cohorts), design-based transparency (clear estimands, assumptions, and diagnostics), and credible inference (accounting for clustering, small samples, and dependence) remain non-negotiable. Regularisation complements these themes by enabling flexible covariate adjustment, accommodating high-dimensional controls, and interfacing cleanly with double-selection and DML estimators that respect design structure. The unifying principle is that strong, design-grounded assumptions do the causal work. Regularisation helps you operationalise those assumptions in complex marketing panels without losing control of variance or model complexity.

### 13.8 Tuning and Implementation

Implementing regularisation in panels requires careful choices about cross-fitting, hyperparameter tuning, and practical details like standardisation. The aim is to balance predictive accuracy with the demands of causal identification, recognising that dependence and clustering reduce the effective sample size for tuning and inference.

**Proposition 13.2 (Penalty Choice under Independence)** *Under independent errors with variance  $\sigma^2$ , iid rows of the within-transformed design matrix, suitably standardised regressors, and a sub-Gaussian-type tail condition on the error process, a theoretical choice of the penalty*

$$\eta = 2\sigma \sqrt{\frac{2 \log(2p/a)}{NT}}$$

*ensures that  $\eta \geq 2\|\tilde{\mathbf{X}}'\varepsilon/(NT)\|_\infty$  with probability at least  $1 - a$ . This bound, combined with the restricted eigenvalue condition, yields standard high-dimensional rates and support-recovery guarantees.*

In most marketing panels, the independence assumptions behind this expression are unrealistic. Serial correlation within units and common shocks across units mean that the effective sample size is smaller than  $NT$ , so penalties calibrated to  $NT$  will tend to under-penalise.

**Scaling Rule under Clustering.** When dependence is predominantly within clusters (for example, units), the effective sample size for high-dimensional inference and tuning is the number of independent clusters  $G$ , not the total number of observations  $NT$ . Theory suggests that penalty formulas should scale as

$$\eta \asymp \sigma \sqrt{\frac{\log p}{G}},$$

up to constants that depend on the unknown long-run covariance structure within clusters. This is a scaling heuristic, not a fully proved theorem for general clustered panels. Formal results require additional assumptions on cluster size growth, mixing, and dependence decay that are rarely verifiable in marketing data. In practice, blocked cross-validation with unit-level folds provides a data-driven choice of  $\eta$  that automatically adapts to the unknown dependence structure and is the primary practical calibration tool. If residual cross-sectional dependence is large, compare unit-blocked and time-blocked tuning as a sensitivity check and treat large differences in selected penalties as evidence that dependence matters for selection.

### Blocking and Cross-Fitting

Blocking and cross-fitting are designed to respect dependence and to prevent information leakage. Folds should be constructed at the unit or time level, and in DML settings the same fold structure should govern both nuisance training and hyperparameter tuning.

Unit-level blocking partitions units into  $K$  folds (typically  $K = 5$  or  $10$ ), trains lasso on  $K - 1$  folds of units (using all time periods for those units), and validates on the held-out fold. This is appropriate when the number of units  $N$  is large and within-unit serial correlation is the main dependence.

Time-level blocking partitions periods into folds, trains on  $K - 1$  folds of periods (all units in those periods), and validates on the held-out periods. This is useful when  $T$  is large and within-period cross-sectional dependence dominates.

Two-dimensional blocking partitions both units and periods into blocks, training on blocks that exclude the target block in both dimensions. It is more robust to complex dependence but more computationally expensive and leaves fewer observations per validation block. Confirm the choice of blocking with residual dependence diagnostics (Section 13.9).

In orthogonalised DML, cross-fitting further partitions the sample into folds for nuisance estimation. Hyperparameter tuning (for example,  $\lambda$  selection) should be nested within these training folds: nuisance learners are tuned using only the data designated for their training, and evaluated on distinct folds, to avoid optimistic validation due to information leakage. Concretely, any cross-validation used to tune nuisance models is performed within each cross-fitting training fold, using only that fold's training data. Validation for tuning and the held-out fold for orthogonalisation are kept disjoint to avoid information leakage that would violate Neyman orthogonality.

## Avoiding Temporal Leakage

In this chapter there are two distinct leakage concerns. In forecasting problems, leakage occurs when models are trained on data that occur after the validation observations in calendar time. In causal DiD and event-study designs, the more serious leakage occurs when post-treatment observations for treated cohorts contribute to the models used to select controls or tune penalties for those cohorts.

To prevent this, restrict training data for each cohort  $g$  and period  $t$  to pre-treatment periods ( $s < g$ ) for cohort- $g$  units and to not-yet-treated or never-treated units up to period  $t$ . This is a non-negotiable rule: for a given cohort  $g$  and post-treatment period  $t \geq g$ , neither nuisance training nor hyperparameter tuning may use any post-treatment observations for cohort- $g$  units.

In addition, restrict covariates to their pre-treatment values for cohort  $g$  (or baseline summaries computed strictly pre- $g$ ). All nuisance models for  $\tau(g, t)$  and  $\theta_k$  are tuned exclusively on pre-treatment data for that cohort and appropriate comparison units. This aligns with the temporal logic of DiD and event-study identification (Chapters 4 and 5) and with the cross-fitting logic of DML (Chapter 12).

## Hyperparameter Tuning

Hyperparameter tuning balances parsimony and fit. The penalty  $\eta$  controls the trade-off: larger  $\eta$  yields sparser models (higher bias, lower variance), smaller  $\eta$  yields denser models (lower bias, higher variance).

The usual practice is to select  $\eta$  by minimising out-of-sample prediction error measured via blocked cross-validation.

For each candidate  $\eta$  in a logarithmically spaced grid between  $\eta_{\max}$  and  $\eta_{\min}$ , train the lasso on training folds, validate on held-out folds, compute the mean squared error, and choose the penalty that minimises average validation error. Two useful choices are  $\eta_{\min}$  and  $\eta_{1se}$ , the largest penalty within one standard error of the minimum. When predictive performance is comparable,  $\eta_{1se}$  is often preferred for causal work because it yields simpler models with more stable selection.

## Stability Paths

Stability paths complement cross-validation by showing how variable selection evolves as  $\eta$  varies. Plot the set of selected controls as a function of  $\eta$ . Controls that enter early and remain selected over a wide range of penalties are more stable than those that appear only for very small  $\eta$  or flicker in and out.

Reporting stability paths alongside cross-validation curves helps distinguish robust from fragile selections. Analysts can focus interpretation on controls that are stable across both folds and penalties, and treat marginal selections with more caution.

## Model Parsimony and Interpretability

Model parsimony and interpretability justify preferring simpler models when prediction accuracy is similar. Marketing practitioners value models that are easy to explain (a modest number of controls, each with a clear rationale) and that align with domain knowledge (selected controls are plausible confounders or key drivers).

The  $\eta_{1se}$  rule operationalises this preference, trading a small increase in validation error for a sparser specification. Ex ante restrictions on the candidate control set—limiting attention to clearly pre-treatment, plausibly exogenous variables—further support interpretability and identification.

## Prediction versus Identification

Tension between prediction and identification, highlighted by Breiman [2001], is acute in high dimensions. Prediction-focused tuning would include any covariate that improves validation error, even if it is a collider, mediator, or post-treatment variable. Identification-focused tuning constrains the candidate set to variables that respect the design, even if excluding some predictors slightly hurts forecast performance.

The solution is to define the candidate set with identification in mind—pre-treatment, plausibly exogenous controls—and then use regularisation to choose among them. Validation error remains the tuning criterion, but only within a design-respecting library of controls. Transparent reporting of which variables were eligible for selection and which were excluded ex ante is essential.

## Practical Penalty Grids

Practical penalty grids use logarithmic spacing to cover a broad range of penalties efficiently. Start from

$$\eta_{\max} = \max_j \frac{|\tilde{X}'_j \tilde{Y}|}{NT},$$

where  $\tilde{X}_j$  is the  $j$ th within-transformed control and  $\tilde{Y}$  the within-transformed outcome; this definition is tied to the objective's scaling convention. Under clustering, penalty selection should still be evaluated via blocked folds and interpreted using  $G$ -based heuristics. Set  $\eta_{\min}$  as a fixed fraction of  $\eta_{\max}$  (for example,  $0.01 \eta_{\max}$  or  $0.001 \eta_{\max}$ , depending on expected sparsity) and generate a grid of  $L$  logarithmically spaced values between  $\eta_{\max}$  and  $\eta_{\min}$ . This covers the range from the null model to a nearly saturated model, allowing cross-validation to locate a suitable compromise.

## Standardisation and Scaling

Standardisation ensures that penalties are applied uniformly across controls measured on different scales. After any within or factor-residualisation, standardise each control column  $\tilde{X}_j$  by its empirical standard deviation: replace  $\tilde{X}_j$  with  $\tilde{X}_j / \hat{\sigma}_j$ , where  $\hat{\sigma}_j^2$  is the sample variance of  $\tilde{X}_j$ . This puts all controls on a comparable scale so that  $\eta |\beta_j|$  has the same meaning across  $j$ . When using cross-validation or cross-fitting, standardisation parameters (means, standard deviations) must be estimated within each training fold and then applied to that fold's validation data, rather than computed on the full sample. This prevents temporal and cross-fold leakage that would otherwise bias validation error and violate the independence assumptions underlying cross-fitting.

Once the lasso coefficients have been estimated, transform them back to the original scale for interpretation by dividing by  $\hat{\sigma}_j$ . This simple standardisation improves numerical stability, aligns with the RE assumptions used in theory, and avoids the complexities of bespoke between-cluster variance formulas. Interpret coefficients as within-transformed (demeaned) relationships unless you explicitly reconstruct levels.

## Handling Interactions and Lags

Handling interactions and lags requires care to respect hierarchical structure and to keep the candidate library manageable. Interactions such as  $D_{it} \times X_{j,it}$  capture effect modification and should be considered only once main effects are included. Group lasso can select interactions jointly with their main effects, ensuring that an interaction is not included without its components. In most causal specifications,  $\tau$  and core design terms (treatment indicators, event-time dummies, key lags of  $D_{it}$ ) must remain unpenalised.

Lags of covariates and treatment (distributed effects) are often grouped by variable: group lasso can select all lags of a given covariate jointly, while hierarchical lasso can enforce that lower-order lags are included

before higher-order lags. Theory and prior evidence should guide which families of interactions and lags are generated in the first place; blindly including all possible combinations can overwhelm limited panels even with strong penalties.

A critical identification constraint applies: interactions and lags are constructed only from pre-treatment, plausibly exogenous covariates (and, when the estimand permits, pre-treatment lags of  $D_{it}$ ). We never generate interactions or lags involving post-treatment variables or descendants of  $D_{it}$  in the causal graph from Chapter 2. Clear reporting of which interactions and lags are eligible and which are ultimately selected helps readers assess whether the resulting dynamic specification aligns with economic intuition and identification constraints.

## 13.9 Diagnostics and Robustness

Credible use of regularisation requires careful diagnostics. We must assess covariate balance, overlap, residual dependence, and the stability of our results, and we must integrate these checks with the design-diagnostics workflow from Chapter 17. This section sketches a diagnostic playbook tailored to high-dimensional panels.

Post-selection balance improvement asks whether the selected controls actually reduce covariate imbalance between treated and control units. Compute standardised mean differences (SMDs) for key covariates before and after regression adjustment on the selected controls and fixed effects. In panels, you should state whether SMDs are computed at the cell level ( $i, t$ ) or on unit-level aggregates (for example, pre-treatment means by unit). When clustering is by unit, unit-level aggregates often better reflect the effective sample size  $G$ .

For each covariate  $X_j$ , define

$$\text{SMD}_j^{\text{before}} = \frac{\bar{X}_{j,\text{treat}} - \bar{X}_{j,\text{control}}}{\sqrt{(s_{j,\text{treat}}^2 + s_{j,\text{control}}^2)/2}},$$

where means and variances are computed in the raw data. Then residualise  $X_j$  with respect to the selected controls and fixed effects and recompute SMDs using the residuals (denoted by the “adj” superscript). Balance improves if  $|\text{SMD}_j^{\text{after}}| < |\text{SMD}_j^{\text{before}}|$  for most covariates. Common thresholds such as  $|\text{SMD}| < 0.1$  or  $|\text{SMD}| < 0.25$  are rough heuristics. In high-dimensional settings, many covariates will exceed any fixed threshold by chance, and different covariates matter differently for confounding. Focus on covariates with strong substantive links to treatment assignment and outcomes rather than mechanically “passing” a global threshold across hundreds of variables. If many covariates still have  $|\text{SMD}_j^{\text{after}}| > 0.25$ , the selected controls may not adequately control for confounding. Double selection, richer nuisance models, or alternative designs (for example, propensity-score weighting or synthetic control) may be preferable. When clustering is by unit, it is often informative to compute balance diagnostics at or aggregated to the unit level (for example, unit-level averages) to reflect the effective sample size  $G$ .

In DiD and event-study settings, check balance in pre-periods and possibly by cohort and period, not just overall. This links directly to the design-diagnostics chapter and ensures that conditional parallel trends are plausible within the subgroups that drive identification.

### Overlap Checks

Overlap checks examine whether treated and comparison units remain comparable after selection. Estimate propensity scores using the selected controls and fixed effects—for example, a logistic regression of  $D_{it}$  on the selected  $X_{S,it}$ ,  $\alpha_i$ , and  $\lambda_t$ , or a more flexible ML-based propensity model—and plot the distributions of predicted propensities for treated and control units. Compute an overlap statistic such as the integral of the minimum of the two density estimates.

FE logit can be numerically unstable (separation, incidental parameter bias when  $T$  is small). Treat it as a rough overlap diagnostic. Consider simpler propensity models on pre-treatment summaries or ML classifiers

without fixed effects as complementary checks. Covariates entering the propensity model are restricted to pre-treatment, plausibly exogenous variables as per the design sections.

Flag regions where one group dominates (for example, treated units concentrated at very high propensities, controls at very low propensities). Overlap checks can reveal severe support violations but cannot prove the absence of support problems outside the observed data. Consider trimming observations with extreme propensities (for example, below 0.05 or above 0.95) or restricting attention to an overlap region identified by these plots. Report effective sample sizes (numbers of treated and control observations in the overlap region) and discuss how trimming changes the target estimand relative to the full sample. When clustering is by unit, it is often informative to compute overlap diagnostics at or aggregated to the unit level to mirror the effective sample and dependence structure.

## Residual Dependence Checks

Residual dependence diagnostics assess whether the selected model has removed enough structure for your chosen clustering scheme to be credible. For serial dependence, compute autocorrelations of residuals within units,  $\text{Corr}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{i,t-k})$ , for lags  $k = 1, 2, \dots$ . Large within-unit autocorrelation motivates clustering by unit (or HAC if dependence spans both dimensions).

For cross-sectional dependence, compute average correlations of residuals across units within periods,  $\text{Corr}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{jt})$  for  $i \neq j$ . Substantial cross-correlation suggests that time clustering, two-way clustering, or spatial HAC adjustments are warranted. ACF and cross-correlation diagnostics motivate but do not mathematically guarantee the adequacy of a given clustering scheme. These are heuristics. The appropriate variance estimator depends on the clustering unit implied by your identifying variation and the dominant dependence structure. Reporting these diagnostics alongside estimated treatment effects clarifies why a particular variance estimator (unit, time, two-way, or HAC) was chosen.

## Inclusion Frequency and Stability

Inclusion frequency and stability across folds quantify how often each control is selected and how sensitive selection is to sample splits. For each control  $j$ , compute the inclusion frequency  $f_j = (\text{number of folds where } j \text{ is selected})/K$ . A control with  $f_j = 1$  is selected in all folds (highly stable). One with  $f_j \approx 0$  is never selected. Values strictly between zero and one indicate instability.

Plot inclusion frequencies against covariate indices or names, marking highly stable controls and those that are only sporadically selected. Treat highly stable controls as core adjustment variables, and treat unstable controls as candidates for sensitivity analysis. Report results both with and without them, or focus interpretation on specifications where unstable controls are excluded or penalised more heavily (for example, using  $\eta_{1se}$ ).

## Sensitivity to Penalty Choice

Sensitivity to penalty choice asks whether estimated treatment effects change materially as  $\eta$  varies over a plausible range. Estimate  $\hat{\tau}$  for a grid of penalties (for example,  $\eta_{\min}$ ,  $\eta_{1se}$ , and a few intermediate points) and plot  $\hat{\tau}(\eta)$  against  $\eta$ .

If  $\hat{\tau}$  is stable across penalties within the range where validation error is near its minimum, conclusions are robust to tuning choices. If  $\hat{\tau}$  varies widely (for example, ranging from 0.10 to 0.25), the choice of  $\eta$  matters, and you should report estimates for multiple penalties, discuss which settings are most defensible based on cross-validation and domain knowledge, and interpret the range as reflecting model uncertainty rather than sampling noise alone.

## Triangulation against Simpler Designs

Triangulation compares regularised estimates to estimates from simpler designs with different identifying assumptions: a DiD specification without high-dimensional controls, event studies with low-dimensional controls, synthetic control, or factor models. If estimates agree (for example, within one standard error), this suggests that high-dimensional adjustment is not driving the conclusions. If they differ substantially, you should examine whether the difference plausibly reflects better adjustment for observed confounding or whether it might arise from overfitting, collider adjustment, or fragile variable selection. Disagreements should trigger an assumption-level discussion, not only a modelling discussion.

Reporting multiple estimates and discussing why they agree or disagree is central to design-based transparency. Regularisation should sharpen, not overturn, credible designs without explanation.

## Integration with Design-Diagnostics Workflow

Finally, integrate these checks with the broader design-diagnostics workflow from Chapter 17. Use event-study leads as a pre-trend diagnostic for (conditional) parallel trends. Plot event-time effects with confidence intervals and check that pre-treatment estimates cluster around zero. Conduct placebo diagnostics on never-treated units or placebo intervention dates. Assess overlap and support in covariates and propensities, both before and after regularisation. Compare regularised estimates to simpler designs as part of triangulation.

Diagnostic plots and tables—balance charts, propensity-score overlap plots, residual-dependence summaries, inclusion-frequency plots, and sensitivity paths—should accompany any high-dimensional causal analysis. They allow readers to see not just point estimates, but the design, overlap, dependence structure, and model-selection stability that support those estimates.

### 13.10 Inference

Inference for regularised panel estimators must account for clustering induced by panel dependence, for the additional uncertainty introduced by selection, and for multiplicity when testing many outcomes or moderators. The tools developed in Chapter 16—cluster-robust variance estimators, wild cluster bootstrap, and joint confidence bands—apply here as well. The details depend on the estimator.

In this chapter, clean influence-function representations arise for orthogonal-score DML estimators, for debiased lasso under high-dimensional regularity conditions, and for post-double-selection OLS under the maintained post-selection conditions discussed in Section 13.4. These representations rely on the high-dimensional sparsity, overlap, dependence, and design-alignment assumptions in Section 13.7. Without those, neither post-selection nor debiased constructions deliver reliable inference.

#### Cluster-Robust Standard Errors

Cluster-robust standard errors for double selection and debiased lasso adjust for within-cluster correlation by aggregating score contributions within clusters. Let clusters be indexed by  $c = 1, \dots, G$ , with cluster membership sets  $\mathcal{C}_c \subseteq \{1, \dots, N\}$ . For post-double-selection OLS (on the union of selected controls), inference applies a CRVE to the selected model, under the post-selection conditions in Section 13.4 and with clustering aligned to the dependence structure.

For debiased lasso, the cluster-level influence contribution has the form

$$\Psi_c = \sum_{i \in \mathcal{C}_c} \sum_t \hat{r}_{it} \hat{\varepsilon}_{it} / \left( \frac{1}{NT} \sum_{i,t} \hat{r}_{it}^2 \right),$$

where  $\hat{r}_{it}$  are residualisation residuals from Section 13.5 and  $\hat{\varepsilon}_{it}$  are residuals from the debiased outcome equation. Cluster-robust standard errors are obtained by computing the empirical variance of  $\Psi_c$  across clusters and using  $\sqrt{G}$ -based asymptotics, as in Chapter 16.

#### Small-Sample Inference

When the number of clusters  $G$  is small (for example, fewer than 30 markets or products), asymptotic cluster-robust approximations can be fragile. Wild cluster bootstrap provides a finite-sample alternative. There are two common variants. An influence-function bootstrap treats cluster-level influence contributions  $\Psi_c$  as pseudo-observations and reweights them with random multipliers (such as Rademacher weights). A residual-based wild cluster bootstrap instead reweights residuals and re-fits the estimator.

In post-selection settings there is a choice. A conservative approach re-runs the entire selection and estimation procedure (lasso, double selection, or debiased lasso) inside each bootstrap replication, fully reflecting selection uncertainty but increasing computational cost. A lighter-weight diagnostic approach holds the se-

lected model fixed and bootstraps only the post-selection estimator. Chapter 16 discusses these variants in detail.

When  $G$  is small and selection is volatile, re-selection bootstrap is a conservative practice, not a guarantee. It still relies on the maintained cluster sampling structure for the chosen bootstrap scheme. When you use re-selection bootstrap, report how often selected sets change across bootstrap draws as an additional selection-stability diagnostic.

## Multiple-Testing Considerations

Multiple testing arises naturally when regularisation is used to estimate effects for many outcomes, subgroups, moderators, or event-time lags. For example, an analyst may produce a panel of 20 outcome effects or a vector of heterogeneous treatment effects across many customer segments. Testing each effect at the 5 per cent level without adjustment inflates the chance of spurious findings.

Classical Bonferroni correction divides the nominal significance level by the number of tests, controlling the family-wise error rate at the cost of power. False-discovery-rate (FDR) control via Benjamini–Hochberg offers a less conservative alternative by controlling the expected proportion of false discoveries among rejected hypotheses. Stepdown procedures such as Romano–Wolf improve on Bonferroni by exploiting the joint dependence structure of test statistics, often estimated via bootstrap.

In the context of this chapter, it is often more informative to report joint confidence bands or joint tests for vectors of treatment effects—such as event-time profiles or moderator-specific effects—using the joint-inference machinery from Chapter 16. All joint tests and confidence bands should use the same clustering and dependence adjustments chosen for the base estimator (unit, time, two-way, or HAC), as detailed in Chapter 16. A practical approach is to report (i) unadjusted p-values for individual effects, (ii) joint tests (for example, that all lagged effects are zero), and (iii) adjusted p-values or bands for the collection. Treat individual effects as exploratory when the joint test rejects.

## Situating within Modern Panel Frameworks

Situating regularised inference within modern panel frameworks [Arkhangelsky and Imbens, 2024] clarifies its role. The Arkhangelsky–Imbens survey emphasises three pillars: credible identification (cohort–time designs, robust aggregation), credible inference (clustering, small-sample adjustments, multiplicity), and transparent diagnostics (balance, overlap, pre-trends, placebo tests).

Regularisation complements these pillars rather than replacing them. It enables flexible covariate adjustment while keeping identification anchored in the DiD, event-study, factor, and DML designs from earlier chapters. It interfaces cleanly with cluster-robust and bootstrap-based inference via influence functions for post-selection and debiased estimators. It also encourages richer diagnostic reporting—on balance, overlap, selection stability, and sensitivity to tuning—that sits alongside standard pre-trend and placebo checks.

The unifying principle is that careful panel causal inference requires both strong, design-based identification and honest uncertainty quantification. Regularisation supplies high-dimensional estimation tools within that framework. It does not exempt us from the clustering, small-sample, and multiple-testing concerns that the modern panel literature has worked hard to make explicit.

## 13.11 Marketing Applications

Regularisation methods are particularly valuable in marketing applications where many potential confounders are available—rich demographics, competitor actions, seasonality, and platform signals—and where parsimony aids interpretation and communication. This section sketches how the tools in this chapter appear in five settings: MMM-style panels with rich covariates, competitor controls in pricing studies, high-resolution calendar and catalogue features in retail, creative attributes in advertising effectiveness, and brand premiums in commodity categories. Full workflows and empirical case studies appear in Chapter 18.

### MMM-Style Panels with Rich Covariates

Panel variants of media mix modelling arise when estimating the association between multiple marketing channels (TV, digital, print, radio, social media) and sales using store-by-time data with hundreds of controls. Consider a retailer analysing a panel of 200 stores over 104 weeks. The outcome is weekly sales,  $Y_{it}$ . Treatments are channel intensities (for example, GRPs) collected in a vector  $D_{it}$ . Controls  $X_{it}$  include demographics, competitive density, seasonality dummies, search trends, social-media signals, and lagged outcomes.

In a regression specification, the channel coefficients are regression parameters such as  $\beta_{\text{TV}}$  and  $\beta_{\text{digital}}$ . Interpreting these coefficients as causal channel effects requires a credible assignment mechanism. A selection-on-observables story would assume unconfoundedness of channel intensities conditional on  $X_{it}$ ,  $\alpha_i$ , and  $\lambda_t$ , and adequate overlap. In many platform settings this is a demanding assumption because budgets, bids, and targeting respond to expected demand. Treat this as a sensitivity and specification exercise unless you can point to plausibly exogenous variation. Experiments or credible instruments are often preferable. See Section 16.10 for instrumental-variable designs.

A regularised implementation uses double selection with unit-level blocking. For each channel component of  $D_{it}$ , run lasso of  $Y_{it}$  on  $X_{it}$  (excluding that channel) and lasso of that channel on  $X_{it}$ , with penalties chosen by unit-blocked cross-validation. Take the union of selected controls across outcome and treatment models, and run OLS of  $Y_{it}$  on the channel, the selected controls, and store fixed effects  $\alpha_i$  and week fixed effects  $\lambda_t$ , with cluster-robust standard errors. In practice one can also estimate a joint DML model with a multivariate  $D_{it}$ , as discussed in Chapter 12; the separate-channel description here is schematic.

In an illustrative example, TV might generate a moderate positive estimated association with sales, digital a smaller one, and social media an estimate whose confidence interval includes zero. Selected controls could include demographics, competitive density, and seasonality. Diagnostics would show improved covariate balance, adequate overlap, and stable selection across folds. Any causal interpretation remains conditional on the maintained identification assumptions.

## Competitor Controls in Pricing Studies

Competitor controls in pricing studies illustrate group lasso for structured covariate groups. Consider a panel of 500 stores over 52 weeks, with prices and sales for a focal brand and 10 competing brands. The outcome is log sales for the focal brand. The regressor of interest is its own log price. Controls  $X_{it}$  include the 10 competitor prices, demographic variables, and week-of-year dummies. A common log-demand regression takes the form

$$\log Q_{it} = \beta_p \log P_{it} + X'_{it}\beta + \alpha_i + \lambda_t + \varepsilon_{it}.$$

Here  $\beta_p$  is a regression slope. Interpreting  $\beta_p$  causally requires an exogenous source of price variation and a defensible exclusion restriction (for example, cost shifters), as discussed in Section 16.10. Without such variation, the slope is generally not identified as an elasticity from observational pricing panels.

Group lasso applies a group penalty to the block of competitor prices, encouraging either joint inclusion or joint exclusion of all competitor-price coefficients. A regularised workflow partitions stores into folds, trains group lasso on training folds, validates on held-out folds, and selects the penalty that minimises blocked-CV error. In an illustrative setting the selected model might include all competitor prices (group selected) and a small set of demographic and seasonality variables. The resulting price-slope estimate  $\hat{\beta}_p$  is negative and stable across penalties, and balance and overlap diagnostics look satisfactory. These estimates support pricing simulations and revenue forecasts as regression-based sensitivity analyses. A causal elasticity interpretation still requires an exogenous source of price variation.

## High-Resolution Calendar and Catalogue Features

High-resolution calendar and catalogue features in retail arise when estimating campaign or promotion effects using daily or weekly data with rich calendar indicators (day-of-week, week-of-year, month-of-year, holidays, weather) and catalogue features (SKU attributes, individual promotions, cross-promotions). In a panel of 1,000 SKUs over 365 days with 200 such features per SKU-day, lasso or elastic net can select a subset of these controls when estimating the effect of a particular promotion variable  $D_{it}$  on sales  $Y_{it}$ .

A plausible outcome is that around 20 features—day-of-week and holiday dummies, temperature, and a handful of promotion flags—enter the model. The estimated promotion effect is robust across penalties and folds, and the selected controls line up with retail intuition about what drives sales.

Any causal interpretation still depends on the underlying design assumptions (for example, quasi-random timing of promotions across SKUs). You should support that claim with falsification diagnostics such as pre-trends, placebo promotion dates, or balance on pre-period outcomes. Control selection does not validate quasi-random timing.

## Creative Attributes in Advertising Effectiveness

Creative attributes in advertising effectiveness provide a natural use case for hierarchical lasso. Suppose we observe 100 DMAs over 52 weeks, with 50 advertising campaigns and 200 creative attributes (message themes, image types, formats, lengths) for each campaign–week. The outcome is weekly conversions; treatments capture campaign exposure and creative intensity. The analyst wants to know which creative elements are associated with higher campaign effects, conditional on overall exposure.

Hierarchical lasso enforces a structure in which a campaign indicator must enter before its creative attributes, and main effects must enter before interactions. A regularised model regresses outcomes on campaign indicators, creative attributes, and controls  $X_{it}$  (with campaign and key main effects unpenalised, creative attributes penalised in groups), using blocked CV to tune penalties. In a stylised example, the procedure might highlight a subset of message themes (“quality”), image types (“lifestyle”), and formats (“30-second video”) as consistently associated with stronger campaign effects.

Absent randomisation of creatives across comparable markets, interpret these as correlates of performance conditional on the model, not as causal effects of creative attributes. In many settings, creatives are themselves chosen in response to performance, which can reverse the causal direction.

## Brand Premiums in Commodity Categories

Brand premiums in commodity categories illustrate how regularisation can support brand-equity measurement in high-dimensional scanner data. Consider a panel with brands  $b = 1, \dots, B$ , stores  $s = 1, \dots, S$ , and weeks  $t = 1, \dots, T$ . Let  $Q_{bst}$  denote the quantity sold of brand  $b$  in store  $s$  and week  $t$ , and  $P_{bst}$  its shelf price. A log-demand specification is

$$\log Q_{bit} = \alpha_i + \lambda_t + \delta_b + \beta_p \log P_{bit} + X'_{bit}\theta + \varepsilon_{bit},$$

where  $\alpha_i$  and  $\lambda_t$  are store and week fixed effects,  $X_{bit}$  collects high-dimensional controls (promotions, displays, distribution, local demographics, competitor prices), and  $\delta_b$  is a brand-specific intercept. Normalising the supermarket own label to  $\delta_{\text{own}} = 0$ , the coefficient  $\delta_b$  captures the average log-demand shift for brand  $b$  relative to the own label, after controlling for price and covariates.

Mapping  $\delta_b$  into willingness-to-pay is a back-of-the-envelope calculation under a maintained demand model. It relies on a constant-elasticity interpretation, stable substitution patterns, and an exogenous source of price variation for  $\beta_p$ . This chapter does not validate those structural assumptions. It discusses regularisation as an estimation tool given a specified model.

Under a maintained constant-elasticity interpretation, one can solve for a compensating price change that holds expected demand constant. For small changes, an approximate premium in level terms is

$$\Delta p_b \approx -\frac{\delta_b}{\beta_p} \bar{p},$$

where  $\bar{p}$  is a reference category price. A positive  $\delta_b$  then corresponds to a higher implied willingness-to-pay for brand  $b$  than for the otherwise similar own label. When these conditions fail,  $\delta_b$  should be read as a reduced-form brand effect rather than a literal willingness-to-pay premium.

In high-dimensional versions of this model,  $X_{bst}$  may include many brand-size combinations, store characteristics, and competitor interactions. Group lasso or sparse group lasso can treat brand indicators (or brand-size blocks) as structured groups, shrinking small or indistinct brands' intercepts towards the own-label baseline while retaining distinct premiums for major brands. Double selection ensures that both price and brand indicators remain in the model whenever they act as confounders for each other, even if a purely predictive lasso would drop some of them. The resulting  $\delta_b$  estimates provide a disciplined, design-compatible measure of brand premiums, grounded in revealed-preference data and robust to high-dimensional controls.

## Summary

These examples illustrate the versatility of regularisation in marketing. Double selection controls for confounding flexibly while supporting valid inference; group lasso selects structured groups of controls (competitor prices, media channels, creative attributes, brand indicators) in a way that encodes domain knowledge; and hierarchical lasso enforces logical hierarchies (campaigns before creatives, main effects before interactions, lower-order lags before higher-order lags).

Across all these settings, regularisation is subordinate to the identification strategies developed earlier in the book. It manages high-dimensional controls in a principled way so that panel designs—DiD, event studies, factor models, and DML—can be implemented in rich marketing data without losing interpretability or control over variance. Chapter 18 revisits brand equity from a dynamic stock perspective in media-mix settings, linking these cross-sectional premiums to long-run pricing power.

## 13.12 Workflow Checklist

This section provides a compact, reproducible protocol for conducting regularisation analyses in marketing panels. The workflow integrates design, partialling, tuning, diagnostics, inference, and reporting, so that conclusions are both credible and transparent.

### Step 1: Define Target Estimand and Design

Clarify the target estimand (for example ATE, ATT, cohort–time effects  $\tau(g, t)$ , or event-time effects  $\theta_k$ ) and the design providing identification (DiD, event study, unconfoundedness conditional on controls, factor-based designs, or DML with orthogonal scores). Document the estimand mathematically and explain its business relevance. For staggered adoption, specify whether the main target is cohort–time effects  $\tau(g, t)$ , event-time effects  $\theta_k$ , or an overall ATT defined as a cohort-weighted average.

### Step 2: Decide on Fixed Effects or Factor Partialling

Decide how to partial out unit and time heterogeneity before regularisation. In simple settings with moderate  $N$  and  $T$ , within-transformation that removes unit and time fixed effects (Chapter 2) is often a reasonable starting point for additive unobserved heterogeneity, but it does not by itself justify parallel trends or unconfoundedness. When untreated outcomes exhibit rich common-shock structure that violates simple parallel trends, factor residualisation (Chapter 8) may be more appropriate: estimate latent factors, residualise outcomes (and, where required, treatments and controls), and apply regularisation to the residualised model. Document and justify the chosen partialling strategy in light of the untreated-potential-outcome structure.

### Step 3: Choose Regulariser and Blocking Scheme

Select the regularisation method based on the control structure: lasso for sparse, roughly independent controls, elastic net when controls are highly correlated, group lasso for grouped controls (lags, channels, competitors, brand blocks), and hierarchical lasso for nested structures (campaigns and creatives, main effects and interactions, lower- and higher-order lags). These methods provide the core building blocks for double selection and DML (Sections 13.3–13.4).

Choose the blocking scheme for cross-validation (unit-level, time-level, or two-dimensional) in line with the dominant dependence pattern (Section 13.8). Verify the choice using residual dependence diagnostics (Section 13.9). In DML settings, use the same fold structure for cross-fitting and for tuning nuisance learners, to avoid information leakage between training and validation sets. Within each cross-fitting training fold,

tune nuisance models using only that fold’s training data. Do not reuse held-out folds for tuning, to avoid leakage that would violate Neyman orthogonality.

#### Step 4: Run Double Selection or DML with Selection

For double selection (Section 13.4), run lasso of outcomes on controls (excluding the treatment), run lasso of treatment on controls, take the union of selected controls, and run OLS of outcomes on treatment and the union of controls, with unit and time fixed effects and cluster-robust standard errors. Penalise only nuisance coefficients—not the treatment effect  $\tau$  or core design variables such as event-time indicators.

For DML with selection (Chapter 12), partition the data into folds via dependence-respecting cross-fitting, train outcome and propensity-score lasso or elastic-net models on out-of-fold data, construct Neyman-orthogonal scores, and solve the sample moment conditions for the treatment effect. Define the score at the level of the independent sampling unit (typically the unit  $i$  as a cluster) and compute inference using cluster-aggregated influence contributions.

Define the candidate covariate library  $X_{it}$  ex ante as pre-treatment, plausibly exogenous variables that are not descendants of  $D_{it}$  in the causal graph of Chapter 2. Regularisation then selects within this library but never adds post-treatment or clearly endogenous variables. Document the penalties used (from blocked cross-validation or theory-based scaling), the selected controls (which variables enter the nuisance functions), and the rationale for any exclusions.

#### Step 5: Conduct Post-Selection or Debiased Inference with Cluster-Robust SEs

For post-double-selection OLS, compute cluster-robust standard errors by aggregating influence contributions at the unit level (or via two-way clustering when both unit and time dependence matter), as described in Section 13.5 and Chapter 16. This is often the default choice when the number of clusters  $G$  is moderate and sparsity is plausible.

When the goal is asymptotically normal inference for a parameter in a very high-dimensional model and sparsity conditions are strong, debiased lasso (Section 13.5) provides an alternative. Construct the bias-corrected estimator via residualisation of the within-transformed treatment on controls and compute cluster-robust standard errors using cluster-level influence contributions.

For small  $G$  (for example, fewer than 30 clusters), use wild cluster bootstrap to build a bootstrap distribution of the estimator. When selection is unstable or  $G$  is very small (for example,  $G < 20$ ), re-run selection inside each bootstrap replication and treat fixed-model bootstraps as diagnostic only. Report point estimates, standard errors, confidence intervals, and p-values, clearly indicating whether they rely on asymptotic cluster-robust formulas or on bootstrap-based critical values.

### Step 6: Validate via Diagnostics and Sensitivity

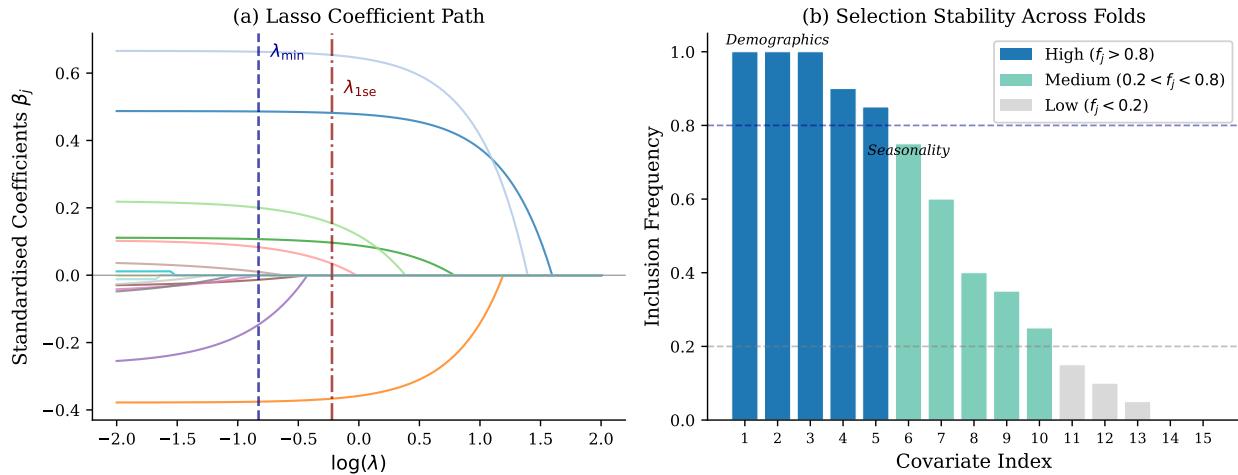
Validate the regularised analysis using the diagnostics from Section 13.9 and Chapter 17. Assess post-selection balance (SMDs before and after adjustment), overlap (propensity-score distributions and common support), residual dependence (autocorrelation and cross-correlation of residuals), inclusion frequency (how often each control is selected across folds), and sensitivity to penalty choice (how  $\hat{\tau}$  varies across  $\eta$ ).

Diagnostics can reveal severe violations of the assumptions (for example, poor balance, weak overlap, unstable selection), but they cannot prove that unconfoundedness, conditional parallel trends, or design alignment hold. When diagnostics flag problems—poor balance, weak overlap, strong residual dependence, unstable selection—adapt the design: trim or restrict to an overlap region, adjust the clustering scheme, refine the covariate library, or simplify the specification. Compare regularised estimates to simpler designs (two-way fixed-effects DiD, synthetic control, low-dimensional event studies) for triangulation, and interpret divergences in light of identification and overfitting concerns.

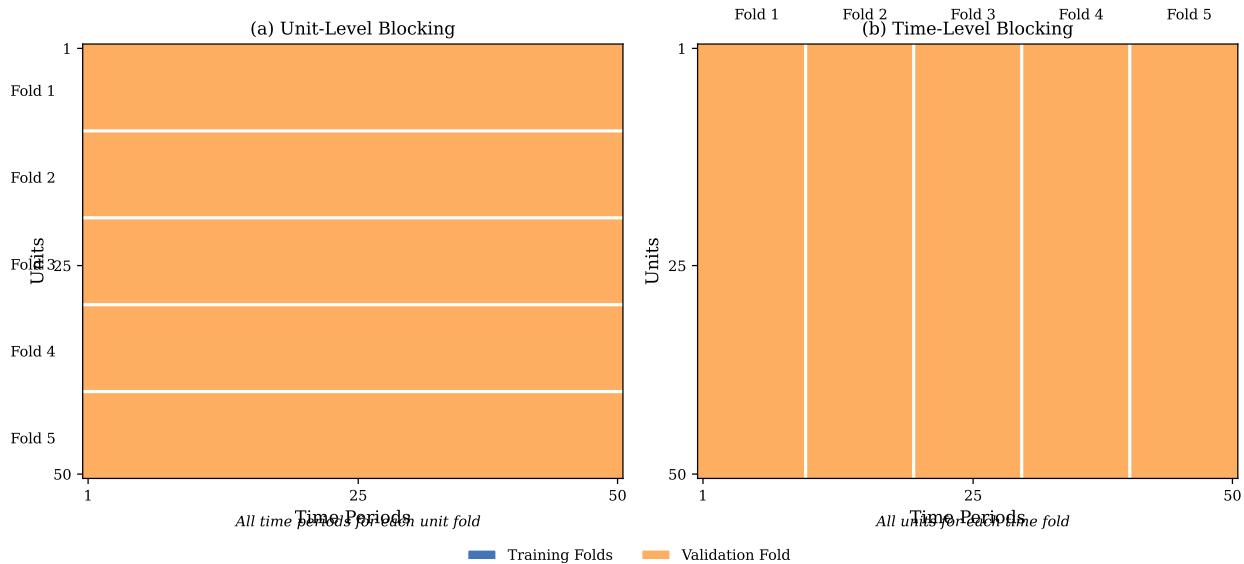
### Step 7: Report Findings with Replication Materials

Report treatment-effect estimates with confidence intervals, the set of selected controls (and their interpretation), diagnostics (balance, overlap, selection stability, sensitivity), and comparisons to simpler estimates. State which penalty was used (for example,  $\eta_{1se}$  from blocked cross-validation), which blocking and clustering schemes were adopted, and how estimates vary with alternative penalties or designs. Provide replication materials—data, scripts, and documentation—so that others can verify the analysis and explore alternative specifications. When raw data cannot be shared, provide synthetic or de-identified replication datasets and full code for the analysis pipeline.

By following this workflow, practitioners can conduct regularisation analyses that balance flexibility and parsimony, respecting identification requirements while managing high-dimensional controls. Covariate selection remains disciplined by design-based logic; inference accounts for dependence and selection uncertainty; and diagnostics provide transparent evidence on the credibility of conclusions. The result is causal evidence that can withstand scrutiny and inform strategic decisions.



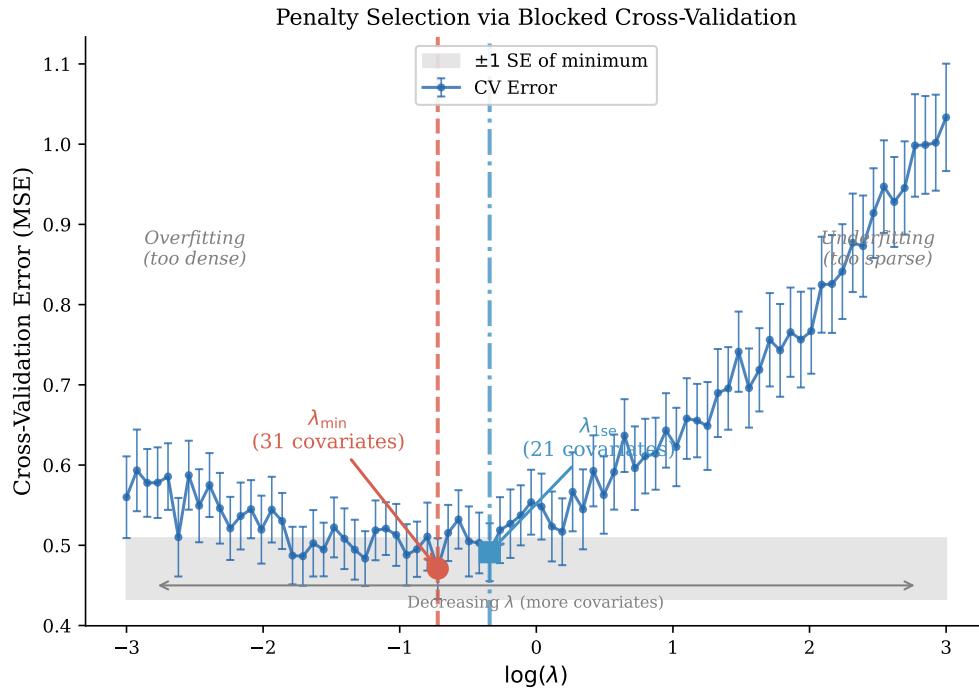
**Fig. 13.1** Lasso/elastic net coefficient paths and inclusion frequencies. The figure summarises how coefficients and selected sets change as the penalty varies.



**Fig. 13.2** Blocked cross-validation schematics for panels (by unit and by time). Use blocking to respect dependence and avoid leakage across folds.

#### Box 13.1: High-Dimensional Controls Checklist

1. Define the estimand and the design that provides identification.
2. Restrict the candidate covariate library to design-admissible, pre-treatment variables.
3. Partial out fixed effects or factors and choose a dependence-respecting blocking scheme.
4. Tune penalties inside blocked folds and, for DML, nest tuning inside cross-fitting.
5. Estimate the target using double selection or orthogonal-score methods, keeping core design terms unpenalised.
6. Do inference with clustering aligned to the identifying variation, using wild bootstrap when  $G$  is small.
7. Report sensitivity to tuning and the key diagnostics (balance, overlap, dependence, selection stability).



**Fig. 13.3** Validation error across penalty values under blocked cross-validation. The figure illustrates the stability–fit trade-off when choosing  $\eta$ .

**Table 13.1** Mapping of Method to Assumptions, Tuning, and Use-Cases. Lasso, elastic net, and group lasso often serve as building blocks for double selection or DML rather than as final estimators; see Sections 13.3–13.4.

Method	Key Assumptions	Tuning and Inference	Use-Cases
Lasso	Approximate sparsity, overlap, dependence-aware blocking	Blocked CV, cluster-robust SE, $\eta_{1se}$ rule	Sparse controls, high $p$ , interpretability priority
Elastic net	Approximate sparsity, overlap, dependence-aware blocking	Blocked CV, cluster-robust SE, tune $\rho$ and $\eta$	Correlated controls, stability over pure lasso
Group lasso	Approximate group sparsity, overlap, dependence-aware blocking	Blocked CV, cluster-robust SE, group penalties scaled by size	Grouped controls (lags, channels, competitors)
Double selection (Section 13.4)	Approximate sparsity in outcome and treatment models, overlap, no colliders	Union of lasso selections, OLS with cluster-robust SE	Causal targets, reduces omitted-variable bias
Debiased lasso (Section 13.5)	Approximate sparsity, overlap, residualisation sparsity	Residualisation step, cluster-robust SE or wild bootstrap	Asymptotically normal inference, small $s$ relative to $G$
Sparse DiD (Section 13.6)	Conditional parallel trends, no post-treatment controls, design alignment	Blocked CV, cohort-time structure, cluster-robust SE	Augment DiD with many controls while preserving identification



## Chapter 14

# Continuous and Nonlinear Panel Models

Marketing panels often combine treatments measured on a continuum with outcomes that have limited support. Prices move along a grid, advertising campaigns have budgets and intensities, discounts come in percentages, and customers make binary choices, purchase counts, or spend amounts censored at zero. These features change both the estimands we target and the models we can use for estimation. We maintain the book's notation: treatment intensity is  $D_{it} \in \mathcal{D} \subseteq \mathbb{R}$  and potential outcomes are  $Y_{it}(d)$  or, in dynamic settings,  $Y_{it}(\underline{d}_i^t)$ . This chapter develops a panel framework for such settings. We formalise dose-response functions for continuous and multi-valued treatments, state the identification conditions under which they are causal objects (see Section 14.9), and then describe estimators, including DML-style implementations from Chapter 12, that target these objects under those conditions. We also introduce nonlinear panel models for discrete or limited outcomes and link their effect summaries back to the potential-outcomes notation from Chapter 2 and the dynamic estimands of Chapter 10. Unless stated otherwise, we treat units  $i$  as independent sampling units and allow serial dependence within each unit. When outcomes and treatment are clustered (for example by market or geo cell), we index clusters by  $c = 1, \dots, G$  and treat  $G$  as the effective sample size for inference.

## 14.1 Motivation and Setup

Many marketing interventions are not binary. A price change is a movement along a price grid, an advertising campaign has an intensity or budget, and a discount comes as a percentage rather than a yes–no flag. Outcomes are often nonlinear as well: customers make binary adoption choices, purchase counts are integers with many zeros, and spending is censored at zero. These features of the data require tools that go beyond simple linear models with binary treatments and continuous outcomes.

The potential-outcomes framework from Chapter 2 extends to this setting by indexing potential outcomes by treatment doses or paths. For a static continuous treatment with dose  $d$ , we write  $Y_{it}(d)$  for the outcome that would be observed for unit  $i$  at time  $t$  under that dose. Here  $Y_{it}(d)$  refers to the period- $t$  outcome. When an application summarises outcomes over a window (for example, cumulative revenue from  $t$  to  $t+K$ ), we define that aggregation explicitly and keep the primitive object  $Y_{it}(d)$  in the background. For dynamic settings, Chapter 10 indexes potential outcomes by entire treatment paths  $Y_{it}(\underline{d}_i^t)$ . When we use a scalar summary (for example, a cumulative or average dose over a horizon), we define it explicitly as a function of  $\underline{d}_i^t$  and state the horizon. Causal interpretation remains conditional on the identification assumptions in Section 14.9.

As always, causal interpretation rests on a credible assignment mechanism. For continuous treatments, the key requirements are conditional independence or a design-based analogue, together with overlap and support in the dose region relevant for the estimand. You cannot learn  $\mu(d)$  at dose levels that are never or rarely observed in the relevant subpopulation, and extrapolation is a modelling choice rather than a design.

Panel structure adds two practical hazards. First, you must avoid bad controls. Any covariate that is itself affected by  $D_{it}$ , including many “engagement” and “demand” measures, can act as a mediator or collider and distort causal comparisons. Second, you need stability, in the sense that untreated potential outcomes and the assignment mechanism do not drift in ways that invalidate comparisons across dose changes.

For nonlinear outcomes, link functions and distributional assumptions govern how we parameterise conditional means and compute interpretable effect summaries. They do not supply identification, which still comes from the assignment mechanism.

The methods in this chapter complement the design-based approaches in earlier chapters. When a staggered rollout or experiment secures identification at the design level, continuous-treatment and nonlinear panel models let you ask richer questions about marginal responses to intensity, saturation points, and outcome distributions, while remaining anchored in the same potential-outcomes and panel-causal framework used throughout the book (see also Arkhangelsky and Imbens [2024]). Throughout, we link estimands tightly to their identification conditions, estimation strategies, and inference tools. We emphasise diagnostics for overlap and support in the dose space, sensitivity to modelling and tuning choices, and reconciliation with simpler design-based benchmarks.

## 14.2 Dose–Response Estimands

When treatment is continuous rather than binary, the key causal objects become functions that describe how outcomes move with treatment intensity, rather than a single average treatment effect. The basic building block is the average dose–response function.

**Definition 14.1 (Average Dose–Response Function)** Let  $D_{it} \in \mathcal{D} \subseteq \mathbb{R}$  denote a continuous treatment, and write  $d$  for a realised dose. The Average Dose–Response Function (ADRF) is

$$\mu(d) = \mathbb{E}[Y_{it}(d)], \quad d \in \mathcal{D},$$

where  $Y_{it}(d)$  is the potential outcome for unit  $i$  at time  $t$  under dose  $d$ . The ADRF maps treatment intensities to average outcomes across the population. Unless stated otherwise, the expectation is taken over the distribution of unit–time cells  $(i, t)$  in the target population. When past doses matter,  $Y_{it}(d)$  is shorthand for a restriction of the dynamic primitive  $Y_{it}(\underline{d}_i^t)$ , and we state the required no-anticipation or no-carryover conditions explicitly.

If an application uses an aggregated outcome (for example, cumulative revenue over a horizon), we define that aggregation explicitly and keep the primitive potential outcome notation  $Y_{it}(d)$  in the background.

**Definition 14.2 (Conditional Dose–Response Function)** The Conditional Dose–Response Function (CDRF) conditions the ADRF on observed covariates  $X_{it}$ :

$$\mu(d | x) = \mathbb{E}[Y_{it}(d) | X_{it} = x], \quad d \in \mathcal{D}, \quad x \in \mathcal{X}.$$

The CDRF describes how expected outcomes at dose  $d$  vary across unit–time cells with covariate profile  $X_{it} = x$ . The ADRF is recovered by averaging over the covariate distribution:  $\mu(d) = \mathbb{E}[\mu(d | X_{it})]$ .

**Definition 14.3 (Marginal and Average Partial Effects)** The marginal treatment effect at dose  $d$  is the derivative of the ADRF,

$$\tau(d) = \frac{\partial \mu(d)}{\partial d},$$

which captures the incremental change in the expected outcome from a small increase in treatment intensity around  $d$ . The average partial effect (APE) over a dose range  $[d_0, d_1]$  is

$$\text{APE}(d_0, d_1) = \frac{\mu(d_1) - \mu(d_0)}{d_1 - d_0} = \frac{1}{d_1 - d_0} \int_{d_0}^{d_1} \tau(u) du.$$

A conditional marginal effect is defined analogously as  $\tau(d | x) = \partial \mu(d | x) / \partial d$ .

We reserve  $\tau(d)$  for marginal effects in this chapter. In staggered-adoption settings elsewhere in the book,  $\tau(g, t)$  denotes a cohort–time effect.

In dynamic settings, we also study path-dependent marginal responses (ACR) that differentiate with respect to a component of the treatment path. Section 14.8 defines this object formally.

**Definition 14.4 (Treatment Effect Contrasts and Optimal Dose)** For any two doses  $d_1, d_0 \in \mathcal{D}$ , the average treatment effect of dose  $d_1$  relative to  $d_0$  is

$$\text{ATE}(d_1, d_0) = \mu(d_1) - \mu(d_0) = \mathbb{E}[Y_{it}(d_1) - Y_{it}(d_0)].$$

For policy optimisation, suppose that applying dose  $d$  incurs a cost  $c(d)$  in outcome units (for example, expected margin). A decision rule based on the ADRF chooses

$$d^* = \arg \max_{d \in \mathcal{D}} \{\mu(d) - c(d)\},$$

and interpreting the resulting  $d^*$  causally requires that  $\mu(d)$  is identified over the decision-relevant dose region and that  $c(d)$  is specified in the same units as  $\mu(d)$  (see Section 14.9). When the optimum is interior and  $\mu$  and  $c$  are differentiable, a first-order condition is  $\tau(d^*) = c'(d^*)$ .

## Connecting to Event Studies and Dynamics

Event-time and dynamic-multiplier summaries from Chapters 5 and 10 remain central in marketing applications. When treatment intensity changes at identified event times, for example when an advertising budget steps up or a price discount deepens, the relevant object is a dose contrast  $\mu(d_1) - \mu(d_0)$  between the post-change and pre-change doses (Definition 14.4), possibly indexed by horizon. You can then summarise the implied path of effects over event time and compute multipliers, linking the continuous-treatment perspective in this chapter back to the dynamic estimands and impulse responses developed earlier.

## 14.3 Identification

Dose-response estimands are identified in panels only under explicit assumptions. The most common routes are selection on observables, implemented through the generalised propensity score, parallel-trends restrictions adapted to intensity, and factor-model or instrumental-variable strategies when selection on unobservables is a concern. This section summarises the main identification routes. Detailed conditions and estimators follow in later sections and chapters.

### Identification via Unconfoundedness

The first approach assumes continuous-treatment unconfoundedness. After conditioning on an admissible covariate set, dose variation is independent of potential outcomes. In the continuous-treatment setting this is encoded through the generalised propensity score (GPS).

**Definition 14.5 (Generalised Propensity Score)** Let  $D_{it}$  denote a continuous treatment and  $d$  a realised dose. Conceptually, the generalised propensity score (GPS) is the conditional density of the treatment given covariates and additive unit and time effects,

$$r(d | X_{it}, \alpha_i, \lambda_t) = f_{D|X,\alpha,\lambda}(d | X_{it}, \alpha_i, \lambda_t),$$

where  $f_{D|X,\alpha,\lambda}(\cdot | \cdot)$  denotes the conditional density of  $D_{it}$  evaluated at  $d$ . The GPS generalises the binary propensity score  $e(X_{it}) = \mathbb{P}(D_{it} = 1 | X_{it})$  to continuous or multi-valued treatments, allowing us to summarise the information in  $(X_{it}, \alpha_i, \lambda_t)$  that is relevant for the dose assignment. This is the panel analogue of the GPS introduced by Hirano and Imbens [2004]. In practice,  $\alpha_i$  and  $\lambda_t$  are not observed. We operationalise this conditioning by using observed proxies for  $\alpha_i$  (for example, correlated-random-effects summaries such as unit means of pre-treatment covariates or outcomes) and flexible time controls for  $\lambda_t$ , or by transforming the data to remove additive fixed effects before estimating the dose model.

When treatment affects outcomes with lags, a static statement such as  $Y_{it}(d) \perp\!\!\!\perp D_{it} | X_{it}, \alpha_i, \lambda_t$  is shorthand for a restriction on the dynamic primitive  $Y_{it}(d_i^t)$ . In such settings, identification typically requires either a no-carryover restriction that makes  $Y_{it}(d)$  well defined as a period- $t$  object, or a sequential ignorability condition that rules out unmeasured time-varying confounding.

Under the continuous-treatment unconfoundedness and overlap assumptions in Section 14.9, the GPS has a balancing property analogous to the binary case.

**Proposition 14.1 (GPS Balancing)** Suppose that for all  $d$  in an overlap region  $\mathcal{D}_0 \subseteq \mathcal{D}$ ,

$$Y_{it}(d) \perp\!\!\!\perp D_{it} | X_{it}, \alpha_i, \lambda_t$$

and that overlap holds on  $\mathcal{D}_0$  in the sense that  $r(d | X_{it}, \alpha_i, \lambda_t)$  is not vanishingly small for covariate profiles in the target population. In practice,  $\mathcal{D}_0$  is defined by trimming sparse tails of the dose distribution. Define

the scalar GPS evaluated at the realised dose as  $R_{it}(D_{it}) = r(D_{it} | X_{it}, \alpha_i, \lambda_t)$ . Then within strata of this scalar, the realised dose is independent of the covariates (and fixed effects) that drive selection. In particular, conditional on  $R_{it}(D_{it})$ , we have

$$D_{it} \perp\!\!\!\perp (X_{it}, \alpha_i, \lambda_t),$$

so that for any function  $g$ ,

$$\mathbb{E}[g(X_{it}, \alpha_i, \lambda_t) | D_{it} = d, r(d | X_{it}, \alpha_i, \lambda_t) = r] = \mathbb{E}[g(X_{it}, \alpha_i, \lambda_t) | r(d | X_{it}, \alpha_i, \lambda_t) = r].$$

This balancing property underpins covariate-balance diagnostics and GPS-based adjustment in continuous-dose settings.

**Theorem 14.1 (GPS Identification)** Recall that the average dose-response function (ADRF) is  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  for doses  $d$  in  $\mathcal{D}$ , as defined in Section 14.2. Under continuous-treatment unconfoundedness and overlap,  $\mu(d)$  is identified by the regression-on-GPS representation

$$\mu(d) = \mathbb{E}\left[\mathbb{E}[Y_{it} | D_{it} = d, r(d | X_{it}, \alpha_i, \lambda_t)]\right].$$

Equivalently, kernel-based inverse-probability formulations approximate  $\mu(d)$  as

$$\mu(d) \approx \frac{\mathbb{E}[Y_{it} K_h(D_{it} - d)/r(D_{it} | X_{it}, \alpha_i, \lambda_t)]}{\mathbb{E}[K_h(D_{it} - d)/r(D_{it} | X_{it}, \alpha_i, \lambda_t)]},$$

where  $K_h(\cdot)$  is a kernel with bandwidth  $h$  that localises around  $d$ . As  $h \rightarrow 0$  and  $Nh \rightarrow \infty$  under standard smoothness and support conditions, with  $N$  independent units and serial dependence handled via unit-level clustering as in Chapter 16, these expressions converge to the ADRF for interior doses. If  $T$  grows with  $N$ , the effective sample size and bandwidth conditions must be adjusted accordingly. These representations express  $\mu(d)$  as a functional of observable quantities under unconfoundedness and overlap. Estimators in Section 14.4 implement these representations with estimated nuisance functions and cross-fitting.

Trimming to define an overlap region  $\mathcal{D}_0$  changes the target. In practice, we interpret the identified object as the ADRF restricted to  $d \in \mathcal{D}_0$ . Outside  $\mathcal{D}_0$ ,  $\mu(d)$  is not identified without extrapolation.

For asymptotics and inference, the relevant independent sampling unit is whatever you can credibly treat as independent. In many marketing panels that is the unit  $i$ , but in geo and market panels it is often the cluster  $c = 1, \dots, G$ . In the clustered case,  $G$  is the effective sample size, and inference should use a cluster-robust variance estimator or a valid cluster bootstrap.

## Identification via Parallel Trends

A second route adapts the parallel-trends logic from Chapter 4 to continuous treatments that vary over time within units. The idea is to treat changes in intensity, rather than binary on/off switches, as the “treatment” in a difference-in-differences design.

**Assumption 64 (Continuous-Dose Parallel Trends)** For all doses  $d$  in  $\mathcal{D}$  and all periods  $t > 1$ , conditional on  $(X_{it}, \alpha_i, \lambda_t)$ ,

$$\mathbb{E}[Y_{it}(d) - Y_{i,t-1}(d) | D_{it}, X_{it}, \alpha_i, \lambda_t] = \mathbb{E}[Y_{it}(d) - Y_{i,t-1}(d) | X_{it}, \alpha_i, \lambda_t].$$

That is, outcome trends under any fixed dose  $d$  are mean-independent of the realised dose path, conditional on observables and fixed effects.

Under Assumption 64, comparisons of outcome changes across units with different dose changes identify contrasts such as  $\mu(d_1) - \mu(d_0)$  or local marginal effects  $\tau(d) = \partial\mu(d)/\partial d$ . The estimand is typically a difference in dose-response levels or a derivative, not the full ADRF at a single dose. We return to dynamic, event-time versions of this idea in Section 14.8.

## Identification via Factors or Instruments

When selection on observables is implausible, we can instead exploit structure in unobservables or exogenous sources of variation in treatment intensity.

Factor models, as developed in Chapter 8, posit that untreated potential outcomes follow a low-rank structure,

$$Y_{it}(0) = \alpha_i + \lambda_t + \sum_{r=1}^R \lambda_{ir} f_{tr} + \varepsilon_{it},$$

where  $\alpha_i$  and  $\lambda_t$  are additive fixed effects,  $\lambda_{ir}$  are unit-specific loadings,  $f_{tr}$  are common factors, and  $\varepsilon_{it}$  is idiosyncratic noise. Identification relies on the factor-stability assumption: loadings  $\lambda_{ir}$  are invariant to treatment, and  $\mathbb{E}[\varepsilon_{it} | D_{it}, \{f_{tr}\}_{r=1}^R, \{\lambda_{ir}\}_{r=1}^R] = 0$ . Under appropriate rank conditions ( $R \ll \min(N, T)$ ), we can estimate the factor structure from untreated cells and project out these unobserved components before analysing the relationship between dose and outcomes. For continuous  $D_{it}$ , this approach identifies the ADRF  $\mu(d)$  only if the factor structure extends to all potential outcomes  $Y_{it}(d)$  with the same loadings, a strong assumption that should be stated explicitly.

Instrumental-variable strategies for continuous doses require an instrument  $Z_{it}$  that shifts  $D_{it}$  while satisfying relevance ( $Z_{it}$  affects  $D_{it}$ ), exclusion ( $Z_{it}$  affects outcomes only through  $D_{it}$ ), and an independence or exogeneity condition conditional on observables and fixed effects. In linear-dose models where  $D_{it}$  enters the outcome equation linearly,

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + X'_{it} \gamma + \varepsilon_{it},$$

standard two-stage least squares identifies the coefficient  $\beta$ , which corresponds to a constant marginal effect  $\tau(d) = \beta$ . For nonlinear dose-response, a single instrument generally identifies only a local derivative or weighted average of marginal effects, not the entire ADRF curve. Cost shifters, regulatory changes, or platform-side allocation rules can sometimes serve as instruments in marketing panels. Chapter 16 and Section 16.10 spell out these conditions and their implementation in detail.

Across all three strategies, the key is to be explicit about which identification route you are using for a given application, to align estimation methods with that route, and to use diagnostics such as balance diagnostics, placebo falsification checks, and robustness to alternative designs to assess whether the chosen assumptions are remotely plausible in the marketing context at hand. Section 14.11 and Chapter 17 provide concrete guidance on these diagnostic procedures.

## 14.4 Estimation Strategies for Continuous Treatments

We now turn from estimands and identification to concrete estimators for continuous treatments. We begin with outcome-regression and GPS-based estimators for the ADRF, then show how to combine them into doubly robust and DML estimators suitable for panel settings.

### GPS and Outcome Regression

A practical starting point is to combine a model for the conditional dose density,  $r(d | X_{it})$ , with a model for the conditional outcome,  $m(d, X_{it})$ . Conceptually, identification conditions are stated conditional on  $(X_{it}, \alpha_i, \lambda_t)$ , as in Section 14.3. Operationally, this means that the covariate vector used in nuisance estimation should include admissible proxies for  $\alpha_i$  (for example, correlated-random-effects summaries such as unit means of pre-treatment covariates or outcomes) and flexible time controls for  $\lambda_t$ , or else you should transform the data to remove additive unit and time effects before estimating nuisance functions.

**Definition 14.6 (Outcome Regression for ADRF)** The outcome regression function is

$$m(d, X) = \mathbb{E}[Y_{it} | D_{it} = d, X_{it} = X],$$

so that the ADRF can be expressed as  $\mu(d) = \mathbb{E}_X[m(d, X)]$ . An outcome-regression estimator replaces  $m$  with an estimate  $\hat{m}$  and averages over the covariate distribution:

$$\hat{\mu}^{\text{OR}}(d) = \frac{1}{NT} \sum_{i,t} \hat{m}(d, X_{it}),$$

where  $N$  is the number of units,  $T$  the number of periods, and  $\hat{m}(d, X)$  is estimated using flexible methods in  $d$  and  $X$  (for example, splines, series, or machine learning). Here  $X_{it}$  denotes the admissible feature set used for adjustment, including any fixed-effect proxies and time controls used in the identification argument. Unless stated otherwise, this estimator targets the cell-average ADRF, averaging over unit-time cells. If an application targets a unit-level ADRF, define the unit-level aggregation explicitly and average over units  $i$ .

**Definition 14.7 (GPS-Weighted Estimator)** Given an estimate  $\hat{r}(d | X)$  of the generalised propensity score (GPS), a GPS-weighted inverse-probability estimator of  $\mu(d)$  takes the form

$$\hat{\mu}^{\text{IPW}}(d) = \frac{\sum_{i,t} Y_{it} K_h(D_{it} - d) / \hat{r}(D_{it} | X_{it})}{\sum_{i,t} K_h(D_{it} - d) / \hat{r}(D_{it} | X_{it})},$$

where  $K_h(\cdot)$  is a bounded, symmetric kernel with  $\int K(u) du = 1$  and bandwidth  $h$  that localises around  $d$ . Stabilised weights divide  $\hat{r}(D_{it} | X_{it})$  by an estimate of the marginal dose density  $\hat{f}(D_{it})$  to reduce variance when the marginal dose distribution is uneven.

Outcome-regression estimators are efficient when  $m(d, X)$  is well specified but can be biased when it is not. GPS-weighted estimators are robust to misspecification of  $m$  but sensitive to errors and extreme values in  $\hat{r}$ . Doubly robust estimators combine the two.

**Definition 14.8 (Doubly Robust ADRF Estimator)** The doubly robust estimator for  $\mu(d)$  combines outcome regression and GPS weighting:

$$\hat{\mu}^{\text{DR}}(d) = \frac{1}{NT} \sum_{i,t} \left[ \hat{m}(d, X_{it}) + \frac{K_h(D_{it} - d)}{\hat{r}(D_{it} | X_{it})} \{Y_{it} - \hat{m}(D_{it}, X_{it})\} \right].$$

The term in brackets is an orthogonal score for  $\mu(d)$  that combines imputation with an inverse-GPS residual correction localised around dose  $d$ .

**Theorem 14.2 (Double Robustness)** *Under the continuous-treatment unconfoundedness and overlap assumptions from Section 14.9, and under standard smoothness and bandwidth conditions ( $h \rightarrow 0$ ,  $Nh \rightarrow \infty$  as  $N \rightarrow \infty$ , with  $N$  the number of units), the estimator  $\hat{\mu}^{\text{DR}}(d)$  has the following properties for each interior dose  $d$ :*

1. *If  $\hat{m}(d, X)$  converges in probability to the true conditional mean  $m_0(d, X)$ , then  $\hat{\mu}^{\text{DR}}(d)$  converges to  $\mu(d)$  even if  $\hat{r}$  is misspecified.*
2. *If  $\hat{r}(d | X)$  converges to the true GPS  $r_0(d | X)$ , then  $\hat{\mu}^{\text{DR}}(d)$  converges to  $\mu(d)$  even if  $\hat{m}$  is misspecified.*
3. *If both nuisance functions are correctly specified and estimated with sufficient accuracy,  $\hat{\mu}^{\text{DR}}(d)$  is efficient under correct specification and appropriate regularity conditions.*

*The score in Definition 14.8 is Neyman-orthogonal to first-order perturbations in  $m$  and  $r$ , which underpins the double robustness result and the debiasing argument. Chapter 12 develops the general theory of orthogonal scores.*

## Double or Debiased ML for Continuous Doses

Double machine learning (DML), introduced in Chapter 12, extends these ideas by using cross-fitting to decouple nuisance estimation from score evaluation. Orthogonal scores such as the one in Definition 14.8 then support valid inference even when  $\hat{m}$  and  $\hat{r}$  are estimated with flexible learners.

**Definition 14.9 (DML Estimator for ADRF)** Partition the independent sampling units into  $K$  folds, as in Section 12.3. When units  $i$  are independent, let  $\mathcal{I}_k \subset \{1, \dots, N\}$  denote the set of unit indices in fold  $k$ . When dependence is clustered, partition clusters  $c = 1, \dots, G$  and let  $\mathcal{I}_k$  denote the set of units in the held-out clusters.

For each fold  $k$ :

1. Estimate nuisance functions  $\hat{m}^{(-k)}(d, X)$  and  $\hat{r}^{(-k)}(d | X)$  using all observations  $(i, t)$  that are not in the evaluation fold.

2. For each observation  $(i, t)$  in the evaluation fold, construct the orthogonal score

$$\psi_{it}(d) = \hat{m}^{(-k)}(d, X_{it}) + \frac{K_h(D_{it} - d)}{\hat{r}^{(-k)}(D_{it} | X_{it})} \{Y_{it} - \hat{m}^{(-k)}(D_{it}, X_{it})\}.$$

Bandwidth  $h$  is part of the estimator design and should be treated as a tuning parameter. Its choice trades bias and variance, and sensitivity to  $h$  should be reported.

The DML estimator of  $\mu(d)$  averages these cross-fitted scores over all unit-time observations,

$$\hat{\mu}^{\text{DML}}(d) = \frac{1}{NT} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \sum_{t=1}^T \psi_{it}(d).$$

In panels, inference is based on the influence function aggregated over the independent sampling unit. When units are independent,  $\Psi_i(d) = \sum_t \psi_{it}(d)$ . When clusters are independent, aggregate further to  $\Psi_c(d) = \sum_{i \in \mathcal{C}_c} \sum_t \psi_{it}(d)$  and treat  $G$  as the effective sample size.

**Assumption 65 (Nuisance-Rate Conditions for Continuous DML)** Let  $N$  denote the number of independent units and treat  $T$  as fixed or moderate. Suppose nuisance estimators are accurate enough that, with cross-fitting, the remainder term in the orthogonal-score expansion of  $\hat{\mu}^{\text{DML}}(d)$  is  $o_P(N^{-1/2})$  for each interior dose  $d$ . In many DML settings this can be ensured by  $N^{-1/4}$ -type rates for each nuisance estimator. The exact requirements depend on how  $r(d | X)$  is estimated and on the localisation scheme.

**Theorem 14.3 (ADRF Inference with DML)** *Under the continuous-treatment identification assumptions from Section 14.3 and Assumption 65, the DML estimator satisfies asymptotic normality for each interior dose  $d$ , with scaling determined by the independent sampling unit.*

When units are independent,

$$\sqrt{N} (\hat{\mu}^{\text{DML}}(d) - \mu(d)) \xrightarrow{d} \mathcal{N}(0, V(d)),$$

where  $V(d) = \mathbb{E}[\Psi_i(d)^2]$  and  $\Psi_i(d) = \sum_t \psi_{it}(d)$  is the unit-level influence contribution. When clusters  $c = 1, \dots, G$  are independent, replace  $N$  by  $G$  and use the cluster-level influence contributions  $\Psi_c(d)$ .

Cluster-robust variance estimators based on the sample analogue of  $V(d)$  yield valid pointwise confidence intervals for  $\mu(d)$ . Uniform inference over ranges of  $d$  typically requires bootstrap or multiplier methods, as discussed in Chapter 16.

## Local and Global Marginal Effects

Smooth estimates of  $\mu(d)$ , obtained via series, splines, or flexible ML methods, provide marginal effects  $\tau(d) = \partial\mu(d)/\partial d$  and average partial effects over ranges of doses. In skewed or highly concentrated dose distributions, partial effects at the mean dose can be misleading. A more informative summary is the average partial effect over an interval,

$$\text{APE}(d_0, d_1) = \frac{1}{d_1 - d_0} \int_{d_0}^{d_1} \tau(u) du = \frac{\mu(d_1) - \mu(d_0)}{d_1 - d_0},$$

which can be estimated by evaluating  $\hat{\mu}(d)$  over a grid of doses and taking finite differences. Reporting confidence bands over a grid of economically relevant doses, rather than a single point estimate at the mean, gives a more complete picture of the dose-response relationship.

Credible marginal effects and partial effects require adequate support in a neighbourhood of  $d$ . Extrapolating  $\hat{\mu}(d)$  outside the support of observed  $D_{it}$  is dangerous and common in marketing applications where doses such as GRPs or ad spend are highly skewed. Before reporting marginal effects, diagnose support using GPS or dose-density plots: if few observations fall near  $d$ , the kernel localisation will be driven by extrapolation rather than data, and the resulting estimates will be unreliable.

These estimators revive the prediction–identification tension highlighted by Breiman [2001]. Learners that forecast outcomes well do not automatically deliver credible counterfactuals. Orthogonality and cross-fitting protect inference against first-order nuisance errors under the stated identification assumptions. They do not protect against violations of unconfoundedness, poor overlap, or a misspecified target population.

## 14.5 Nonlinear Panel Outcome Models

When outcomes are not continuous, we still want to link them to the potential-outcomes framework and to the estimands defined earlier in the chapter. Binary choices, counts, and censored outcomes require nonlinear mean functions and link functions that respect their support. This section covers fixed-effects logit for binary outcomes, fixed-effects Poisson for counts, the computation of average marginal effects for causal interpretation, and brief remarks on interactive fixed effects and profit-aware classification.

Throughout, causal interpretation requires an identification assumption about the assignment mechanism. A convenient sufficient condition is continuous-treatment unconfoundedness. For doses  $d$  in the reporting region,

$$Y_{it}(d) \perp\!\!\!\perp D_{it} \mid X_{it}, \alpha_i, \lambda_t$$

together with overlap and stability conditions as in Section 14.9. The effects  $\alpha_i$  and  $\lambda_t$  are conceptual and are operationalised via admissible proxies or transformations. Here  $D_{it}$  denotes the causal variable of interest, and additional observed controls belong in  $X_{it}$ .

**Definition 14.10 (Conditional Fixed-Effects Logit)** For binary outcomes  $Y_{it} \in \{0, 1\}$ , the two-way fixed-effects logit model specifies

$$P(Y_{it} = 1 \mid D_{it}, X_{it}, \alpha_i, \lambda_t) = \Lambda(\alpha_i + \lambda_t + D_{it}\tau + X'_{it}\beta),$$

where  $\Lambda(z) = e^z/(1 + e^z)$  is the logistic CDF,  $\alpha_i$  are unit fixed effects,  $\lambda_t$  are time fixed effects that absorb common shocks such as platform-wide changes in targeting or seasonality,  $\beta_D$  is the slope on  $D_{it}$ , and  $\beta$  collects other coefficients. Under unconfoundedness and correct model specification,  $\beta_D$  indexes how the conditional log-odds varies with  $D_{it}$ , holding  $X_{it}$  and the fixed effects constant. For causal reporting, translate this parameter into outcome-scale contrasts such as  $\mu(d_1) - \mu(d_0)$  or marginal effects computed from the fitted model.

The conditional likelihood, obtained by conditioning on the sufficient statistic  $\sum_t Y_{it}$  for  $\alpha_i$ , eliminates the unit-level incidental parameters:

$$\mathcal{L}_i^{\text{cond}}(\beta_D, \beta) = P(\{Y_{it}\}_{t=1}^T \mid \sum_{t=1}^T Y_{it}, \{D_{it}, X_{it}\}_{t=1}^T).$$

Units with  $\sum_t Y_{it} = 0$  or  $T$  contribute no information about  $\beta_D$  and drop out of the conditional likelihood. This changes the effective target population. Identification comes from units whose outcomes vary over time, and you should state whether that restriction is substantively acceptable. In practice, time effects  $\lambda_t$  are often absorbed by including period dummies in  $X_{it}$  or by using within-period comparisons.

**Proposition 14.2 (Incidental-Parameters Bias)** *In many nonlinear fixed-effects models with  $N \rightarrow \infty$  and fixed  $T$ , the maximum likelihood estimator for  $(\alpha_1, \dots, \alpha_N, \beta_D, \beta)$  is inconsistent for  $\beta_D$  because the number of incidental parameters  $\alpha_i$  grows with  $N$ . Bias is typically of order  $1/T$  in short panels and does not vanish as  $N \rightarrow \infty$  with fixed  $T$ .*

For fixed-effects logit, the conditional MLE that conditions on  $\sum_t Y_{it}$  removes  $\alpha_i$  from the likelihood and yields a consistent estimator of  $\beta_D$  under correct logit specification and the unconfoundedness and overlap conditions above. For probit models, no exact conditioning suffices to eliminate  $\alpha_i$ , so bias-corrected or jackknife estimators are required when  $T$  is small. The Poisson fixed-effects estimator in Theorem 14.4 is a notable exception: the exponential-family structure preserves consistency for  $(\beta_D, \beta)$  even when  $T$  is fixed.

**Definition 14.11 (Fixed-Effects Poisson)** For count outcomes  $Y_{it} \in \{0, 1, 2, \dots\}$ , the two-way fixed-effects Poisson model specifies the conditional mean

$$\mathbb{E}[Y_{it} | D_{it}, X_{it}, \alpha_i, \lambda_t] = \exp(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta).$$

The corresponding log-likelihood is

$$\ell(\beta_D, \beta, \{\alpha_i\}, \{\lambda_t\}) = \sum_{i,t} \left[ Y_{it}(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta) - \exp(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta) \right],$$

up to terms that do not depend on the parameters. In practice, time effects  $\lambda_t$  are often absorbed by including period dummies in  $X_{it}$ .

**Theorem 14.4 (Poisson Fixed-Effects Consistency)** Suppose the conditional mean is correctly specified as

$$\mathbb{E}[Y_{it} | D_{it}, X_{it}, \alpha_i, \lambda_t] = \exp(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta),$$

with the unconfoundedness and overlap conditions from earlier chapters holding for  $D_{it}$ . Then:

1. The fixed-effects Poisson pseudo-MLE (pseudo-MLE because we base inference on the Poisson likelihood even when the conditional variance need not equal the mean) for  $(\beta_D, \beta)$  is consistent as  $N \rightarrow \infty$  with  $T$  fixed, despite the presence of many  $\alpha_i$ , because of the exponential-family structure.
2. The conditional variance of  $Y_{it}$  may differ from the mean (overdispersion). In that case the Poisson pseudo-MLE remains consistent for  $(\tau, \beta)$  but is no longer fully efficient.
3. Cluster-robust (sandwich) standard errors with clustering at the unit level account for overdispersion and within-unit serial dependence, as in Chapter 16.

For nonlinear models we often interpret  $\beta_D$  through marginal effects on the outcome scale. Under unconfoundedness and correct model specification, model-implied marginal effects can be interpreted as derivatives of  $\mu(d)$ , or of a conditional dose-response function, averaged over the sample distribution.

**Definition 14.12 (Average Marginal Effect)** For a nonlinear mean function

$$\mathbb{E}[Y_{it} | D_{it}, X_{it}, \alpha_i, \lambda_t] = g(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta),$$

with scalar treatment component  $D_{it}$ , the average marginal effect (AME) of treatment is the average derivative of the conditional mean with respect to  $D_{it}$ :

$$\text{AME} = \mathbb{E} \left[ \frac{\partial g(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta)}{\partial D_{it}} \right] = \mathbb{E} [g'(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta) \cdot \beta_D],$$

where the expectation is taken over the empirical distribution of  $(i, t)$  in the sample. Write  $\eta_{it} = \alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta$ . For logit,  $g(z) = \Lambda(z)$  and  $g'(z) = \Lambda(z)(1 - \Lambda(z))$ , so the marginal effect at  $(i, t)$  is  $\Lambda(\eta_{it})\{1 - \Lambda(\eta_{it})\}\beta_D$  and

$$\text{AME}_{\text{logit}} = \mathbb{E}[\Lambda(\eta_{it})\{1 - \Lambda(\eta_{it})\}\beta_D].$$

For Poisson,  $g(z) = e^z$  and  $g'(z) = e^z$ , giving

$$\text{AME}_{\text{Poisson}} = \mathbb{E}[\exp(\eta_{it})\beta_D].$$

In practice these expectations are evaluated at fitted parameters. Reporting AMEs, or discrete contrasts such as  $\mu(1) - \mu(0)$  where applicable, provides a more interpretable measure of effect size than raw coefficients in nonlinear models. When  $D_{it}$  is binary,  $\mu(1) - \mu(0)$  coincides with the average treatment effect on the probability or count scale, depending on the outcome.

## Nonlinear Panels with Interactive Fixed Effects

Interactive fixed-effects models extend two-way fixed effects by allowing unit and time effects to interact through a low-rank factor structure. In potential-outcomes terms, we can write the untreated potential outcome for binary or count  $Y_{it}$  as

$$\mathbb{E}[Y_{it}(0) | X_{it}, \{f_{tr}\}_{r=1}^R, \{\lambda_{ir}\}_{r=1}^R] = g\left(\sum_{r=1}^R \lambda_{ir} f_{tr} + X'_{it}\beta\right),$$

where  $g$  is the logit or Poisson mean function, and  $\{f_{tr}\}$  and  $\{\lambda_{ir}\}$  form a low-rank factor structure as in Chapter 9. In a logit or Poisson panel, the latent index can thus contain both observed covariates and an unobserved matrix of factors that capture common shocks with heterogeneous unit-specific loadings.

These models are attractive in marketing when latent demand or category conditions evolve in ways that a single time effect or observed covariates cannot capture. The main challenge is computational: the fixed-effects estimator is defined as the solution to a high-dimensional, non-convex likelihood problem.

Zeleneev and Zhang [2025] propose a practical two-step procedure. First, solve a convex, nuclear-norm-regularised problem to obtain a low-rank approximation to the unobserved factor structure and coefficient vector. Second, run local optimisation of the original nonlinear likelihood starting from this preliminary estimate. They show that, under the same low-rank and factor-coverage assumptions as in Chapter 9, the resulting estimator is asymptotically equivalent to the full fixed-effects estimator. Their guarantees rely on strong factor-coverage and regularity conditions. In applied work you should treat the nuclear-norm step as an initialisation heuristic and stress-test sensitivity to rank and penalty choices.

For applied work, the message is operational but conditional. If you want to fit a logit or Poisson panel with interactive fixed effects, nuclear-norm tools provide a tractable way to initialise the problem before using a standard gradient-based optimiser. Identification of treatment effects rests on the same factor-coverage and rank conditions as in the linear case: the factor structure for  $Y_{it}(0)$  must be stable across treatment status,

and the factors must be recoverable from untreated cells. Computation alone does not cure violations of those assumptions.

**Table 14.1** Outcome-link decision guide and incidental-parameter cautions

Outcome type	When appropriate	Pitfalls and remedies
Binary (logit/probit)	Probabilities in $[0, 1]$ with panel fixed effects	Incidental parameters with short $T$ . Conditional fixed-effects logit avoids incidental-parameter bias by conditioning on $\sum_t Y_{it}$ . Use shrinkage or bias-correction for probit. Report marginal effects on the probability scale.
Count (Poisson)	Non-negative counts with possible exposure offsets	Overdispersion. Use cluster-robust variance at the unit level. Negative binomial models with fixed effects face more demanding incidental-parameter and identification constraints than Poisson. See Proposition 14.2.
Limited (tobit-like)	Censoring or truncation present	Incidental-parameter bias with fixed effects. Prefer semiparametric or distribution-free approaches and validate with sensitivity checks.

### Practitioner’s Note: Profit-Aware Classification

Standard classification models, such as logistic regression or gradient-boosted trees, typically optimise statistical accuracy: they weight false positives and false negatives symmetrically. In marketing that symmetry rarely holds. The cost of spamming a disinterested user may be small relative to the lost margin from missing a likely purchaser.

We can align classifiers more closely with business objectives by redefining the loss function. Babii et al. [2024] discuss reweighting classification losses to align prediction with an explicit profit or cost objective. You specify the profit or cost associated with each outcome—such as the profit from a successful retention offer versus the cost of sending the offer—and use these to weight observations in the training loss. The algorithm then prioritises errors that are most damaging to the bottom line.

This is a predictive, not a causal, adjustment: the reweighting changes the loss function used to train the classifier, not the assignment mechanism that determines which users receive offers. The model estimates conditional purchase or response probabilities under the existing assignment mechanism, but trains the classifier to make decisions that are profit-optimal given those predictions. For causal questions—such as the effect of an offer on purchase probability—you must still rely on the design and identification strategies developed elsewhere in the book.

In practice, you can embed profit-aware classifiers as candidate targeting rules within an experiment or uplift-modelling design and then use the causal machinery from Chapters 3 and 12.6 to evaluate their

incremental profit. The classifier determines who to target. The causal framework determines whether that targeting actually works.

## 14.6 Handling Excess Zeros: Hurdle and Zero-Inflated Models

Marketing panels frequently exhibit extreme zero-inflation. A retailer tracking ten million customers over fifty-two weeks may find that 90 per cent of customer-week observations record zero purchases. Even with fixed effects, a single log-linear mean function for  $Y_{it}$  given  $D_{it}, X_{it}, \alpha_i$ , and  $\lambda_t$  often struggles to capture such patterns. In these cases it is useful to distinguish between different zero-generating mechanisms and to estimate separate effects on participation (any purchase) and intensity (how much is bought, conditional on purchase).

This section develops two-part (hurdle) models and zero-inflated models for panel data, with particular attention to the causal interpretation of treatment effects on participation versus intensity. These models should be viewed as parameterisations of the outcome process layered on top of the identification strategies in earlier chapters, not as substitutes for design.

**Assumption 66 (Panel Unconfoundedness with Zero Handling)** For doses  $d$  in an overlap region  $\mathcal{D}_0 \subseteq \mathcal{D}$ ,

$$Y_{it}(d) \perp\!\!\!\perp D_{it} \mid X_{it}, \alpha_i, \lambda_t,$$

with overlap and stability conditions as in Sections 2.4 and 14.9. When we work with the participation indicator  $Y_{it}^{\text{ext}}(d) = \mathbf{1}\{Y_{it}(d) > 0\}$ , it inherits the same identification route. Hurdle, ZIP, and three-stage likelihoods do not relax these conditions. They are outcome models under the same identification assumptions.

### Why a Single Poisson Mean Can Be Misleading

The fixed-effects Poisson model (Definition 14.11) specifies

$$\mathbb{E}[Y_{it} \mid D_{it}, X_{it}, \alpha_i, \lambda_t] = \exp(\alpha_i + \lambda_t + D_{it}\beta_D + X'_{it}\beta) = \mu_{it}.$$

Poisson pseudo-MLE remains consistent for  $\beta_D$  when this conditional mean is correctly specified, even if the conditional variance differs from the mean. In heavily zero-inflated panels, a single log-linear mean can be a poor approximation. A two-part mixture is a modelling device that can fit the distribution better, but it does not create identification. One way to motivate it is to imagine two latent states: those who are effectively not at risk of purchase in that period, and those who are at risk but happen not to buy.

**Definition 14.13 (Structural vs Sampling Zeros)** Let  $Z_{it} \in \{0, 1\}$  indicate whether unit  $i$  at time  $t$  is “at risk” of a positive outcome.

Structural zeros correspond to  $Z_{it} = 0$ : outcomes are constrained to be zero over the dose range under consideration (for example, a customer who has not yet entered the category or an out-of-stock store). Sampling zeros correspond to  $Z_{it} = 1$ : the potential outcome  $Y_{it}(d)$  can be positive, but we observe  $Y_{it} = 0$  in this period. Standard count models conflate these two sources of zeros.

In most applications the distinction between structural and sampling zeros is not identified from the data without strong functional-form or design restrictions. We treat  $Z_{it}$  as a modelling device rather than an observable feature.

## The Two-Part (Hurdle) Model

Hurdle models treat participation and intensity as distinct components. A natural causal object for participation is the potential participation indicator  $Y_{it}^{\text{ext}}(d) = \mathbf{1}\{Y_{it}(d) > 0\}$  and its dose-response function  $\mu^{\text{ext}}(d) = \mathbb{E}[Y_{it}^{\text{ext}}(d)]$ . The overall dose-response remains  $\mu(d) = \mathbb{E}[Y_{it}(d)]$ . You can also define an intensive-margin object  $\mu^+(d) = \mathbb{E}[Y_{it}(d) | Y_{it}^{\text{ext}}(d) = 1]$ , but this is not the same as  $\mathbb{E}[Y_{it} | Y_{it} > 0, D_{it} = d]$ . The latter conditions on a post-treatment event and is descriptive unless you add assumptions that rule out selection.

**Definition 14.14 (Hurdle Model)** A hurdle model specifies two linked components:

1. **Participation (extensive margin):** a binary model for any purchase,

$$P(Y_{it} > 0 | D_{it}, X_{it}, \alpha_i^{(1)}, \lambda_t) = \Lambda(\alpha_i^{(1)} + \lambda_t + D_{it}\beta_D^{\text{ext}} + X'_{it}\beta_1),$$

where  $\Lambda(\cdot)$  is the logistic CDF.

2. **Intensity (intensive margin):** a zero-truncated count model for positive purchases,

$$Y_{it} | Y_{it} > 0, D_{it}, X_{it}, \alpha_i^{(2)}, \lambda_t \sim f(y | \mu_{it}, \phi),$$

with  $\mu_{it} = \exp(\alpha_i^{(2)} + \lambda_t + D_{it}\beta_D^{\text{int}} + X'_{it}\beta_2)$  and  $f$  a zero-truncated Poisson or negative-binomial density.

We allow the participation and intensity equations to have distinct unit effects  $\alpha_i^{(1)}$  and  $\alpha_i^{(2)}$  to capture the possibility that the unobserved drivers of extensive and intensive behaviour differ. The unconditional mean factors as

$$\mathbb{E}[Y_{it}] = P(Y_{it} > 0) \cdot \mathbb{E}[Y_{it} | Y_{it} > 0].$$

This representation yields two slope parameters:  $\beta_D^{\text{ext}}$  in the participation equation and  $\beta_D^{\text{int}}$  in the intensity equation. Under Assumption 66 and correct model specification,  $\beta_D^{\text{ext}}$  parameterises how the purchase probability varies with dose, while the intensity equation parameterises  $\mathbb{E}[Y_{it} | Y_{it} > 0, D_{it} = d, \dots]$  and thus describes basket size among observed buyers. For causal reporting, treat the unconditional mean  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  and the participation dose-response  $\mu^{\text{ext}}(d) = \mathbb{E}[\mathbf{1}\{Y_{it}(d) > 0\}]$  as primary targets. Conditional-on-buyers contrasts require additional assumptions because treatment can change participation.

**Proposition 14.3 (Decomposition of Total Effect)** Let  $\pi_{it} = P(Y_{it} > 0)$  and  $\mu_{it}^+ = \mathbb{E}[Y_{it} | Y_{it} > 0]$ . Under differentiability in  $D_{it}$ , the derivative of the unconditional mean with respect to treatment can be decomposed pointwise in  $(i, t)$  as

$$\frac{\partial \mathbb{E}[Y_{it}]}{\partial D_{it}} = \underbrace{\frac{\partial \pi_{it}}{\partial D_{it}} \mu_{it}^+}_{\text{Extensive margin}} + \underbrace{\pi_{it} \frac{\partial \mu_{it}^+}{\partial D_{it}}}_{\text{Intensive margin}}.$$

This decomposition separates the effect of treatment on the probability of any purchase from its effect on average purchase size among those who buy. Averaging this expression over units and time yields an analogous decomposition for the derivative of the ADRF  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  into extensive and intensive components.

**Estimation and Selection Bias.** The hurdle decomposition is attractive for interpretation but requires care in estimation. A naive approach would estimate the extensive margin on  $\mathbf{1}\{Y_{it} > 0\}$  and the intensive margin on the subset  $Y_{it} > 0$ . When treatment affects participation, conditioning on  $Y_{it} > 0$  is conditioning on a post-treatment variable and induces selection bias.

For example, suppose a discount draws marginal, low-spending customers into the store. The treated group with  $Y_{it} > 0$  now includes these new low-spenders alongside loyal high-spenders, while the control group with  $Y_{it} > 0$  still consists mainly of loyal high-spenders. Comparing mean basket sizes may suggest a negative intensive effect—even if every individual’s spending rises—because of the composition shift. This is a textbook case of collider bias.

A pragmatic strategy is to estimate three related summaries and be explicit about their interpretation:

First, estimate a model for the unconditional mean using PPML as a quasi-likelihood,

$$\mathbb{E}[Y_{it} | D_{it}, X_{it}, \alpha_i, \lambda_t] = \exp(\alpha_i + \lambda_t + D_{it}\beta_D^{\text{mean}} + X'_{it}\beta),$$

and, under Assumption 66, interpret the fitted curve and its contrasts as an approximation to  $\mu(d)$  on the outcome scale.

Second, estimate an **extensive-margin effect** using a linear probability model or conditional logit for  $\mathbf{1}\{Y_{it} > 0\}$ , with unit fixed effects and cluster-robust standard errors. In an LPM specification,

$$\mathbf{1}\{Y_{it} > 0\} = \alpha_i + D_{it}\beta_D^{\text{ext}} + X'_{it}\beta_1 + \varepsilon_{it},$$

$\beta_D^{\text{ext}}$  is directly interpretable as a change in purchase probability.

Third, estimate a **conditional basket-size regression** on  $Y_{it} > 0$  using PPML,

$$\mathbb{E}[Y_{it} | Y_{it} > 0, D_{it}, X_{it}, \alpha_i, \lambda_t] = \exp(\alpha_i + \lambda_t + D_{it}\beta_D^{\text{cond}} + X'_{it}\beta_2),$$

which answers “how does average basket size among buyers change?” but should be reported as descriptive unless there is strong reason to believe treatment does not affect participation. In practice, treating the conditional regression as causal requires either design-based justification (for example, an experiment in which  $D_{it}$  is random among verified active customers) or strong assumptions that treatment shifts spending but not the participation decision.

Formally correct estimation of the intensity equation works with the zero-truncated likelihood. PPML on the restricted sample  $\{(i, t) : Y_{it} > 0\}$  is a descriptive robustness tool. Treat it as causal only under a design that fixes participation or under assumptions that rule out treatment effects on participation.

**Software.** In R, the `fixest` package offers `feols()` for the LPM and `fepois()` for PPML with high-dimensional fixed effects and clustered standard errors. In Stata, `reghdfe` handles the LPM, and `ppmlhdfe` implements PPML. Other languages and packages can be used as long as they support fixed effects and cluster-robust inference.

## Zero-Inflated Models

Zero-inflated models take the structural-vs-sampling-zero distinction from Definition 14.13 seriously by modelling a mixture of an “always-zero” component and a standard count distribution that can also produce zeros.

**Definition 14.15 (Zero-Inflated Poisson (ZIP))** A zero-inflated Poisson (ZIP) model specifies

$$P(Y_{it} = 0) = \omega_{it} + (1 - \omega_{it})e^{-\mu_{it}}, \quad P(Y_{it} = y) = (1 - \omega_{it}) \frac{e^{-\mu_{it}} \mu_{it}^y}{y!}, \quad y \geq 1,$$

where  $\omega_{it} = \Lambda(\alpha_i^{(1)} + \lambda_t + D_{it}\gamma_1 + X'_{it}\delta_1)$  is the probability of belonging to the always-zero group and  $\mu_{it} = \exp(\alpha_i^{(2)} + \lambda_t + D_{it}\gamma_2 + X'_{it}\delta_2)$  is the Poisson mean for at-risk units.

In practice, the separation between the inflation probability  $\omega_{it}$  and the Poisson mean  $\mu_{it}$  is often weakly identified when both equations use similar covariates and fixed effects. Estimates of  $\gamma_1$  and  $\gamma_2$  can therefore be sensitive to functional-form choices and starting values.

**Hurdle vs ZIP.** Hurdle models treat all zeros as arising from the participation process and then model positive counts separately. ZIP models explicitly partition the population into always-zero units and at-risk units. Hurdles are natural when zeros mainly reflect deliberate non-purchase. ZIP models are natural when some units are truly incapable of positive outcomes (for example, inactive customers or stockouts).

In panels with unit fixed effects in both the inflation and count equations, ZIP models face the same incidental-parameters problem as other nonlinear FE models: as  $N \rightarrow \infty$  with  $T$  fixed, the many  $\alpha_i^{(1)}$  and  $\alpha_i^{(2)}$  induce bias in  $\gamma$  and  $\delta$ . Even aside from incidental parameters, finite-sample performance can be poor when the mixture components are weakly separated. Bias-corrected estimators exist, but simulation evidence suggests they require moderately long panels before coverage stabilises. For short panels, pooled ZIP with cluster-robust standard errors or correlated random-effects specifications may be preferable, albeit at the cost of stronger distributional assumptions. Design-based identification should not rest on zero-inflation structure alone.

## Upper-Bound Censoring

Upper-bound censoring arises when observed outcomes are capped at a maximum. A loyalty programme may cap points at 100. A survey may top-code income at £150,000. A promotion may limit purchases to ten units. Let  $Y_{it}^*$  denote the latent uncensored outcome and  $c$  the censoring threshold, and define

$$Y_{it} = \min(Y_{it}^*, c).$$

In potential-outcomes terms, we model latent outcomes  $Y_{it}^*(d)$  and assume that, conditional on  $(X_{it}, \alpha_i, \lambda_t)$ , the censoring rule  $Y_{it} = \min(Y_{it}^*(d), c)$  does not depend on  $d$  beyond its effect on  $Y_{it}^*(d)$ .

A censored-count likelihood contribution takes the form

$$\mathcal{L}_{it} = \begin{cases} f(Y_{it} | \mu_{it}) & \text{if } Y_{it} < c, \\ 1 - F(c - 1 | \mu_{it}) & \text{if } Y_{it} = c, \end{cases}$$

where  $f$  and  $F$  are the density and CDF for the latent count model.

Fixed-effects Tobit and censored-count models suffer from the same incidental-parameters problem as other nonlinear FE models when  $T$  is small. In practice:

If censoring affects only a small share of observations, a useful robustness check is to treat censored values as uncensored and see whether estimates move materially. This does not recover the latent uncensored effect but provides a check on whether your main conclusions are driven by a small censored tail or by the bulk of the distribution. When censoring is more prevalent, correlated random-effects (for example, Mundlak–Chamberlain devices) or pooled estimation with cluster-robust standard errors are safer than FE Tobit.

## The Three-Stage Approach

When data exhibit both excess zeros and a non-trivial mass at the ceiling—as in a panel with 90 per cent zeros, one per cent at the cap, and nine per cent in between—a three-stage decomposition may be appropriate.

**Definition 14.16 (Three-Stage Hurdle with Censoring)** Decompose the outcome into three components:

1. **Stage 1:** the probability of any purchase,  $P(Y_{it} = 0)$  versus  $P(Y_{it} > 0)$ .
2. **Stage 2:** among purchasers, the probability of hitting the ceiling,  $P(Y_{it} = c | Y_{it} > 0)$ .
3. **Stage 3:** among interior outcomes, the distribution of  $Y_{it}$  conditional on  $0 < Y_{it} < c$ , modelled with a doubly truncated count model.

The full likelihood factors accordingly, with separate slope parameters  $\beta_D^{\text{ext}}, \beta_D^{\text{cap}}, \beta_D^{\text{int}}$  for extensive, saturation, and intensive margins.

This decomposition can illuminate whether an intervention mainly brings customers into the store, pushes them to the cap, or increases interior spending. It is, however, data-hungry. Under panel unconfoundedness,

consistent estimation of the stage-specific slopes requires sufficient within-unit variation in both treatment intensity and the stage-specific indicators (any purchase, hitting the cap, interior outcomes). In many marketing panels, the ceiling event is so rare that the Stage-2 regression is effectively underpowered.

**Practical guidance.** For most marketing applications, a two-stage hurdle that distinguishes only between “any purchase” and “how much among purchasers” suffices, treating censoring as part of the intensive margin or as a robustness issue. Reserve explicit three-stage models for settings where the cap is economically central and where you have enough variation to estimate all three margins reliably. A simple rule of thumb is that if only a handful of units ever hit the cap, Stage-2 parameters are not empirically identified in a useful sense.

## Marketing Interpretation

The extensive–intensive decomposition answers a core business question: does an intervention create new customers or increase spending among existing ones?

In the randomised-loyalty-programme example, suppose a three-stage model yields the following approximate results.

The programme increases the probability of any purchase from about 12 per cent to 15 per cent (an extensive-margin gain of roughly three percentage points). Conditional on purchase, it raises interior basket sizes by around 10–15 per cent, with little change in the probability of hitting the cap. The programme thus works primarily by bringing customers into the store and modestly increasing their spending. A natural operational implication, conditional on a credible design for  $D_{it}$ , is that marketing spend should focus on reach—bringing new customers into the store—rather than deepening already-large baskets. Detailed three-stage decompositions can inform such decisions, but they should always be interpreted alongside the simpler unconditional-mean estimates and the identification checks from earlier chapters.

**Table 14.2** Model selection for zero-inflated and censored outcomes. All recommendations presuppose that identification for the causal effect of  $D_{it}$  has been secured through design or unconfoundedness assumptions. Choosing a richer outcome model cannot compensate for violations of those assumptions.

Data pattern	Recommended model	Key assumptions and caveats
Moderate zeros (30–60%)	Fixed-effects Poisson or negative binomial	Overdispersion handled by robust SE. Zeros mainly sampling zeros.
Excess zeros (>70%)	Two-stage hurdle (participation + intensity)	Zeros reflect participation decisions. Distinct extensive and intensive effects.
Excess zeros with structural interpretation	Zero-inflated Poisson/NegBin	Some units are effectively always zero. Incidental-parameter bias in short $T$ . Weak identification risk.
Upper-bound censoring (>5% at ceiling)	Censored count model or three-stage approach	Tobit-style likelihood. FE inconsistent in short $T$ . Consider CRE or pooled models.
Zeros and ceiling both prevalent	Three-stage hurdle	Data-hungry. Reserve for economically meaningful caps and well-powered panels.

## 14.7 Duration Models for Takeoff in Panels

Marketing innovations often have a distinct takeoff point at which adoption or sales growth accelerates. Duration models offer a natural framework for analysing this time-to-takeoff, modelling the hazard of transitioning from a pre-takeoff state to a post-takeoff state. In this section we summarise basic hazard models and explain how they complement the panel causal designs developed elsewhere in the book.

**Definition 14.17 (Hazard and Survival)** Let  $T_i > 0$  denote a duration of interest for unit  $i$ —for example, the time from product launch to takeoff in market  $i$ . To study outcomes in event time relative to takeoff, define an event-time index  $k = t - T_i$ . To avoid confusion with treatment adoption times  $G_i$  used elsewhere in the book, we do not reuse  $G_i$  for takeoff times. Event-time plots around outcome-defined takeoff events are descriptive by default, because defining  $T_i$  using outcomes can induce mechanical pre-patterns. Causal interpretation requires an external design that shifts takeoff timing.

The survival function is

$$S(t) = \mathbb{P}(T_i > t),$$

the probability that unit  $i$  has not yet experienced takeoff by time  $t$ . The hazard function is the instantaneous rate of transition at time  $t$ :

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t \mid T_i \geq t)}{\Delta t} = \frac{f(t)}{S(t)},$$

where  $f(t) = -S'(t)$  is the density. The cumulative hazard  $H(t) = \int_0^t h(s) ds = -\log S(t)$  summarises the accumulated risk up to time  $t$ .

In panel applications, each unit  $i$  (for example, a country, market, or product) is observed over discrete time periods, and  $T_i$  is constructed from the panel time index by identifying the first period at which a takeoff criterion is met. For weekly data,  $T_i$  might be the number of weeks from launch until the first week that meets a pre-specified takeoff rule. In discrete time, with periods indexed by  $t = 1, \dots, T$ , we work with the discrete hazard

$$h_t = \mathbb{P}(T_i = t \mid T_i \geq t),$$

and  $S(t) = \mathbb{P}(T_i > t)$  is understood as the survivor function over these discrete periods.

**Definition 14.18 (Cox Proportional Hazards)** The Cox proportional hazards model specifies the hazard of takeoff as

$$h(t \mid X_i) = h_0(t) \exp(X_i' \beta),$$

where  $h_0(t)$  is an unspecified baseline hazard and  $\exp(X_i' \beta)$  is a hazard ratio capturing how observed characteristics  $X_i$  shift the risk of takeoff. With time-varying covariates derived from the panel, the model becomes

$$h(t \mid X_i(t)) = h_0(t) \exp(X_i(t)' \beta),$$

where  $X_i(t)$  may include evolving market conditions or marketing actions.

The Cox partial likelihood eliminates  $h_0(t)$  and estimates  $\beta$  using only the relative risk structure. If  $t_i$  are observed event times and  $\mathcal{R}(t)$  is the risk set at time  $t$ , the partial likelihood is

$$\mathcal{L}(\beta) = \prod_{i: T_i=t_i} \frac{\exp(X_i(t_i)' \beta)}{\sum_{j \in \mathcal{R}(t_i)} \exp(X_j(t_i)' \beta)}.$$

For a causal interpretation of  $\beta$  when  $X_i(t)$  includes treatments derived from the panel (for example, marketing spend  $D_{it}$ ), a causal interpretation requires defining duration potential outcomes under treatment paths, for example  $T_i(d_i^T)$ , and then stating sequential or path-based unconfoundedness and non-informative censoring assumptions. These conditions parallel the unconfoundedness and identification discussion in Sections 2.4 and 14.3.

To make the causal target explicit, you can define dose-specific survival and hazard objects such as  $S_d(t) = \mathbb{P}(T_i(d_i^T) > t)$  for a policy that sets  $d$  over a stated horizon. Interpreting hazard ratios causally requires overlap for the relevant treatment paths and a censoring mechanism that does not depend on unobserved determinants of  $T_i(d_i^T)$ , conditional on the stated controls.

When marketing actions respond to evolving adoption risk (for example, heavier spend in markets that appear slow to take off), the standard Cox assumptions are violated and hazard coefficients mix causal effects with this feedback. Addressing such dynamics typically requires structural models or marginal structural methods beyond the scope of this chapter.

**Definition 14.19 (Shared Frailty)** Shared-frailty models introduce unit-level random effects to capture unobserved heterogeneity in baseline risk. A simple specification is

$$h(t | X_i, \nu_i) = h_0(t) \exp(X_i' \beta + \nu_i),$$

where  $\nu_i$  is a unit-specific frailty term, often modelled as  $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$  or with  $\exp(\nu_i)$  gamma-distributed. Frailty plays the same role as a random intercept in panel regression, absorbing persistent unobserved factors that shift the baseline hazard across units.

From a causal perspective, shared frailty is a modelling assumption. Consistent interpretation of  $\beta$  as a within-unit causal effect generally requires strong restrictions on the dependence between  $\nu_i$  and  $X_i(t)$ , or a correlated-frailty (CRE-style) specification that allows dependence through unit-level means. If this is violated,  $\beta$  conflates causal effects with unobserved heterogeneity. Correlated-frailty analogues that include unit-level means of time-varying covariates in  $X_i$  can partially relax this, mirroring the correlated random-effects approaches in Section 2.5.

Frailty does not, by itself, address time-varying unobservables or dynamic feedback between marketing actions and adoption.

## Integration with Panel Causal Designs

Classic diffusion studies use hazard models to document systematic patterns in takeoff timing across products and countries, emphasising the roles of economic conditions, culture, and innovativeness [Golder and Tellis, 1997, Tellis et al., 2003, Hauser et al., 2007]. These studies primarily use hazard models descriptively to characterise takeoff timing. Causal identification typically requires an external shock to marketing actions or launch timing. In this book, hazard models play a complementary role to the panel causal designs developed in earlier chapters.

Hazard models answer the “when” question: how long does it take for a market or product to reach a takeoff point, and which observed characteristics predict shorter or longer durations? Under strong unconfoundedness—meaning no unmeasured determinants of  $T_i(d_i^T)$  beyond  $X_i(t)$  and any allowed unit heterogeneity—and non-informative censoring—meaning no dependence of censoring on unobserved takeoff propensity, conditional on the stated controls—coefficients in Cox or frailty models can be interpreted as causal effects of covariates on the timing of takeoff. More often in marketing, they are best treated as descriptive summaries of how takeoff risk co-moves with observed market and strategy variables.

Panel causal designs answer the “what happens after” question. Once takeoff events are identified—possibly using hazard-based definitions of  $T_i$ —you can analyse outcomes in event time  $k = t - T_i$  and report dynamic summaries indexed by  $k$ . When the takeoff event is outcome-defined, these event-time profiles are descriptive by default. To interpret post-takeoff changes causally, you need a design that shifts takeoff timing or a separate treatment assignment mechanism that is plausibly exogenous.

When the design supports it, event-time analyses can still use the event-study tools from Chapter 5 to trace pre-event diagnostics and post-event dynamics on outcomes such as sales, margins, or market share. Synthetic control and factor methods from Chapters 6 and 8 can provide counterfactual paths when a credible comparison group exists.

Taken together, duration models and panel designs give a fuller picture of marketing innovations. Hazard models help explain and forecast when takeoff occurs across units. Panel causal methods, grounded in explicit designs and assumptions, quantify what changes after takeoff and by how much. Analysts should be clear about which role a duration model is playing in a given study—descriptive forecasting or causal inference on timing—and should make the corresponding assumptions explicit.

Inference for duration models is typically based on independent units  $i$ . When outcomes are clustered (for example by market group), treat clusters  $c = 1, \dots, G$  as the independent sampling units and use cluster-robust variance estimators or a valid cluster bootstrap.

In most observational marketing panels, the conditions for causal hazard interpretation are hard to justify. Treat hazard models for takeoff as descriptive tools for forecasting and pattern discovery unless the design (for example, staggered randomisation of launch timing) clearly supports a causal interpretation.

## 14.8 Dynamics with Continuous Intensity

Marketing responses to continuous treatments such as prices and advertising expenditure are rarely instantaneous. A change in price today can influence sales over several periods. An advertising burst can have delayed and decaying effects. Xiao and Wu [2025] define a path-specific marginal effect and then average it over the empirical distribution of treatment paths. We use the same interpretation here. As in the rest of the book, modelling dynamics does not create identification. Causal interpretation requires restrictions on how dose paths are assigned, especially when marketing intensity responds to past outcomes or shocks.

**Definition 14.20 (Average Causal Response)** Let  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$  denote the history of a continuous treatment for unit  $i$  up to time  $t$ , as in Section 2.1. Define the path-specific marginal response at time  $t$  to a perturbation in the current dose as

$$\frac{\partial \mathbb{E}[Y_{it}(\underline{d}_i^t)]}{\partial d_{it}},$$

where  $Y_{it}(\underline{d}_i^t)$  is the potential outcome under the treatment path  $\underline{d}_i^t$ . The Average Causal Response at time  $t$  averages this derivative over the distribution of treatment histories in the target population. Equivalently, fixing the past path  $\underline{d}_i^{t-1}$  and varying only the current dose  $d_{it}$ , the path-specific response is a marginal effect of the ADRF

$$\mu_t(d_{it}, \underline{d}_i^{t-1}) = \mathbb{E}[Y_{it}(\underline{d}_i^{t-1}, d_{it})]$$

with respect to  $d_{it}$ . The Xiao framework averages these path-specific derivatives over the empirical distribution of paths. Lagged ACRs—responses to perturbations in  $d_{i,t-s}$  for  $s > 0$ —generalise this idea to impulse-response profiles.

**Assumption 67 (Linearity and No Differential Confounding for ACR)** The following conditions hold:

- (i) A linear distributed-lag causal model governs potential outcomes:  $Y_{it}(\underline{d}_i^t)$  responds linearly to past doses over the relevant horizon.
- (ii) There is no differential confounding in dose innovations. Conditional on an admissible information set that includes observed covariates, unit effects, and time effects, innovations in  $D_{it}$  are mean independent of innovations in counterfactual outcomes. This rules out feedback in which unobserved demand shocks shift both outcomes and future marketing intensity.
- (iii) Support for identifying variation is adequate over the reporting region. In practice, this requires that the distribution of dose innovations is not degenerate in the periods and units used for identification.

**Theorem 14.5 (Identification via Generalised TWFE)** *Under Assumption 67, the coefficient  $\beta_s$  in a distributed-lag TWFE regression can be interpreted as a weighted average of lag- $s$  marginal causal responses over unit-time cells with identifying variation in  $D_{i,t-s}$  (Xiao and Wu [2025]).*

*In this setting, Xiao and Wu [2025] show that appropriately weighted TWFE estimators can be written as convex combinations of unit-level marginal effects  $\partial Y_{it}(\underline{d}_i^t)/\partial d_{i,t-s}$ . Specifically, the coefficient  $\beta_s$  in a correctly specified distributed-lag TWFE regression equals a weighted average of unit-time marginal effects*

$\partial Y_{it}(\underline{d}_i^t)/\partial d_{i,t-s}$  over units and periods with variation in  $D_{i,t-s}$ . Under the sufficient conditions in Xiao and Wu [2025], weights are non-negative and sum to one. When those conditions fail, weights can be negative and  $\beta_s$  may be a distorted mixture of different marginal effects, much as TWFE can deliver non-convex weights in the binary staggered-adoption case.

A practical implementation uses linear distributed-lag models with continuous regressors and unit and time effects,

$$Y_{it} = \sum_{s=0}^L \beta_s D_{i,t-s} + \alpha_i + \lambda_t + \varepsilon_{it},$$

where  $D_{it}$  is the realised treatment intensity in period  $t$ ,  $\alpha_i$  and  $\lambda_t$  are unit and time effects, and  $\beta_s$  measure the association between outcomes and lagged doses. This regression is not automatically causal. If marketing intensity responds to lagged outcomes or shocks,  $\beta_s$  is a forecasting coefficient unless identification is secured by design (randomised intensity shocks, valid instruments, or credible exogeneity restrictions). Under Assumption 67, each  $\beta_s$  can be interpreted as an average causal response at lag  $s$ , analogous to an impulse-response function.

Inference should respect dependence in the panel. If units are independent, cluster standard errors by unit. If dependence is clustered, treat clusters  $c = 1, \dots, G$  as the independent sampling units and use cluster-robust variance estimators or a valid cluster bootstrap.

## Bridging Static and Dynamic Estimands

The ACR framework links the static dose-response objects from Section 14.2 to the dynamic binary estimands of Chapter 10. Static ADRFs  $\mu(d)$  and marginal effects  $\tau(d)$  describe how long-run or horizon-specific outcomes vary with treatment intensity, without explicit time dynamics. Dynamic binary panels describe how outcomes evolve after discrete treatment adoption through event-time profiles and impulse responses.

For continuous treatments, ACRs play the same role as dynamic ATTs and impulse responses: they trace how outcomes respond over time to changes in dose. The long-run effect of a permanent one-unit increase in dose is

$$\sum_{s=0}^L \beta_s.$$

Following Section 10.2, the corresponding long-run multiplier is

$$\text{LRM} = \frac{\sum_{s=0}^L \beta_s}{\beta_0}.$$

Long-run ROI from continuous ad-spend models corresponds to the long-run effect scaled by cost. Under Assumption 67, these cumulative responses can be interpreted as average causal effects of a sustained change in intensity. In practice, the same caveats apply as in the binary case: credible design or strong unconfoundedness assumptions are needed, and diagnostics such as pre-trend checks and sensitivity analyses remain essential.

In summary, dynamic dose–response models extend the panel toolkit to settings where treatment intensity varies continuously and effects propagate over time. They should be used when the design and identification assumptions support interpreting distributed-lag coefficients as average causal responses, not merely as forecasting coefficients in high-dimensional regressions. Absent credible identifying variation in  $D_{it}$ —for example, from experiments, instruments, or plausibly exogenous policy shifts—distributed-lag coefficients should be treated as forecasting coefficients, not ACRs.

## 14.9 Assumptions

The core assumptions for continuous and nonlinear panel methods mirror those in earlier chapters. They remain technical, but each should be motivated by the substantive design of your study and supported by diagnostics and institutional knowledge.

**Assumption 68 (SUTVA and Exposure Mapping for Continuous Doses)** Potential outcomes may depend on own dose and on others' doses only through the specified exposure mapping. There is no interference beyond  $h_i(\cdot)$ .

Formally, let  $D_{it}$  denote the own continuous dose for unit  $i$  at time  $t$ , and let  $h_i(D_{-i,t})$  denote an exposure mapping that summarises others' treatments for unit  $i$  at time  $t$ . Spillover-aware potential outcomes can be written as  $Y_{it}(d, h)$  or  $Y_{it}(d_{it}, h_i(D_{-i,t}))$ . Baseline analyses assume that in the designs considered either (i)  $h_i(D_{-i,t})$  is held fixed at its observed path (no interference beyond the exposure mapping) or (ii) spillovers are explicitly modelled using the framework of Chapter 11. When spillovers are plausible, exposure mappings and associated assumptions must be specified alongside the dose-response estimands.

**Assumption 69 (Unconfoundedness for Continuous Treatments)** For all  $d \in \mathcal{D}$ , continuous-treatment unconfoundedness holds at the unit-time level:

$$Y_{it}(d) \perp\!\!\!\perp D_{it} \mid X_{it}, \alpha_i, \lambda_t.$$

Here  $X_{it}$  denotes observed covariates,  $\alpha_i$  unit effects (or CRE proxies), and  $\lambda_t$  time effects or flexible time trends.

When outcomes depend on past doses,  $Y_{it}(d)$  is shorthand for a restriction of the dynamic primitive  $Y_{it}(d_i^t)$ . In that case, a static unconfoundedness statement is not enough by itself. Identification typically requires either a no-carryover restriction that makes  $Y_{it}(d)$  a well-defined period- $t$  object, or a sequential ignorability condition that rules out unmeasured time-varying confounding and feedback from outcomes to future doses.

The generalised propensity score (GPS)

$$r(d \mid X_{it}, \alpha_i, \lambda_t) = f_{D|X,\alpha,\lambda}(d \mid X_{it}, \alpha_i, \lambda_t)$$

the conditional density of  $D_{it}$  evaluated at dose  $d$ , exists and is well defined on the reporting region  $\mathcal{D}_0 \subseteq \mathcal{D}$ . Unconfoundedness is the continuous analogue of the selection-on-observables assumption in Chapter 2. It rules out unobserved time-varying confounders correlated with changes in intensity after conditioning on  $(X_{it}, \alpha_i, \lambda_t)$ .

**Assumption 70 (Overlap and Support in Dose)** There exists an overlap region  $\mathcal{D}_0$  defined by trimming such that estimated conditional densities are not vanishingly small on  $\mathcal{D}_0$  for the target population. In words, for each covariate profile and subgroup in which we report dose-response estimates, there is sufficient variation in  $D_{it}$  to support comparisons across doses. In practice,  $\mathcal{D}_0$  is an overlap region obtained by trimming extreme doses or covariate profiles with GPS values below a pre-specified threshold. All reported effects are understood

to be local to  $\mathcal{D}_0$ . When tails of the dose distribution are sparse, estimation is accompanied by trimming to this overlap region and by transparent reporting of the resulting support, as in Chapter 17.

Trimming is not an innocuous implementation detail. It changes the estimand from  $\mu(d)$  over all  $d \in \mathcal{D}$  to an overlap-restricted object over  $d \in \mathcal{D}_0$ . Outside  $\mathcal{D}_0$ , any reported curve relies on extrapolation.

**Assumption 71 (Parallel Trends with Intensity)** When using an intensity-based parallel-trends restriction, we require that innovations in dose  $\Delta D_{it}$  are mean independent of innovations in potential outcomes after conditioning on  $(X_{it}, \alpha_i, \lambda_t)$ .

More concretely, consider two units with different intensity paths. In the absence of differential changes in  $D_{it}$ , their counterfactual outcomes would have followed parallel trends conditional on  $(X_{it}, \alpha_i, \lambda_t)$ . Equivalently, for any fixed current dose  $d \in \mathcal{D}_0$ , we require

$$\mathbb{E}[\Delta Y_{it}(\underline{d}_i^{t-1}, d) | X_{it}, \alpha_i, \lambda_t, \Delta D_{it}] = \mathbb{E}[\Delta Y_{it}(\underline{d}_i^{t-1}, d) | X_{it}, \alpha_i, \lambda_t]$$

so that changes in intensity do not predict changes in counterfactual outcomes after controlling for  $(X_{it}, \alpha_i, \lambda_t)$ . When you build event-time analyses around discrete intensity shocks (for example, step-changes in budgets), define the shock time explicitly and avoid aggregations that mix heterogeneous shocks across units, following the cohort structure of Chapter 5 and the aggregation cautions in modern panel frameworks [Arkhangelsky and Imbens, 2024].

**Assumption 72 (Factor or IV Alternatives)** When identification uses factor models or instrumental variables rather than pure selection-on-observables, the corresponding structural conditions hold.

In factor-based designs (Chapter 8), untreated potential outcomes admit a low-rank representation driven by a small number of common shocks with heterogeneous unit loadings, and the factor-coverage and rank conditions needed for identification are satisfied. In IV-based designs for continuous doses (Chapter 16 and Section 16.10), instruments  $Z_{it}$  shift  $D_{it}$  (relevance), affect outcomes only through  $D_{it}$  (exclusion), and are independent of potential outcomes conditional on covariates and fixed effects (exogeneity). Rank and relevance conditions are checked empirically, and exclusion restrictions are justified by research design and domain knowledge.

For continuous doses, a single scalar instrument  $Z_{it}$  generally identifies a local effect or a low-dimensional functional of the ADRF rather than the entire curve  $\mu(d)$ . Recovering  $\mu(d)$  over a range of  $d$  requires either a continuum of instruments or strong functional-form restrictions. Linear-dose IV models identify average causal derivatives when the structural function is linear in  $D_{it}$ , connecting back to the linear setups in Chapter 16.

These assumptions are not new. They specialise the global framework of Chapters 2, 4, 8, and 16 to continuous and nonlinear outcomes. The estimation and DML methods in this chapter change how we fit models and compute dose-response functions, but they do not relax the substantive requirements for identification. Those requirements must still be defended with institutional detail, diagnostics, and sensitivity analyses. A sophisticated DML or machine-learning implementation of a misspecified identification strategy is still wrong. The burden of justification lies with Assumptions 68–72, not with the choice of learner.

When you report uncertainty, state the independent sampling unit. In many panels it is the unit  $i$ . In geo and market panels it is often the cluster  $c = 1, \dots, G$ , with  $G$  the effective sample size. Use variance estimators and resampling procedures that match that dependence structure.

## 14.10 Tuning and Implementation

Implementing dose–response estimators in panels requires a series of design choices that sit at the intersection of identification and prediction. These include how to represent the dose dimension, how to stabilise weights and trim to overlap regions, and how to structure cross-validation without leaking post-treatment information. This section summarises practical guidance, building on the more general tuning discussions in Chapters 12 and 17.

### Basis and Model Specification

Basis functions for the dose dimension—price, discount depth, advertising spend—should be flexible enough to capture realistic curvature but not so rich that they overfit sparse tails. Low-order splines or series expansions in the dose, with knots placed at quantiles of the observed distribution, are natural defaults. Degrees of freedom can be selected by cross-validation confined to training folds, as in Section 12.8, rather than by fitting ever-more-flexible bases on the full sample. Place knots and choose basis complexity within the overlap region  $\mathcal{D}_0$ . Outside  $\mathcal{D}_0$ , do not interpret fitted curvature as causal.

Interactions between dose and key moderators (store format, device type, baseline demand) are valuable when theory or prior evidence suggests clear effect modification, but they should be introduced parsimoniously. In practice this often means interacting dose with a small set of hand-picked moderators rather than with the full high-dimensional  $X_{it}$ . Generalised propensity-score models can be parametric (for example, normal or log-normal regressions for  $D_{it}$  given  $X_{it}$ ). In practice, a tractable default is a parametric conditional density model for  $D_{it} | X_{it}$  (with transformations for skewed doses). If you use ML methods, state explicitly how you obtain a conditional density (not just a mean), and diagnose calibration by comparing predicted vs empirical dose distributions within folds. In either case, the covariate set should respect identification constraints by excluding post-treatment variables and leakage-prone features. For example, including contemporaneous click-through rates or realised conversions in a GPS model for display impressions would leak post-treatment information, since these outcomes are themselves affected by past exposure.

When both outcome and GPS models include basis functions of  $D_{it}$ , choose compatible degrees of smoothness so that the implied ADRF and the conditional density  $r(d | X_{it})$  describe the same level of curvature in the dose dimension. Extremely rough GPS models paired with overly smooth outcome models (or vice versa) tend to produce unstable DR/DML scores. A practical diagnostic is to plot the DR score contributions  $\psi_{it}(d)$ : if a small number of observations dominate because of extreme  $1/\hat{r}(\cdot)$ , the estimated curve is effectively extrapolation.

At minimum, complement prediction-based tuning with stability diagnostics for the causal target. Check sensitivity of  $\hat{\mu}(d)$  to plausible ranges of basis complexity, trimming thresholds, and bandwidths. If conclusions change materially, treat the design as fragile.

## Weighting and Trimming

Stabilised weights based on the GPS—ratios of conditional to marginal densities—can improve precision by shrinking very large raw inverse-probability weights towards one. However, extreme weights still arise in regions of poor overlap. Truncation of weights at pre-specified percentiles (for example, capping at the 99th percentile) or trimming observations with the most extreme propensities reduces variance but changes the target estimand to the overlap region. Analysts should make this explicit by describing the overlap population (for example, in terms of ranges of  $D_{it}$  and key covariates) and by avoiding language that suggests global generality. Describe the resulting overlap population in terms of dose ranges and key covariates, and report the fraction trimmed.

Overlap diagnostics from Section 14.4 and Chapter 17 should drive these decisions. Before trimming or truncating, examine the distribution of estimated GPS values by treatment intensity and covariate subgroup. Graphical displays of the effective dose support before and after trimming should be standard. When tails are very sparse, report estimates on coarser dose grids or combine extreme regions into bins, and interpret any remaining estimates in those regions as exploratory. Always document trimming rules, the fraction of observations removed, and the resulting support for dose-response curves.

## Cross-Validation Design

Cross-validation for nuisance learners in continuous-dose DML must respect the same dependence and leakage constraints as in the binary-treatment case. Use folds defined on the independent sampling unit. When units are independent, a conservative default is unit-level folds, so all periods for a unit belong to one fold. When outcomes are clustered (for example by market), define folds at the cluster level and include all observations from a cluster in the same fold.

Add time blocking only when it is required to avoid serial leakage in forecasting-style nuisance learners. Units assigned to a fold should not contribute observations to the training sets used to predict their outcomes. Time-based folds should avoid mixing future information into nuisance estimates, especially when features include lagged outcomes or rolling aggregates. Treat any feature constructed from realised outcomes as post-treatment unless you can justify it as pre-treatment relative to the causal contrast being estimated.

In staggered-adoption panels, optionally impose time gaps between training and validation windows to reduce serial leakage.

Hyperparameter tuning for outcome-regression and GPS models should take place entirely within the training folds defined by the cross-fitting scheme. If you use an inner cross-validation loop, use the same split unit as in the outer loop, unit or cluster, and add time blocks only when needed to prevent leakage. Validation scores that ignore the panel structure—for example, random shuffling of unit-time observations—will tend to be optimistically biased in the presence of serial correlation and staggered adoption.

In short, the same principles that govern tuning in binary DML apply here. Basis choice and learner complexity are chosen to balance flexibility and stability. Weighting and trimming are anchored in overlap

diagnostics and reported transparently. Cross-validation is nested within a dependence-respecting cross-fitting design that keeps post-treatment information out of nuisance training. No amount of clever tuning can compensate for misspecified identification assumptions or poor overlap.

## 14.11 Diagnostics and Design Considerations

Diagnostics for continuous treatments mirror those in Chapter 17, but must be adapted to the dose–response setting. The goal is to assess whether the overlap and unconfoundedness conditions from Section 14.9 are even remotely plausible, and to quantify how sensitive estimated dose–response functions are to modelling and design choices. These diagnostics are heuristics. They can reveal fragility and obvious failures, but they do not certify identification.

### Overlap and Balance

Begin by examining overlap in the dose dimension. Plot the empirical distribution of realised doses  $D_{it}$  overall and by key subgroups (for example, cohorts, markets, or treatment regimes). Then examine overlap in the generalised propensity score (GPS)  $r(d | X_{it}, \alpha_i, \lambda_t)$ . Plot estimated GPS values by dose strata, and report the proportion of observations whose GPS exceeds a minimum threshold in the regions of the dose space where you intend to estimate effects. Define an overlap region  $\mathcal{D}_0 \subseteq \mathcal{D}$  where estimated GPS values exceed a minimum threshold and covariate balance is acceptable. All reported ADRFs and marginal effects are local to  $\mathcal{D}_0$ . State the trimming rule explicitly (for example, drop observations with estimated GPS below the 1st percentile within dose bins) and report the resulting support.

Balance checks extend the diagnostics to continuous doses. After any adjustment (weighting, residualisation, or matching) built on the GPS, within narrow dose bins (or within GPS strata evaluated at the bin midpoints), compare covariate distributions across observations in neighbouring bins. Report standardised mean differences for key covariates as a function of dose, following Chapter 17. Good balance on observed covariates is necessary but not sufficient for unconfoundedness. It shows that, conditional on the GPS or dose band, the design approximates a randomised experiment on observables, but unobserved confounding remains a substantive concern. Persistent imbalance in important predictors—especially in high-dose or low-dose regions—signals weak overlap or misspecified nuisance models and should trigger design revisions or more conservative trimming.

### Sensitivity and Robustness

Dose–response estimates should be stress-tested against both modelling and design choices. On the modelling side, vary basis functions for the dose (for example, different spline orders or knot placements), bandwidths in kernel-based estimators, and trimming or weight-truncation thresholds. Plot ADRF or marginal-effect estimates across these specifications, highlighting where substantive conclusions are stable and where they depend on particular tuning choices.

On the design side, run placebo checks tailored to continuous treatments. As a falsification check, estimate a dose–response relationship where no effect is plausible (for example, replace  $Y_{it}$  with a strictly pre-treatment

outcome or define a pseudo-dose from a pre-period variable). Large or systematic effects in these placebos suggest residual confounding, leakage, or misspecified nuisances. Restrict analysis to regions with stronger overlap—such as mid-range doses or cohorts with richer support—and check whether conclusions persist. Where possible, compare continuous-dose results with simpler designs that rely on coarsened treatments (for example, high vs low intensity) to see whether patterns align. Agreement with coarsened-treatment benchmarks is reassuring but not dispositive. Disagreement is a warning sign that the continuous-dose curve is driven by extrapolation or model artefacts.

Construct placebos to avoid leakage. If you define pseudo-doses or placebo outcomes using realised post-treatment variables, you can create mechanical “effects” even when the causal effect is zero. Treat placebo diagnostics as stress tests of the full pipeline, including feature construction and nuisance estimation.

Throughout, report diagnostics alongside main estimates: dose and GPS distributions, balance summaries, trimmed sample descriptions, and sensitivity plots. This allows readers to see not just point estimates but also the regions of the dose space, covariate distributions, and modelling choices that support those estimates, and to judge whether the continuous-treatment assumptions from Section 14.9 are credible in the marketing context at hand.

## 14.12 Inference

Inference for continuous-treatment models follows the principles laid out in Chapter 16, with two main adaptations for the dose–response setting. Influence functions now index dose and, often, lag, and we care about joint uncertainty over whole curves rather than just scalar parameters.

### Standard Errors and Clustering

For ADRF and ACR estimators, the basic building block is the orthogonal score defined in Section 14.4, whose Neyman-orthogonality with respect to nuisance functions  $m$  and  $r$  ensures that first-order bias from estimating these functions does not contaminate the influence functions  $\psi_{it}(d)$  (see Chapter 12). For a fixed dose  $d$ , let  $\psi_{it}(d)$  denote the score contribution at the unit–time level. The unit-level influence function is then

$$\Psi_i(d) = \sum_t \psi_{it}(d),$$

which aggregates score contributions across time for unit  $i$ . Inference must treat the independent sampling unit as the cluster. In many applications that is the unit  $i$ . In geo and market panels, it is often the cluster  $c = 1, \dots, G$ , with  $G$  the effective sample size. Pointwise standard errors at dose  $d$  use the sample variance of the influence contributions across the independent units. Cluster-robust estimators formed from these contributions yield asymptotically valid variance estimates when the number of independent clusters is large.

When the number of independent clusters is modest (for example, fewer than 30 markets or products), asymptotic normal approximations can be fragile. In such cases, apply wild cluster bootstrap or block-bootstrap procedures to the influence contributions, following Chapter 16. When using multiplier or wild bootstrap on influence functions, keep the cross-fitted nuisance estimates and fold assignments fixed and resample only the influence contributions. If you instead re-estimate nuisances in each bootstrap draw, state the procedure explicitly and check that it respects the cross-fitting design. The continuous-treatment setting changes the form of the influence contributions, but not the underlying clustering logic.

### Confidence Bands for Dose–Response Curves

For dose–response curves, uncertainty is inherently joint: we care whether the entire estimated function lies in a plausible region, not just whether a single point estimate is different from zero. Pointwise confidence intervals at each grid point  $d$  underestimate this joint uncertainty when read as a statement about the whole curve.

To address this, report confidence bands that control coverage over the entire dose grid or over economically relevant regions of the dose space. Uniform bands derived from multiplier or wild-bootstrap procedures—applied to the vector of influence functions  $\{\Psi_i(d_1), \dots, \Psi_i(d_K)\}$  over a grid  $d_1, \dots, d_K$ —ensure that, with

high probability (for example, 95 per cent), the true curve is contained between the lower and upper band at all grid points simultaneously. Uniform bands are defined over the reported grid  $\{d_1, \dots, d_K\}$ . Pre-specify this grid and a small set of primary contrasts to avoid implicit multiple testing. Treat dense grids and many subgroup curves as exploratory unless you control multiplicity with joint bands.

Pointwise intervals can still be useful for visualisation, but do not read them as joint statements about the whole curve.

In many marketing applications, you will estimate dose–response curves by subgroup (for example, device type or region) or over lags in ACR profiles. In those cases, the same joint-inference tools from Chapter 16 apply: treat the stacked vector of estimates as a single object, construct a cluster-robust covariance matrix from influence functions, and use either analytic joint bands or bootstrap-based critical values to control for multiple comparisons. This is a direct application of the joint-inference results in Chapter 16: replace scalar parameters by the vector of ADRF or ACR evaluations, and construct Wald-type or bootstrap joint bands using the cluster-robust covariance matrix of the corresponding influence functions. Reporting these bands alongside point estimates helps readers see not just where effects appear large or small, but also which features of the dose–response surface are statistically well supported by the data.

## 14.13 Marketing Applications

The tools in this chapter were motivated by familiar marketing problems where intensity and nonlinearity matter. Here we briefly sketch how the main estimands and models plug into those problems. We point forward to the detailed case studies in Chapter 18.

Dynamic advertising with continuous spend is the prototypical use case. Static binary designs answer whether an ad campaign matters on average. Continuous ADRFs  $\mu(d)$  and ACR profiles from Sections 14.2 and 14.8 answer how outcomes respond to different spend levels. They also answer how those responses unfold over time. In the ad-spend case studies in Chapter 18, we treat weekly or daily spend as  $D_{it}$ . We estimate  $\mu(d)$  over a grid of spend levels. We use distributed-lag models to recover short-run and long-run marginal effects. A cumulative response  $\sum_{s=0}^L \beta_s$  summarises the total outcome change from a sustained one-unit increase in dose over the horizon. Converting this into ROI requires a separate mapping from dose units (for example, £ of spend) to costs. It also requires a mapping from outcome units (for example, revenue) to profit. This is subject to the same identification and diagnostic checks as in the binary dynamic-treatment chapters.

Price changes illustrate how nonlinear outcome models and continuous doses interact. Rather than coding a price cut as a binary dummy, we treat price (or discount depth) as  $D_{it}$ . We estimate either log-linear or semi-parametric dose-response functions for sales or contribution margins. Fixed-effects logit and Poisson models from Section 14.5 provide binary and count analogues where the estimands are marginal effects. If you want elasticities, define the estimand explicitly. For example,  $\partial \log \mu(d) / \partial \log d$  or an arc elasticity over  $[d_0, d_1]$ . State the outcome scale (sales vs log sales vs margin). In Chapter 18 we revisit these models in the context of retail pricing and promotional depth. We emphasise how identification depends on high-frequency quasi-random variation or instruments rather than on the nonlinear link alone.

Zero-inflated purchase and engagement outcomes—common in loyalty, app-usage, and campaign-response data—call for the hurdle and zero-inflated models in Section 14.6. There, the key estimands are extensive and intensive-margin effects. These are changes in  $P(Y_{it} > 0)$  and in  $\mathbb{E}[Y_{it} | Y_{it} > 0]$  as functions of  $D_{it}$ . The three-stage decompositions for capped outcomes sharpen these margins further into participation, saturation, and interior intensity. Stage-specific causal interpretation requires additional structure. This can be principal-stratum estimands (units that would participate under both doses) or a design that fixes participation ex ante. Absent that structure, stage-specific regressions should be reported as descriptive decompositions. In Chapter 18, we use these decompositions to interpret loyalty-programme and capped-incentive designs. We make clear when each stage can be treated causally and when it is descriptive.

Across these applications, continuous and nonlinear panel methods extend, rather than replace, the design-based frameworks in earlier chapters. They allow you to ask marginal and dynamic questions. How much to spend? How deep to discount? How behaviour changes at intensive versus extensive margins? Once a credible identification strategy has been chosen, these questions can be answered. The panel causal designs in Chapter 18 therefore reuse the estimands, models, and diagnostics from this chapter. ADRFs and ACRs are used for continuous treatments. Nonlinear FE logit and Poisson target binary and count outcomes. Hurdle and zero-inflated models address excess zeros and caps. The common thread is that design and diagnostics come first. For ad spend, design might mean randomised budget shocks or a credible instrument. For pricing, it might mean cost-shifter IVs or quasi-random variation. For engagement, it might mean a randomised

targeting rule. Continuous and nonlinear models then translate those designs into interpretable dose-response and dynamic effects that speak directly to marketing decisions. The choice of nonlinear link or continuous ADRF/ACR model never substitutes for credible identifying variation. Quasi-random timing, instruments, or carefully argued unconfoundedness remain the linchpins of causal interpretation.

## 14.14 Workflow Checklist

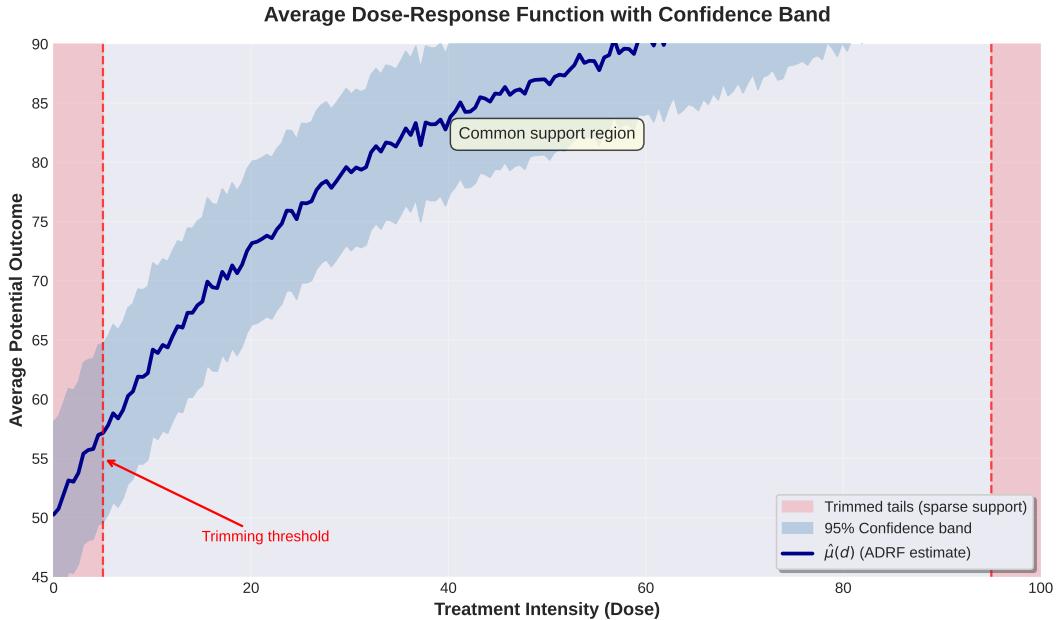
The following protocol supports reproducible analysis from estimand to decision.

### Box 14.1: Continuous-Treatment Analysis Checklist

A rigorous analysis of continuous treatments begins by defining the estimand—an ADRF, specific dose contrasts, or marginal effects (Section 14.2)—and stating the design assumptions (Section 14.9) that identify it. First, write the estimand in the book’s notation, specify the overlap region  $\mathcal{D}_0$  and population (for example, units and time windows) to which it applies, and explain its business meaning. Next, inspect the dose distributions and assess GPS overlap (Sections 14.4 and 14.11), planning for trimming, binning, or coarser grids in sparse regions. Use these diagnostics to decide where in the dose space effects are credibly identified and restrict primary reporting to that overlap region.

Then select an appropriate estimator—GPS-weighted, outcome regression, doubly robust, or DML (Section 14.4)—and choose a basis for the dose that balances flexibility with parsimony. Tune nuisance learners using cross-validation within dependence-respecting folds (Section 14.10), ensuring that nuisance functions are trained only on pre-treatment or out-of-fold data to prevent leakage. Tuning choices can stabilise estimates within the identified region but cannot repair violations of unconfoundedness or overlap.

Finally, validate the model with diagnostics including covariate-balance checks, overlap inspections, placebo checks, and sensitivity to basis and bandwidth choices (Section 14.11). Conduct inference using clustered or wild cluster bootstrap methods on the relevant influence functions (Section 14.12), and report confidence bands for the ADRF and derived marginal effects. If you map results to decisions, state the objective and the cost model explicitly (for example, profit per incremental unit of spend) and restrict decisions to the overlap-supported region.

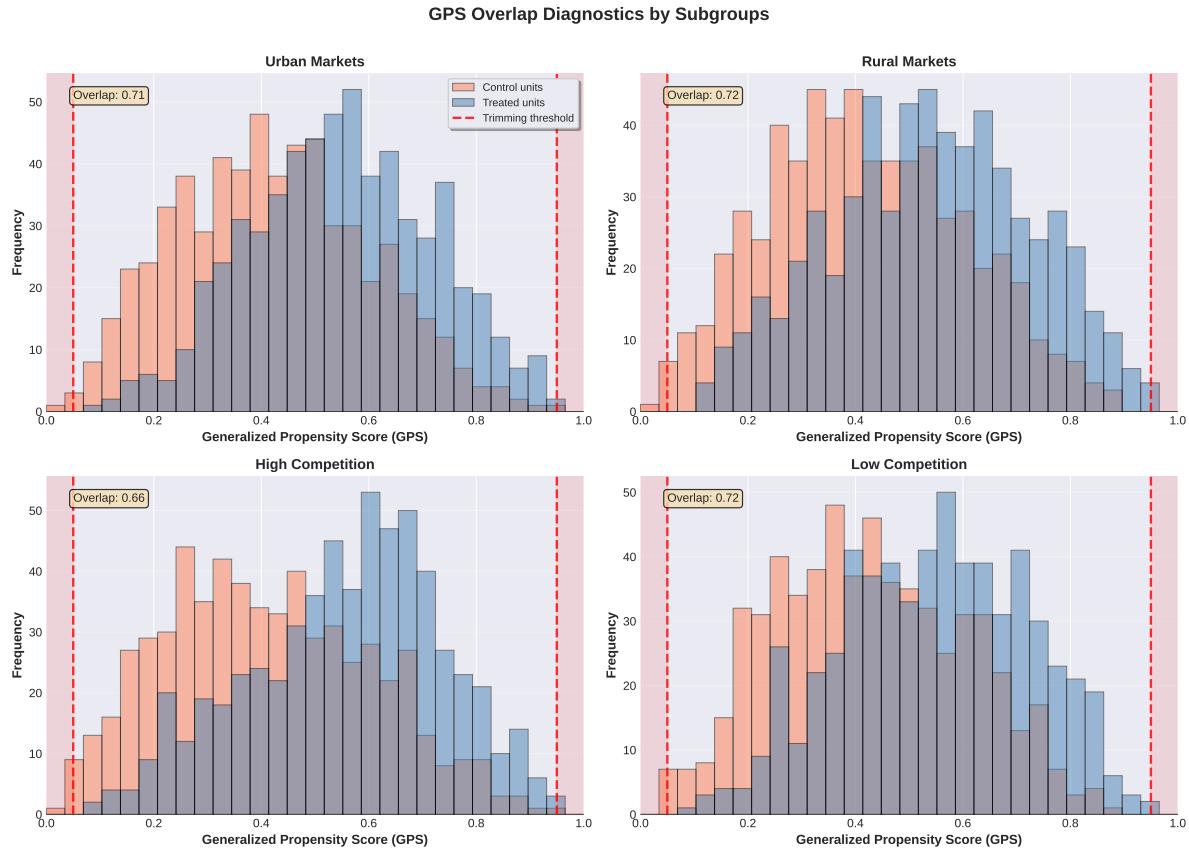


**Fig. 14.1** Average dose-response function with confidence band

The ADRF  $\hat{\mu}(d)$  is plotted over the reporting region with 95% confidence bands. Shaded tail regions denote doses excluded by the trimming rule.

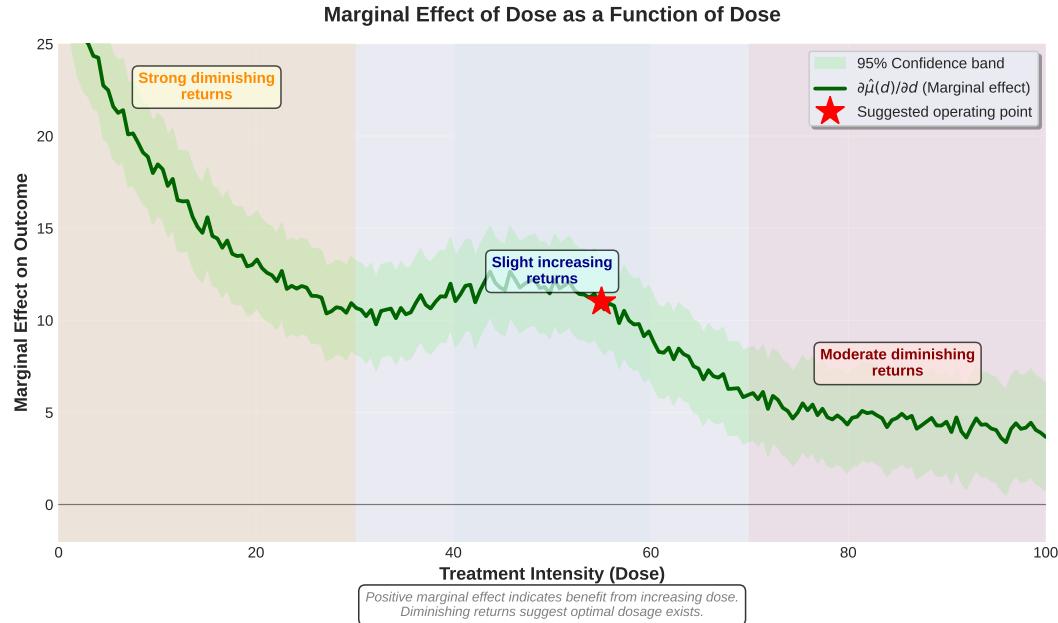
**Table 14.3** Estimator to assumptions, tuning, and use-cases

Estimator	Key assumptions	Tuning and diagnostics	Use-cases
GPS-weighted / OR	Unconfoundedness, overlap	Basis in dose, stabilised weights, overlap and balance checks	Moderate $X$ , clear support. Simple ADRFs $\mu(d)$
DR (GPS + OR)	Unconfoundedness, overlap. Model double robustness	Cross-fitted nuisances, trimming, sensitivity to outcome and GPS link	Guard against misspecification in either GPS or outcome model
DML (continuous)	Orthogonal scores, overlap, fold independence. Cluster-level inference	Blocked cross-fitting, learner stability checks, joint bands over dose	High-dimensional $X$ , flexible ADRFs $\mu(d)$ and ACR profiles
Factor / IV	Factor structure or valid instruments. Rank, relevance, and exclusion	Rank tests, placebo exposures, over-identification checks (when applicable)	When selection-on-observables is doubtful. Leverage common shocks or exogenous shifts in intensity to identify local or low-dimensional functionals of $\mu(d)$ (for example, linear-price elasticities), recognising that a single instrument rarely identifies the full ADRF



**Fig. 14.2** Overlap diagnostics for the GPS by dose bins and subgroups

*Stacked histograms show the distribution of estimated GPS values within dose bins and subgroups. Dashed lines mark trimming thresholds.*



**Fig. 14.3** Marginal effect of dose as a function of dose

The derivative  $\partial\hat{\mu}(d)/\partial d$  is plotted with confidence bands. Any operating point requires an explicit objective and cost curve and should be justified separately.

**Part VII**

**Validity, Inference, and Diagnostics**



## Chapter 15

# Threats to Validity in Marketing Panels

Marketing panels sit at the intersection of business strategy, platform engineering, and data collection. Algorithms decide which impressions are shown, privacy policies restrict what can be logged, and calendars inject seasonality and shocks into every time series. Those forces create recurrent threats to the validity of causal estimates.

This chapter catalogues those threats and links them directly to the identification assumptions from Chapter 2. We show how seasonality, platform policy changes, algorithmic confounding, and measurement error violate unconfoundedness, parallel trends, or SUTVA in concrete ways. We then lay out a practical playbook of diagnostics and sensitivity analyses, aligned with the design-based principles in Chapters 17 and 16. These diagnostics are heuristics rather than pass-fail checks, and they cannot rescue a design whose identifying assumptions are implausible.

## 15.1 Motivation and Scope

Platform algorithms, privacy policies, and seasonal forces shape what we observe in marketing panels. These features generate recurrent threats to the validity of causal estimates.

A threat to validity is any feature of the data or design that biases our estimates, inflates uncertainty, or causes the identified quantity to differ from the causal effect we care about. In the language of Chapter 2, threats operate by breaking the link between the target estimand and the functional that is actually identified under a maintained set of assumptions. When that link fails, the quantity identified by the design drifts away from the effect we care about, creating identification bias that does not vanish with more data.

Threats also differ by design class. A geo experiment and a staggered-adoption DiD face different failure modes, as do a single treated unit design, a common shock design, a continuous-treatment design, or an interference setting. Throughout, we keep the analysis design-first. We start from the estimand at risk and the assumptions required for identification, then describe how marketing data can break those assumptions.

It also helps to separate internal and external validity. Internal validity is about whether the design identifies the intended estimand in the study sample. External validity is about whether that estimand transports to a different time, market, or platform environment. This chapter concentrates on internal validity, and it flags external-validity threats when the same forces that break identification also limit transportability.

We advocate anticipating these threats and integrating diagnostics into research design from the outset rather than treating them as an afterthought. This chapter maps threats to the assumptions they violate, explains the bias pathways they create, and presents mitigation strategies that range from redesigning the study to reporting sensitivity analyses. We draw on the identification foundations in Chapter 2 and the robust inference methods in Chapter 16. Our perspective connects modern panel frameworks for causal inference—see, for example, Imbens and Rubin [2015], Hernán and Robins [2020], Rosenbaum [2002], Manski [2003], Arkhangelsky and Imbens [2024]—with the specific threats that arise in marketing panels.

## 15.2 A Taxonomy of Bias Sources

To diagnose threats rigorously, we classify the sources of bias by the identification assumptions they violate. Different estimation strategies rely on different assumptions, and a given threat can manifest differently depending on the design. This taxonomy provides a roadmap for sensitivity analysis tailored to your chosen strategy.

**Definition 15.1 (Taxonomy of Bias Sources)** Let  $\hat{\delta}$  be an estimator targeting a causal estimand  $\delta^*$ . Threats to validity arise when the identification assumptions underlying the estimator fail. We group those threats into four canonical categories, organised by the assumption they most directly attack.

*Confounding bias*,  $B_{\text{confound}}$ , arises when unobserved variables affect both treatment and outcome. In selection-on-observables designs (matching, regression adjustment), this is a failure of unconfoundedness. In difference-in-differences designs, time-varying unobservables that differentially affect treated and control units show up as parallel-trends violations. In marketing panels, algorithmic budget allocation that targets high-response users or markets is a central source of confounding bias.

*Parallel-trends bias*,  $B_{\text{PT}}$ , is specific to DiD and synthetic control methods. It arises when treated and control units would have followed different outcome trajectories absent treatment, even after conditioning on observed covariates and fixed effects. Seasonality and holiday shocks that differ across treated and control markets are common sources of parallel-trends violations in marketing settings.

*Measurement bias*,  $B_{\text{meas}}$ , stems from error in measuring treatment status, outcomes, or key covariates. It affects all designs but has method-specific consequences: classical attenuation in regressions, misclassification in DiD cell means, or non-classical error that violates exclusion restrictions in IV. Impression logging bugs, cookie deletion, and panel churn are pervasive sources of measurement bias in digital marketing data.

*Spillover bias*,  $B_{\text{spill}}$ , occurs when one unit's treatment affects another unit's outcome, violating SUTVA. It is particularly relevant in networked and geographic settings. Competitive responses—such as a rival retailer matching a price cut—and geographic spillovers across adjacent markets are common sources of spillover bias in marketing panels.

These labels are schematic rather than mutually exclusive. A single underlying problem, such as time-varying unobservables interacting with seasonality, can appear under different headings depending on the design and the estimand.

The goal of research design is to eliminate or neutralise the threats relevant to your chosen method. When elimination is impossible, sensitivity analysis must bound their plausible magnitude.

### Identification Bias versus Estimation Bias

A fundamental distinction separates two sources of error: identification failure and finite-sample estimation error.

**Definition 15.2 (Identification Bias)** Let  $\delta^*$  denote the target estimand (for example, ATT or ATE) and let  $\tilde{\delta}$  denote the population quantity that is actually identified by a given design under its maintained

assumptions. The *identification bias* is

$$\Delta_{\text{id}} = \delta^* - \tilde{\delta}.$$

This gap is zero when assumptions hold and the identified functional coincides with the target. Threats to validity create  $\Delta_{\text{id}} \neq 0$  by causing the identified quantity to diverge from the causal effect of interest. Identification bias persists even with infinite data.

**Definition 15.3 (Estimation Bias)** Let  $\hat{\delta}$  denote a finite-sample estimator of the identified functional  $\delta$ . The *estimation bias* is

$$\Delta_{\text{est}} = \mathbb{E}[\hat{\delta}] - \tilde{\delta}.$$

This reflects finite-sample properties of the estimator—regularisation, sample selection, small-sample corrections—and vanishes asymptotically for consistent estimators. It is the focus of Chapter 16 and the post-selection methods in Chapter 13.

Under the maintained model, we can write total bias as the sum of these two components,

$$\mathbb{E}[\hat{\delta}] - \delta^* = \Delta_{\text{id}} + \Delta_{\text{est}}.$$

Chapters 17 and 16 give you tools for reducing  $\Delta_{\text{est}}$  through better estimators and uncertainty quantification. The rest of this chapter is primarily about identifying, diagnosing, and bounding  $\Delta_{\text{id}}$ —the part of the error that does not vanish with more data. We deploy placebo diagnostics, calibration exercises, and Rosenbaum-type sensitivity analyses to probe designs and bound the plausible magnitude of identification bias.

### 15.3 Parallel Trends Violations

The parallel trends assumption is the cornerstone of difference-in-differences (DiD) and related designs (see Chapter 4 for the formal treatment). In marketing panels this assumption is frequently threatened by seasonality, calendar effects, and diverging competitive trajectories.

**Definition 15.4 ( $\delta$ -Violation of Parallel Trends)** Let  $Y_{it}(0)$  denote the potential outcome under no treatment, and let  $G_i$  denote the adoption time with  $G_i = \infty$  for never treated. In this two-period illustration we collapse to an ever-treated comparison for readability. The underlying treatment path remains  $D_{it} = \mathbf{1}\{t \geq G_i\}$  as defined in Chapter 4. In the simplest two-period DiD setup, parallel trends holds if, for a post-treatment period  $t$  and a pre-treatment period  $t' < t$ ,

$$\mathbb{E}[Y_{it}(0) - Y_{it'}(0) | G_i < \infty] = \mathbb{E}[Y_{it}(0) - Y_{it'}(0) | G_i = \infty].$$

A  $\delta$ -bounded violation allows these untreated trends to differ by at most  $\delta$ :

$$\left| \mathbb{E}[Y_{it}(0) - Y_{it'}(0) | G_i < \infty] - \mathbb{E}[Y_{it}(0) - Y_{it'}(0) | G_i = \infty] \right| \leq \delta.$$

The parameter  $\delta$  quantifies the maximum allowable deviation from parallel trends [Rambachan and Roth, 2023].

When parallel trends does not hold exactly, we can no longer point-identify the treatment effect. However, if we can bound the violation by  $\delta$  in this sense, we can construct partial-identification bounds on the ATT in the simple two-period setting.

**Theorem 15.1 (DiD Bias Bounds)** *In a two-period, two-group DiD design, under a  $\delta$ -bounded parallel-trends violation as in Definition 15.4, the probability limit of the DiD estimator satisfies*

$$\text{ATT} - \delta \leq \text{plim } \hat{\tau}^{\text{DiD}} \leq \text{ATT} + \delta,$$

where ATT is the true average treatment effect on the treated in this two-period setting. The bound follows from the fact that, in the two-period setup, the DiD estimand differs from ATT only through the difference in untreated trends between treated and control units, which is at most  $\delta$  in absolute value. Inverting this relationship, the identified set for ATT given the DiD estimate is

$$\mathcal{T}(\delta) = [\hat{\tau}^{\text{DiD}} - \delta, \hat{\tau}^{\text{DiD}} + \delta].$$

Reporting  $\mathcal{T}(\delta)$  for a range of  $\delta$  values provides transparent sensitivity analysis to bounded violations of parallel trends. See Rambachan and Roth [2023] for formal derivations and extensions to multi-period designs.

In more complex multi-period or staggered-adoption designs, the mapping between  $\delta$  and bias involves the entire path of violations across time and cohorts. Chapter 5 discusses dynamic DiD and event-study specifications, where these ideas extend to bounds on the whole event-time profile. As a rough calibration,

the scale of  $\delta$  should be interpreted relative to outcome variability: if you are unwilling to rule out average trend deviations of economically meaningful magnitude, you should expect DiD estimates to be uncertain within roughly that band.

A practical question is how to choose  $\delta$ . Data-driven approaches use pre-treatment trends to calibrate plausible post-treatment deviations [Rambachan and Roth, 2023].

**Proposition 15.1 (Calibrating  $\delta$  from Pre-Trends)** *Let  $\hat{\delta}_{\text{pre}}$  denote the maximum observed pre-treatment trend difference,*

$$\hat{\delta}_{\text{pre}} = \max_{t < t_0} \left| (\bar{Y}_t^{\text{treat}} - \bar{Y}_{t-1}^{\text{treat}}) - (\bar{Y}_t^{\text{control}} - \bar{Y}_{t-1}^{\text{control}}) \right|.$$

$\bar{Y}_t^{\text{treat}}$  and  $\bar{Y}_t^{\text{control}}$  denote the sample means of  $Y_{it}$  at time  $t$  in the treated and control groups respectively, and  $t_0$  is the first post-treatment period. If we assume that post-treatment trend deviations are bounded by a multiple of pre-treatment deviations, setting  $\delta = c \cdot \hat{\delta}_{\text{pre}}$  for some multiplier  $c \geq 1$  provides a data-driven sensitivity bound.  $c = 1$  treats the worst pre-period deviation as a bound for post-period deviations.  $c > 1$  allows for the possibility that divergences are worse after treatment. In practice, compute  $\hat{\delta}_{\text{pre}}$  from pre-treatment data and report identified sets for  $\delta \in \{\hat{\delta}_{\text{pre}}, 2\hat{\delta}_{\text{pre}}\}$ . This “honest” approach avoids using diagnostic checks for parallel trends both to validate and to rely on the parallel-trends assumption.

## Calendars, Seasonality, and Event Interference

In practice, calendar effects are the most common source of parallel-trends violations  $B_{\text{PT}}$  in marketing applications, and hence of identification bias in DiD designs. Holidays, pay cycles, and major events shift demand in ways that are unrelated to the intervention but may correlate with its timing. If treated and control units experience these calendar shocks differently, DiD estimates will conflate the treatment effect with calendar-driven variation.

We can mitigate these effects through design and specification. Balanced pre- and post-treatment windows, rich fixed effects for holidays and other recurring events, and flexible seasonality controls all help. Event-time alignment, as discussed in Chapter 5, ensures that we compare cohorts at the same point in their treatment journey, which reduces bias from seasonality and staggered adoption.

Event interference arises when concurrent campaigns, competitor actions, or macroeconomic shocks overlap with the study window. Isolating clean windows is sometimes feasible but may sacrifice external validity. High-dimensional controls (Chapter 13) can adjust for observable concurrent exposures, while designs that absorb common shocks via unit and time fixed effects (Chapter 4) or latent factors (Chapter 8) reduce bias from unobserved common trends. Cohort-specific event-time profiles and aggregation-robust DiD estimators from Chapter 4 provide additional tools for diagnosing and mitigating these violations.

## 15.4 Omitted Variable Bias and Sensitivity

When randomisation is imperfect (or absent), unobserved confounders may drive both treatment assignment and outcomes, creating omitted-variable bias  $B_{\text{confound}}$ .

**Definition 15.5 (Omitted-Variable Bias)** Consider the true relationship after applying the within transformation to remove unit and time fixed effects:

$$Y_{it}^{\text{within}} = \tau D_{it}^{\text{within}} + \gamma U_{it}^{\text{within}} + \varepsilon_{it}^{\text{within}},$$

where  $U_{it}$  is an unobserved confounder and the superscript denotes demeaning by unit and time means. The short regression omitting  $U_{it}^{\text{within}}$  yields

$$\hat{\tau}^{\text{short}} \xrightarrow{P} \tau + \gamma \cdot \delta_U,$$

where  $\delta_U = \text{Cov}(D_{it}^{\text{within}}, U_{it}^{\text{within}})/\text{Var}(D_{it}^{\text{within}})$  is the slope from regressing the within-transformed confounder on the within-transformed treatment. The omitted-variable bias is

$$B_{\text{OVB}} = \gamma \cdot \delta_U.$$

This  $B_{\text{OVB}}$  is a concrete representation of the confounding bias  $B_{\text{confound}}$  from Definition 15.1 in a linear setting. Bias is large when the confounder strongly affects outcomes ( $|\gamma|$  large) and is strongly correlated with treatment ( $|\delta_U|$  large). This classical formula, introduced in Chapter 2, remains the conceptual foundation for sensitivity analysis even when the exact linear assumptions are relaxed.

To assess robustness to hidden bias, we use sensitivity parameters that quantify how strong a confounder would need to be to explain away a result.

**Definition 15.6 (Sensitivity Parameter  $\Gamma$ )** Rosenbaum's sensitivity parameter  $\Gamma \geq 1$  bounds the odds ratio of treatment assignment for units with identical observed covariates [Rosenbaum, 2002]:

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(D_i = 1 | X_i, U_i)/\mathbb{P}(D_i = 0 | X_i, U_i)}{\mathbb{P}(D_j = 1 | X_j, U_j)/\mathbb{P}(D_j = 0 | X_j, U_j)} \leq \Gamma,$$

For all pairs  $(i, j)$  with  $X_i = X_j$  but potentially  $U_i \neq U_j$ . For clarity, we write this definition in a cross-sectional, unit-level form with a single binary treatment  $D_i$ . In panel settings with once-treated-always-treated designs, apply it to unit-level adoption indicators and suitable summaries of  $X_{it}$ . When  $\Gamma = 1$ , assignment depends only on observed covariates (no hidden bias). As  $\Gamma$  increases, larger departures from random assignment within strata are permitted.

**Proposition 15.2 (Rosenbaum Bounds)** *Given overlap and a sensitivity parameter  $\Gamma$ , bounds on the true treatment-effect p-value satisfy*

$$p^-(\Gamma) \leq p_{\text{true}} \leq p^+(\Gamma),$$

where  $p^-(\Gamma)$  and  $p^+(\Gamma)$  are the minimum and maximum p-values over all assignments consistent with  $\Gamma$ . The critical value  $\Gamma^*$  is the smallest  $\Gamma$  such that  $p^+(\Gamma^*) > \alpha$ . Reporting  $\Gamma^*$  indicates how much hidden bias

*is required to overturn the finding. These bounds do not recover an unbiased point estimate when  $\Gamma > 1$ . They quantify how much hidden bias is required before the study becomes inconclusive at level  $\alpha$ .*

In practice, these bounds are computed by re-evaluating sharp-null randomisation statistics over the set of treatment assignments consistent with  $\Gamma$ . They are most naturally applied when treatment is binary and assigned once at the unit level (for example, store-level adoption designs). In high-frequency ad-exposure settings, aggregate to store-week or user-campaign cells and treat the resulting binary indicator as the treatment variable when applying Rosenbaum-style analyses. Extensions to high-frequency dynamic treatments require additional structure.

An alternative, popular in economics, uses the stability of coefficients when controls are added to bound the bias from remaining unobservables.

**Definition 15.7 (Coefficient-Stability Approach)** Following Oster [2019], assume proportional selection: the relationship between treatment and unobservables is proportional to the relationship between treatment and observables. Define the proportionality parameter

$$\kappa = \frac{\text{Selection on unobservables}}{\text{Selection on observables}}.$$

The bias-adjusted treatment effect under proportional selection is

$$\tau^{\text{adj}}(\kappa) = \hat{\tau}_{\text{long}} - \kappa(\hat{\tau}_{\text{short}} - \hat{\tau}_{\text{long}}) \cdot \frac{R_{\text{max}}^2 - R_{\text{long}}^2}{R_{\text{long}}^2 - R_{\text{short}}^2},$$

where  $\hat{\tau}_{\text{short}}$  and  $\hat{\tau}_{\text{long}}$  are treatment-effect estimates from regressions without and with controls,  $R_{\text{short}}^2$  and  $R_{\text{long}}^2$  are their respective  $R^2$  values, and  $R_{\text{max}}^2$  is the hypothetical  $R^2$  if all confounders were observed. In practice, economists often set  $R_{\text{max}}^2$  slightly above  $R_{\text{long}}^2$  (for example,  $R_{\text{max}}^2 = \min\{1, R_{\text{long}}^2 + 0.03\}$ ) and report  $\tau^{\text{adj}}(\kappa)$  for  $\kappa \in \{1, 2\}$ . This proportional-selection assumption is strong and not testable from the data alone, so  $\tau^{\text{adj}}(\kappa)$  should be interpreted as a sensitivity curve rather than a point-identified estimand.

## Algorithmic Confounding and Targeting Feedback

Algorithmic confounding is the primary source of omitted-variable bias in digital markets. Platforms use optimisation algorithms to target treatments based on predictions of user behaviour, creating selection on unobservables whenever the platform's model uses information unavailable to the researcher. Treated and control groups may differ systematically in ways that remain unobserved even after conditioning on all recorded covariates.

We can mitigate algorithmic confounding by design and estimation. Clustered randomisation and switchback designs limit the optimiser's ability to fine-tune assignments at the individual level. Orthogonalised machine-learning methods from Chapter 12 allow us to incorporate observable platform predictions as controls while protecting against regularisation bias, but they do not eliminate bias from genuinely unobserved features inside the optimiser. These methods protect against overfitting and regularisation bias conditional

on observed covariates, but they cannot fix violations of unconfoundedness. The tension that Breiman [2001] highlighted between prediction and identification remains: better prediction does not guarantee better causal identification.

Budget pacing redistributes spend over time to smooth delivery against capacity or budget constraints, and bandit learning adapts targeting rules as data accumulate. Both introduce intertemporal feedback that violates standard identification when interpreted as exogenous shocks. Event-time analyses (Chapter 5) and dynamic models (Chapter 10) can accommodate these adaptive mechanisms, provided the timing and rules are documented and explicitly modelled. Treating algorithm-driven dose changes as if they were randomised invites bias and should be avoided, unless the allocation rules are randomised by design and fully documented.

## 15.5 Measurement Error

Measurement error  $B_{\text{meas}}$  is a pervasive threat in marketing panels. Our measures of exposure (for example, ad impressions, discount depth) and outcomes (for example, sales, conversions) are imperfect. Measurement error can enter through the treatment  $D_{it}$ , the outcome  $Y_{it}$ , or covariates  $X_{it}$ , and its consequences depend on the design and estimator. We distinguish between classical error, which often attenuates estimates in simple linear regressions, and non-classical error, which can introduce bias in any direction.

**Definition 15.8 (Classical Measurement Error)** Let  $D_{it}^*$  denote the true treatment or exposure. Potential outcomes are  $Y_{it}(d)$  for dose level  $d$ , and we observe  $Y_{it} = Y_{it}(D_{it}^*)$ . Let  $D_{it}$  be the mismeasured regressor used in the working panel specification (typically after removing unit and time fixed effects or factors). Classical measurement error assumes

$$D_{it} = D_{it}^* + \eta_{it}$$

where  $\eta_{it}$  is measurement error with  $\mathbb{E}[\eta_{it}] = 0$ ,  $\text{Cov}(\eta_{it}, D_{it}^*) = 0$ , and  $\text{Cov}(\eta_{it}, \varepsilon_{it}) = 0$ . In the simplest linear regression with a single mismeasured regressor, these conditions imply attenuation toward zero [Fuller, 2009]. With fixed effects, additional regressors, or nonlinear estimators, the direction and magnitude of bias need not follow the textbook attenuation formula.

**Proposition 15.3 (Attenuation Factor)** *Under classical measurement error in treatment with reliability ratio*

$$\lambda = \frac{\text{Var}(D_{it}^*)}{\text{Var}(D_{it})} = \frac{\text{Var}(D_{it}^*)}{\text{Var}(D_{it}^*) + \text{Var}(\eta_{it})} \in (0, 1]$$

the OLS estimator satisfies

$$\hat{\delta}^{\text{OLS}} \xrightarrow{p} \lambda \cdot \delta^*.$$

The attenuation factor  $\lambda < 1$  implies that estimates are biased toward zero, with attenuation bias

$$B_{\text{attenuation}} = (1 - \lambda) \cdot \delta^*.$$

This  $B_{\text{attenuation}}$  is a specific component of the measurement bias  $B_{\text{meas}}$  from Definition 15.1 under classical error. When validation data provide an estimate of  $\text{Var}(\eta_{it})$ , we can compute  $\lambda$  and correct the estimate by dividing by  $\lambda$ . In richer panel specifications with additional regressors or fixed effects, the exact attenuation formula may not hold, but the reliability-ratio logic still describes the main force pushing estimates towards zero when error is roughly classical. In non-linear models the exact multiplicative form may not hold, but the reliability-ratio intuition remains useful for approximate debiasing.

In marketing data, measurement error is often non-classical.

**Definition 15.9 (Non-Classical Measurement Error)** Non-classical measurement error violates one or more of the classical conditions. Common forms include:

*Differential error:* the error distribution differs by outcome value or treatment status (for example, high-activity users generate more trackable impressions).

*Endogenous error:* the error correlates with the true treatment,  $\text{Cov}(\eta_{it}, D_{it}^*) \neq 0$  (for example, capping or throttling at platform limits). Endogenous error of this form behaves like an additional unobserved confounder and contributes to the confounding bias component  $B_{\text{confound}}$ .

*Systematic bias:* the error depends on covariates,  $\mathbb{E}[\eta_{it} | X_{it}] \neq 0$  (for example, under-reporting of offline conversions in certain regions).

Non-classical error can bias estimates in any direction, not necessarily toward zero. Platform-specific measurement choices—such as attribution windows, viewability thresholds, and deduplication rules—often induce non-classical error.

When the reliability ratio  $\lambda$  is unknown but classical error is a reasonable approximation, we can bound the treatment effect under assumptions about the plausible range of measurement error.

**Proposition 15.4 (Measurement-Error Bounds)** *Suppose validation data or domain knowledge suggest that the reliability ratio lies in  $[\lambda_{\min}, \lambda_{\max}]$  with  $0 < \lambda_{\min} \leq \lambda_{\max} \leq 1$ . If we know that the true effect  $\delta^*$  is non-negative, so that attenuation pushes estimates towards zero from above, then under classical measurement error the identified set for  $\delta^*$  is*

$$\mathcal{T}(\lambda_{\min}, \lambda_{\max}) = \left[ \frac{\hat{\delta}}{\lambda_{\max}}, \frac{\hat{\delta}}{\lambda_{\min}} \right].$$

*When the sign of  $\delta^*$  is unknown, the interval endpoints may flip and the set should be defined without imposing an order. Reporting bounds for plausible ranges of  $\lambda$  (for example,  $\lambda \in [0.7, 1.0]$  when measurement is believed to be reasonably accurate) connects point estimates to a partial-identification perspective.*

## Attribution Misalignment

Attribution misalignment is a specific form of measurement threat. Platform attribution rules—such as last-click, first-touch, or view-through attribution—define outcomes in ways that depend on treatment paths and on other exposures. These definitions generally do not coincide with the counterfactual-based estimands we seek in causal inference. For example, “last-click conversions in a 7-day window” defines a different  $Y_{it}(d)$  object than “incremental revenue over a 28-day horizon,” and treating one as a proxy for the other introduces measurement bias.

The best practice is to use stable, external outcomes when possible (for example, revenue recorded in internal systems rather than platform-reported conversions). When platform-reported conversions must be used, be transparent about the attribution rules, relate them to the potential-outcomes estimand, and run sensitivity analyses that vary attribution windows or rules. Attribution-induced measurement error is one reason you might only be willing to bound the true effect rather than point-identify it. Chapter 19 returns to these issues in more detail and links attribution choices to the partial-identification framework in Section 15.6.

## 15.6 Partial Identification

When point identification is impossible due to threats such as measurement error or parallel-trends violations, we turn to partial identification [Manski, 2003]. Rather than abandoning causal inference when assumptions are questionable, we report the range of treatment effects consistent with the data under weaker assumptions.

**Definition 15.10 (Identified Set)** Under a set of maintained assumptions  $\mathcal{A}$ , the identified set for a scalar parameter  $\delta$  is

$$\mathcal{T}_{\mathcal{A}} = \{\delta : \text{there exists a data-generating process satisfying } \mathcal{A} \text{ that generates the observed data and has parameter } \delta\},$$

where the data-generating process refers to the joint distribution of observables and potential outcomes. In this chapter,  $\delta$  typically corresponds to a treatment effect such as ATT or ATE. Point identification occurs when  $\mathcal{T}_{\mathcal{A}}$  is a singleton. Partial identification occurs when  $\mathcal{T}_{\mathcal{A}}$  is an interval or, more generally, a non-degenerate subset of the parameter space. In this chapter we focus on interval-valued identified sets for scalar treatment effects and report  $\mathcal{T}_{\mathcal{A}}$  as bounds when point identification fails.

The partial-identification framework unifies the sensitivity analyses developed earlier. The  $\delta$ -bounded parallel-trends violation in Definition 15.4 yields an identified set  $\mathcal{T}_{\mathcal{A}(\delta)}$ , which we previously denoted by  $\mathcal{T}(\delta)$ , around the DiD estimate. The measurement-error bounds in Proposition 15.4 yield an identified set  $\mathcal{T}_{\mathcal{A}(\lambda)}$ , which we previously denoted by  $\mathcal{T}(\lambda_{\min}, \lambda_{\max})$ , around an attenuated estimate. In each case, we trade the precision of a point estimate for the credibility of bounds that relax untestable assumptions.

**Proposition 15.5 (Hierarchical Bounds)** *For nested assumption sets  $\mathcal{A}_1 \supset \mathcal{A}_2 \supset \mathcal{A}_3$  (where  $\supset$  denotes progressively stronger restrictions on the model space), the corresponding identified sets satisfy*

$$\mathcal{T}_{\mathcal{A}_1} \supseteq \mathcal{T}_{\mathcal{A}_2} \supseteq \mathcal{T}_{\mathcal{A}_3}.$$

*Stronger assumptions yield tighter bounds. Moving up this hierarchy corresponds to ruling out more data-generating processes. Any point that survives all three layers is consistent with both the data and the strongest assumptions.*

In practice it is useful to present results as a hierarchy. For example, you might report  $\mathcal{T}_{\mathcal{A}_1}$  as a wide set under minimal assumptions, such as sign restrictions on effects (for example, non-negative price elasticities on average) and very weak structure. You can then report  $\mathcal{T}_{\mathcal{A}_2}$  as a narrower set under bounded violations, such as  $\delta$ -bounded trends from Section 15.3 or reliability ratios in  $[\lambda_{\min}, \lambda_{\max}]$  from Section 15.5. Finally, you can report  $\mathcal{T}_{\mathcal{A}_3}$  as a point estimate (or tight interval) under the full identifying assumptions—parallel trends, no hidden confounding, classical measurement. Transparent reporting of this hierarchy allows readers to see how conclusions depend on the strength of the maintained assumptions and to form their own judgement about the credibility of each layer. In practice, we recommend displaying all three layers side by side in tables and figures so that readers can see how conclusions vary as assumptions are relaxed.

## 15.7 Structural Breaks

Marketing panels often rely on data from platforms that constantly evolve, creating structural breaks that threaten parameter stability. A structural break occurs when model parameters shift at some point in time—the treatment effect, the baseline level, or the error variance changes. This is distinct from unit roots and other forms of nonstationarity discussed in Section 2.5 and Appendix B. Structural breaks are the primary concern in short- $T$  marketing panels, where we cannot reliably diagnose unit roots but can often detect and model discrete parameter shifts.

**Definition 15.11 (Parameter Stability)** A panel model exhibits structural stability if the parameters  $\psi_t = (\tau_t, \beta_t, \sigma_t^2)$  are constant over the estimation window,

$$\psi_t = \psi \quad \text{for all } t \in \{1, \dots, T\}.$$

Here  $\tau_t$  denotes the time-specific regression coefficient on the treatment regressor  $D_{it}$  in the working specification, not a cross-period causal estimand such as the ATT. A structural break at time  $t^*$  occurs if

$$\psi_t = \begin{cases} \psi^{(1)} & t < t^* \\ \psi^{(2)} & t \geq t^* \end{cases}, \quad \psi^{(1)} \neq \psi^{(2)}.$$

The break magnitude is  $\|\psi^{(2)} - \psi^{(1)}\|$  for any chosen norm (typically Euclidean). Breaks may affect the treatment effect  $\tau_t$ , the coefficients on covariates  $\beta_t$ , or the error variance  $\sigma_t^2$ . For causal interpretation, breaks in  $\tau_t$  and in the mapping from observables to outcomes are of primary concern.

**Definition 15.12 (Chow Diagnostic Statistic)** For a candidate break point  $t^*$  in a linear specification with homoskedastic, serially uncorrelated errors, the Chow diagnostic statistic is

$$F_{t^*} = \frac{(SSR_{\text{pooled}} - SSR_{\text{split}})/k}{SSR_{\text{split}}/(NT - 2k)},$$

where  $SSR_{\text{pooled}}$  is the sum of squared residuals from the pooled model,  $SSR_{\text{split}} = SSR_1 + SSR_2$  is the sum from separate regressions before and after  $t^*$ , and  $k$  is the number of parameters. Under the null of no break and the classical error assumptions,  $F_{t^*} \sim F(k, NT - 2k)$ .

**Proposition 15.6 (Sup-F Diagnostic)** When the break point  $t^*$  is unknown, the sup-F statistic is

$$\sup F = \max_{t^* \in [\underline{t}, \bar{t}]} F_{t^*},$$

where  $[\underline{t}, \bar{t}]$  excludes boundary regions with insufficient observations. Critical values can be taken from Andrews [1993] or obtained via bootstrap. Large values suggest that a structural break exists somewhere in the sample, although in clustered panel settings the reference distribution should be treated as approximate and supported by simulation-based critical values. In marketing panels with clustering and serial correlation, we

*recommend computing sup- $F$  statistics with cluster-robust residuals and calibrating critical values by wild bootstrap (Chapter 16).*

## Platform Policy Changes

Platform changes—such as an algorithm update or a new user interface—create structural breaks. An algorithm update can alter who sees an ad, changing the effective treatment effect and introducing new confounding. A change in how conversions are measured can break the comparability of outcomes over time.

We can detect these breaks using break diagnostics such as the Chow and sup- $F$  statistics or more general changepoint methods, and mitigate them by estimating effects on stable sub-periods or by modelling breaks explicitly. External data can serve as negative controls: outcomes that should not be affected by treatment but are affected by platform changes reveal confounding or measurement shifts. If negative-control outcomes move at the break while the treated outcome also shifts, this is evidence that identification has changed rather than the treatment effect alone. Platform-reported metrics often lack counterfactual interpretation. Without break diagnostics and an explicit potential-outcomes mapping, these metrics should be treated as noisy surrogates that at best support bounds, not clean point estimates (Section 15.6).

## 15.8 Spillovers and SUTVA Violations

Spillovers—cross-unit effects—violate the Stable Unit Treatment Value Assumption (SUTVA), leading to spillover bias  $B_{\text{spill}}$ . In marketing, spillovers are pervasive: treated users share promoted content with control users on social platforms, price changes at one retailer affect competitors, and advertising in one geography attracts customers from nearby regions.

**Definition 15.13 (Spillover Bias)** Under SUTVA, potential outcomes depend only on a unit's own treatment. With an exposure mapping  $h_i(D_{-i,t})$  summarising others' treatments (Chapter 11), SUTVA implies  $Y_{it}(d, h_i(D_{-i,t})) = Y_{it}(d)$  for all exposure levels. When SUTVA fails because neighbours' treatments matter, we can write the probability limit of the naive ATT estimator schematically as

$$\hat{\delta}^{\text{naive}} \xrightarrow{P} \underbrace{\delta_{\text{own}}}_{\text{own effect}} + \underbrace{\delta_{\text{spill}}}_{\text{spillover component}},$$

where

$$\delta_{\text{own}} = \mathbb{E}[Y_{it}(1, e) - Y_{it}(0, e)]$$

is the direct effect of treating unit  $i$  holding exposure fixed, and  $\delta_{\text{spill}}$  captures the contribution of changes in exposure to outcomes for treated units. This decomposition is conceptual. Without additional structure on interference,  $\delta_{\text{own}}$  and  $\delta_{\text{spill}}$  are not point identified from naive ATT estimators. In general, both components depend on the structure of interference and the exposure mapping (Chapter 11).

**Proposition 15.7 (Spillover Bounds)** Suppose spillovers are uniformly bounded in magnitude: for all exposure levels  $e$  and  $e'$  in a given interference set,

$$|Y_{it}(d, e) - Y_{it}(d, e')| \leq \bar{s}.$$

Then the own effect satisfies

$$\hat{\delta}^{\text{naive}} - \bar{s} \leq \delta_{\text{own}} \leq \hat{\delta}^{\text{naive}} + \bar{s}.$$

Under this bound, the identified set for the own effect is the interval  $[\hat{\delta}^{\text{naive}} - \bar{s}, \hat{\delta}^{\text{naive}} + \bar{s}]$ , in the same sense as the partial-identification framework in Section 15.6. The bound  $\bar{s}$  can be calibrated from empirical evidence on spillover decay (for example, effects that decline with geographic distance or network hops) or from designs with buffer zones that create physical separation between treated and control units. A practical starting point is to use the maximum observed effect in buffer zones or at the boundary of the interference set as an upper bound on  $\bar{s}$ .

This bound is intentionally conservative. In many marketing networks, a single scalar  $\bar{s}$  will be crude, but it illustrates how partial-identification logic extends to interference: the naive ATT is only informative about the direct effect to the extent that plausible spillovers are small. For example, in a geo experiment with distance bands, one might set  $\bar{s}$  equal to the largest observed effect difference between units at the inner and outer edges of a buffer zone.

The spillover framework in Chapter 11 formalises exposure mappings that summarise neighbours' treatments and provides estimation and diagnostic tools for partial interference. Clustered designs with buffers or exclusion zones physically separate treated and control units to reduce contamination, and network-based designs define interference sets explicitly. In threats-to-validity terms, the key steps are to articulate where SUTVA is likely to fail, to design experiments and panels that limit spillovers where possible, and to be explicit about how much spillover you are willing to tolerate when interpreting naive ATT estimates.

## 15.9 Small-Sample and Dependence Corrections

Building on the cluster-robust methods in Chapter 16, we address the small- $G$  problem. Our statistical inference often relies on large-sample asymptotics that assume many independent clusters. When the number of clusters  $G$  is small (for example, fewer than 30 to 50), asymptotic cluster-robust standard errors can be misleading.

**Definition 15.14 (Effective Sample Size)** Under a simple exchangeable within-cluster dependence model with  $G$  clusters indexed by  $c = 1, \dots, G$ , each of size  $n_c$ , and intra-cluster correlation  $\rho$ , a heuristic effective sample size is

$$N_{\text{eff}} = \frac{N}{1 + (n_{\text{avg}} - 1)\rho}$$

where  $N = \sum_{c=1}^G n_c$  is the total number of observations and  $n_{\text{avg}} = N/G$  is the average cluster size. In practice  $\rho$  must be estimated or bounded. Even noisy estimates are useful for conveying the effective loss of information from within-cluster dependence. When  $\rho$  is large (strong within-cluster dependence),  $N_{\text{eff}}$  is close to  $G$ . Inference should then be calibrated as if the sample size were on the order of  $G$ , not  $N$ .

This design-effect formula is an approximation, but it captures the key message: what matters for cluster-robust inference in panels is the number of independent clusters, not the total number of unit-time cells. In marketing panels,  $G$  is often the number of markets, regions, or products, not the number of stores or user-time cells. This affects the variance and distribution of  $\hat{\delta}$  rather than identification, but it is a first-order threat to valid uncertainty quantification (see the estimation-bias discussion in Section 15.2).

**Proposition 15.8 (Wild Cluster Bootstrap and Small  $G$ )** *When the number of clusters  $G$  is small, the wild cluster bootstrap provides a practical route to more reliable inference than purely asymptotic cluster-robust standard errors [Cameron et al., 2008]. The basic procedure is to compute the cluster-robust estimate  $\hat{\delta}$  and its cluster scores from the original data, then, for each bootstrap iteration, reweight cluster-level residuals with random multipliers (for example, Rademacher weights taking values  $\{-1, +1\}$  with equal probability), recompute the estimator on the reweighted data, and form confidence intervals from the empirical distribution of the bootstrap estimates.*

*Under broadly similar regularity conditions to those used for cluster-robust variance estimators, simulation and theory show that the wild cluster bootstrap can improve coverage when  $G$  is small by approximating the finite-sample distribution of  $\hat{\delta}$ . Chapter 16 provides detailed guidance on choice of weights, studentisation, and implementation.*

Recognising small- $G$  problems and using bootstrap-based corrections when appropriate is part of treating clustered dependence as a first-order threat to valid inference rather than a technical footnote. When  $G$  is below roughly 50, and especially below 30, treat asymptotic CRVEs with caution and prefer wild cluster bootstrap confidence intervals. Small- $G$  should be flagged explicitly at the design and reporting stages, not relegated to a footnote.

## 15.10 Diagnostics: A Practical Playbook

Diagnosing threats before finalising estimates is essential. Chapter 17 sets out a comprehensive diagnostic workflow. Here we summarise the key checks most directly connected to the threats in this chapter and to the bias taxonomy in Section 15.2.

Assumption checks target identification bias. Each diagnostic constrains a component of the bias vector ( $B_{\text{confound}}, B_{\text{PT}}, B_{\text{meas}}, B_{\text{spill}}$ ) from Definition 15.1. Pre-trend diagnostics in event studies speak to parallel-trends violations  $B_{\text{PT}}$ . Pre-period deviations calibrate  $\delta$  for the identified set  $\mathcal{T}(\delta)$  in Theorem 15.1. Balance and overlap checks speak to confounding  $B_{\text{confound}}$ . Stability of estimates across specifications speaks to both omitted variables and non-classical measurement error  $B_{\text{meas}}$ . Comparisons of internal versus external outcomes or pre/post logging changes inform plausible ranges for the reliability ratio  $[\lambda_{\min}, \lambda_{\max}]$  in Proposition 15.4. These diagnostics cannot prove that assumptions hold, but they can flag gross violations and calibrate the sensitivity parameters that define identified sets.

Influence diagnostics identify fragile designs. Leave-one-out analyses by unit, time, or cohort reveal whether results hinge on a small number of clusters or periods, connecting back to the leverage and influence measures in Chapter 16. When a single market or time block drives most of the influence function, estimates are fragile even if standard errors look small. Large cluster-level influence points to effective  $G$  being even smaller than the raw number of clusters and should trigger the small- $G$  corrections of Section 15.9.

Design-robustness checks address method-specific threats. In synthetic control and generalised synthetic control (Chapters 6 and 7), donor-weight stability and donor-balance plots ensure that no single donor dominates and that pre-treatment fit is genuinely strong. In event studies, support-by- $k$  checks verify that each event-time bin has sufficient treated and comparison observations, so that dynamic profiles are not driven by a thin tail of cohorts. For spillover concerns, buffer-zone plots and geographic decay analyses constrain  $\bar{s}$  in the spillover bounds of Proposition 15.7.

Structural and seasonal diagnostics target breaks and calendar confounding. Changepoint scans and structural-break diagnostics (Section 15.7) flag periods where platform changes, policy shifts, or measurement changes alter the data-generating process. Seasonality checks and calendar plots catch holiday effects, pay cycles, and other recurring patterns that threaten parallel trends and overlap.

Often the most valuable diagnostic is the simplest: plot your data. Time series of outcomes for treated and control groups, with vertical lines marking intervention dates, reveal pre-trends, seasonality, outliers, and structural breaks more directly than any formal diagnostic statistic. Running these diagnostics ex ante, documenting them in a pre-analysis plan, and reporting results alongside main estimates supports transparent inference [Angrist and Pischke, 2010]. The diagnostic playbook in Chapter 17 provides templates for implementing these visual and formal checks in a reproducible way.

## 15.11 Sensitivity Analyses

Sensitivity analyses quantify how robust conclusions are to violations of key assumptions. For parallel trends, bounded-violation frameworks vary the allowable departure from exact parallel trends and report the identified set  $\mathcal{T}(\delta)$  from Theorem 15.1. For measurement error, we report  $\mathcal{T}(\lambda_{\min}, \lambda_{\max})$  from Proposition 15.4. For unobserved confounding, bounding strategies ask how strong a hidden confounder must be to overturn the main conclusion (Proposition 15.2). In each case, the object of interest is an identified set—a special case of  $\mathcal{T}_{\mathcal{A}}$  in Definition 15.10—rather than a single point estimate.

Specification curves systematically vary design choices—control sets, time windows, donor pools—while holding the estimand and core identification strategy fixed, and then plot the resulting distribution of estimates [Simonsohn et al., 2020]. Formally, the curve explores variation in estimators  $\hat{\delta}^{(m)}$  that all target the same identified functional  $\tilde{\delta}$  under a fixed assumption set  $\mathcal{A}$ . When the resulting distribution is tight and the sign is stable, robustness is high. When results change sign or magnitude dramatically, fragility is revealed. Sensitivity is then reported alongside the preferred specification, with a clear explanation of why that specification is preferred.

Robustness to sample and design choices takes several forms. Alternative windows restrict attention to subsamples with stronger pre-trend evidence or exclude transitional policy periods (targeting  $B_{PT}$  and structural breaks). Alternative donor sets remove near neighbours, restrict to similar units, or reweight by recent similarity (targeting spillover and contamination). Bounded-effect analyses present worst-case and best-case scenarios under explicit assumptions about the magnitude of omitted-variable bias or measurement error (targeting  $B_{confound}$  and  $B_{meas}$ ).

All sensitivity results should be reported alongside main estimates, with clear links to the assumptions that have been relaxed and to the identified sets they imply. Doing so gives readers a complete picture of robustness and makes explicit how substantive conclusions depend on the strength of the maintained assumptions. For general approaches to sensitivity and partial identification see, for example, Manski [2003], Rosenbaum [2002], Arkhangelsky and Imbens [2024].

## 15.12 Marketing-Specific Checklists

Marketing contexts introduce distinct threat profiles that go beyond generic causal-inference concerns. Platform dependencies, fraud, and creative dynamics create identification challenges specific to advertising and customer analytics, and they map directly into the bias taxonomy from Section 15.2.

### Platform Policy Shifts

Apple’s App Tracking Transparency policy, cookie deprecation, and similar platform changes alter both measurement and targeting. They threaten measurement invariance by changing which events and users are observed ( $B_{\text{meas}}$ ), they threaten overlap by changing who can be targeted at all ( $B_{\text{confound}}$ ), and they may introduce structural breaks (Section 15.7). Teams should anticipate these shifts, document dates in a timeline, and favour aggregated or externally verifiable outcomes that are less susceptible to platform-specific drift.

App Tracking Transparency and cookie changes are canonical settings where we should treat treatment effects as partially identified and use the measurement-error and partial-identification frameworks from Sections 15.5 and 15.6. When measurement changes are unavoidable, sensitivity analyses around the transition date—narrowing windows to stable sub-periods, treating pre- and post-change estimands separately, and bounding effects under alternative measurement assumptions—help preserve credibility. The recommended sensitivity parameter is the reliability-ratio range  $[\lambda_{\min}, \lambda_{\max}]$ .

### Fraud and Quality Issues

Bot traffic, invalid clicks, and non-viewable impressions contaminate marketing data and often introduce non-classical measurement error: error that depends on treatment, outcome, or platform optimisation. Fraud-induced mismeasurement typically breaks the classical conditions in Definition 15.8 and lands us in the non-classical regime of Definition 15.9, where bias can go in any direction. Vendors may supply quality flags, but these rules evolve and may themselves be endogenous.

Negative controls using outcomes that should not respond to treatment can detect unmeasured quality drift. Sensitivity to inclusion or exclusion of flagged observations bounds the influence of quality screens on substantive conclusions and can be interpreted as constraining plausible  $\lambda$  ranges or as bounding additional bias terms. The recommended sensitivity parameter is the reliability-ratio range  $[\lambda_{\min}, \lambda_{\max}]$  under alternative quality-filtering rules.

## Creative Fatigue and Saturation

Repeated exposures reduce marginal responsiveness, a phenomenon known as creative fatigue. Separating creative effects from spend effects requires event-time designs (Chapter 5) to track decay after creative refreshes and dose-response analyses (Chapter 14) to estimate diminishing returns by frequency. In the language of Chapter 10, creative fatigue shows up as event-time profiles  $\theta_k$  that decline with  $k$  and as path-dependent potential outcomes  $Y_{it}(\underline{d}_i^t)$  where early high-frequency doses change later responsiveness.

Ignoring saturation conflates level and dynamic effects and can bias both estimated treatment paths and budget allocation decisions. From the threats perspective, creative fatigue is a dynamic mis-specification problem that belongs alongside parallel-trends and stationarity violations, not just a marketing-operations concern. When short-run  $\theta_0$  is used as if it were a long-run effect, this is an identification-bias problem: the estimand itself is wrong, not just the estimator. The recommended sensitivity parameter is the event-time horizon  $K$  over which  $\theta_k$  is aggregated.

### 15.13 Assumptions and Failure Modes

This section consolidates the core assumptions underlying valid causal inference in marketing panels. Each assumption maps to a threat category from this chapter and to a specific bias pathway. When an assumption fails, the corresponding component of the bias taxonomy activates. Stating assumptions explicitly clarifies what must hold for estimates to be credible and what sensitivity analyses to report when assumptions are questionable. These assumptions specialise the global framework in Chapter 2 to the marketing-panel context.

**Assumption 73 (SUTVA and Limited Interference)** Potential outcomes depend on a unit's own treatment only through a well-defined exposure mapping, with spillovers absent or explicitly modelled. When spillovers are plausible, the estimand must distinguish own effects from spillover effects, and designs should impose structure such as partial interference (interference within clusters but not across them), buffers, or clustered assignment [Hudgens and Halloran, 2008]. Failure activates  $B_{\text{spill}}$ . The main tools are the spillover bounds in Proposition 15.7 and the partial-interference designs in Chapter 11.

**Assumption 74 (Stationarity and Structural Stability)** The mapping from treatments and covariates to potential outcomes is stable over the estimation window, or time-varying parameters are modelled explicitly. Structural breaks and nonstationarity invalidate pooled estimates and require subsample analysis, time-varying factor methods, or explicit break modelling (Chapters 8 and Section 15.7). Failure activates  $B_{\text{PT}}$  and structural-break bias. The main tools are sup-F diagnostics (Proposition 15.6) and factor models.

**Assumption 75 (Measurement Invariance)** Exposure and outcome definitions remain stable over time and across treated and control states. Platform policy changes and algorithmic confounding can violate this assumption by changing what is recorded or how it is attributed, requiring reconciliation, external outcomes, or partial-identification bounds for treatment effects when measurement shifts cannot be fully adjusted. Failure activates  $B_{\text{meas}}$ . The main tools are the classical and non-classical measurement-error frameworks in Definitions 15.8 and 15.9 and the  $\lambda$ -bounds in Proposition 15.4.

**Assumption 76 (Overlap and Support)** Treated and control groups share common support on covariates and, when applicable, on doses. Algorithmic targeting and endogenous delivery can erode overlap, concentrating treatment in regions of covariate space with few or no controls. Diagnostics and remedies include propensity-score weighting, trimming, clustered randomisation, and being explicit that the estimand is restricted to the overlap region. Failure activates  $B_{\text{confound}}$  and extrapolation bias. The main tools are GPS weighting and trimming (Chapter 14).

**Assumption 77 (Unconfoundedness or Valid Design-Based Variation)** Identification requires either (i) unconfoundedness conditional on controls and fixed effects (selection-on-observables designs), or (ii) valid design-based variation such as natural experiments, instruments, or quasi-random shocks that satisfy their own exclusion and relevance conditions. These are *alternative* identification routes with different failure modes. Prediction-focused machine learning can threaten identification when used naively [Breiman, 2001]. Orthogonalised approaches in Chapter 12 address estimation error and regularisation bias conditional on a valid design, but they do not repair violations of unconfoundedness or exclusion. Failure activates  $B_{\text{confound}}$ . The main tools are Rosenbaum  $\Gamma$  (Definition 15.6) and Oster-style coefficient-stability analysis (Definition 15.7).

## 15.14 Workflow Checklist

The following protocol consolidates threat assessment, diagnostics, and sensitivity into a reportable workflow, organised around the bias taxonomy and assumptions in Sections 15.2 and 15.13.

Box 15.1: Threats-to-Validity Workflow Checklist

**Map threats.** Given your estimand and design class, identify calendar effects, platform policy dates, sources of algorithmic confounding, measurement-error channels, spillover pathways, structural breaks, and small- $G$  dependence for the specific marketing setting. Map each to affected assumptions (SUTVA, measurement invariance, stationarity, overlap, independence) and to the corresponding bias components ( $B_{\text{confound}}$ ,  $B_{\text{PT}}$ ,  $B_{\text{meas}}$ ,  $B_{\text{spill}}$ ) in Section 15.2. Table 15.1 provides a template for this mapping.

**Align estimand and design.** State the estimand (ATT, ATE, dose-response function, etc.) and the assignment mechanism or design class that is meant to identify it. Choose a design (DiD, event study, SC/SDID, factor, DML) that is credible given the threats. When identification is only partial, acknowledge this explicitly and use the partial-identification framework in Section 15.6 to report bounds rather than point estimates. Document all maintained assumptions.

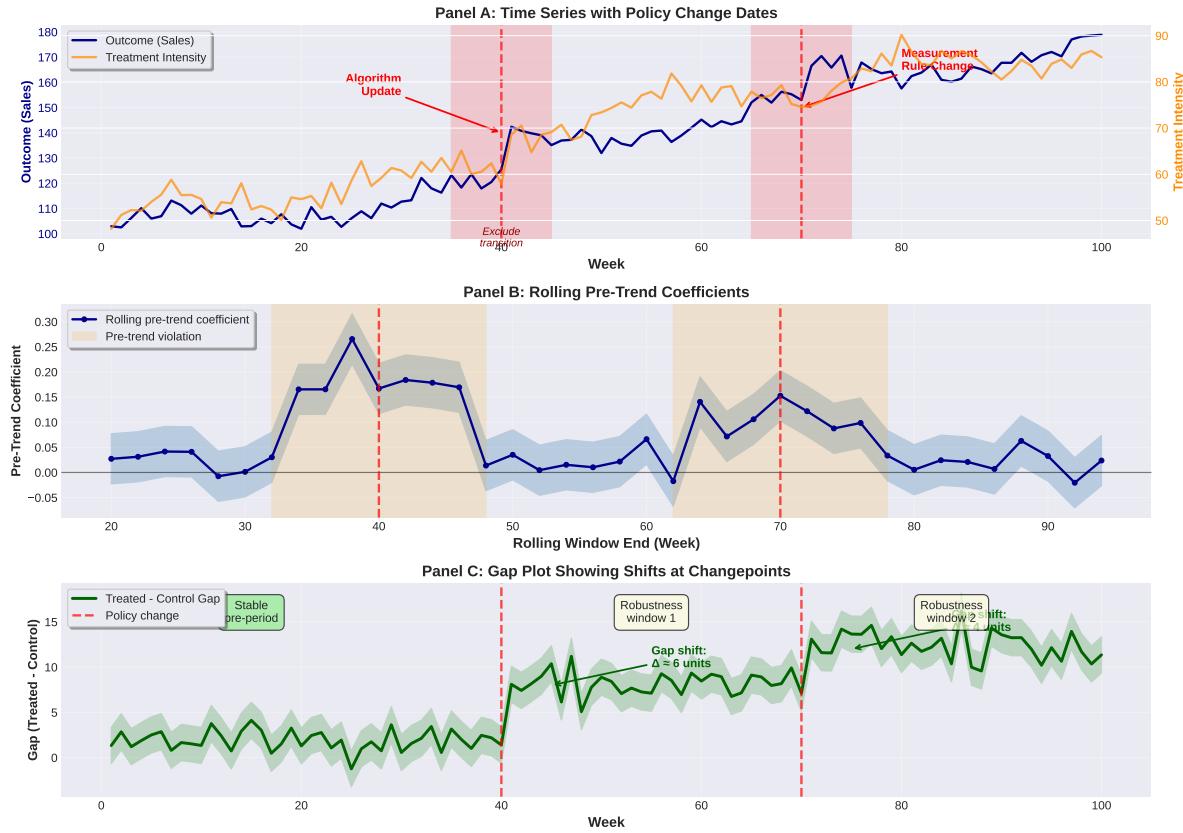
**Document platform and calendar changes.** Maintain a timeline of policy updates, algorithm changes, and major calendar events. Flag windows where stability may be compromised.

**Plan diagnostics ex ante.** Pre-specify pre-trend diagnostics, overlap checks, leave-one-out diagnostics, and changepoint scans. Register these in a design document.

**Choose robust inference.** State the independent sampling unit, which is often a cluster indexed by  $c = 1, \dots, G$ . Use wild cluster bootstrap (Section 15.9) or randomisation inference when it applies, following the guidance in Chapter 16. Account for serial and spatial dependence via clustering, and treat  $G$  as the effective sample size. When  $G$  is below roughly 50, and especially below 30, treat asymptotic CRVEs with caution and report finite-sample caveats explicitly.

**Run sensitivity analyses.** Vary windows, donor sets, controls, and trimming rules. Report the identified sets  $\mathcal{T}(\delta)$  for parallel-trends violations,  $\mathcal{T}(\lambda_{\min}, \lambda_{\max})$  for measurement error, and spillover bounds from Proposition 15.7. Present specification curves (Section 15.11) alongside main estimates. Assess sensitivity to measurement error and attribution rules using the frameworks in Section 15.5.

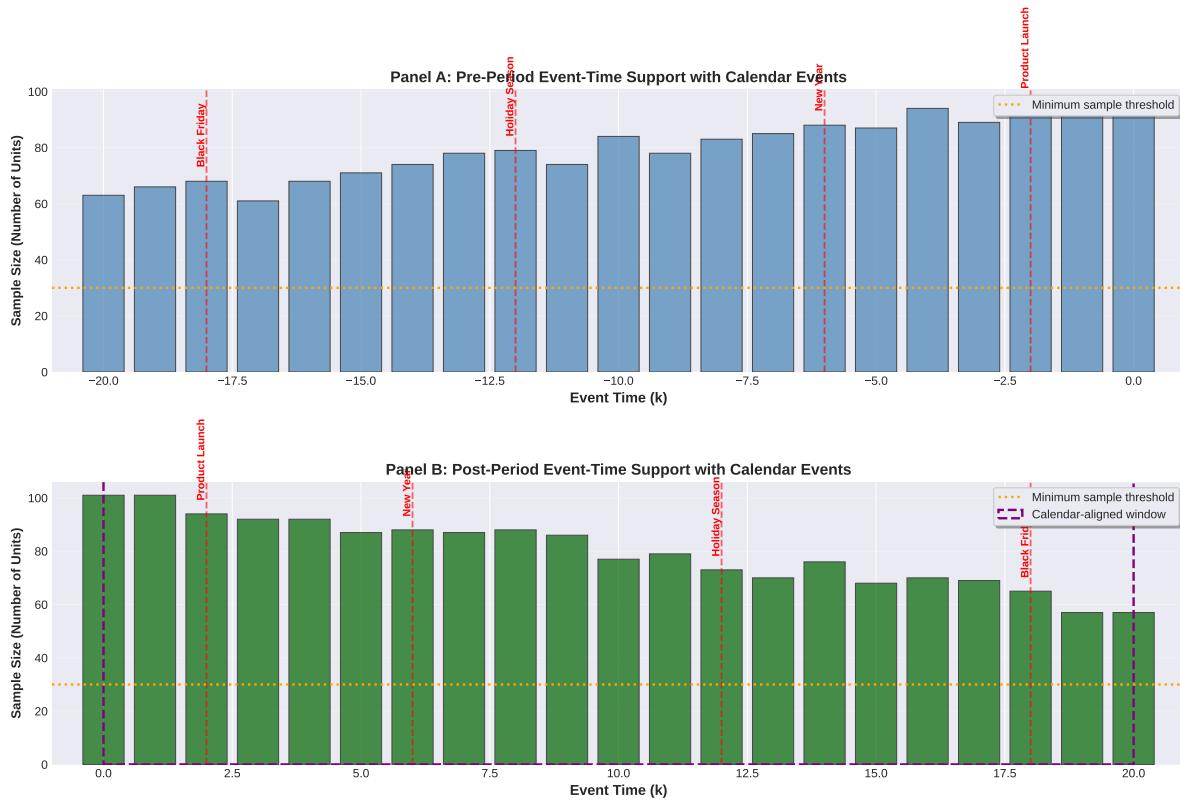
**Report transparently.** Present main and sensitivity results with clear statements of assumptions, threats, diagnostics, and limitations. Include timelines and changepoint annotations. Link estimands to platform metrics where relevant, and be explicit about which bias components remain unaddressed and which sensitivity parameters were varied.



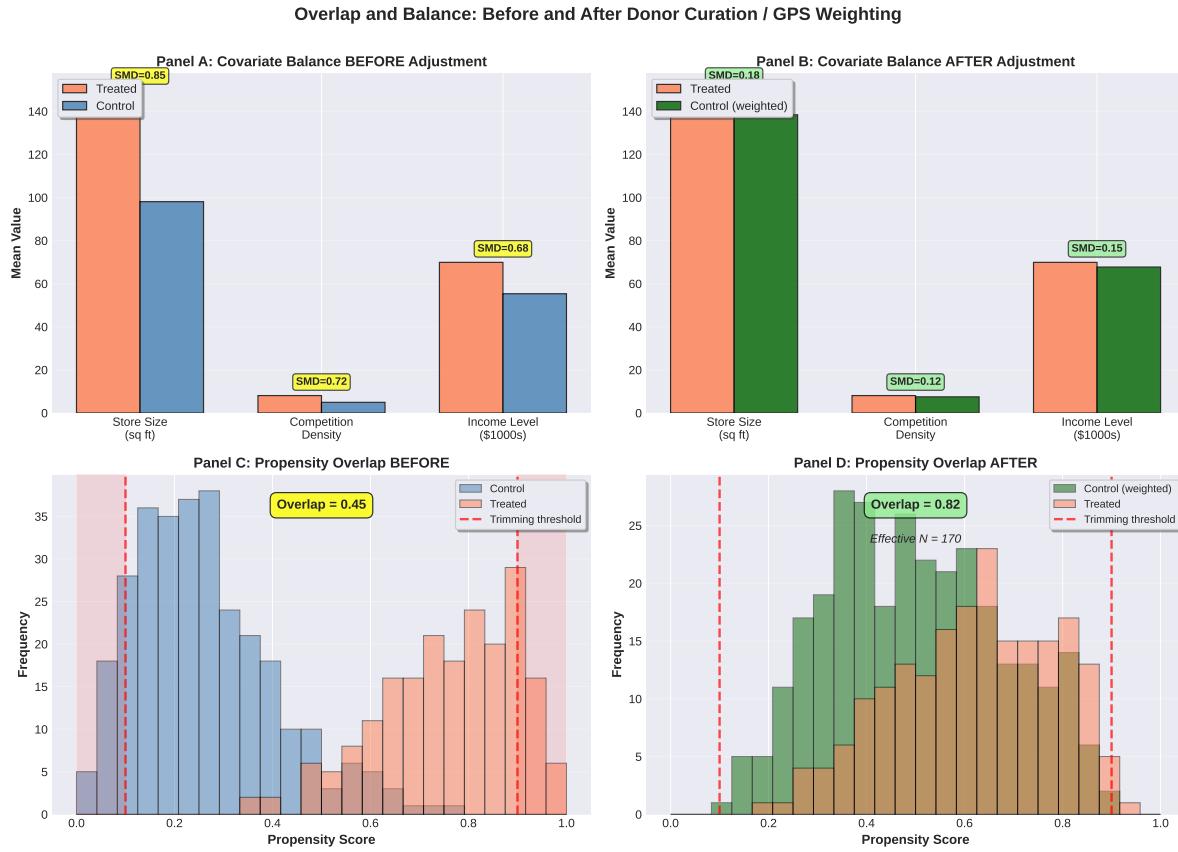
**Fig. 15.1** Changepoint diagnostics around policy or algorithm updates

Panel A shows time series of outcome (sales) and treatment intensity with vertical red lines marking policy change dates (algorithm update at week 40, measurement rule change at week 70). Red shaded regions indicate transition periods to exclude. Panel B displays rolling pre-trend coefficients showing violations near changepoints (orange shaded regions). Panel C presents gap plots (treated minus control differences) revealing discrete shifts at policy changes. Robustness windows are identified for stable estimation. These diagnostics target structural breaks and measurement invariance ( $B_{\text{meas}}$ ). See Section 15.7.

## Event-Time Support and Seasonality Alignment

**Fig. 15.2** Event-time support and seasonality overlay in pre and post windows

Panel A shows pre-period event-time support (sample sizes by relative event time  $k$ ) with calendar events marked by red dashed lines (Black Friday, Holiday Season, New Year, Product Launch). Panel B shows post-period support with aligned calendar events. Orange dotted lines mark minimum sample size thresholds. Purple dashed box highlights the calendar-aligned window ensuring comparable seasonal patterns across treated and control groups. Adequate support at all event times is essential for valid inference. These diagnostics target parallel-trends violations ( $B_{PT}$ ). See Section 15.3.



**Fig. 15.3** Overlap and balance before and after donor curation or GPS weighting

*Top panels show covariate balance (store size, competition density, income level) before (Panel A) and after (Panel B) adjustment. Standardised mean differences (SMDs) improve substantially after weighting (from 0.68–0.85 to 0.12–0.18).*

*Bottom panels show propensity score distributions before (Panel C) and after (Panel D) adjustment. Overlap statistic increases from 0.45 to 0.82. Red dashed lines mark trimming thresholds at 0.1 and 0.9. Effective sample size after trimming is reported. Improved balance and overlap reduce confounding bias ( $B_{\text{confound}}$ ). See Section 15.4.*

**Table 15.1** Threat to assumptions affected, diagnostics, and mitigation strategies

Threat	Assumptions affected	Diagnostics	Mitigation
Calendar/seasonality	Parallel trends, stationarity	Pre-trends, balanced windows, placebo diagnostics	Calendar FE, flexible seasonality, event-time alignment
Platform policy change	Measurement invariance, stability	Changepoint scans, rolling diagnostics	Robustness windows, external outcomes, partial-identification bounds
Algorithmic confounding	Independence, overlap	Balance, GPS overlap, negative controls	Clustered randomisation, orthogonalised ML, switchbacks
Measurement error	Measurement invariance	Validation data, external benchmarks	Bounding, instrumented variation, sensitivity
Spillovers/interference	SUTVA	Spatial diagnostics, donor contamination checks	Buffers, clustered design, spillover models
Nonstationarity/breaks	Stationarity, parallel trends	Gap plots, rolling pre-trends, break diagnostics	Shorten windows, time-varying factors, subsample analysis
Small-G/dependence	Inference validity	Cluster counts, residual autocorrelation	Wild bootstrap, randomisation inference, finite-sample caveats



## Chapter 16

# Inference and Uncertainty Quantification

Point estimates alone cannot guide marketing decisions. When a campaign appears to lift sales by 5%, the critical question is not whether the number is positive but whether the uncertainty around it is small enough to justify action. This chapter develops the statistical machinery for answering that question in panel-data settings, where standard inference methods routinely fail. We treat estimation error and uncertainty around  $\hat{\tau}$ . Identification assumptions from Chapter 15 are taken as given by the design.

We begin with cluster-robust variance estimators and establish when asymptotic approximations suffice. For settings with few clusters, we introduce the wild cluster bootstrap and randomisation inference as finite-sample alternatives. We then develop conformal prediction methods for distribution-free uncertainty quantification and establish protocols for multiplicity corrections and joint confidence bands. The chapter concludes with inference after machine-learning selection and diagnostics for weak instruments, tying together the designs and estimands developed in earlier chapters.

## 16.1 Motivation and Scope

Point estimates tell only half the story. In marketing applications, where decisions involve significant capital allocation, quantifying uncertainty is as critical as estimating the effect size itself. A campaign lift of 5% is actionable if the confidence interval is [4%, 6%], but useless if it is [-2%, 12%].

Many default standard errors assume i.i.d. sampling or, in panel settings, rely on large numbers of independent clusters for their asymptotic justification. In marketing panels, outcomes for the same unit  $i$  are often serially correlated, and units can share common shocks within larger groupings such as markets, platforms, or regions. When dependence may be arbitrary within clusters  $c = 1, \dots, G$ , the effective sample size for asymptotics is driven by the number of independent clusters  $G$ , not the number of unit-time cells  $NT$ . The cluster definition should match the level at which treatment is assigned or as-if randomised. When you ignore this structure, standard errors and confidence intervals can be badly miscalibrated, with large size distortions and poor coverage. Bertrand et al. [2004] show that conventional difference-in-differences standard errors can over-reject the no-effect null at rates exceeding 50% when the nominal size is 5%, producing spurious discoveries.

This chapter builds a hierarchy of inference tools for such settings. Throughout, we treat the identification strategy and target estimand (Chapter 15 and earlier method chapters) as given, and focus on quantifying uncertainty around the resulting estimator. We start with cluster-robust asymptotic approximations when  $G$  is large, move to bootstrap and randomisation methods when  $G$  is small, and introduce mosaic permutation procedures for settings where a local exchangeability approximation is more plausible than independent clusters (Section 16.3). We then discuss inference after high-dimensional selection and related machine-learning steps.

## 16.2 Unified Variance Estimation

Variance estimation in panels with few treated units (for example, synthetic control and geo-experiments) has traditionally been split between regression-based cluster-robust variance estimators and placebo-based methods. Almeida et al. [2025] show that several variance estimators used in regression and synthetic-control-style placebos can be written as weighted averages of squared residuals, differing only in which residuals receive weight. Let  $\hat{Y}_{it}(0)$  denote an estimated counterfactual outcome (from synthetic control or a related method) and define the residual  $\hat{\varepsilon}_{it} = Y_{it} - \hat{Y}_{it}(0)$ . In this section, we form residuals on a control set  $\mathcal{C}$  of untreated unit-time cells. Interpreting  $\hat{\varepsilon}_{it}^2$  as informative about local noise requires that  $\hat{Y}_{it}(0)$  is trained without using post-treatment outcomes for the target cell and that untreated residuals are comparable to treated-cell noise under the design. This framework clarifies which residuals are most informative about the variance at the target cell  $(i^*, t^*)$ . The marginal estimator corresponds to the classical homoskedastic OLS variance, while UP and TP align with placebo-based methods widely used in synthetic control and geo-experiments. For applications to synthetic control, see Chapter 6. For geo-experiments in marketing, see Section 18.3.

**Definition 16.1 (Unified Variance Estimators)** Let  $(i^*, t^*)$  denote the target unit-time pair and let  $\mathcal{C}$  denote a set of untreated control cells used to form residuals. Define  $\hat{\varepsilon}_{it} = Y_{it} - \hat{Y}_{it}(0)$  for  $(i, t) \in \mathcal{C}$ . Write  $\mathcal{C}_{t^*} = \{i : (i, t^*) \in \mathcal{C}\}$  and  $\mathcal{T}_{i^*} = \{t : (i^*, t) \in \mathcal{C}\}$ . We define four variance estimators for a generic estimator  $\hat{\theta}$  of a target effect  $\theta_0$ :

1. **Marginal (M):**  $\hat{V}_M = \frac{1}{|\mathcal{C}|} \sum_{(i,t) \in \mathcal{C}} \hat{\varepsilon}_{it}^2$ , motivated by homoskedasticity ( $\text{Var}(\varepsilon_{it}) = \sigma^2$  for all  $i, t$ ).
2. **Unit-Placebo (UP):**  $\hat{V}_{UP} = \frac{1}{|\mathcal{C}_{t^*}|} \sum_{i \in \mathcal{C}_{t^*}} \hat{\varepsilon}_{it^*}^2$ , motivated by time-specific heteroskedasticity ( $\text{Var}(\varepsilon_{it})$  depends on  $t$ ).
3. **Time-Placebo (TP):**  $\hat{V}_{TP} = \frac{1}{|\mathcal{T}_{i^*}|} \sum_{t \in \mathcal{T}_{i^*}} \hat{\varepsilon}_{i^*t}^2$ , motivated by unit-specific heteroskedasticity ( $\text{Var}(\varepsilon_{it})$  depends on  $i$ ).
4. **Conditional (C):**  $\hat{V}_C = \sum_{(i,t) \in \mathcal{C}} w_{it} \hat{\varepsilon}_{it}^2$ , where weights satisfy  $w_{it} \geq 0$  and  $\sum_{(i,t) \in \mathcal{C}} w_{it} = 1$  and are estimated using only untreated information to target the variance relevant for  $(i^*, t^*)$ . For example, one can model  $\mathbb{E}[\varepsilon_{it}^2 | X_{it}, i, t]$  on untreated cells and then use the fitted values to construct a weighted average. Any machine-learning scheme should use sample-splitting so that the same residuals are not used both to learn weights and to aggregate.

The labels above describe which untreated residuals are treated as informative about the variance at  $(i^*, t^*)$ . Coverage still requires additional conditions, including that the residuals entering the average are comparable for the target problem and sufficiently independent for the inferential approximation used later in the chapter.

**Table 16.1** Unified variance estimators for panel causal effects

Estimator	Definition	Valid Under
Marginal (M)	Average of all squared residuals in control set $\mathcal{C}$	Homoskedasticity
Unit-Placebo (UP)	Average of squared residuals for control units at treated time $t^*$	Time-specific heteroskedasticity
Time-Placebo (TP)	Average of squared residuals for treated unit $i^*$ in control times	Unit-specific heteroskedasticity
Conditional (C)	Weighted average targeting local volatility (requires modelling and sample-splitting)	Both dimensions vary

### Practitioner's Guide: Choosing a Variance Estimator

The inferential target is the sampling variability of your estimator  $\hat{\theta}$ , which is typically driven by the number of independent sampling units. In marketing panels that is often the cluster count  $G$  when dependence may be arbitrary within clusters  $c = 1, \dots, G$ . This section uses untreated residuals  $\hat{\varepsilon}_{it}$  as variance proxies. That logic requires two conditions. First,  $\hat{Y}_{it}(0)$  must be trained without post-treatment leakage for the target problem. Second, untreated residuals must be comparable to the treated setting under the design.

Choose the residual set to match the dominant source of heteroskedasticity. Use **TP** when volatility differs mainly across units or clusters (for example, markets of very different size). Use **UP** when volatility differs mainly across time (for example, holiday weeks). Use **M** only when a homoskedastic approximation is defensible. Use **C** when both dimensions vary, but report the feature set, tuning choices, and the sample-splitting scheme used to learn weights.

The main failure modes are small effective  $G$ , a few influential units or clusters, heavy tails, and non-comparability of untreated residuals due to spillovers or structural breaks. In reporting, state the sampling unit used for inference, the control set  $\mathcal{C}$  used to form residuals, and a sensitivity check across at least two variance estimators.

## 16.3 Mosaic Permutation Tests

Cluster-robust inference assumes independence across clusters (e.g., states). In marketing, spatial spillovers often violate this assumption. Demand shocks propagate across geographic boundaries, and platform interventions spill over to adjacent markets (Section 18.11, Section 18.13). Spector et al. [2025] introduce mosaic permutation procedures, which relax the independence assumption to *local exchangeability*.

**Definition 16.2 (Local Exchangeability)** Partition the panel into blocks  $B_1, \dots, B_K$  (for example, unit-time tiles) and define a neighbourhood system  $\mathcal{N}$  that groups blocks expected to share dependence (for example, adjacent weeks or geographically contiguous markets). Local exchangeability means that, within each neighbourhood, the distribution of the block array is invariant to permutations given the complement. Making this condition operational requires specifying which blocks are permuted and what information is held fixed. This is a weakening of independence. Invariance is required only within local neighbourhoods, and the neighbourhood structure is a modelling decision that should reflect substantive knowledge about how shocks and spillovers propagate.

### Procedure: Mosaic Permutation

To construct a permutation reference distribution for a sharp null of no effect on treated cells (for example,  $H_0 : Y_{it}(1) = Y_{it}(0)$  for all treated  $(i, t)$  under the assumed exposure mapping) without assuming full independence:

1. **Define Resolution:** Divide the panel into a grid (mosaic) of small blocks (unit-time tiles). Define neighbourhoods that reflect the expected structure of dependence (for example, adjacent weeks or geographically contiguous markets).
2. **Obtain Residuals:** Compute residual blocks  $\hat{\varepsilon}_{it} = Y_{it} - \hat{Y}_{it}(0)$  under the null model. If  $\hat{Y}_{it}(0)$  is estimated, train the prediction rule without using the target blocks to be permuted (for example, via sample-splitting).
3. **Restricted Permutation:** Permute residuals only within their local neighbourhood. For each permutation, add the permuted residuals back to the fitted values and re-estimate  $\hat{\theta}$ .
4. **Statistic:** Compute the statistic  $T^*$  on the locally permuted data.
5. **Inference:** Reject  $H_0$  if the observed  $T$  falls in the tails of the mosaic permutation distribution.

Under local exchangeability and an appropriately defined sharp null, and provided the statistic construction respects the relevant permutation invariances, mosaic permutation can deliver finite-sample-valid p-values. When residual blocks are constructed from an estimated model without sample-splitting, treat the resulting permutation distribution as an approximation rather than an exact reference distribution.

*Remark 16.1 (When to Use Mosaic Permutation)* Mosaic permutation is often attractive when:

1. **Cluster boundaries are arbitrary.** Geographic clusters (DMAs, postcodes) rarely align with economic boundaries. Spillovers cross borders.

2. **Few clusters.** With fewer than 20–30 clusters, CRVE-based normal approximations can be unreliable. Mosaic permutation can be a useful alternative when a local exchangeability approximation is defensible for the dependence structure.
3. **Spatial or temporal dependence.** When adjacent units or time periods are correlated (for example, geo-experiments with regional spillovers), local exchangeability is more plausible than full independence.

For standard cluster designs with clear cluster boundaries and no strong cross-cluster spillovers, the wild cluster bootstrap (Section 16.4) is simpler. For geo-experiments (Section 18.3) and transport network analyses (Section 18.9), mosaic permutation often provides more credible inference than naive clustering.

## 16.4 Bootstrap and Resampling

When the number of clusters is small (typically fewer than 50), asymptotic approximations can be poor. The wild cluster bootstrap provides a refinement by resampling the residuals while preserving the cluster structure. This is essential for geo-experiments (Section 18.3) where the number of markets is often 10–40.

**Definition 16.3 (Wild Cluster Bootstrap)** The wild cluster bootstrap for inference on  $H_0 : \theta = \theta_0$  proceeds in four steps:

1. Estimate the restricted model under  $H_0$  and obtain residuals  $\hat{\varepsilon}_{it}$ .
2. For  $b = 1, \dots, B$ , draw independent weights  $w_c^{(b)}$  (Rademacher or Webb-type) for each cluster  $c = 1, \dots, G$ .
3. Construct bootstrap residuals  $\tilde{\varepsilon}_{it}^{(b)} = w_{c(i)}^{(b)} \cdot \hat{\varepsilon}_{it}$  and outcomes  $Y_{it}^{(b)} = \hat{Y}_{it}^{H_0} + \tilde{\varepsilon}_{it}^{(b)}$ , where  $c(i)$  denotes the cluster containing unit  $i$ .
4. Re-estimate the unrestricted model to obtain  $\hat{\theta}^{(b)}$  and compute the bootstrap t-statistic  $t^{(b)}$ .

Here  $c$  indexes clusters (for example, markets), and the same weight  $w_c^{(b)}$  is applied to all observations in cluster  $c$ . The bootstrap p-value is the proportion of bootstrap statistics exceeding the observed statistic:  $\hat{p} = \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{|t^{(b)}| \geq |t^{\text{obs}}|\}$ .

For extremely small numbers of clusters (fewer than 10), the choice of weights matters.

**Definition 16.4 (Webb Six-Point Distribution)** Webb-type six-point weights

$$w_i \in \{-\sqrt{3/2}, -1, -\sqrt{1/2}, \sqrt{1/2}, 1, \sqrt{3/2}\}$$

with equal probabilities 1/6 can improve finite-sample performance relative to Rademacher weights when the number of clusters is very small ( $G < 10$ ) [MacKinnon and Webb, 2017]. The weights satisfy  $\mathbb{E}[w_i] = 0$ ,  $\mathbb{E}[w_i^2] = 1$ , and  $\mathbb{E}[w_i^3] = 0$ , matching moments with the standard normal.

The theoretical justification for the bootstrap is its ability to provide a higher-order approximation to the finite-sample distribution.

**Assumption 78 (Clustered Errors)** Errors are independent across clusters (for example, markets or geo-regions) but may be arbitrarily correlated within a cluster over time. Cluster sizes may differ, and the number of clusters  $G$  grows with the sample. Asymptotics are in the number of independent clusters  $G$ , not the number of unit-time cells.

**Theorem 16.1 (Asymptotic Refinement)** *Under Assumption 78 and additional conditions that ensure the bootstrap reproduces the distribution of a studentised statistic, the wild cluster bootstrap can provide asymptotic refinement [Cameron et al., 2008]:*

1. *The bootstrap distribution of  $t^{(b)}$  consistently estimates the finite-sample distribution of  $t^{\text{obs}}$ .*
2. *Coverage error of bootstrap confidence intervals can be reduced from order  $G^{-1/2}$  to  $G^{-1}$ .*

For  $G < 10$ , randomisation inference is preferred (Section 16.5). Simulation evidence suggests that when  $G \geq 10$ , rejection rates are close to nominal levels  $\alpha$ .

In time series settings without clear cluster definitions, block bootstrapping captures serial dependence.

**Definition 16.5 (Moving Block Bootstrap)** For time series with serial dependence, the moving block bootstrap captures correlation by resampling blocks of data. Choose a block length  $\ell$  and form overlapping blocks  $B_j = (Y_j, \dots, Y_{j+\ell-1})$ . A common heuristic sets  $\ell$  on the order of  $T^{1/3}$  [Künsch, 1989]. Draw blocks with replacement and concatenate them to form a bootstrap sample of length  $T$ .

Block length trades bias (longer blocks capture more dependence) against variance (shorter blocks allow more resampling variability). Block-length choice can materially affect coverage, and practitioners should check robustness to different block lengths. In panels with both cross-sectional and serial dependence, block bootstrap alone is not a full solution. Cluster bootstrap or mosaic permutation (Section 16.3) may be required. Block bootstrap is particularly relevant for media mix modelling (Section 18.5) where weekly or monthly data exhibit strong serial correlation.

*Remark 16.2 (Inference Method Selection)* The choice of inference method depends on the number of clusters  $G$ :

1.  $G \geq 50$ : CRVE-based normal approximations are often adequate, but sensitivity to influential clusters and cluster imbalance remains important.
2.  $10 \leq G < 50$ : Compute CRVEs but base p-values and confidence intervals on the wild cluster bootstrap distribution.
3.  $G < 10$ : Randomisation inference (Section 16.5) provides exact p-values, provided the assignment mechanism satisfies Assumption 79.
4. **Time series without clusters:** Block bootstrap or HAC standard errors.

In marketing applications with geographic units,  $G$  is often in the 10–40 range, making the wild cluster bootstrap, with carefully chosen weights and studentised statistics, the default choice. For very small  $G$  or highly structured experiments, randomisation-based inference that replays the actual assignment mechanism remains the most credible option.

## 16.5 Randomisation Inference

For design-based causal inference, particularly with synthetic control (Chapter 6) or small- $N$  difference-in-differences, randomisation inference (RI) can deliver exact p-values without relying on large-sample asymptotics. The logic is design-based. If we replay the same assignment mechanism under a sharp null, how unusual is the observed statistic relative to its randomisation distribution?

**Definition 16.6 (Fisher Randomisation Test)** Let  $\mathcal{W}$  be the set of feasible treatment assignment matrices  $\mathbf{D}$  under the experimental design, and let  $\mathbf{D}^{\text{obs}}$  be the realised assignment. The randomisation p-value for test statistic  $T(\mathbf{Y}, \mathbf{D})$  is

$$p_{\text{rand}} = \frac{1}{|\mathcal{W}|} \sum_{\mathbf{d} \in \mathcal{W}} \mathbf{1}\{|T(\mathbf{Y}, \mathbf{d})| \geq |T(\mathbf{Y}, \mathbf{D}^{\text{obs}})|\},$$

For simplicity we write  $Y_i(1), Y_i(0)$  for an aggregated outcome (for example, a post-period mean). In panels, you should state the aggregation window and define the statistic as a functional of  $\{Y_{it}(\cdot)\}_t$  over that window. Under the sharp null  $H_0 : Y_i(1) = Y_i(0)$  for all  $i$  (or the corresponding panel sharp null over the aggregation window), the potential outcomes are fixed and only assignment varies. Randomisation inference is exact for sharp nulls. With heterogeneous effects, the test remains a falsification check for “no effects anywhere”, not a direct test of an average effect such as ATT.

Under a sharp null and a correctly encoded assignment mechanism, RI delivers exact finite-sample size control.

**Assumption 79 (Known assignment mechanism for randomisation inference)** The set of feasible treatment assignments  $\mathcal{W}$  is known and correctly specified. Permutations respect the design (geo-stratification, switchback periods, partial interference structure). Assignment is unconfounded within the permutation class. Violations of this requirement—for example, ignoring capacity constraints or algorithmic throttling in a switchback—invalidate exactness. If platform logic or capacity constraints break the assumed randomisation, resulting p-values may be anti-conservative even if the test is implemented correctly. Many marketing designs are only approximately random. RI is exact only to the extent that the design can be encoded in  $\mathcal{W}$ .

**Theorem 16.2 (Exact Size Control)** *Under Assumption 79 (known assignment mechanism) and the sharp null hypothesis:*

1. *The randomisation p-value is exactly uniformly distributed:  $P(p_{\text{rand}} \leq \alpha) = \alpha$  for all  $\alpha \in [0, 1]$ .*
2. *No distributional assumptions on outcomes are required.*
3. *The test controls size exactly in finite samples for any  $G$  or  $N$ .*

*When  $|\mathcal{W}|$  is large, Monte Carlo approximation with  $B$  random draws yields  $\hat{p}_{\text{rand}}$  with simulation error  $O(B^{-1/2})$ .*

The power of the test depends on the choice of statistic.

**Table 16.2** Test statistics for randomisation inference

Statistic	Formula	Properties
Difference in means	$T_{\text{diff}} = \bar{Y}_{\text{treat}} - \bar{Y}_{\text{control}}$	Unstudentised, simple but less powerful
t-statistic	$T_t = (\bar{Y}_{\text{treat}} - \bar{Y}_{\text{control}})/\hat{s}_e$	Studentised, better power under heterogeneity
Rank statistic	$T_{\text{rank}} = \sum_{i:D_i=1} R_i$	Robust to outliers and heavy-tailed or ordinal outcomes, distribution-free

**Definition 16.7 (Choice of Test Statistic)** For randomisation inference, Table 16.2 summarises common test statistics and their properties.

*Remark 16.3 (Randomisation Inference in Marketing Applications)* Randomisation inference is often most useful for:

1. **Synthetic control** (Chapter 6): Permute treatment across donor units to generate placebo distribution.
2. **Geo-experiments with few markets** (Section 18.3): When  $G < 10$ , RI provides exact inference where bootstrap fails.
3. **Switchback experiments** (Section 18.11): Permute treatment periods within the design structure.
4. **Single-unit interventions**: When only one unit is treated, classical asymptotic inference is not available. RI is often the most defensible route to p-values when a credible assignment mechanism can be encoded.

The key requirement is that the permutation class  $\mathcal{W}$  correctly reflects the experimental design.

## 16.6 Conformal and Distribution-Free Methods

When our goal is prediction, such as constructing counterfactual trajectories in synthetic control (Chapter 6), conformal inference provides rigorous finite-sample *marginal* coverage guarantees for prediction intervals under exchangeability, without parametric distributional assumptions [Shafer and Vovk, 2008, Lei et al., 2018].

**Assumption 80 (Exchangeability)** The sequence of observations  $(X_1, Y_1), \dots, (X_n, Y_n), (X_{\text{new}}, Y_{\text{new}})$  is exchangeable, meaning its joint distribution is invariant to permutations of indices. This holds for i.i.d. data. Extensions exist for some dependent settings (for example, martingale conformal methods), but we focus on the standard i.i.d./exchangeable case.

**Definition 16.8 (Conformity Score)** For a prediction model  $\hat{\mu}(X)$  and observation  $(X_i, Y_i)$ , the conformity score measures how well the observation conforms to the model:

$$s_i = |Y_i - \hat{\mu}(X_i)|,$$

or more generally  $s_i = \mathcal{S}(X_i, Y_i, \hat{\mu})$  for some score function  $\mathcal{S}$ . Low scores indicate good conformity. High scores indicate the observation is unusual relative to the model.

We use these scores to calibrate the uncertainty of future predictions.

**Definition 16.9 (Split Conformal Prediction)** Split conformal prediction constructs prediction intervals by splitting the data into a training set  $\mathcal{I}_{\text{train}}$  and a calibration set  $\mathcal{I}_{\text{cal}}$ . First, fit the model  $\hat{\mu}$  on  $\mathcal{I}_{\text{train}}$ . Then, compute conformity scores  $s_i = |Y_i - \hat{\mu}(X_i)|$  for all  $i \in \mathcal{I}_{\text{cal}}$ .

Find the  $(1 - \alpha)(1 + 1/|\mathcal{I}_{\text{cal}}|)$ -quantile of the calibration scores, denoted  $\hat{q}$ . For a new point  $X_{\text{new}}$ , the prediction interval is  $[\hat{\mu}(X_{\text{new}}) - \hat{q}, \hat{\mu}(X_{\text{new}}) + \hat{q}]$ .

This construction guarantees marginal coverage at level  $1 - \alpha$  under Assumption 80.

**Theorem 16.3 (Marginal Coverage)** *Under Assumption 80 (exchangeability of the calibration sample and the new observation), the conformal prediction interval satisfies*

$$P(Y_{\text{new}} \in \hat{C}(X_{\text{new}})) \geq 1 - \alpha,$$

for any sample size. The probability is over the joint distribution of calibration data and new observation. The coverage is marginal (averaging over  $X_{\text{new}}$ ) rather than conditional on  $X_{\text{new}}$ . No distributional assumptions beyond exchangeability are required. This distribution-free guarantee is the defining feature of conformal prediction.

Exchangeability is most natural when the calibration data and new points are drawn from the same population in an i.i.d. fashion. In panel settings, exchangeability typically fails because units differ systematically and outcomes exhibit serial correlation. If you use conformal anyway, you should define an exchangeability unit (for example, donor units rather than time periods), ensure calibration does not use post-treatment

information for the target unit, and report sensitivity to alternative calibration sets. Conformal methods are therefore best suited to problems where you can construct approximately exchangeable calibration sets, for example across donor units in synthetic control or across stores in a forecasting application. The resulting guarantee, when it is credible, is marginal over that population rather than conditional on a fixed unit's characteristics.

*Remark 16.4 (Conformal Inference in Marketing)* Conformal methods are valuable in marketing when:

1. **Synthetic control counterfactuals** (Chapter 6): Constructing prediction intervals for the counterfactual trajectory  $\hat{Y}_{i^*t}(0)$  without assuming normality. Calibration is across donor units or placebo periods, which are treated as approximately exchangeable.
2. **Demand forecasting with uncertainty**: When point predictions inform decisions (pricing, inventory), conformal intervals provide marginal coverage guarantees under exchangeability.
3. **Heterogeneous treatment effects**: Conformal methods can provide prediction intervals for *model-based* CATE predictions. These are uncertainty statements about the predictor's future performance, not a substitute for identification of a causal estimand.

The key limitation is that coverage is *marginal* (averaging over prediction points), not conditional. In settings with substantial heterogeneity, intervals may be too wide for some observations and too narrow for others. Conditional conformal methods exist but require additional assumptions. When applying conformal methods to heterogeneous treatment effects or CATE models, one typically conformalises residuals from the CATE predictor and interprets the resulting intervals as marginal coverage for future units drawn from the same covariate distribution, not as exact uncertainty bands for a given realised unit.

## 16.7 Aggregated Estimands and Joint Inference

Often we are interested in functions of parameters, such as cumulative effects or dynamic paths. This is essential for event studies (Chapter 5) and dynamic treatment effect estimation where we report trajectories rather than single coefficients.

In clustered panels, the effective sample size for asymptotics is typically the number of independent clusters  $G$  rather than the total number of unit-time observations. We therefore write asymptotics in  $\sqrt{G}$  throughout.

**Theorem 16.4 (Delta Method for Variance Propagation)** *Let  $\hat{\theta} \in \mathbb{R}^k$  be an asymptotically normal estimator with  $\sqrt{G}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ , where  $G$  is the number of independent clusters used for inference. For a differentiable function  $g : \mathbb{R}^k \rightarrow \mathbb{R}^m$ ,*

$$\sqrt{G}(g(\hat{\theta}) - g(\theta_0)) \xrightarrow{d} \mathcal{N}(0, J\Sigma J'),$$

where  $J = \nabla g(\theta_0)$  is the  $m \times k$  Jacobian matrix (derivatives of outputs with respect to inputs). For event-time aggregation, write  $\theta_k = \sum_g w_g \tau(g, g+k)$ , where  $\tau(g, t)$  is the cohort-time estimand and  $\mathbf{w}$  is the vector of cohort weights. The variance formula

$$\text{Var}(\hat{\theta}_k) = \mathbf{w}' \hat{\Sigma}_\tau \mathbf{w}$$

is understood with  $\hat{\Sigma}_\tau$  as the estimated cluster-robust covariance matrix of the stacked  $\hat{\tau}(g, g+k)$ . The cohort-time effects  $\tau(g, g+k)$  correspond to the dynamic estimands developed in Chapters 5 and 4.

When examining an entire trajectory of effects, pointwise confidence intervals can be misleading. If we report 20 pointwise 95% intervals, we should expect at least one spurious excursion outside the band by chance. Joint confidence bands solve this.

**Definition 16.10 (Simultaneous Confidence Band)** For vector of estimators  $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_K)'$  with covariance  $\hat{\Sigma}$ , a  $(1 - \alpha)$  simultaneous confidence band is:

$$\mathcal{C}_{1-\alpha} = \{(\theta_1, \dots, \theta_K) : \max_k |\hat{\theta}_k - \theta_k| / \hat{s}\hat{e}_k \leq c_{1-\alpha}\},$$

where  $\hat{s}\hat{e}_k = \sqrt{\hat{\Sigma}_{kk}}$  is the cluster-robust standard error from the diagonal of  $\hat{\Sigma}$ , and  $c_{1-\alpha}$  is chosen so that  $P(\theta_0 \in \mathcal{C}_{1-\alpha}) = 1 - \alpha$  jointly for all  $k$ . Equivalently, pointwise intervals are  $[\hat{\theta}_k \pm c_{1-\alpha} \cdot \hat{s}\hat{e}_k]$ .

**Proposition 16.1 (Constructing Uniform Bands)** *The critical value  $c_{1-\alpha}$  for simultaneous coverage can be obtained by the methods in Table 16.3.*

Bootstrap-based bands are often tight when correlation among  $\hat{\theta}_k$  is substantial, which is common in event-study settings where adjacent time periods are correlated. In high-correlation settings such as event studies, Bonferroni bands are often overly conservative because they ignore cross-time correlation. Bootstrap-based bands that resample cluster-level influence functions, using the wild cluster bootstrap from Section 16.4, can be tighter while still controlling the maximum statistic when the resampling scheme respects the dependence structure and the statistic is appropriately studentised [Montiel Olea and Plagborg-Møller, 2019].

**Table 16.3** Methods for constructing uniform confidence bands

Method	Critical Value	Properties
Bonferroni	$c_{1-\alpha}^{\text{Bonf}} = z_{1-\alpha}/(2K)$	Conservative but simple, ignores correlation
Scheffé	$c_{1-\alpha}^{\text{Sch}} = \sqrt{\chi^2_{K,1-\alpha}}$	Simultaneous coverage for linear combinations under an approximate multivariate normal limit
Bootstrap	(1 - $\alpha$ )-quantile of $\max_k  t_k^{(b)} $	Can be tight and accounts for correlation when resampling respects dependence

### 16.7.1 Inference for Partially Identified Bounds

In some designs the target effect is not point identified but belongs to an identified set characterised by linear constraints. Synthetic Parallel Trends is a leading example [Liu, 2025]. It defines the counterfactual post-treatment trend as a weighted average of control trends for any convex weights that match pre-trends and then treats all such weights as admissible.

The treatment effect is then bounded by the minimum and maximum values attainable over this weight set. Inference must therefore deliver a confidence set that covers the entire identified set with high probability rather than a single point.

Operationally this is achieved by combining the linear programming representation of the identified set with resampling. Moments that summarise pre-trends and post-trends are estimated at a rate determined by the number of independent sampling units, which in panel applications is often the cluster count  $G$  rather than the number of unit-time cells. The mapping from those moments to the identified set is directionally differentiable but not linear, so asymptotic theory relies on tools for functionals of this type (Hadamard directional derivatives rather than standard delta-method linearisation) [van der Vaart, 2000]. The standard delta method fails because the functional is non-smooth, and ignoring this leads to coverage distortions. Specialised inference procedures are required. See Manski [2003] for foundational treatment.

In practice we invert the decision rule over candidate values of the treatment effect, using bootstrap critical values for the underlying moment statistics, and collect all values that are not rejected. The resulting interval is often wider than a conventional confidence interval but it remains valid under a much weaker set of assumptions and makes clear how much of the uncertainty comes from disagreement across admissible weighting schemes.

*Remark 16.5 (Joint Inference in Marketing Event Studies)* Joint confidence bands are essential for marketing applications:

1. **Pre-trend diagnostics:** When assessing parallel trends (Section 17.4), joint bands control the family-wise error rate across all pre-treatment periods. Using pointwise pre-trend diagnostics at 5% without a joint correction recreates the multiple-comparisons problem that uniform bands are designed to address.
2. **Dynamic effect visualization:** Event-study plots (Figure 18.10) should show uniform bands, not pointwise intervals, to avoid overstating significance.

3. **Cumulative effects:** When aggregating effects over time (e.g., total incremental revenue from a campaign), the delta method propagates uncertainty correctly.

For marketing applications in Chapter 18, bootstrap-based uniform bands are often a sensible default for event-study visualisations when the resampling scheme matches the dependence structure.

## 16.8 Inference after Selection and ML

Using machine learning to select controls or nuisance parameters introduces additional uncertainty. We use Double Machine Learning (DML) to separate the selection step from the inference step. This section covers inference for the DML estimators introduced in Chapter 12.

**Theorem 16.5 (Double/Debiased ML Inference)** *Under Assumptions in Chapter 12 (orthogonality, rate conditions, cross-fitting), the DML estimator in clustered panels satisfies:*

$$\sqrt{G}(\hat{\theta}^{DML} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V),$$

where clusters are indexed by  $c = 1, \dots, G$ . Let  $O_{it} = (Y_{it}, D_{it}, X_{it})$  denote the observed data for unit  $i$  in period  $t$ , and let  $\psi(O_{it}, \theta, \eta)$  denote the unit-time score contribution for the orthogonal moment defining  $\theta_0$ . The cluster-level influence contribution is  $\Psi_c(\theta, \eta) = \sum_{i \in C_c} \sum_t \psi(O_{it}, \theta, \eta)$ . The cluster-robust variance estimator is:

$$\hat{V} = \frac{1}{G} \sum_{c=1}^G \Psi_c(\hat{\theta}, \hat{\eta})^2,$$

aggregating squared cluster-level scores across folds.

This formulation aligns DML inference with the broader cluster-robust framework of this chapter. DML corrects for regularisation and selection effects in estimating nuisance components, but does not fix violations of unconfoundedness or weak-overlap problems. Those remain identification threats, not estimation ones.

Recent work on direct debiased machine learning shows that these orthogonal scores can be constructed systematically through a Riesz representer estimated by Bregman-Riesz regression [Kato, 2025b]. In average treatment effect settings the Riesz representer coincides with a density-ratio between treated and control covariate distributions, and Riesz regression becomes a particular direct density-ratio estimator [Kato, 2025a].

Treating the weight function as a density-ratio links the stability of estimated weights directly to the finite-sample behaviour of the influence-function estimator. Weight diagnostics therefore play two roles. They can reveal overlap problems, which are identification threats. They can also reveal unstable influence-function estimation, which is a finite-sample inference threat. The interpretation depends on whether weight instability reflects thin support or model and regularisation artefacts.

**Proposition 16.2 (Variance from Cross-Fitting Randomness)** *Let  $\hat{\theta}^{(r)}$  denote the DML estimate from the  $r$ -th random partition into folds,  $r = 1, \dots, R$ . A pragmatic variance decomposition used to assess split sensitivity is [Chernozhukov et al., 2018, Fuhr and Papies, 2024]:*

$$\bar{\theta} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)}, \quad \hat{V}_{total} = \frac{1}{R} \sum_{r=1}^R \hat{V}^{(r)} + \frac{1}{R-1} \sum_{r=1}^R (\hat{\theta}^{(r)} - \bar{\theta})^2,$$

where the first term is the average within-split variance and the second term is the between-split variance capturing randomness from cross-fitting.

Increasing  $R$  reduces split sensitivity at the cost of higher computation. Moderate repetition, for example tens of splits, can reduce split sensitivity, but the appropriate  $R$  depends on computation budget and estimator stability.

*Remark 16.6 (DML Inference in Marketing Applications)* DML inference is relevant for marketing when:

1. **CATE estimation** (Section 12.5): When using ML to estimate heterogeneous treatment effects across customer segments, DML provides valid confidence intervals for segment-level effects.
2. **High-dimensional controls** (Chapter 13): When many covariates are available (purchase history, demographics, browsing behaviour), ML-based control selection requires DML for valid inference. This includes both the regularisation chapter's double-selection setups and the DML chapter's flexible nuisance models. DML inference is the common influence-function layer on top of either approach.
3. **Propensity score estimation**: When the propensity score is estimated with flexible ML methods, DML ensures that uncertainty in the propensity model is properly accounted for.

The key practical concern is weight stability: if estimated propensity weights are extreme, influence function variance inflates and confidence intervals become unreliable. Diagnostics for weight dispersion (Section 17.3) should accompany all DML analyses. In all these applications, diagnostic plots for the estimated weights—propensity scores, inverse-probability weights, or Riesz weights—should be treated as first-class inference diagnostics, on par with balance and overlap checks, because they reveal whether the influence-function approximation will behave well in finite samples.

## 16.9 Multiplicity and Multiple Testing

When you assess many hypotheses—such as effects on multiple outcomes or across multiple subgroups—you must control the error rate for the family of decisions. This arises frequently in marketing when analysing heterogeneous treatment effects (Chapter 12) or assessing effects across customer segments. In event-study settings, the natural “family” is the collection of pre-period coefficients (for assessing parallel trends) or the entire dynamic trajectory. The joint bands from Section 16.7 provide a convenient way to implement FWER control over time.

**Definition 16.11 (FWER Control)** For a family of  $m$  null hypotheses  $H_1, \dots, H_m$ , the familywise error rate is:

$$\text{FWER} = P(\text{reject at least one true } H_j).$$

A procedure controls FWER at level  $\alpha$  if  $\text{FWER} \leq \alpha$  regardless of which hypotheses are true. Bonferroni correction rejects  $H_j$  if  $p_j \leq \alpha/m$ . Holm’s stepdown [Holm, 1979] rejects  $H_{(j)}$  (ordered by p-value) if  $p_{(j)} \leq \alpha/(m - j + 1)$ .

Sometimes controlling the proportion of mistakes is more appropriate than avoiding any mistake.

**Definition 16.12 (FDR Control)** The false discovery rate is the expected proportion of false rejections among all rejections:

$$\text{FDR} = \mathbb{E} \left[ \frac{V}{R \vee 1} \right],$$

where  $V$  is the number of false rejections and  $R$  is the total number of rejections. The Benjamini–Hochberg procedure [Benjamini and Hochberg, 1995]:

1. Order p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$ .
2. Find largest  $k$  such that  $p_{(k)} \leq k\alpha/m$ .
3. Reject  $H_{(1)}, \dots, H_{(k)}$ .

This controls FDR at level  $\alpha$  under independence and under certain positive dependence conditions, often summarised as positive regression dependence on a subset (PRDS). In marketing panels with strong common shocks, those conditions may fail. In that case, Benjamini–Hochberg is best framed as exploratory unless its dependence assumptions are defended.

For correlated comparisons, we can gain power by using resampling.

**Definition 16.13 (Resampling-Based Stepdown)** The Romano–Wolf procedure [Romano and Wolf, 2005] accounts for dependence among studentised statistics:

1. Compute observed studentised statistics  $t_1, \dots, t_m$  and p-values  $p_1, \dots, p_m$ .
2. Generate bootstrap samples  $b = 1, \dots, B$  using the same dependence-respecting scheme used for base inference, for example the wild cluster bootstrap over clusters  $c = 1, \dots, G$  or mosaic permutation when local exchangeability is assumed.
3. For each  $b$ , compute  $t_j^{(b)}$  under the null (centring at null values).

4. At step  $s$ , let  $\mathcal{H}_s$  denote the current set of unrejected hypotheses. Compute the adjusted p-value for  $H_{(s)}$  as the fraction of bootstrap samples where  $\max_{j \in \mathcal{H}_s} |t_j^{(b)}| \geq |t_{(s)}|$ .
5. Reject if adjusted p-value  $\leq \alpha$ . Remove the rejected hypothesis and repeat.

This provides FWER control while exploiting correlation to improve power over Bonferroni. The bootstrap resampling in step 2 must respect the panel's dependence structure (clustered or mosaic resampling, as discussed in Section 16.4). In practice, we use the same cluster-level (or mosaic) resampling scheme as in Section 16.4, and the max  $|t|$  distribution is exactly the object used for the bootstrap-based joint bands in Section 16.7.

*Remark 16.7 (Choosing Between FWER and FDR Control)* The choice between FWER and FDR depends on the research context. FWER control (Bonferroni, Holm, Romano-Wolf) is appropriate when any false positive is costly, such as launching an ineffective campaign based on spurious significance. It suits confirmatory analysis with a small number of pre-specified hypotheses and business or regulatory decision implications. FDR control (Benjamini-Hochberg) is appropriate for exploratory screening of many hypotheses, such as identifying which customer segments respond to a campaign. It tolerates some false positives in exchange for greater power in large-scale testing where Bonferroni would reject nothing.

A mixed strategy is often appropriate, but it requires pre-specifying primary estimands and families of hypotheses. Otherwise, multiplicity adjustments become post hoc and hard to interpret. Use FWER-controlled bands for pre-trend checks and primary outcomes where false positives have direct business consequences. Use FDR-controlled procedures for exploratory subgroup searches where the goal is hypothesis generation.

*Remark 16.8 (Multiple Testing in Marketing)* Common marketing scenarios requiring multiplicity adjustment:

1. **Subgroup analysis:** Assessing treatment effects across customer segments (high-value, new, churned) inflates false positive rates without correction.
2. **Multiple outcomes:** Assessing effects on revenue, retention, engagement, and satisfaction simultaneously.
3. **Multiple time periods:** Event-study coefficients across many pre- and post-periods create a family of decisions. Rather than applying Bonferroni or Holm mechanically to each time-period  $t$ -statistic, it is often more natural to use the joint-band machinery from Section 16.7, which directly controls the maximum  $t$ -statistic across  $k$  and yields uniform confidence bands for the whole path.
4. **A/B experiment variants:** Comparing multiple treatment arms against control.

Romano-Wolf is preferred when statistics are correlated (e.g., outcomes measured on the same customers), as it exploits the correlation structure for tighter inference.

## 16.10 Instrumental Variables in Marketing

Instrumental variables (IV) methods address *endogeneity*, situations where the treatment variable is correlated with unobserved determinants of the outcome. In marketing, endogeneity arises frequently. Prices respond to demand, budgets respond to sales, and targeting algorithms select high-propensity users. When selection-on-observables is implausible and no credible design based on randomisation or parallel trends is available, IV can sometimes identify a causal parameter under strong additional assumptions.

*Remark 16.9 (Scope of IV Coverage)* This section provides conceptual foundations and marketing-specific guidance for instrumental variables. We do not develop the full IV estimation theory, which requires substantial treatment of weak instruments, many instruments, and specification diagnostics. For comprehensive coverage, see Angrist and Pischke [2009] for applied foundations, Andrews et al. [2019] for weak instrument inference, and Londschein [2025] for a recent practitioner-oriented survey. The methods in Chapter 4 through Chapter 6 (DiD, synthetic control) often provide more credible identification in marketing than IV, where exclusion restrictions are difficult to defend.

### When IV Is Needed

IV addresses situations where naive regression fails due to simultaneity or omitted variable bias.

**Definition 16.14 (Endogeneity)** In the structural equation  $Y_i = \alpha + \beta D_i + U_i$ , endogeneity occurs when:

$$\text{Cov}(D_i, U_i) \neq 0.$$

The treatment  $D$  is correlated with unobserved factors  $U$  that also affect the outcome  $Y$ . OLS estimation of  $\beta$  is biased. Here  $\beta$  plays the role of a generic causal effect parameter, analogous to  $\theta_0$  elsewhere in the book. We use  $\beta$  in this section to match the IV literature.

In panels we typically work with  $Y_{it}$  and  $D_{it}$  rather than a single cross-sectional  $Y_i$  and  $D_i$ . The endogeneity condition  $\text{Cov}(D_i, U_i) \neq 0$  should therefore be read as a shorthand for correlation between treatment paths and unobserved determinants of outcomes—often through dynamic feedback and common shocks.

In marketing, endogeneity arises from three sources:

1. **Simultaneity.** Prices are set based on expected demand, and budgets are allocated based on expected returns. The outcome (demand, sales) determines the treatment (price, spend).
2. **Omitted variables.** Unobserved brand strength, market conditions, or consumer preferences affect both marketing decisions and outcomes.
3. **Measurement error.** If the treatment is measured with error, the measurement error is correlated with the regressor, biasing estimates toward zero. When measurement error in  $D$  is classical and a validation-style instrument is available, IV corrects attenuation bias as in the measurement-error framework of Section 15.5.

## The IV Solution

An instrument  $Z$  provides exogenous variation in the treatment  $D$  that is unrelated to the outcome except through  $D$ .

**Definition 16.15 (Instrumental Variable)** A variable  $Z$  is a valid instrument for the effect of  $D$  on  $Y$  if:

1. **Relevance:**  $\text{Cov}(Z, D) \neq 0$ . The instrument is correlated with the treatment.
2. **Exclusion restriction:** the instrument has no direct effect on the outcome and affects the outcome only through its effect on  $D$ . In linear additive models this is often represented as  $\text{Cov}(Z, U) = 0$ , but the substantive claim is a restriction on the causal pathway, not a generic covariance condition.

In panel settings with covariates and fixed effects, the relevance and exclusion conditions are understood conditional on controls and fixed effects, rather than in the unconditional scalar form. In potential-outcomes terms, a common exogeneity condition is

$$Z_{it} \perp \{Y_{it}(d) : d \in \mathcal{D}\} \mid X_{it}, \alpha_i, \lambda_t.$$

Two-stage least squares (2SLS) uses the instrument to isolate exogenous variation in  $D$ :

$$\text{First stage: } \hat{D}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i + \hat{\nu}_i, \tag{16.1}$$

$$\text{Second stage: } Y_i = \alpha + \beta \hat{D}_i + \varepsilon_i. \tag{16.2}$$

The 2SLS estimator  $\hat{\beta}_{2SLS}$  is consistent if both relevance and exclusion hold. In panel applications, the same two-stage logic is applied to within- or FE-transformed regressions with cluster-robust inference (as in Section 16.4) rather than to raw cross-sectional  $i$ -level regressions.

## Marketing Instruments: Examples and Pitfalls

Table 16.4 summarises commonly proposed instruments in marketing and the challenges each faces.

*Remark 16.10 (The Exclusion Restriction Is Rarely Credible in Marketing)* The exclusion restriction requires that the instrument affects the outcome *only* through the treatment. In marketing, this is difficult to defend:

1. **Cost shifters.** Input costs (commodities, shipping) affect prices, but they may also affect product quality, retailer effort, or consumer expectations.
2. **Hausman instruments.** Using prices in other markets as instruments assumes no common demand shocks—implausible for national brands or correlated economic conditions.
3. **Ad pre-emptions.** Major events that pre-empt advertising may directly affect consumer attention, mood, or category salience.
4. **Competitor instruments.** Competitor actions are often strategic responses to the same market conditions affecting the focal firm.

**Table 16.4** Instruments in marketing: examples and exclusion concerns

Instrument	Application	Exclusion Concern
Cost shifters (fuel, commodities)	Price elasticity	Costs may signal quality or affect demand directly
Hausman instruments (prices in other markets)	Price elasticity	Common demand shocks across markets
Ad pre-emptions (sports events, breaking news)	Advertising effects	Events may directly affect consumer mood or attention
Competitor actions (entry, exit, promotions)	Own-brand effects	Strategic response may correlate with demand
Weather	Retail traffic, ad exposure	Weather affects both exposure and purchase propensity
Algorithmic lags (delayed price updates)	Dynamic pricing	Lags may correlate with demand persistence

*Note:* Many of these proposals are better framed as natural experiments or DiD designs rather than as traditional IV, when the timing or geographic variation permits.

Unlike randomised experiments or parallel-trends designs where assumptions can be partially probed with diagnostics, the exclusion restriction is *fundamentally untestable*. Overidentification diagnostics (Hansen J) can detect some violations but have low power and cannot validate the exclusion restriction.

## Key Landmines

Beyond the exclusion restriction, IV estimation faces several practical challenges.

**Weak instruments.** When the first-stage relationship is weak ( $F < 10$ ), 2SLS estimates are biased toward OLS and have unreliable standard errors. See Section 16.11 for diagnostics and weak-instrument-robust inference.

**Many instruments.** Using many instruments (e.g., many cost shifters, many market dummies) can bias 2SLS toward OLS even when each instrument is individually strong. The bias is proportional to the number of instruments relative to sample size.

**Local average treatment effect (LATE).** IV identifies the effect for *compliers*—units whose treatment status changes with the instrument. Interpreting IV estimates as a LATE additionally requires (i) independence of the instrument from potential outcomes, (ii) the exclusion restriction, and (iii) a monotonicity assumption ruling out “defiers” [see Angrist and Pischke, 2009, Chapter 4]. Monotonicity can be dubious in marketing settings where, for example, price increases cause some customers to buy *more* through signalling or stockpiling behaviours. The LATE may differ from the average treatment effect (ATE) if treatment effects are heterogeneous. In marketing, the complier population (for example, customers whose purchases respond to cost-driven price changes) may not be the population of interest. In panels, compliers may be a tiny and

peculiar subset of units—for example, marginal price-sensitive shoppers in one region—which often makes the LATE less relevant for broad marketing policy decisions.

**Exclusion violations.** Even small violations of the exclusion restriction can produce large biases in IV estimates. Sensitivity analysis [for example, Conley et al., 2012, who treat the exclusion restriction as approximately rather than exactly true and derive bounds] can assess robustness to plausible violations, but this is rarely done in marketing applications.

## When to Use IV in Marketing

Given the stringent requirements, IV is often a method of last resort in marketing because exclusion restrictions are hard to defend. It can still be appropriate when the instrument's causal pathway is unusually clear and the first stage is strong:

Box 16.3: IV Decision Checklist

**Before using IV, verify:**

1. **Randomisation is impossible.** No experiment (geo, A/B, or switchback) can address the question.
2. **Parallel trends fail.** DiD or synthetic control are not credible due to differential trends or confounding shocks.
3. **Selection-on-observables fails.** Propensity score methods require conditioning on all confounders, which is implausible.
4. **A credible instrument exists.** You can articulate why the instrument satisfies both relevance and exclusion.
5. **First stage is strong.** The effective F-statistic exceeds critical values (see Section 16.11). All IV results should be reported with weak-IV-robust intervals (for example, Anderson–Rubin or conditional likelihood ratio), not just conventional 2SLS confidence intervals.
6. **LATE is interpretable.** The complier population is economically meaningful for the business question.

If any condition fails, reconsider the identification strategy. In many cases, an honest acknowledgement of limited identification and reliance on design-based panel methods (DiD, synthetic control, factor models) is preferable to IV with a dubious exclusion restriction.

### 16.10.1 Connection to Chapter 18 Applications

Several marketing applications in Chapter 18 use or discuss IV. Price elasticity estimation (Section 18.7) employs cost shifters and promotion timing as instruments. Dynamic pricing applications (Section 18.9)

consider fuel costs, competitor capacity, and algorithmic lags. Digital attribution (Section 18.4) treats IV as an alternative when natural experiments are unavailable. CLV attribution (Section 18.8) uses channel assignment instruments for selection correction. In each case, IV is presented as one of several candidate designs, and readers should prioritise designs with testable assumptions (DiD, synthetic control) where available. The tables in Chapter 18 note the exclusion restriction required and the common concerns. Analysts should approach these applications with appropriate scepticism about instrument validity.

## 16.11 Weak Instrument Diagnostics

When using instrumental variables (Section 16.10), weak identification occurs when the instrument is only weakly correlated with the treatment, rendering standard inference unreliable. This section provides diagnostic tools for detecting weak instruments and robust inference methods when weakness is suspected.

We focus on diagnostic tools that are robust to heteroskedasticity and clustering, recognising that in panel data the relevant notion of ‘first-stage strength’ is not captured by the textbook homoskedastic  $F$  statistic. In robust settings, practitioners rely on weak-identification diagnostics such as the Kleibergen–Paap rk statistic [Kleibergen and Paap, 2006] and on effective  $F$  statistics with heteroskedasticity- and cluster-robust calibration [Montiel Olea and Pflueger, 2019].

**Definition 16.16 (Kleibergen–Paap rk Statistic)** For IV regression with multiple endogenous variables and clustering, the Kleibergen–Paap rk statistic provides a rank diagnostic for the first-stage coefficient matrix [Kleibergen and Paap, 2006]. For a single endogenous regressor, practitioners often summarise strength via an effective  $F$  statistic constructed using a robust covariance for first-stage coefficients. One convenient form is:

$$F_{\text{eff}} = \frac{\hat{\pi}' \hat{V}_{\pi}^{-1} \hat{\pi}}{k_z},$$

where  $\hat{\pi}$  is the first-stage coefficient vector,  $\hat{V}_{\pi}$  is a heteroskedasticity- and cluster-robust covariance matrix for  $\hat{\pi}$  consistent with the clustering structure used in the outcome equation, and  $k_z$  is the number of instruments. Stock–Yogo critical values apply under homoskedasticity [Stock and Yogo, 2005]. For robust and clustered settings, Montiel Olea–Pflueger’s effective  $F$  and associated critical values are the appropriate reference [Montiel Olea and Pflueger, 2019].

When instruments are weak, we base inference on procedures that remain valid under weak identification.

**Definition 16.17 (Anderson–Rubin Inference)** Anderson–Rubin inference for  $H_0 : \beta = \beta_0$  in IV regression is based on the reduced-form statistic:

$$\text{AR}(\beta_0) = \frac{(Y - X\beta_0)' P_Z (Y - X\beta_0)/k_z}{(Y - X\beta_0)' M_Z (Y - X\beta_0)/(n - k_z - k_x)},$$

where  $P_Z = Z(Z'Z)^{-1}Z'$  is the projection onto instruments and  $M_Z = I - P_Z$ . This formula is the classical homoskedastic version. The  $F(k_z, n - k_z - k_x)$  reference distribution does not carry over to clustered panels. In clustered panels, the AR statistic is typically studentised using a cluster-robust residual covariance and calibrated via wild bootstrap or other robust resampling. Size control comes from these robust critical values, not from the classical  $F$ . Anderson–Rubin confidence sets are obtained by inverting the Anderson–Rubin procedure with cluster-robust or bootstrap-based critical values, preserving correct size even when instruments are weak.

**Implementation Note.** Critical values for weak-instrument diagnostics and weak-IV inference are available from several standard sources. The **Kleibergen–Paap rk statistic** (implemented in standard software

such as `ivreg2` in Stata, `ivreghdfe`, and equivalents in R and Python) reports the rk statistic with robust errors, typically compared against Stock-Yogo (2005) critical values for homoskedastic designs. For robust effective  $F$ -statistics with size-corrected thresholds, practitioners should consult the **Montiel Olea-Pflueger (OP)** effective  $F$ , available via packages such as `weakiv` and its equivalents. The Stock-Yogo tables remain the standard reference for maximal size distortion and relative bias under homoskedasticity.

In clustered panels, define  $k_z$  as the number of excluded instruments and  $k_x$  as the number of included controls in the relevant first-stage regression. The symbol  $n$  in the classical AR display denotes the number of observations in that regression. In this chapter, the effective sample size for inference is typically the number of independent clusters  $G$ . That is why AR critical values should be obtained from a dependence-respecting resampling scheme rather than from a homoskedastic  $F$  reference distribution.

*Remark 16.11 (Weak Instruments and Marketing Applications)* Many marketing instruments are weak in practice. Cost shifters may explain only a small fraction of price variation. Ad pre-emptions are infrequent. While the classical rule-of-thumb is  $F > 10$ , in robust settings practitioners should compare the effective  $F$  to Montiel Olea-Pflueger critical values rather than relying on a universal threshold. When the effective  $F$ -statistic falls below the appropriate critical value, do not rely on 2SLS point estimates. Options include:

1. **Report reduced-form effects only.** The effect of  $Z$  on  $Y$  is identified even when instruments are weak.
2. **Use Anderson-Rubin confidence intervals.** These have correct coverage regardless of instrument strength.
3. **Reconsider the identification strategy.** When both exclusion and strength are questionable, reduced-form designs (DiD, synthetic control) or partial-identification bounds are generally preferable to relying on 2SLS at all.

For the applications in Chapter 18, Tables 18.9, 18.6, and 18.11 note exclusion concerns alongside each proposed instrument. These concerns compound when instruments are also weak.

## Chapter 17

# Design and Diagnostics

Credible causal claims in marketing panels rest on two pillars: designs that make identification assumptions plausible and diagnostics that make those assumptions transparent. Diagnostics cannot rescue a bad design. They can, however, reveal when key implications fail and when conclusions hinge on a small number of independent clusters  $c = 1, \dots, G$  rather than on the naive  $NT$  intuition. This chapter develops a disciplined diagnostic workflow that links design choices to identification, estimation, and inference. We show how to select and curate control and donor groups for DiD, event studies, and synthetic control, how to assess overlap, balance, pre-trends, and placebo behaviour, how to use influence and leave-one-out sensitivity and weight-dispersion diagnostics and specification curves to detect fragile designs, and how to implement sensitivity analyses for parallel-trends violations, unobserved confounding, and measurement shifts. The goal is a practical playbook you can layer on top of any design in the book, making assumptions explicit and stability visible.

## 17.1 Motivation and Scope

Credible causal claims rest on good design and disciplined diagnostics. Start by stating an estimand and the assignment mechanism meant to identify it. Diagnostics then probe observable implications of the identifying assumptions. They can reveal fragile designs, but they cannot certify identification. In marketing panels, dependence is the rule rather than the exception, so you should interpret diagnostics with the sampling unit in mind. Uncertainty often hinges on a small number of independent clusters  $c = 1, \dots, G$ , even when you have many unit-time observations.

We assemble a practical playbook that applies across many methods in this book, emphasising pre-specification of diagnostics, triangulation across designs, and clear guidance on what to do when diagnostics raise concerns. Chapter 15 catalogues threats to validity and links them to identification assumptions. Chapter 16 develops the corresponding inference tools. Here we turn those ideas into a concrete diagnostic process for DiD and event studies, synthetic-control-style imputation, factor-based imputation, and DML-based estimators. For DML, we treat nuisance-model training, cross-fitting, and overlap checks as part of the diagnostic protocol and we flag the risks from post-treatment tuning and information leakage.

## 17.2 Control Selection and Donor Curation

The credibility of any causal estimate depends on the quality of the comparison group. For difference-in-differences and event studies, control units should be screened on pre-treatment outcomes and covariates under a pre-specified rule to make parallel trends (Chapter 4) plausible. Adaptive screening followed by standard inference is a form of selection that can overstate precision. Exclusions must also respect the identification logic of the design and must not introduce post-treatment controls or colliders. Large pre-treatment differences or divergent trends are red flags that may require reconsidering the control group or the design itself.

For synthetic control methods (Chapter 6), this process is called donor curation. Units are excluded from the donor pool when they are contaminated by spillovers from the treated unit (Chapter 11), affected by concurrent policies, or subject to measurement shifts. In a regional pricing experiment by a grocery retailer, stores in adjacent markets may experience spillovers through competitive price-matching algorithms or consumer cross-shopping. Such stores belong to the same interference set and should be excluded from donor pools to avoid SUTVA violations. Stores undergoing concurrent renovations or loyalty-programme launches cannot serve as clean controls. Documenting these exclusion rules is essential for transparent reporting. Regularisation can operationalise donor selection, but you should still report donor-weight concentration and leave-one-out sensitivity so that dependence on particular donors is explicit.

Factor-space coverage (Chapter 8) requires that donors span the latent structure governing untreated outcomes. Conceptually, if treated and donor units load on different factors, imputation fails. In practice, we use pre-period fit metrics and rank-selection criteria as diagnostics. Poor pre-period fit is consistent with inadequate factor coverage, but it can also reflect noise, misspecification, over-regularisation, or instability.

**Definition 17.1 (Factor Space Coverage)** Let  $\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{iR})'$  denote the factor loadings for unit  $i$  under the model in Chapter 8. The donor pool  $\mathcal{J}$  provides adequate coverage if:

$$\min_{\mathbf{w} \in \Delta^{|\mathcal{J}|}} \left\| \boldsymbol{\lambda}_{i^*} - \sum_{j \in \mathcal{J}} w_j \boldsymbol{\lambda}_j \right\| < \epsilon,$$

for tolerance  $\epsilon$ . Pre-period fit error is consistent with coverage failure, but it can also reflect noise, misspecification, or instability.

*Remark 17.1 (Systematic Donor Curation)* Donor exclusions should be documented systematically. Begin by excluding units plausibly exposed under a pre-specified exposure mapping, including a defensible buffer choice, and report sensitivity to alternative mappings or buffers. Next, remove units experiencing concurrent treatments during the estimation window, since their outcomes reflect multiple interventions. Data quality issues or measurement coverage changes warrant further exclusions. Units with fundamentally different factor loadings can also be problematic, but pre-period fit should be treated as a diagnostic warning rather than a proof of non-coverage. Report the initial donor universe, exclusion criteria, and remaining donor count. Sensitivity analysis over donor exclusions and buffer choices should be part of the multiverse and specification-curve exercises later in this chapter.

When coverage is inadequate, hybrid methods that blend factors with synthetic control weights or augment with high-dimensional controls can improve fit (Chapter 7).

### 17.3 Overlap and Balance Diagnostics

Overlap is a necessary condition for making treated and control units comparable without heavy extrapolation. If they do not share common support on key covariates, any comparison will rely on extrapolation beyond support. In panel settings, propensity scores and overlap diagnostics should be computed at the level of the design, such as stores, markets, or cohorts, and conditional on whatever fixed effects or factors are maintained. When treatment is assigned at the unit or cohort level, define  $X_i$  as a pre-treatment summary or vector and estimate  $e(X_i) = \mathbb{P}(D_i = 1 | X_i)$  over an explicit pre-period window. When treatment varies at the unit-time level, use the cell-level score  $e(X_{it}) = \mathbb{P}(D_{it} = 1 | X_{it})$ . We assess overlap by plotting the distribution of propensity scores (Chapter 12) for treated and control groups.

**Definition 17.2 (Overlap Statistics)** For propensity scores  $e(X_i)$  in treated and control groups, overlap is quantified by two statistics. The coefficient of overlap is  $OVL = \int \min\{f_1(e), f_0(e)\} de$ , where  $f_1(e)$  and  $f_0(e)$  are the propensity score densities for treated and control groups. The common support fraction is  $N_{\text{common}}/N$ , where  $N_{\text{common}}$  counts units with  $e(X_i) \in [\epsilon, 1 - \epsilon]$  for threshold  $\epsilon$ .

**Definition 17.3 (Common Support)** Units  $i$  satisfy the common support condition if:

$$0 < e(X_i) < 1, \quad \text{or more stringently} \quad \epsilon < e(X_i) < 1 - \epsilon,$$

for trimming threshold  $\epsilon > 0$ . The common support set is:

$$\mathcal{S}_\epsilon = \{i : \epsilon \leq e(X_i) \leq 1 - \epsilon\}.$$

Estimates restricted to  $\mathcal{S}_\epsilon$  reduce the impact of extreme weights, but they also change the estimand. When OVL or the common-support fraction is low, trimming to  $\mathcal{S}_\epsilon$  and reporting how estimates change across plausible thresholds connects overlap diagnostics to the partial-identification and sensitivity analyses in Chapter 15. In such cases, be explicit that the target estimand is  $\text{ATT}_{\text{overlap}}$  or  $\text{ATE}_{\text{overlap}}$  on the overlap region rather than on the full population.

Where there is no overlap, we may need to trim the sample or re-weight our observations. For the continuous treatments discussed in Chapter 14, overlap concerns the generalised propensity score  $r(d | X_{it}, \alpha_i, \lambda_t)$  and the support of  $D_{it}$  given the conditioning set.

Exposure overlap in spillover-aware designs (Chapter 11) maps the distribution of neighbours' treatment intensities. Buffers and exclusion zones reduce extreme exposures, and sensitivity to buffer radii assesses robustness. Stabilised weights or inverse-probability-of-treatment weights can improve balance, but extreme weights signal fragility and invite either trimming or clustered assignment to reduce dependence.

Covariate balance reports means and standardised differences for key covariates before and after weighting or selection.

**Definition 17.4 (Standardised Mean Difference)** The standardised mean difference (SMD) for covariate  $X_j$  between treated and control groups is:

$$\text{SMD}_j = \frac{\bar{X}_{j,1} - \bar{X}_{j,0}}{\sqrt{(s_{j,1}^2 + s_{j,0}^2)/2}},$$

where  $\bar{X}_{j,d}$  and  $s_{j,d}^2$  are the mean and variance of covariate  $j$  in group  $d \in \{0,1\}$ . Here  $D_i$  denotes the treatment indicator at the same assignment level as  $e(X_i)$ . After weighting with IPW weights  $\hat{\omega}_i$ :

$$\text{SMD}_j^{\text{weighted}} = \frac{\sum_i \hat{\omega}_i D_i X_{ij} / \sum_i \hat{\omega}_i D_i - \sum_i \hat{\omega}_i (1 - D_i) X_{ij} / \sum_i \hat{\omega}_i (1 - D_i)}{\sqrt{(s_{j,1}^2 + s_{j,0}^2)/2}}.$$

As a heuristic, analysts often view  $|\text{SMD}|$  below 0.1 as small. Treat thresholds as context-dependent diagnostics rather than pass/fail rules. After weighting, one can also compute weighted variances to form variance-ratio diagnostics separately. Both SMDs and variance ratios matter for assessing balance.

*Remark 17.2 (Balance Improvement from Weighting)* For inverse propensity weighted estimators with weights  $\hat{\omega}_i = D_i/\hat{e}(X_i) + (1 - D_i)/(1 - \hat{e}(X_i))$ , if  $\hat{e}(X) = e(X)$  is correctly specified, then  $\text{SMD}_j^{\text{weighted}} \xrightarrow{P} 0$  for all covariates  $j$ . The variance ratio  $\text{VR}_j = s_{j,1}^2 / s_{j,0}^2$  should approach 1 after weighting. Good balance on covariates is necessary but not sufficient for unconfoundedness (Chapter 2). In marketing panels, unobserved demand shocks and platform signals often remain, so balance should be interpreted as design-check evidence rather than proof of causal validity. Report  $\text{SMD}_j$  before and after weighting for all key covariates.

Pre-period outcome balance is especially diagnostic because outcomes summarise many unobserved factors. Residualising outcomes against covariates or fixed effects and then checking balance on residuals aligns with the design and connects directly to pre-trend diagnostics (Chapter 4) and the synthetic parallel trends ideas in Chapter 6. Balance improvement after propensity weighting or double selection (Chapters 12–13) shows that the adjustment aligns observed covariates with the design’s target. It is suggestive evidence, not a guarantee, about confounding.

## 17.4 Pre-Trends and Placebo Designs

Pre-trend diagnostics are a central probe of the parallel-trends assumption (Chapter 4), but they have limited power and should not be treated as a pass/fail gate. In an event study (Chapter 5), pre-treatment leads should be small relative to a pre-specified tolerance, which you can assess using joint bands, equivalence-style diagnostics, and joint Wald statistics. If you proceed only when pre-trends “look good,” you are conditioning on a random variable and can induce pre-test bias (Definition 17.20).

**Definition 17.5 (Joint Pre-Trend Diagnostic)** For event-study coefficients  $\hat{\theta}_k$  at relative event times  $k < 0$  (pre-treatment), the joint Wald test statistic is:

$$W = \hat{\theta}_{\text{pre}}' \hat{\Sigma}_{\text{pre}}^{-1} \hat{\theta}_{\text{pre}},$$

where  $\hat{\theta}_{\text{pre}} = (\hat{\theta}_{-K}, \dots, \hat{\theta}_{-1})'$  is the vector of pre-treatment coefficients and  $\hat{\Sigma}_{\text{pre}}$  is their estimated covariance matrix. Under  $H_0$ : parallel trends (all  $\theta_k = 0$  for  $k < 0$ ):

$$W \xrightarrow{d} \chi^2(K),$$

where  $K$  is the number of pre-treatment periods tested. The F-version is  $F = W/K$ .

Because pre-treatment coefficients are estimated with clustered dependence and often with relatively few clusters, the  $\chi^2(K)$  reference for  $W$  should be interpreted through cluster-robust or bootstrap-based critical values rather than through textbook homoskedastic approximations. In practice, pre-trend diagnostics, TOST equivalence checks, and joint bands for pre-period coefficients should all be implemented using the clustered and resampling tools from Chapter 16.

### Procedure: TOST Pre-Trend Equivalence

Rather than failing to reject the null of no pre-trends, equivalence testing assesses whether pre-trends are practically negligible. First, specify an equivalence bound  $\delta$  (e.g., choose  $\delta$  in business-relevant units such as a small percentage of baseline sales). Second, test  $H_{01} : \theta_k \leq -\delta$  versus  $H_{A1} : \theta_k > -\delta$ . Third, test  $H_{02} : \theta_k \geq \delta$  versus  $H_{A2} : \theta_k < \delta$ . Conclude equivalence if both nulls are rejected at level  $\alpha$ . This provides positive evidence for parallel trends rather than merely failing to find evidence against. When testing equivalence across multiple pre-periods, use joint-band or multiple-testing adjustments (Section 16.7 and Section 16.9) rather than interpreting unadjusted per-period equivalence tests in isolation (add citation).

For synthetic control methods (Chapter 6), we use placebo-style diagnostics. An *in-space* placebo applies the method to each donor unit in turn, generating a distribution of placebo effects to which we compare the main estimate. This yields a permutation-style reference distribution. It is exact only under strong exchangeability assumptions, otherwise treat it as a diagnostic benchmark rather than a literal randomisation test. An *in-time* placebo uses a pseudo-treatment date in the pre-treatment period to check whether the

method produces spurious effects and guards against design misuse over time. In evaluating a television advertising campaign's effect on regional sales, for example, in-space placebos would apply synthetic control to untreated regions, while in-time placebos might use a pseudo-treatment date six months before the campaign launch to detect whether the method spuriously identifies effects when none should exist.

**Definition 17.6 (Root Mean Squared Prediction Error)** For synthetic control with treated unit  $i^*$  and pre-treatment periods  $t = 1, \dots, T_0$ :

$$\text{RMSPE}_{\text{pre}} = \sqrt{\frac{1}{T_0} \sum_{t=1}^{T_0} (Y_{i^*t} - \hat{Y}_{i^*t}^{\text{SC}})^2},$$

where  $\hat{Y}_{i^*t}^{\text{SC}} = \sum_{j \in \mathcal{J}} \hat{w}_j Y_{jt}$  is the synthetic control prediction and  $\mathcal{J}$  denotes the donor pool. For post-treatment periods  $t = T_0 + 1, \dots, T$ :

$$\text{RMSPE}_{\text{post}} = \sqrt{\frac{1}{T - T_0} \sum_{t=T_0+1}^T (Y_{i^*t} - \hat{Y}_{i^*t}^{\text{SC}})^2}.$$

The RMSPE ratio  $\text{RMSPE}_{\text{post}}/\text{RMSPE}_{\text{pre}}$  measures the magnitude of post-treatment deviation relative to pre-treatment fit.

**Definition 17.7 (Placebo Inference)** For in-space placebos, apply synthetic control to each donor unit  $j \in \mathcal{J}$  as if it were treated. Define the placebo score  $R_j = \text{RMSPE}_{\text{post},j}/\text{RMSPE}_{\text{pre},j}$  and the treated score  $R_{i^*} = \text{RMSPE}_{\text{post},i^*}/\text{RMSPE}_{\text{pre},i^*}$ . The rank-based placebo p-value is:

$$p_{\text{placebo}} = \frac{1 + \sum_{j \in \mathcal{J}} \mathbf{1}\{R_j \geq R_{i^*}\}}{|\mathcal{J}| + 1},$$

the fraction of placebo units with scores at least as large as the treated unit, with a finite-sample adjustment that avoids zero p-values. If pre-treatment fit is poor (large  $\text{RMSPE}_{\text{pre},j}$ ), exclude unit  $j$  from the placebo distribution or report sensitivity to the inclusion threshold.

Negative controls use outcomes or covariates that should not respond to treatment under the maintained hypothesis. We distinguish *negative-control outcomes*—variables that should not respond to treatment—from *negative-control exposures or groups*—units or treatments that should not be affected. In marketing, platform metrics or seemingly unaffected geographies often serve in these roles. If negative controls show spurious effects, unobserved confounding or measurement error is likely. Negative controls are especially useful for detecting platform algorithm changes or data quality shifts that might otherwise masquerade as treatment effects.

## 17.5 Support, Exposure, and Contamination

We must verify that we have adequate data and support at each point in the analysis. For event studies (Chapter 5), this means tabulating the number of units and cohorts that contribute to each event-time bin and checking that treatment and control units overlap in the relevant covariate space. If only a few observations contribute to a particular bin, or if they come from highly atypical units, the estimate for that period will be noisy and unstable, and may be highly sensitive to modelling and weighting choices. Thin support at high  $|k|$  increases reliance on functional-form and regularisation choices, which can behave like extrapolation in finite samples. Treat far-from-zero  $k$  as descriptive unless support is strong. The solution is often to bin sparse periods together or to restrict attention to a window where support is strong, trading temporal precision for statistical power and representativeness.

Exposure mapping (Definition 11.3 in Chapter 11) defines treatment intensity for unit  $i$  as a function of own treatment and an exposure mapping  $h_i(D_{-i,t})$  that summarises neighbours' treatments, such as  $\{D_{jt} : j \in \mathcal{N}_i\}$ . Common specifications include distance-based radii (exposure decays with geographic distance), network adjacency (exposure proportional to treated neighbours), and market-overlap matrices (exposure proportional to competitive overlap). In practice, the exposure mapping should be chosen and justified before seeing post-treatment outcomes as much as possible, to avoid post hoc mapping choices that overfit the observed spillover pattern.

Buffers physically separate treated and control units by excluding units within a threshold distance (Chapter 11). Re-estimating with varying buffer radii assesses sensitivity to plausible spillover pathways, recognising that buffers also change the analysis population and can shift the estimand. Treat buffers as a form of trimming and report how results change across defensible buffer choices, consistent with the overlap and trimming discussion in Section 17.3. When spillovers are expected but unmeasured, diagnostics for partial interference and design adjustments are necessary.

Contamination arises when control units receive indirect treatment through competitive responses, supply-chain linkages, or word-of-mouth. In a retail promotion study, for example, control stores may experience contamination if customers shift purchases from promoted stores to nearby untreated locations, or if competitors respond with defensive price-matching. Diagnostic checks include examining outcome changes in controls coincident with treatment rollout, inspecting pre-post trends in donor pools for synthetic control and SDID (Chapters 6–7), and checking geographic or network proximity. When contamination is detected, donor redesign, buffer inclusion, or switching to a spillover-aware estimand (Chapter 11) can reduce the risk of bias. Units that show clear outcome breaks coincident with treatment rollout, despite being nominally control, may warrant scrutiny for reclassification into an interference set or exclusion under a documented rule. Such redesign decisions should be documented explicitly and folded into the specification-curve and sensitivity analyses later in this chapter, rather than treated as ad hoc fixes.

## 17.6 Influence and Stability

Leave-one-out diagnostics detect observations that disproportionately influence estimates. By re-estimating the effect while excluding one unit, time period, or cohort at a time, we identify influential observations. If results change dramatically when a single element is excluded, the findings may be fragile.

**Definition 17.8 (Leave-One-Out Diagnostic)** Let  $\hat{\tau}$  be the treatment effect estimate using all data and  $\hat{\tau}_{(-i)}$  the estimate excluding unit  $i$ . The leave-one-out influence is:

$$\text{DFBETA}_i = \hat{\tau} - \hat{\tau}_{(-i)}.$$

For leave-one-cohort-out in staggered designs with cohort  $g$ :

$$\text{DFBETA}_g = \hat{\tau} - \hat{\tau}_{(-g)}.$$

Units or cohorts with large  $|\text{DFBETA}|$  are influential. Report the influence profile showing  $\hat{\tau}_{(-i)}$  for all  $i$ .

Because panel estimators are built from influence-function contributions, these leave-one-out diagnostics should be interpreted alongside the influence aggregation used in cluster-robust inference (Chapter 16). In clustered panels, prioritise leave-one-cluster-out diagnostics and report the full influence profile rather than relying on universal thresholds. Large  $|\text{DFBETA}_i|$  is a warning that the design is fragile and that conclusions may hinge on idiosyncratic behaviour of particular markets or cohorts, not an automatic licence to drop observations until results stabilise.

**Proposition 17.1 (Standardised Influence)** *To compare influence across units on a common scale, standardise by the estimated variance:*

$$C_i = \frac{(\hat{\tau} - \hat{\tau}_{(-i)})^2}{\hat{V}(\hat{\tau})},$$

analogous to Cook's distance in regression. Treat any cut-off as a heuristic. In clustered panels, what matters is whether a small number of clusters dominate the influence-function sum, so leave-one-cluster-out diagnostics are often more informative than leave-one-unit-out summaries. For donor influence in synthetic control:

$$C_j^{\text{donor}} = \frac{(\hat{\tau} - \hat{\tau}_{(-j)}^{\text{SC}})^2}{\hat{V}(\hat{\tau})},$$

where  $\hat{\tau}_{(-j)}^{\text{SC}}$  re-estimates synthetic control excluding donor  $j$  from the pool.

Weight dispersion and leverage diagnostics inspect the concentration of synthetic control or SDID (Chapter 7) weights and time weights. If one or two donors receive nearly all weight, the counterfactual is fragile to those donors' idiosyncrasies. In evaluating a brand's market exit using synthetic control, for instance, if a single donor region receives 70% of the weight, the counterfactual depends almost entirely on that region's idiosyncratic shocks—a competitor's local promotion or a regional economic downturn could masquerade as an effect of the exit. Entropy or Herfindahl indices quantify such concentration.

**Definition 17.9 (Donor Weight Concentration)** For synthetic control weights  $\hat{w} = (\hat{w}_j)_{j \in \mathcal{J}}$  over donor pool  $\mathcal{J}$ , concentration is measured by three statistics. The Herfindahl-Hirschman Index is  $\text{HHI} = \sum_{j \in \mathcal{J}} \hat{w}_j^2 \in [1/|\mathcal{J}|, 1]$ , with  $1/|\mathcal{J}|$  indicating equal weights (minimum concentration) and 1 indicating all weight on one donor. Entropy is  $H = -\sum_{j \in \mathcal{J}} \hat{w}_j \log \hat{w}_j \in [0, \log |\mathcal{J}|]$ , with high entropy indicating dispersed weights. The effective number of donors is  $N_{\text{eff}}^{\text{donors}} = 1/\text{HHI}$ , the number of equally weighted donors that would produce the same HHI. Low  $N_{\text{eff}}^{\text{donors}}$  is a warning sign, especially when the donor pool is large and heterogeneous. The effective number should be interpreted relative to the size and diversity of the donor pool.

In SDID and related methods (Chapter 7), time weights and cohort weights also concentrate and should be examined jointly with donor weights. An effective number of donors close to one in a large, heterogeneous donor pool suggests that the counterfactual is effectively a single-donor model. Whether this is problematic depends on substantive similarity between that donor and the treated unit; in all cases, donor redesign and perturbation checks should make the dependence on particular donors transparent.

Visualising weights over time reveals whether time weights are stable or shift abruptly, signalling potential structural breaks. Regularisation via ridge or elastic net penalties can spread weights more evenly, though at the cost of potentially worse pre-period fit. Donor redesign—excluding high-weight donors and re-estimating—is a robustness diagnostic for dependence on individual donor influence.

Perturbation checks vary estimation windows, donor sets, or tuning parameters within defensible bounds and report the range of estimates. Tight ranges are consistent with stability within the chosen robustness set.

**Definition 17.10 (Perturbation Interval)** For a set of defensible design variations  $\mathcal{V}$  (e.g., varying window lengths, donor sets, trimming thresholds), the perturbation interval is:

$$\mathcal{I}_{\text{pert}} = \left[ \min_{v \in \mathcal{V}} \hat{\tau}_v, \max_{v \in \mathcal{V}} \hat{\tau}_v \right].$$

The perturbation ratio is:

$$R_{\text{pert}} = \frac{\max_{v \in \mathcal{V}} \hat{\tau}_v - \min_{v \in \mathcal{V}} \hat{\tau}_v}{\widehat{\text{SE}}(\hat{\tau}_{\text{primary}})},$$

where  $\hat{\tau}_{\text{primary}}$  is the pre-specified primary estimate. Treat  $R_{\text{pert}}$  as a descriptive stability diagnostic rather than a rule with a universal threshold. The perturbation interval  $\mathcal{I}_{\text{pert}}$  is descriptive and should be reported alongside confidence intervals, not used to adjust them. A small  $R_{\text{pert}}$  and narrow inference bands both point to robustness. A small  $R_{\text{pert}}$  but wide bands suggests imprecision. A large  $R_{\text{pert}}$  suggests design-driven fragility.

Wide ranges or sign changes signal that conclusions depend sensitively on design choices, and additional triangulation or sensitivity reporting is warranted.

## 17.7 Specification Curves and the Multiverse

Specification curves offer a systematic way to summarise robustness. By varying defensible design choices—control group, time windows, covariate sets—and plotting the distribution of resulting estimates, we can see whether conclusions hinge on a single specification. In evaluating promotional lift across retail chains, for instance, the specification family might vary the control group (non-promoted chains versus non-promoted regions within promoted chains), the time window (weekly versus monthly aggregation), and the covariate set (with or without weather controls). If promotional lift estimates range from 3% to 8% across all defensible specifications and nearly all estimates share the same sign, that pattern is suggestive that conclusions are not driven by a single modelling choice. It does not, by itself, deliver valid inference.

**Definition 17.11 (Specification Family and Curve)** A specification family  $\mathcal{S}$  is a set of defensible estimation specifications, each defined by choices of control group or donor pool, covariate adjustment set, time window, functional form and regularisation, and trimming or weighting scheme. The specification curve plots estimates  $\{\hat{\tau}_s : s \in \mathcal{S}\}$  sorted by magnitude, along with indicators of which choices each specification embodies.

**Definition 17.12 (Curve Summary Statistics)** For specification family  $\mathcal{S}$  with  $|\mathcal{S}| = S$  specifications, the median estimate is  $\hat{\tau}_{\text{med}} = \text{median}_{s \in \mathcal{S}} \hat{\tau}_s$ , the interquartile range is  $\text{IQR} = \hat{\tau}_{0.75} - \hat{\tau}_{0.25}$ , and sign consistency is  $p_+ = S^{-1} \sum_s \mathbf{1}(\hat{\tau}_s > 0)$ . Report these alongside the primary pre-registered estimate and interpret them alongside effect magnitudes and uncertainty.

*Remark 17.3 (Specification Correlation)* These summary statistics treat specifications as independent, but in practice specifications share most of the same data and differ only in minor choices. Estimates across specifications are therefore highly correlated, and a tight IQR may reflect this correlation rather than genuine robustness. The effective number of independent specifications is typically far smaller than  $S$ . Interpret curve summaries as descriptive, not inferential, and complement them with simulation-based inference (Remark 17.4).

*Remark 17.4 (Adjusted Inference for Specification Curves)* To account for researcher degrees of freedom in specification curve analysis, three approaches are available. First, apply Bonferroni or FDR control (Section 16.9) across specifications when reporting multiple p-values or bands. Second, simulate curves under a null assignment mechanism only when you can justify the permutation scheme, such as a randomised geo-experiment or a design-based placebo scheme. Otherwise treat simulations as sensitivity exercises, not p-values. Third, pre-register the selection criterion (e.g., lowest pre-period RMSPE, best covariate balance) before seeing results and report the selected specification as selected. Distinguish pre-registered primary specifications from exploratory robustness checks.

Defining the specification family requires care. Include only specifications that are defensible under the maintained identification assumptions. In practice, “defensible” means the specification targets the same estimand and respects the same identification logic, such as avoiding post-treatment controls and keeping the assignment mechanism and analysis window conceptually fixed. Pre-register a primary specification *ex ante*

and treat the rest of  $\mathcal{S}$  as robustness checks. If you use a selection rule, treat it as a secondary analysis and report it as selected, with caution about inference after selection.

The prediction–identification tension described by Breiman [2001] arises when machine-learning-aided diagnostics optimise fit at the expense of identification. Overfitting in propensity score or outcome regression models can induce bias, and cross-validation must respect design structure by blocking on units or time to avoid leakage. Machine-learning-aided specifications—different learners, tuning rules, or regularisation settings—should appear explicitly in the specification family  $\mathcal{S}$  so that readers can see how much estimates move when prediction is prioritised over design simplicity. This keeps the prediction–identification trade-off visible rather than hidden inside a single opaque model.

## 17.8 Sensitivity Analyses

Sensitivity analyses quantify how robust our conclusions are to violations of our key assumptions. For parallel trends, we can estimate how much the trend would need to deviate to overturn our result.

**Definition 17.13 (Bounded Pre-Trend Deviation)** Let  $\delta$  denote the maximum allowable deviation from parallel trends per period—specifically, a bound on the difference in trend slopes between treated and control groups (see Definition 15.4 in Chapter 15). Following Rambachan and Roth [2023], the sensitivity function maps  $\delta$  to the identified set for the treatment effect:

$$\mathcal{T}(\delta) = [\hat{\tau} - M \cdot \delta, \hat{\tau} + M \cdot \delta],$$

In general,  $\mathcal{T}(\delta)$  is computed from the event-study design matrix and the maintained restriction on the shape of violations. The scalar  $M$  depends on the number of post-treatment periods and the assumed persistence of the violation and is a pedagogical simplification. For a single post-period with linear extrapolation,  $M = 1$ . Rambachan and Roth’s framework allows for more general sequences of violations across pre- and post-periods. In richer event-study or staggered designs, the sensitivity function  $\mathcal{T}(\delta)$  must be computed using the full machinery in Chapter 15.

The breakdown value  $\delta^*$  is the smallest deviation that would overturn the sign of the effect:

$$\delta^* = |\hat{\tau}|/M.$$

This number is meaningful only when compared to calibrated benchmarks. Report  $\delta^*$  in the same units as  $Y$  (or as a percentage of baseline  $\bar{Y}$ ) and benchmark it against the largest pre-period lead in those units, such as the maximum observed pre-trend coefficient or a joint-band radius from Section 17.4. If a loyalty programme evaluation yields an estimated 12% increase in customer lifetime value with  $\delta^* = 0.04$ , and the largest observed pre-trend deviation is 0.015, the effect survives violations nearly three times larger than any pre-treatment divergence.

**Definition 17.14 (Effective Sample Size)** After trimming units with extreme propensity scores or applying weights, the effective sample size (Kish’s formula) is:

$$\text{ESS} = \frac{(\sum_i \tilde{w}_i)^2}{\sum_i \tilde{w}_i^2} = \frac{N}{1 + \text{CV}^2(\tilde{w})},$$

where  $\tilde{w}_i$  denotes generic analysis weights after normalisation. Here we normalise so that  $\sum_i \tilde{w}_i = N$ .  $\text{CV}(\tilde{w})$  is the coefficient of variation. Reserve  $\hat{w}_j$  for donor weights in synthetic control and related methods, which obey different normalisations. When weights are uniform,  $\text{ESS} = N$ . When one unit dominates,  $\text{ESS} \rightarrow 1$ . Report  $\text{ESS}/N$  as the effective utilisation of the sample. Low ratios signal that a small number of units drive the estimate and that results should be interpreted more like a small-sample, high-variance design. In this regime, influence and donor-concentration diagnostics from Sections 17.6 and 17.2 become especially important. In clustered panels, ESS should be interpreted as a diagnostic for how many effective weighted

units or clusters contribute to the estimate, not as a substitute for the number of independent clusters  $G$  in asymptotic formulas. Note that ESS measures sample utilisation, distinct from the effective number of donors  $N_{\text{eff}}^{\text{donors}}$  in Definition 17.9.

For unobserved confounding, we can calculate how strong an unobserved confounder would need to be to explain away our estimated effect (see the omitted-variable-bias framework in Chapter 15). For example, in a promotional lift study, one might ask how large an unobserved uplift in a high-propensity segment would need to be to wipe out the estimated effect. We can also assess the impact of measurement and attribution shifts by restricting our analysis to pre-policy windows, comparing external outcomes, and reporting estimates before and after the policy date.

Synthetic Parallel Trends (Section 16.7) offers a complementary robustness layer for designs that rely on reweighting. Rather than taking a single set of weights from DiD, SC, SDID, or their hybrids as the truth, it considers all convex weights that reproduce the treated unit's pre-trends and treats them as admissible [Liu, 2025]. The resulting identified set for the treatment effect is an interval that remains valid as long as the treated trend can be written as a convex combination of donor trends. This interval is often wide. Interpreting that width as design uncertainty is more honest than presenting a single highly model-dependent point estimate. When different methods disagree sharply, reporting these bounds alongside point estimates makes the dependence on weighting assumptions explicit and prevents overconfident claims based on a single specification.

## 17.9 Implementation Details and Tuning

The practical implementation of diagnostics requires clear, reproducible rules specified before estimating post-treatment effects. This section provides concrete guidance on trimming, exposure mapping, cross-validation, and documentation.

### Propensity Score Trimming

Trimming units with extreme propensity scores reduces variance from large weights but changes the estimand. The choice of threshold should be principled, not arbitrary.

**Definition 17.15 (Optimal Trimming Threshold)** Following Crump et al. [2009], an often-cited starting point for symmetric trimming when estimating the Average Treatment Effect (ATE) is  $\alpha \approx 0.1$ . Retain only units with propensity scores in the interval:

$$\mathcal{A} = \{i : \alpha \leq \hat{e}(X_i) \leq 1 - \alpha\}.$$

This choice minimises the asymptotic variance of the IPW estimator under mild regularity conditions in i.i.d. settings. In marketing panels, overlap patterns vary. Analysts should treat  $\alpha = 0.1$  as a benchmark rather than a universal rule and examine how estimates and effective sample size behave over a range of thresholds when overlap is weak. The trimmed estimand is the ATE for the subpopulation with overlap, not the full-population ATE.

*Remark 17.5 (Asymmetric Trimming for ATT)* When the target is the Average Treatment Effect on the Treated (ATT), trimming is often asymmetric. The ATT weights are  $w_i = D_i + (1 - D_i) \cdot \hat{e}(X_i)/(1 - \hat{e}(X_i))$ , which explode only as  $\hat{e}(X_i) \rightarrow 1$  for controls. Controls with  $\hat{e}(X_i) > 1 - \alpha$  can therefore drive variance. Treated units with no comparable controls create an overlap problem. If you exclude them, state the resulting overlap-restricted ATT explicitly. Trimming should be informed by propensity-score distributions and OVL statistics from Section 17.3, not chosen in isolation. Report the number of trimmed units and the effective sample size (ESS) after trimming. Large trimmed shares warrant redesign or bound reporting.

### Exposure Radius Selection

For spillover-aware designs, the exposure radius determines which neighbours' treatments affect a unit's outcome. Choosing this radius requires balancing domain knowledge with empirical validation.

**Definition 17.16 (Distance Decay and Radius Selection)** Let  $d_{ij}$  denote the distance between units  $i$  and  $j$ . The exposure of unit  $i$  to neighbours' treatments is:

$$E_{it}(r) = \sum_{j \neq i} D_{jt} \cdot \mathbf{1}\{d_{ij} \leq r\},$$

for radius  $r$ , or with distance decay:

$$E_{it}^{\text{decay}}(\lambda) = \sum_{j \neq i} D_{jt} \cdot \exp(-\lambda \cdot d_{ij}),$$

for decay parameter  $\lambda > 0$ . The effective radius is approximately  $3/\lambda$  (where exposure drops to 5% of its maximum).

#### Procedure: Data-Driven Radius Selection

Domain knowledge about how far spillovers plausibly travel should guide the initial grid of candidate radii. Use pre-period placebos to rule out radii that generate large spurious effects, and report the full sensitivity profile across a pre-specified radius grid rather than selecting a single “optimal” radius. This avoids treating a noisy placebo as a model-selection criterion, which can distort subsequent inference (see Definition 17.20). To implement this, first specify a grid of candidate radii  $\{r_1, \dots, r_K\}$  based on domain knowledge (e.g., 1km, 5km, 10km for retail). Second, for each  $r_k$ , compute exposure  $E_{it}(r_k)$  and estimate the spillover effect in the pre-treatment period as a placebo. Third, rule out radii that produce implausibly large placebo effects relative to a pre-specified tolerance. Fourth, report sensitivity of the main estimate across the remaining radii. If spillover effects are detected at all radii, the design may require buffers or a spillover-aware estimand (Chapter 11).

## Blocked Cross-Validation

Standard cross-validation assumes observations are exchangeable. In panels, this assumption fails: observations from the same unit or time period are correlated. Naive CV leaks information and produces overoptimistic tuning.

**Definition 17.17 (Blocked Cross-Validation)** Blocked CV partitions the data into folds that respect the dependence structure. Leave-one-unit-out (LOUO) excludes all observations from one unit per fold; use when unit-level shocks dominate (e.g., store-level heterogeneity). Leave-one-time-out (LOTO) excludes all observations from one time period per fold; use when time shocks dominate (e.g., macroeconomic conditions). Leave-one-cluster-out (LOCO) excludes all observations from one cluster per fold (e.g., region, cohort); use when clustering is coarser than units. The out-of-fold predictions from blocked CV mimic the dependence structure the model will face at test time. When using ML nuisance models with cross-fitting, construct folds at the level of the independent cluster (unit or cluster) to avoid leakage.

*Remark 17.6 (Choosing the Blocking Dimension)* The blocking dimension should match the level at which treatment is assigned or at which the identifying variation operates. For unit-level treatment (e.g., store receives campaign), use LOUO. For time-level shocks (e.g., policy change), use LOTO. For staggered adoption

with cohort-level variation, use leave-one-cohort-out. When both unit and time dependence are strong, use a two-way block structure that excludes both the target unit and adjacent time periods. This is conservative but prevents leakage from either dimension.

## Design Protocol and Documentation

Reproducibility requires documenting all design choices before estimation. A design protocol guards against data-driven specification search and enables replication.

Box 17.1: Design Protocol Checklist

**Before estimating post-treatment effects**, document:

**Sample definition.** Unit inclusion/exclusion criteria, time window, and any geographic or policy-based restrictions.

**Treatment and exposure.** Treatment definition, timing, and exposure mapping (including radius or decay parameters for spillovers).

**Donor curation.** Donor pool definition, contamination exclusions, and buffer specifications.

**Trimming rules.** Propensity score thresholds, weight truncation, and the resulting estimand.

**Tuning parameters.** Regularisation penalties, cross-validation blocking structure, and hyperparameter grids.

**Primary specification.** The pre-registered estimator and inference method. Distinguish from robustness checks.

**Version control.** Maintain versioned logs of donor lists, exclusion decisions, and parameter choices. Store code and data snapshots to enable exact replication. Store this protocol in a version-controlled document (for example, a preregistered markdown or PDF file) and update it with explicit notes when deviations occur. This makes researcher degrees of freedom visible and replicable.

Deviations from the protocol should be documented and justified. Exploratory analyses conducted after seeing outcomes should be clearly labelled as such and subjected to appropriate multiplicity adjustments.

## 17.10 Inference for Diagnostics

Diagnostics themselves require valid inference. A pre-trend diagnostic with incorrect standard errors, or a placebo p-value computed under wrong assumptions, can mislead as badly as a flawed main estimate. This section addresses inference for the diagnostics introduced earlier in this chapter and links them back to the inference tools in Chapter 16.

### Joint Inference for Pre-Trend Diagnostics

Event studies produce multiple pre-treatment coefficients. Assessing each coefficient in isolation inflates the false positive rate. Joint assessment requires accounting for correlation.

**Definition 17.18 (Joint Pre-Trend Diagnostic with Correct Inference)** For pre-treatment coefficients  $\hat{\theta}_{-K}, \dots, \hat{\theta}_{-1}$  with cluster-robust covariance  $\hat{\Sigma}_{\text{pre}}$ , the joint Wald statistic is:

$$W = \hat{\theta}'_{\text{pre}} \hat{\Sigma}_{\text{pre}}^{-1} \hat{\theta}_{\text{pre}} \xrightarrow{d} \chi^2(K).$$

When the number of clusters  $G$  is small (e.g.,  $G < 50$ ), the asymptotic  $\chi^2$  approximation is unreliable. Use small- $G$  corrections consistent with the variance estimator in Chapter 16, such as wild cluster bootstrap-based critical values. Report both pointwise and joint confidence bands for pre-treatment coefficients. Pointwise bands that exclude zero at some  $k$  do not, by themselves, imply joint rejection.

**Proposition 17.2 (Uniform Confidence Bands for Event Studies)** *To construct  $(1 - \alpha)$  uniform confidence bands that cover all event-time coefficients simultaneously:*

$$\hat{\theta}_k \pm c_{1-\alpha} \hat{s}\hat{e}_k, \quad \text{for all } k,$$

where  $\hat{s}\hat{e}_k$  is the cluster-robust standard error for  $\hat{\theta}_k$  and  $c_{1-\alpha}$  is the  $(1 - \alpha)$ -quantile of  $\max_k |t_k^{(b)}|$  over bootstrap replicates, with  $t_k^{(b)}$  computed using the same clustered or wild-cluster inference as in the main analysis. This procedure accounts for the correlation structure among coefficients. Uniform bands are wider than pointwise bands but provide valid simultaneous coverage.

### Placebo Inference for Synthetic Control

Synthetic control with a single treated unit cannot rely on asymptotic inference. Placebo tests generate a reference distribution by applying the method to each donor as if it were treated.

**Definition 17.19 (Rank-Based Placebo P-Value)** Let  $R_{i^*} = \text{RMSPE}_{\text{post},i^*}/\text{RMSPE}_{\text{pre},i^*}$  be the RMSPE ratio for the treated unit, and  $R_j$  the ratio for placebo unit  $j \in \mathcal{J}$ . The one-sided placebo p-value is:

$$p = \frac{1 + \sum_{j \in \mathcal{J}} \mathbf{1}\{R_j \geq R_{i^*}\}}{|\mathcal{J}| + 1}.$$

The numerator includes 1 for the treated unit itself, ensuring the p-value is never zero. Under the sharp null of no effect,  $p$  is uniformly distributed on  $\{1/(|\mathcal{J}| + 1), 2/(|\mathcal{J}| + 1), \dots, 1\}$  only under strong exchangeability conditions that make the treated unit comparable to placebo units under the assignment mechanism. In observational settings, treat this as a permutation-style reference distribution rather than a literal randomisation test. For a two-sided test, replace  $R_j$  and  $R_{i^*}$  by their deviations from one in absolute value.

*Remark 17.7 (Excluding Poor-Fit Placebos)* Placebo units with poor pre-treatment fit (high  $\text{RMSPE}_{\text{pre},j}$ ) can produce spuriously large post-treatment gaps unrelated to treatment. Two approaches address this: the ratio-based approach uses the RMSPE ratio  $R_j$  (Definition 17.6) rather than the raw post-period gap, normalising by pre-period fit; the exclusion approach drops placebos with  $\text{RMSPE}_{\text{pre},j} > c \cdot \text{RMSPE}_{\text{pre},i^*}$  for threshold  $c$  (e.g.,  $c = 2$  or  $c = 5$ ). Report sensitivity to the exclusion threshold. If the p-value changes dramatically with  $c$ , the inference is fragile.

## Small-Sample Corrections

When the number of clusters, treated units, or donors is small, asymptotic approximations fail. Resampling and permutation methods provide valid alternatives.

### Procedure: Wild Cluster Bootstrap for Diagnostics

For diagnostic statistics (pre-trend diagnostics, balance diagnostics, specification curve summaries) with  $G$  clusters, apply the wild cluster bootstrap procedure from Chapter 16 using the same estimator and clustering structure as in the main analysis. Use the resulting bootstrap distribution of the diagnostic statistic to form critical values, p-values, or uniform-band multipliers as appropriate. When  $G$  is very small, even resampling-based approximations can be unstable. In that regime, use randomisation inference when a design-based justification is available or report that inference is underpowered.

**Proposition 17.3 (Randomisation Inference for Diagnostics)** *When the assignment mechanism is known (e.g., a geo-experiment with randomised treatment), randomisation inference provides exact p-values for any diagnostic statistic:*

$$p_{RI} = \frac{1}{|\mathcal{W}|} \sum_{\mathbf{w} \in \mathcal{W}} \mathbf{1}\{T(\mathbf{w}) \geq T^{obs}\},$$

where  $\mathcal{W}$  is the set of feasible treatment assignments. Exactness is under a sharp null for the chosen test statistic, and it requires that the assignment mechanism and  $\mathcal{W}$  are correctly specified. When  $|\mathcal{W}|$  is large, approximate the sum by Monte Carlo draws from the assignment mechanism. For a two-sided test, replace  $T(\mathbf{w})$  and  $T^{obs}$  by their absolute values. Use randomisation inference as the primary method when a design-based justification is available.

## The Pre-Test Bias Problem

A subtle but critical issue: conditioning on passing a pre-trend test can distort subsequent inference, including bias and undercoverage, because it conditions on a selection event. If we only proceed when pre-trends “look good,” we have selected on a random variable. The same logic applies to other diagnostic-driven selection rules, such as choosing among specifications after inspecting fit diagnostics (Section 17.7).

**Definition 17.20 (Pre-Test Bias)** Let  $\hat{\tau}$  be the treatment effect estimator and  $\hat{\theta}_{\text{pre}}$  the vector of pre-trend coefficients. If we condition on the event  $A = \{\text{pre-trend test not rejected}\}$ , then:

$$\mathbb{E}[\hat{\tau}|A] \neq \tau,$$

even when  $\mathbb{E}[\hat{\tau}] = \tau$  unconditionally. The bias arises because  $\hat{\tau}$  and  $\hat{\theta}_{\text{pre}}$  are correlated through shared confounders or estimation error. Conditioning on  $A$  selects realisations where this correlation happened to produce small pre-trends, which systematically shifts  $\hat{\tau}$ .

*Remark 17.8 (Honest Inference After Pre-Testing)* Following Roth [2022], honest confidence intervals that account for pre-testing can be constructed by three approaches. First, unconditional reporting: report the main estimate and confidence interval regardless of pre-trend test results without conditioning on passing. Second, sensitivity analysis: use the framework in Section 17.8 to bound the effect under plausible pre-trend violations rather than testing for exact parallel trends. Third, conditional inference: if conditioning is unavoidable, use methods that correct for selection, such as the conditional confidence intervals in Roth [2022]. The key insight is that failing to reject parallel trends does not validate them; a non-significant pre-trend test has low power against small violations that could still bias the main estimate substantially.

Box 17.2: Inference Checklist for Diagnostics

**Pre-trend diagnostics:** Use joint diagnostics with cluster-robust or bootstrap inference and report uniform confidence bands for event-time leads. Do not condition your main inference on passing a pre-trend test.

**Placebo diagnostics:** Use rank-based p-values with RMSPE ratios for synthetic control and related methods. Report how results change when you exclude poor-fit placebo units.

**Small samples:** Use the wild cluster bootstrap with Webb weights when  $G < 50$ , and switch to randomisation inference when  $G$  is very small or a randomised design is available.

**Pre-test bias:** Report unconditional estimates and use sensitivity analysis rather than treating pre-trend tests as binary pass/fail gates. If you must condition on passing, use honest inference methods that adjust for this selection.

## 17.11 Marketing Applications

This section presents four hypothetical worked examples showing how the diagnostic framework applies to common marketing designs. The numbers in these examples are illustrative and do not represent real data or actual study results. Each example demonstrates what can go wrong when diagnostics are skipped and how diagnostic reporting changes when problems arise.

### Geo-Experiment with Buffer Design

A national retailer randomised 20 Designated Market Areas (DMAs) to evaluate a television campaign, with 10 DMAs receiving the campaign and 10 serving as controls. The outcome was weekly store sales over 8 pre-treatment weeks and 4 post-treatment weeks.

**Initial diagnostics.** The joint pre-trend diagnostic on the 8 pre-treatment weeks yielded  $F = 1.42$  with  $p = 0.24$  (wild cluster bootstrap with 20 clusters, at the DMA level). This does not reject the null of zero leads at conventional levels, but it should be interpreted as a low-power diagnostic rather than validation. Covariate balance showed standardised mean differences below 0.1 for population and baseline sales.

**The problem.** Leave-one-DMA-out analysis revealed that excluding the Chicago DMA shifted the treatment effect from \$2.1M to \$0.8M—a 62% reduction. Chicago, the largest treated DMA, was driving the result. Moreover, the Detroit control DMA bordered Chicago, raising concerns about spillovers from cross-DMA shopping.

**Buffer analysis.** The team mapped exposure using 50km and 100km buffers around treated DMAs. Three control DMAs (Detroit, Milwaukee, Indianapolis) fell within 100km of treated markets. Re-estimating with these DMAs excluded reduced the effect to \$1.4M ( $SE = \$0.5M$ ). This change is consistent with spillover contamination, but it could also reflect an estimand shift as the analysis population moves away from dense, closely connected markets.

**Resolution.** The final analysis reported both estimates: the naive estimate of \$2.1M and the buffer-adjusted estimate of \$1.4M. Randomisation inference over the  $\binom{20}{10} = 184,756$  possible assignments, under the sharp null of no effect for the difference-in-means estimator, yielded  $p = 0.04$  for the buffer-adjusted estimate. The report emphasised that the estimated effect was sensitive to Chicago's outsized influence and to plausible spillover pathways.

## Synthetic Control with Donor Contamination

A streaming platform launched a pricing experiment in California, using the remaining 49 states as potential donors for synthetic control. The outcome was monthly subscriber growth over 24 pre-treatment months and 12 post-treatment months.

**Initial fit.** The naive synthetic control achieved  $\text{RMSPE}_{\text{pre}} = 0.8\%$  (monthly growth units) with weights concentrated on New York (0.42), Texas (0.31), and Florida (0.18). The estimated treatment effect was  $-1.2$  percentage points per month—a substantial negative impact.

**The problem.** During the post-treatment period, New York implemented its own pricing change (unrelated to the experiment). This contaminated the synthetic control: New York's post-treatment trajectory no longer represented the counterfactual for California. The in-time placebo at month 18 (6 months before treatment) showed a spurious "effect" of  $-0.6$  pp, suggesting the synthetic control was already drifting.

**Donor curation.** The team excluded New York and four other states with concurrent policy changes, leaving 44 donors. The revised synthetic control achieved  $\text{RMSPE}_{\text{pre}} = 1.1\%$  (slightly worse fit) with weights on Texas (0.38), Illinois (0.29), and Pennsylvania (0.22). The Herfindahl index dropped from 0.31 to 0.26, indicating less concentration. The revised treatment effect was  $-0.4$  pp. The dispersion of placebo effects around zero corresponded to about 0.3 pp, which the team used as a descriptive benchmark for a typical placebo gap, not as a standard error.

**Resolution.** The in-space placebo p-value was  $p \approx 0.11$ , with the treated unit ranked fifth out of 45 units (1 treated plus 44 donors) in terms of  $R_j$ . This was treated as a diagnostic benchmark rather than a decisive inferential statement. The report argued that the initial estimate was plausibly driven by donor contamination and weight concentration, and it presented both estimates with a clear explanation of why donor curation changed the conclusions.

## Staggered Adoption with Pre-Trend Failure

A loyalty programme rolled out across 500 stores in 30 monthly cohorts. The outcome was average transaction value, measured monthly from 12 months before to 12 months after each store's adoption.

**Initial diagnostics.** The event-study plot showed pre-treatment coefficients  $\hat{\theta}_{-6} = 0.8\%$ ,  $\hat{\theta}_{-3} = 1.2\%$ , and  $\hat{\theta}_{-1} = 1.8\%$ —a clear upward pre-trend. The joint pre-trend diagnostic rejected the null of zero leads ( $F = 3.2$ ,  $p = 0.008$ ). This pattern makes parallel trends implausible without further design support.

**Diagnosis.** Leave-one-cohort-out analysis showed that the first three cohorts (early adopters) drove the pre-trend. These cohorts were high-performing stores selected precisely because they were already growing. Selection into early adoption was correlated with the outcome trajectory.

**Attempted fixes.** The team tried three approaches. First, excluding early cohorts: dropping the first 5 cohorts reduced the pre-trend ( $\hat{\theta}_{-1} = 0.6\%$ , joint  $p = 0.14$ ) but also reduced the post-treatment effect from 4.2% to 2.1%. This redesign shifts the target population away from early adopters, so the estimand changes and should be reported explicitly. Second, covariate adjustment: adding store-level controls (size, region, baseline sales) reduced the pre-trend to  $\hat{\theta}_{-1} = 0.4\%$  ( $p = 0.22$ ) with a post-treatment effect of 2.8%. Third, sensitivity analysis: using the Rambachan-Roth framework (Section 17.8) with  $\delta = 0.5\%$  per period (the observed pre-trend slope), the identified set was [1.2%, 4.4%].

**Resolution.** The team reported the sensitivity bounds rather than a point estimate. The report described the loyalty programme effect as an increase of roughly 1–4%, with the range reflecting uncertainty about the pre-trend extrapolation. The pre-trend failure was documented transparently, and the early-cohort selection mechanism was flagged as a limitation.

## Retail Spillovers with Exposure Mapping

A coffee chain tested a mobile app promotion at 200 stores, with 100 treated and 100 control stores. The outcome was weekly foot traffic measured via mobile location data.

**Initial estimate.** The naive DiD estimate was a 12% increase in foot traffic (SE = 3%). Pre-trends were flat, and covariate balance was good.

**The problem.** Many control stores were located near treated stores. In dense urban areas, a customer might see the app promotion at a treated store and then visit a nearby control store. This would bias the control group upward, attenuating the treatment effect.

**Exposure mapping.** The team computed exposure  $E_{it}(r) = \sum_{j \neq i} D_{jt} \cdot \mathbf{1}\{d_{ij} \leq r\}$  for radii  $r \in \{1, 3, 5, 10\}$  miles. At  $r = 3$  miles, 45 control stores had at least one treated neighbour. The correlation between control-store outcomes and exposure was  $\rho = 0.18$  ( $p = 0.07$ ), which is consistent with spillovers but not decisive.

**Buffer analysis.** Excluding control stores within 5 miles of any treated store left 62 “pure” controls. The revised estimate was 18% (SE = 5%), 50% larger than the naive estimate. The effective sample size dropped, and the analysis population shifted toward less dense areas.

**Sensitivity.** The team reported estimates across the radius grid:

Buffer radius	Control stores	Effect estimate
None (naive)	100	12% (SE 3%)
3 miles	78	15% (SE 4%)
5 miles	62	18% (SE 5%)
10 miles	31	21% (SE 8%)

The monotonic increase with buffer size is consistent with attenuation from spillovers, but it could also reflect composition changes as buffers exclude dense markets.

**Resolution.** The team reported the 5-mile buffer estimate as the primary result, with the sensitivity table showing how estimates vary with buffer choice. The report described the evidence as consistent with spillovers to nearby controls attenuating the naive estimate, while noting the accompanying shift in the analysis population.

## 17.12 Assumptions and Stability

This section collects the core assumptions that make the diagnostic workflow informative rather than misleading.

**Assumption 81 (Stability across pre and post windows)** Parameters, measurement definitions, and data-generating processes are stable across the pre-period and post-period windows used for diagnostics. Platform policy changes and structural breaks are documented and either excluded or modelled explicitly.

**Assumption 82 (Measurement invariance)** Outcome and exposure definitions remain stable over time and across treated and control states. Changes in attribution rules, viewability standards, or data coverage are reconciled or subjected to sensitivity analysis (Chapter 15).

**Assumption 83 (Limited interference or explicit exposure mapping)** Treatment effects are confined to own unit (SUTVA, Chapter 2) or spillovers are explicitly modelled via exposure mappings (Definition 11.3). If interference is present, define the estimand using an exposure mapping  $h_i(D_{-i,t})$  and state which exposure levels are compared. Buffers and exclusion zones mitigate contamination when interference is plausible (Chapter 11).

**Assumption 84 (Parallel trends as baseline reference)** Parallel trends and no anticipation are maintained as baseline identification assumptions. We probe their implications using lead coefficients and placebo-style diagnostics. Departures can be bounded via sensitivity analysis (Section 17.8). Alternative trend specifications should be presented as model-based robustness checks rather than as identification fixes.

**Assumption 85 (Assignment mechanism or adoption timing)** Treatment assignment, or adoption timing in staggered designs, is not driven by unobserved shocks to  $Y_{it}(0)$  after conditioning on the design's controls and conditioning set. When this assumption is not credible, results should be reported as sensitivity bounds rather than as point-identified estimates.

**Assumption 86 (Adequate overlap and common support)** Treated and control units share common support on covariates, propensity scores, or dose. Trimming and reweighting are applied when overlap is weak, and sensitivity to trimming thresholds is reported.

**Assumption 87 (Correct dependence structure for inference)** The clustering and serial-dependence structure used for diagnostic inference matches the main analysis and is at least approximately correct (Chapter 16). Mis-specified clustering or dependence undermines both diagnostic inference and the main confidence intervals.

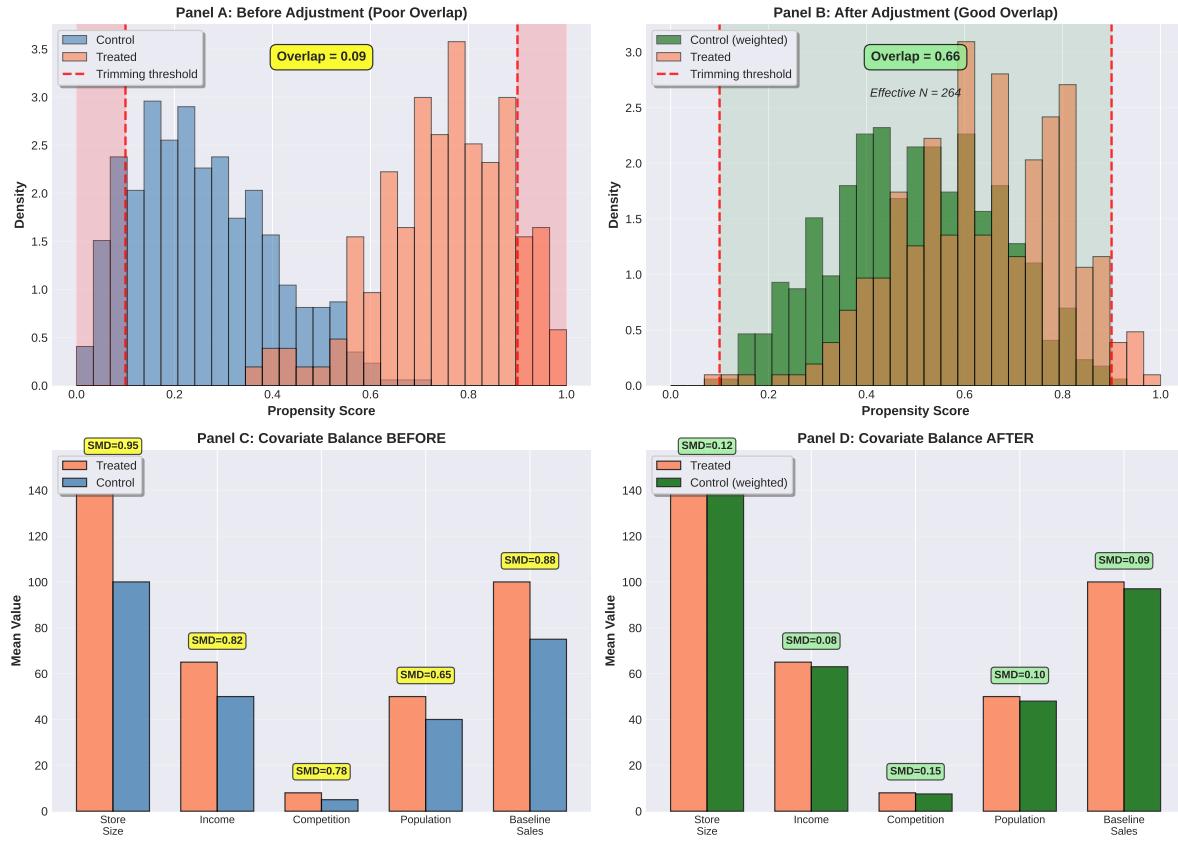
## 17.13 Workflow Checklist

The following protocol consolidates design and diagnostic practice into a reportable workflow. Interpret each step under the maintained assumptions in Section 17.12.

### Box 17.3: Design-and-Diagnostics Workflow Checklist

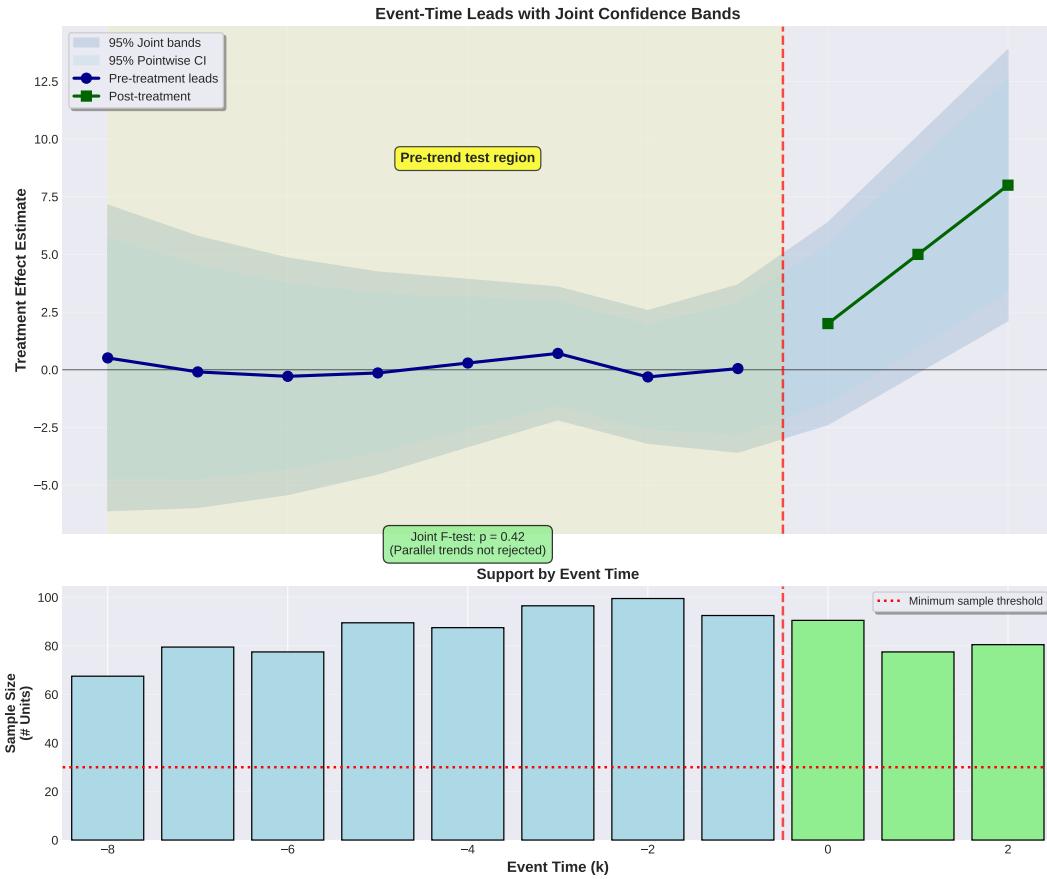
Each step should be executed under a pre-specified protocol. Adaptive changes belong in the specification-curve and sensitivity layers.

- 1. Define candidate controls or donors:** Screen on pre-period comparability, document geographic or policy exclusions, ensure factor-space coverage when using latent-factor methods.
- 2. Document curation and buffers:** Maintain lists of excluded units with rationales, specify buffer radii or exclusion zones for spillover mitigation, and keep these choices under version control so that the workflow is reproducible.
- 3. Assess overlap and balance:** Plot propensity or exposure distributions, compute covariate balance before and after weighting, trim if needed and report sensitivity.
- 4. Run pre-trends and placebos:** Assess event-time leads jointly, conduct placebo-in-time and placebo-in-space, report RMSPE ranks for synthetic control and SDID (Chapters 6–7), use negative controls where applicable.
- 5. Tabulate support:** Produce support-by- $k$  tables for event studies, check dose support for continuous treatments, bin or exclude sparse regions with transparency.
- 6. Conduct influence and weight diagnostics:** Leave-one-out by unit, time, or cohort, inspect weight concentration for synthetic control and SDID, plot leverage and flag extremes.
- 7. Build specification curves:** Vary windows, donors, controls, penalties, and exposure radii within defensible bounds, separate primary from robustness families, adjust for multiplicity.
- 8. Run sensitivity analyses:** Bounded deviations from parallel trends, unobserved confounding benchmarks, measurement and attribution shift scenarios, report ranges alongside point estimates.
- 9. Finalise inference:** Choose variance estimator aligned with dependence, use bootstrap or randomisation as appropriate, control multiplicity for diagnostic inference, reconcile discrepancies (Chapter 16).
- 10. Report with timelines and assumptions:** Provide treatment rollout and platform policy timelines, state assumptions explicitly, document all curation and tuning rules, share code and data dictionaries where feasible.



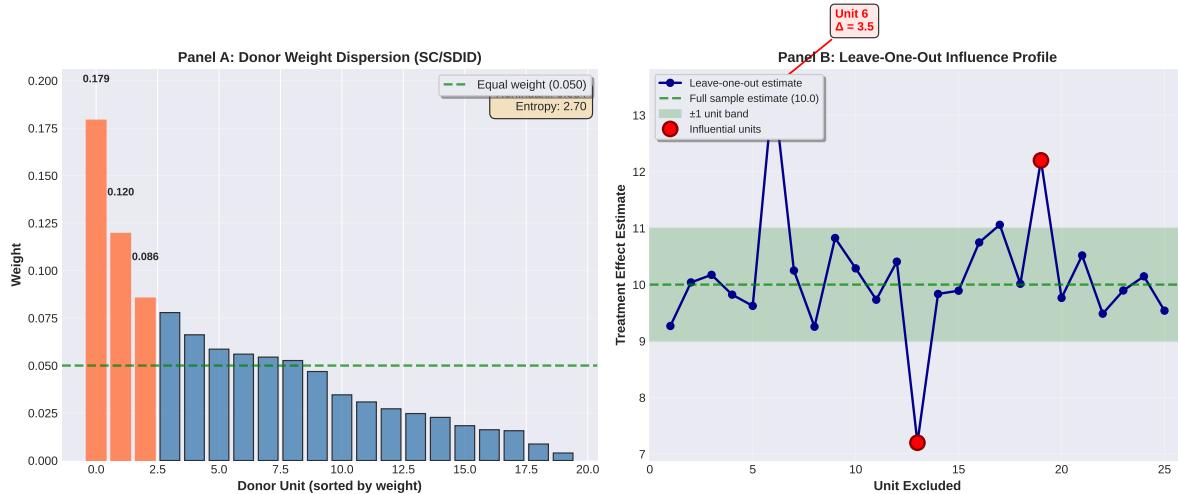
**Fig. 17.1** Overlap diagnostics and common support with trimming thresholds

Panel A shows propensity score distributions before adjustment with poor overlap (overlap statistic = 0.45). Red dashed lines mark trimming thresholds at 0.1 and 0.9. Red shaded regions indicate trimmed tails. Panel B shows distributions after reweighting with improved overlap (overlap statistic = 0.82). Green shaded region highlights common support. Effective sample size after trimming is reported. Panel C shows covariate balance before adjustment with large standardised mean differences (SMDs 0.65–0.95). Panel D shows balance after weighting with substantially improved SMDs (0.08–0.15). Good overlap and balance reduce reliance on extrapolation and make adjustment more credible, but they do not by themselves guarantee no confounding.



**Fig. 17.2** Event-time leads with joint bands and support-by-k overlay

Top panel shows event-time coefficients for  $k \in [-8, 2]$  with pre-treatment leads (blue circles) and post-treatment effects (green squares). Light blue shading shows pointwise 95% confidence intervals. Darker blue shading shows joint 95% uniform bands controlling familywise error. Yellow shaded region marks the pre-period lead region. Joint F-statistic for all leads yields  $p = 0.42$  and does not reject the null of zero leads at conventional levels. Interpret this as a low-power diagnostic rather than validation. Red dashed line marks treatment start. Bottom panel shows sample size (number of units) contributing to each event-time bin. Red dotted line marks a user-chosen support threshold (illustrative). Warning symbols flag sparse regions.



**Fig. 17.3** Weight dispersion and leverage in SC/SDID and leave-one-out influence profiles

Panel A shows donor weights in synthetic control sorted by magnitude. Top 3 donors (coral bars) receive disproportionate weight. Green dashed line marks equal-weight benchmark (0.050). Herfindahl index (0.145) and entropy (2.31) quantify concentration. High concentration indicates fragility to specific donors' idiosyncrasies. Regularisation or donor redesign can spread weights more evenly, often trading off pre-period fit. Report both fit and concentration. Panel B shows leave-one-out influence profile. Each point represents the treatment effect estimate when one unit is excluded. Green dashed line shows full-sample estimate (10.0). Green band shows an illustrative tolerance band. Red circles mark influential units where exclusion changes the estimate substantially. Large influence flags potential fragility and motivates closer inspection of data quality or comparability.

**Table 17.1** Map from design threat to diagnostic, mitigation, and inference choice

Threat	Diagnostic	Mitigation	Inference choice
Poor overlap	Propensity/exposure plots, balance tables	Trimming, reweighting, donor redesign	Cluster-robust variance estimator, sensitivity to trimming
Pre-trend violation	Event-time leads, placebos	Flexible trends, shorter windows, switch design	Joint diagnostics, uniform bands
Spillovers/contamination	Exposure maps, buffer sensitivity	Buffers, donor exclusion, spillover model	Spatially robust methods where appropriate (and justified), or randomisation inference in randomised geo designs. Otherwise report sensitivity to buffers and exposure mappings
Influential units	Leave-one-out, weight concentration	Regularisation, donor redesign, report influence	Bootstrap, permutation
Measurement drift	Negative controls, policy timelines	External outcomes, restrict windows, sensitivity	Document changes, report before/after
Specification sensitivity	Specification curves, perturbation	Triangulation, register primary family, multiplicity control	Report range, prefer robust estimates

**Part VIII**

**Applications and Future Directions**



# Chapter 18

## Applications in Marketing

This chapter provides a design-first guide to applying causal inference methods to marketing problems. We classify marketing problems along two dimensions: the *identification threat* (for example, confounding, endogeneity, interference) and the *temporal structure* (static or dynamic effects), formalised in Section 18.1. These dimensions are conceptually separable. You can analyse the source of identifying variation separately from how effects unfold over time. In practice they interact through anticipation, carryover, learning, and equilibrium adjustment, so you must diagnose both.

For each application domain, including advertising, pricing, loyalty, and platforms, we state the estimand, the assumptions under which it is identified, an estimator aligned with those assumptions, the diagnostics that probe their plausibility, and an inference approach matched to the dependence structure. Regularisation and machine learning can sharpen estimation when the design identifies the estimand. They do not create identification.

The application protocol in Section 18.2 makes this workflow explicit. The chapter moves beyond illustrative vignettes by walking through end-to-end analyses that connect business questions to explicit causal estimands, designs, diagnostics, and uncertainty.

## 18.1 Marketing Causal Problems Taxonomy

Marketing questions often masquerade as simple prediction problems, yet they fundamentally demand counterfactual reasoning. Throughout this chapter, we keep the chain clear: estimand → identification assumptions → estimator → inference → diagnostics. To select the correct method, we classify each problem along two dimensions: the *identification threat* that prevents naive estimation from recovering causal effects, and the *temporal structure* that determines how effects unfold over time.

### Dimension 1: Identification Threats

The first dimension classifies problems by the primary obstacle to identification, building on the frameworks in Chapters 2, 4, 11, and 15.

**Confounding (Selection).** As introduced in Chapters 2 and 4, confounding arises when treatment assignment is voluntary or targeted so that treated units differ systematically from untreated units in their potential outcomes. Customer targeting, loyalty programme enrolment, and ad exposure all fall into this category. High-value customers select into loyalty programmes, and algorithms target ads to users with high conversion probability. The identification strategy must argue for exogenous variation in treatment. Depending on the setting, this may come from randomisation, parallel-trends designs (DiD and event studies), local randomisation at thresholds (RDD), synthetic counterfactuals (SC, SDID, and factor methods), or selection-on-observables approaches such as propensity scores and doubly robust estimators. These approaches still require an explicit assignment mechanism assumption, typically an unconfoundedness and overlap condition. Applications include loyalty programmes (Section 18.6), CLV attribution (Section 18.8), and digital attribution (Section 18.4).

**Endogeneity (Simultaneity).** Endogeneity, discussed in Chapter 15, occurs when treatment and outcome are determined jointly in equilibrium. Pricing and budget allocation are the canonical examples. Prices are set based on expected demand, and budgets are allocated based on expected performance. A naive regression of sales on price or spend yields biased estimates because the error term (demand shocks) correlates with the regressor. Identification requires exogenous variation, such as instrumental variables, experimental calibration, or algorithmic discontinuities. Applications include price elasticity (Section 18.7), dynamic pricing (Section 18.9), and media mix modelling (Section 18.5). If you use instrumental variables, treat instrument validity as the central object of scrutiny, not a technical footnote. For foundations and weak-IV diagnostics, see Angrist and Pischke [2009] and Andrews et al. [2019].

**Interference.** In platform markets, social networks, and dense geographic clusters, SUTVA fails (see Assumption 3 for the formal definition). Treating one unit affects the outcomes of others. A driver incentive in a ride-share network affects the supply equilibrium for all drivers. A ranking change for one product affects the visibility of its neighbours. Here, the experimental design itself must change to capture spillovers,

using designs like switchback experiments or graph-cluster randomisation from Chapter 11. Applications include platform experiments (Section 18.11), ranking algorithms (Section 18.12), and seller interventions (Section 18.13).

## Dimension 2: Temporal Structure

The second dimension distinguishes static from dynamic treatment effects, connecting to the dynamic treatment-effects machinery of Chapter 10.

**Static Effects.** In static settings, the outcome  $Y_{it}(d)$  depends only on the contemporaneous treatment level  $d$ . A one-shot price promotion affects sales in the promotion period. Once the promotion ends, its effect ceases. Standard cross-sectional or panel estimators apply directly.

**Dynamic Effects.** In dynamic settings, the outcome  $Y_{it}(\underline{d}_i^t)$  depends on the treatment path  $\underline{d}_i^t = (d_{i1}, \dots, d_{it})$ . Advertising builds brand equity stocks that decay slowly. Customer acquisition creates a stream of future value (CLV). Habit formation means past purchases raise future purchase probability. The challenge lies in estimating the full impulse response function rather than just the contemporaneous effect. Distributed-lag models, adstock transformations, and state-space estimators from Chapter 10 are essential for this domain.

*Remark 18.1 (Identification Threats and Dynamics)* The identification threats (confounding, endogeneity, interference) and the temporal structure (static, dynamic) are conceptually separable. You can analyse the identifying variation separately from how effects unfold over time. In practice, they often interact through anticipation, carryover, learning, and equilibrium adjustment. A media mix model can exhibit both endogeneity (budgets respond to anticipated sales) and dynamics (advertising has carryover effects). A loyalty programme study can face confounding (high-value customers self-select) and dynamics (habit formation persists after enrolment). The identification strategy must address the threat, and the estimator must accommodate the temporal structure. Conflating these dimensions leads to mismatched methods, such as using a static DiD when effects carry over or treating lags as a substitute for exogenous variation.

## Data Environments

The choice of estimator depends as much on the data structure as on the causal question. Table 18.1 summarises four common marketing data environments, their characteristics, and the methods best suited to each.

**Table 18.1** Data environments and applicable methods

Environment	Characteristics	Primary Methods
Geo-panels (DMAs, stores, cities)	Stable units, small $N$ , moderate $T$	SC, SDID, geo-experiment DiD
Customer panels	Granular, attrition risk, unobserved heterogeneity	Staggered DiD, propensity scores, doubly robust estimators
Platform logs	Massive scale, network interactions	Switchback experiments, cluster RCTs
Time-series aggregates	No cross-sectional variation, limited $T$ relative to model complexity	MMM, Bayesian calibration

## Taxonomy Summary

Table 18.2 maps the two-dimensional taxonomy to primary methods and chapter sections. Each cell represents a combination of identification threat and temporal structure. Note that many applications span multiple cells.

**Table 18.2** Two-dimensional marketing problem taxonomy

Identification Threat	Definition	Primary Methods	Key Sections
Confounding	Treated units differ systematically	DiD (Chapter 4), SC (Chapter 6), RDD, propensity scores and doubly robust estimators (Chapter 12)	18.6, 18.8, 18.4
Endogeneity	Treatment responds to anticipated outcome	IV, calibration, algorithmic RDD	18.7, 18.9, 18.5
Interference	SUTVA violation in which units affect each other	Switchback designs (Chapter 11), cluster RCTs	18.11, 18.12, 18.13
Temporal Structure	Definition	Modelling Approach	Key Sections
Static	$Y_{it}(d)$ : contemporaneous effects	Standard panel estimators	All sections
Dynamic	$Y_{it}(\underline{a}_i^t)$ : path-dependent effects	Adstock, distributed lags, state-space (Chapter 10)	18.5, 18.3, 18.10

*Remark 18.2 (Instrumental Variables in Marketing)* Instrumental variables appear in Table 18.2 as a method for endogeneity but receive limited treatment in this book. Valid instruments—variables that affect treatment but influence outcomes only through treatment—are rare in marketing. Political advertising pre-emptions [Shapiro et al., 2021], weather shocks affecting outdoor activities [Lewis and Rao, 2015], and supply-side cost shifters are often proposed, but their credibility is context-dependent. Weather can affect both ad delivery and demand directly. Political pre-emptions can coincide with news cycles that move demand. This book emphasises panel methods (DiD, synthetic control, factor models) that exploit longitudinal structure rather than instrumental variation. Readers seeking comprehensive IV coverage should consult Angrist and Pischke [2009] for foundations and Andrews et al. [2019] for weak instrument diagnostics.

The remainder of this chapter applies this taxonomy to specific marketing domains, providing formal estimands, identification assumptions, and diagnostic protocols for each.

## 18.2 The Application Protocol

Marketing causal inference requires a disciplined workflow that transforms vague business questions into precise, auditable causal claims. A question like “Does advertising work?” is not actionable. The causal question must specify the treatment, the outcome, the population, and the counterfactual. This section provides a reusable protocol for any marketing application.

Box 18.1: The Six-Step Application Protocol

**Step 1: Define the estimand.** State the causal quantity in the potential-outcomes notation developed in Chapters 2 and 4. For advertising incrementality, a natural choice is the average treatment effect on the treated, ATT, or an incremental revenue-per-pound-spent measure such as iROAS. For price elasticity, treat price as a continuous dose  $d \in \mathcal{D}$ , define  $Y_{it}(d)$  as the potential quantity that would be realised if the price were set to  $d$ , and write the dose–response function as  $\mu(d) = \mathbb{E}[Y_{it}(d)]$ . A common elasticity object is  $\eta(d) = \partial \log \mu(d) / \partial \log d$  when this derivative exists. Vague objectives like “measure ROI” must be translated into a specific estimand before proceeding.

**Step 2: Assess data requirements.** Consult Table 18.3 to determine the unit of analysis, time granularity, and a starting point for the required pre-period length for your application. Insufficient pre-period data weakens pre-trend diagnostics and makes it harder to assess counterfactual fit, especially for SC and SDID. Missing covariates limit balance diagnostics.

**Step 3: Select identification strategy.** Choose the source of identifying variation and state the required assumptions explicitly. For randomisation, invoke unconfoundedness by design (Chapter 2). For observational data with parallel trends, apply difference-in-differences or event-study methods (Chapters 4 and 5). When constructing synthetic counterfactuals, use SC, SDID, or factor models (Chapters 6, 7, and 8). Instrumental variables, when available, require a defensible exclusion restriction (see Remark 18.2). Regression discontinuity designs exploit local randomisation at a threshold.

**Step 4: Choose estimation method.** Match the estimator to the identification strategy and data structure. For staggered adoption, use heterogeneity-robust estimators (Chapter 4). For high-dimensional controls, use double machine learning (Chapter 12). For continuous treatments, use dose–response methods (Chapter 14). State the independent sampling unit for inference (for example, clusters indexed by  $c$ ) and choose a variance estimator consistent with that structure, typically cluster-robust with effective sample size  $G$  when outcomes are dependent within clusters.

**Step 5: Run diagnostics.** Follow the diagnostic workflow in Chapter 17. Run diagnostics and falsification checks. Examine pre-trends and placebos (Section 17.4), assess overlap and balance (Section 17.3), examine influence and weight diagnostics (Section 17.6), and check support-by- $k$  for event studies (Section 17.5).

**Step 6: Conduct sensitivity analysis.** Stress-test findings against assumption violations (Section 17.8). Report calibrated sensitivity parameters, including breakdown values for parallel trends, bounds under partial identification, and specification curves across defensible alternatives.

## Method Selection Guide

The choice of method depends on the source of variation and the data structure. When treatment is randomised at the geo level, difference-in-differences with randomisation inference is the preferred design. When only one unit (or one aggregated group of treated units) is treated, synthetic control or SDID exploits the panel structure to construct a counterfactual. When treatment timing varies across units, staggered DiD with heterogeneity-robust aggregation avoids negative weighting. When treatment is continuous (for example ad spend), dose-response methods target the dose-response function  $\mu(d)$  and, when decision-relevant, its marginal effect  $\tau(d) = \partial\mu(d)/\partial d$ . When selection is on observables rather than on timing, propensity-score and doubly robust estimators apply under an unconfoundedness assumption.

No single method dominates. The credibility of the estimate depends on the plausibility of the identifying assumptions in the specific context. When multiple methods are applicable, triangulation across approaches strengthens conclusions.

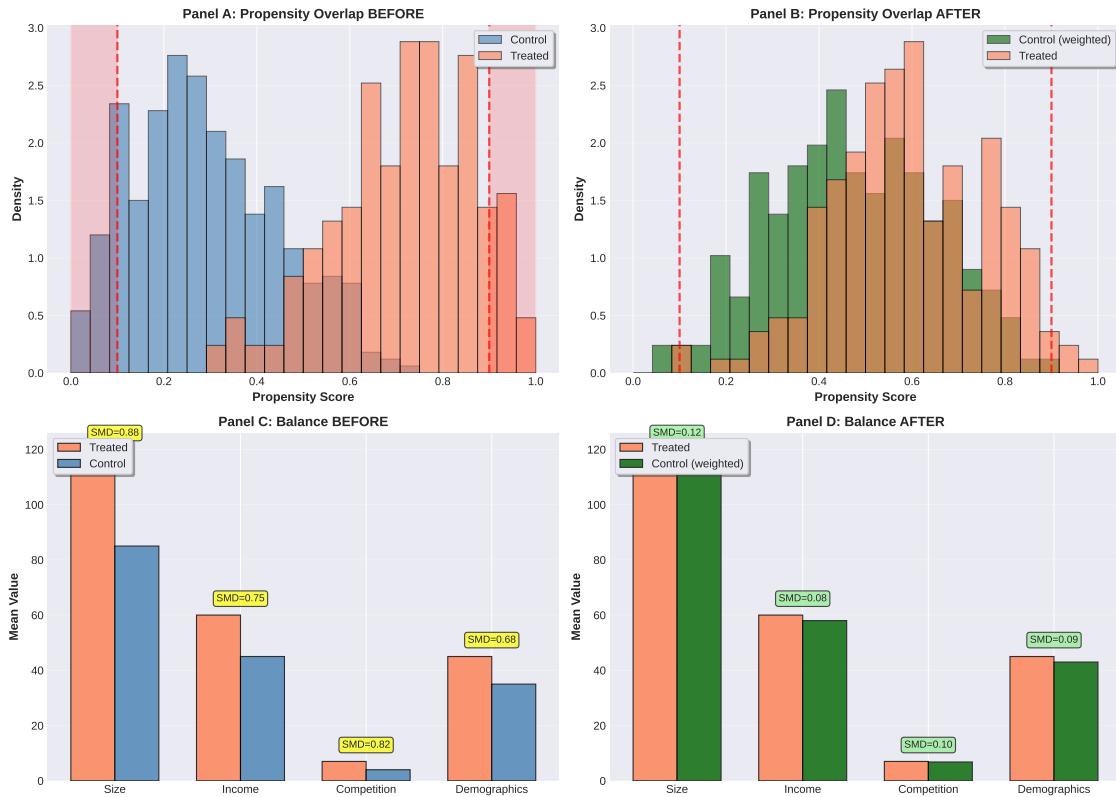
## When Diagnostics Fail

Diagnostics do not always look favourable, and the protocol provides guidance for each failure mode. When pre-trend diagnostics suggest non-parallel trends, a naive DiD is not credible. Alternatives such as SC, SDID, and factor methods change the identifying assumptions, for example by relying on stability of latent factors and counterfactual fit. You should restate the estimand and assumptions rather than treating these approaches as parallel-trends-free fixes. Sensitivity analysis can bound the effect under plausible trend deviations. In some applications, the honest conclusion is that the design fails.

Poor propensity score overlap undermines ATE identification without extrapolation. Options include trimming to the region of common support and reporting the overlap-restricted estimand, changing the target to an ATT-type estimand, or using bounds that acknowledge extrapolation. Changing the estimand changes the target population and must be reported explicitly. When leave-one-out diagnostics reveal that a single unit drives the result, the analyst should report the influence profile transparently, apply regularisation to spread weight, or conclude that the sample is too small for credible inference.

Specification sensitivity poses a different challenge. If the specification curve shows sign changes or wide dispersion, the result depends on arbitrary modelling choices. The honest response is to report the full range rather than a point estimate, pre-register the primary specification to avoid post-hoc selection, or acknowledge that the evidence is inconclusive. The goal throughout is not to force a significant result but to report what the data credibly support.

Table 18.3 provides indicative starting points for data requirements by application domain. The required pre-period length and time granularity depend on variance, seasonality, the number of independent clusters or units, and treatment timing. Verify these requirements before committing to a design.



**Fig. 18.1** Diagnostic protocol: Overlap and balance assessment

(Illustrative.) Left panel shows propensity score distributions for treated (blue) and control (orange) units. The figure shows how you would report overlap and any chosen trimming thresholds. Right panel shows standardised mean differences (SMD) for key covariates before (red) and after (green) inverse-propensity-score weighting. Balance targets and trimming thresholds are design choices and should be justified in context.

**Table 18.3** Indicative data requirements by marketing application domain

Application	Unit	Time Granularity	Pre-Period	Key Variables
Geo-experiment	DMA/postcode	Weekly	26–52 weeks	Sales, demographics
Digital attribution	User-session	Daily	4–8 weeks	Conversions, exposures
Media mix	National	Weekly	2–3 years	Spend, sales, promotions
Loyalty programme	Store/customer	Monthly	12 months	Transactions, enrolment
Price elasticity	Store-SKU	Weekly	52 weeks	Price, quantity, costs
CLV attribution	Customer	Monthly	24 months	Revenue, tenure, channel
Dynamic pricing	Route-day	Daily	90 days	Price, bookings, capacity
Platform experiment	User-market	Hourly/daily	2–4 weeks	Transactions, supply

### 18.3 Advertising Incrementality: Geo-Experiments

The measurement of advertising effectiveness is plagued by the confusion between attribution and incrementality. Attribution credits a sale to a touchpoint, while incrementality measures the causal lift generated by that touchpoint.

#### Estimand Definition

We define incremental lift as an application-specific instance of the average treatment effect on the treated. Let  $c = 1, \dots, G$  index geographies (for example, DMAs), and let  $t = 1, \dots, T$  index weeks. Let  $D_{ct} \in \{0, 1\}$  denote assignment of geography  $c$  in week  $t$  to treatment (ads on) or hold-out (ads off). In a standard geo hold-out, assignment is time-invariant, so  $D_{ct} = D_{c1}$  for all  $t$  in the campaign window. Let  $Y_{ct}(d)$  denote revenue in geography  $c$  at time  $t$  under treatment status  $d \in \{0, 1\}$ , with  $Y_{ct}(1)$  and  $Y_{ct}(0)$  the corresponding potential outcomes. We focus on the extensive margin. Does advertising generate incremental revenue relative to no advertising? The intensive margin, where we study how spend intensity maps into outcomes, requires dose-response methods (Chapter 14).

**Definition 18.1 (Incremental Lift)** Let  $\mathcal{T}_{\text{post}}$  denote the set of post-period weeks used for measurement. The Incremental Lift is the post-period average treatment effect on the treated:

$$\text{ATT}_{\text{post}} = \mathbb{E}[Y_{ct}(1) - Y_{ct}(0) \mid D_{c1} = 1, t \in \mathcal{T}_{\text{post}}],$$

where  $Y_{ct}(1)$  is revenue under ad exposure and  $Y_{ct}(0)$  is counterfactual revenue with zero exposure.

**Definition 18.2 (Incremental Return on Ad Spend)** The business metric of interest is Incremental Return on Ad Spend, defined as the ratio of total incremental revenue over  $\mathcal{T}_{\text{post}}$  to total ad spend over the same horizon:

$$\text{iROAS} = \frac{\sum_{c:D_{c1}=1} \sum_{t \in \mathcal{T}_{\text{post}}} \mathbb{E}[Y_{ct}(1) - Y_{ct}(0) \mid D_{c1} = 1]}{\text{Total Ad Spend}}.$$

We estimate it using the plug-in estimator

$$\widehat{\text{iROAS}} = \frac{\sum_{c:D_{c1}=1} \sum_{t \in \mathcal{T}_{\text{post}}} \widehat{\Delta}_{ct}}{\text{Total Ad Spend}},$$

where  $\widehat{\Delta}_{ct}$  is the estimated incremental revenue (in currency units) for geography  $c$  at time  $t$ . An iROAS of 2.0 means each pound spent generated two pounds of incremental revenue. This differs fundamentally from MMM elasticities or multi-touch attribution weights, which are often correlational rather than causal.

## Identification Strategy

If geographies are randomly assigned to hold-out, identification follows from the assignment mechanism. If assignment is pair-matched rather than fully random, you still need an explicit assumption about the untreated counterfactual. In practice this takes a parallel-trends form.

**Assumption 88 (Randomisation-Based Unconfoundedness)** In a randomised geo hold-out, assignment is independent of potential outcomes:

$$\{Y_{ct}(0), Y_{ct}(1)\}_{t=1}^T \perp\!\!\!\perp D_{c1}.$$

**Assumption 89 (Parallel Trends for Matched-Market Designs)** In the absence of the ad campaign, the trend in untreated potential outcomes is the same for treated and control regions:

$$\mathbb{E}[Y_{ct}(0) - Y_{c,t-1}(0) | D_{c1} = 1] = \mathbb{E}[Y_{ct}(0) - Y_{c,t-1}(0) | D_{c1} = 0], \quad \forall t.$$

This is the geo-specific version of the parallel trends assumption (see Assumption 6 for the general formulation). Pre-period outcomes provide diagnostics (not proof) for the plausibility of parallel trends, but the identifying restriction is still an assumption about the post-period counterfactual.

**Assumption 90 (No Interference Across Regions)** The potential outcome of region  $c$  depends only on its own treatment status, not on spillovers from other regions. Writing spillovers via an exposure mapping  $h_c(D_{-c,t})$ , this restriction takes the form

$$Y_{ct}(d, h) = Y_{ct}(d) \quad \text{for all } h.$$

This assumption fails when ads in treated regions drive sales in control regions (for example, customers travel across boundaries), when platform algorithms redistribute demand geographically, or when competitors respond asymmetrically to the campaign. Violations bias estimates downward if treatment leaks to controls. They can also bias in either direction under strategic responses. Buffer regions, which exclude geographies adjacent to treated areas, provide a partial safeguard. Chapter 11 gives a formal treatment of interference.

## Temporal Structure: Static vs. Dynamic Effects

The taxonomy in Section 18.1 distinguishes static from dynamic treatment effects. Many geo-experiments implicitly assume static effects: the outcome  $Y_{ct}(d)$  depends only on contemporaneous exposure. In practice, advertising often exhibits carryover, with effects that persist or accumulate beyond the exposure period.

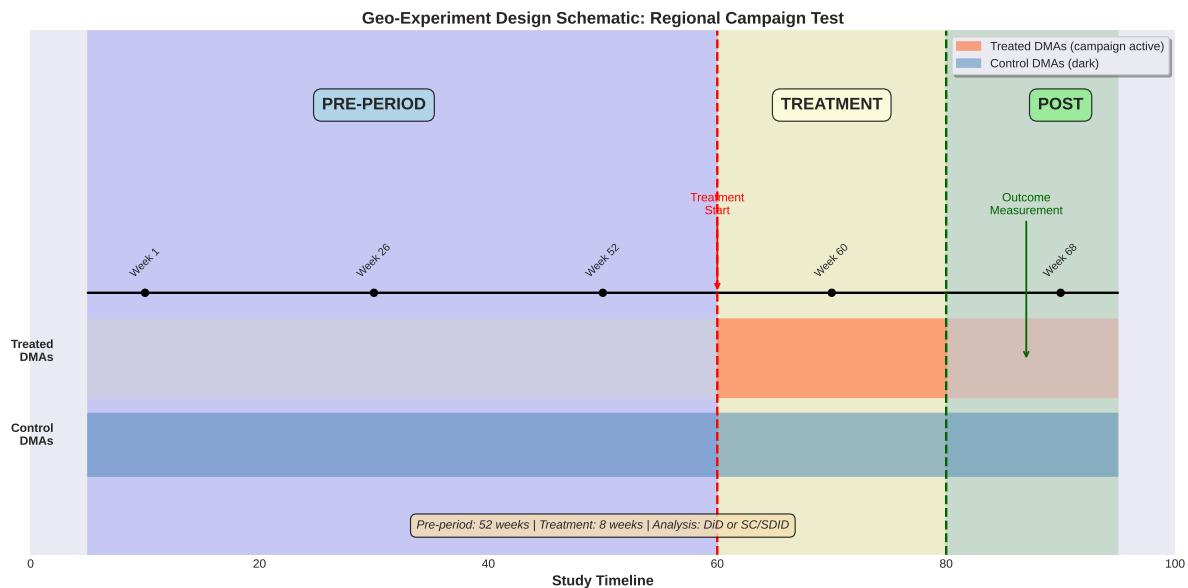
*Remark 18.3 (Carryover and Post-Period Window)* The choice of post-period window materially affects iROAS estimates. Consider two scenarios. For performance advertising such as paid search and retargeting, effects are largely contemporaneous and a short post-period of 4–8 weeks during the campaign captures

most of the lift. For brand advertising such as television and out-of-home, effects accumulate in a brand equity stock that decays slowly. A short post-period underestimates total lift, and extending measurement to 12–26 weeks post-campaign captures delayed conversions and habit formation.

When carryover is substantial, the path-dependent framework  $Y_{ct}(\underline{d}_c^t)$  from Chapter 10 applies. Distributed-lag models or adstock transformations (Section 18.5) can decompose short-run and long-run effects. At minimum, analysts should report iROAS at multiple post-period windows to assess sensitivity.

## Design and Estimation

The preferred design is the geo-experiment. We select a set of geographic units—DMAs, postcode sectors, or store catchments—and randomise them into treatment and control groups. The treatment group receives the ad campaign, while the control group receives no advertising exposure (a “dark” hold-out).



**Fig. 18.2** Geo-experiment design schematic. The timeline shows the division into pre-treatment (baseline) and treatment periods. Geographic units are assigned to treatment (ad exposure) or control (hold-out), allowing for difference-in-differences or synthetic control estimation.

For observational settings where randomisation is impossible, we use an observational geo-panel, exploiting natural variation in ad budgets across regions. Table 18.4 compares the main design options.

*Remark 18.4 (Power and Minimum Detectable Effect)* Geo-experiments operate with small effective sample sizes because the number of independent geographies  $G$  is often modest. With  $G_{\text{treated}} = 5$  DMAs and cluster-level randomisation, the degrees of freedom for inference are far below a typical user-level A/B test. Before committing to a design, compute the minimum detectable effect (MDE) at your planned power target. See

**Table 18.4** Design comparison for advertising incrementality measurement

Design	Advantages	Limitations	When to Use
Randomised hold-out	Benchmark design, unbiased	Opportunity cost, requires scale	New campaigns, large budgets
Pair-matched markets	Balances on pre-period outcomes	Parallel trends still required	Few treated units, no full randomisation
Synthetic control	Single treated or aggregated treated unit OK	Donor-pool quality critical	Market-level pilots, rare events
SDID	Combines SC and DiD strengths	Requires both unit and time weights	Multiple treated units or treated groups

Section 3.7 for definitions and dependence corrections under clustering. Underpowered experiments produce wide confidence intervals and inconclusive results. When MDE exceeds the expected lift, consider pooling more regions, extending the pre-period, or accepting that the design will primarily inform decision-making through uncertainty bounds rather than a single decisive estimate.

The preferred estimators are synthetic control (SC) or synthetic difference-in-differences (SDID). These methods construct a synthetic counterfactual by weighting control units to match the pre-treatment trajectory of the treated units or of an aggregated treated group.

The estimation proceeds in three steps. First, we define the pre-treatment period and validate the quality of the synthetic match. Second, we estimate the synthetic counterfactual for the post-treatment period. Third, we calculate the lift as the difference between the observed and synthetic outcomes.

### Diagnostic Checklist

A valid geo-experiment requires careful diagnostics. The following checklist operationalises the general diagnostic framework from Chapter 17 for the geo-experiment setting.

**Box 18.2: Geo-Experiment Diagnostic Checklist**

- 1. Pre-period balance:** Report standardised mean differences (SMD) on baseline sales, population, and demographics, and interpret magnitudes in context.
- 2. Pre-trend diagnostics:** Examine pre-period trends and placebos. If you use event-time leads, report the lead coefficients and (optionally) a joint F-statistic as a diagnostic, and interpret it cautiously in light of low power and pre-test bias (Section 17.4).
- 3. Synthetic control fit:** Report RMSPE<sub>pre</sub> relative to baseline sales and show the full pre-period fit.
- 4. Weight concentration:** Report the effective number of donors  $N_{\text{eff}} = 1 / \sum_j w_j^2$  as a diagnostic for donor concentration (Section 17.6).
- 5. Placebo analyses:** Run in-space placebos (apply SC to each donor) and in-time placebos (use a pseudo-treatment date). Report where the treated unit falls in the placebo distribution.
- 6. Leave-one-out:** Re-estimate excluding each major donor and report how sensitive results are to donor removal.
- 7. Spillover check:** Check for effects in geographies adjacent to treated regions using buffer analyses (Section 17.5).

### Extended Case Study: Regional Television Campaign

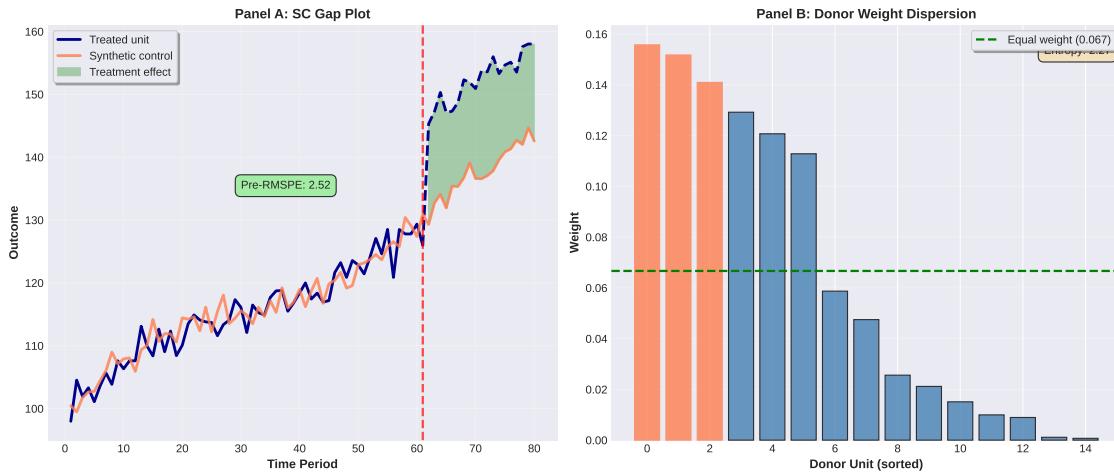
We illustrate the geo-experiment protocol with a retailer running a regional television campaign pilot. The numbers below are illustrative.

**Setting and Data.** A national retailer runs a TV campaign pilot in 5 DMAs (treated) against 45 control DMAs. The data consist of weekly store-level sales aggregated to the DMA level. The pre-period spans 52 weeks, providing a stable baseline. The treatment period is 8 weeks. For estimation, the five treated DMAs are aggregated into a single treated unit, with weights proportional to baseline sales.

**Design.** We apply SDID with penalised synthetic control weights. The regularisation parameter is selected via cross-validation on the pre-period fit.

**Results.** The estimated lift is 8.2% (95% CI: 3.1%, 13.3%), with the interval computed using an inference procedure aligned with the design (for example, an in-space placebo distribution for synthetic estimators or randomisation inference for randomised hold-outs). Total incremental revenue is £2.4 million against campaign spend of £1.0 million, yielding  $\widehat{\text{iROAS}} = 2.4$ .

**Diagnostics.** Pre-period fit is strong (for example, RMSPE<sub>pre</sub> is 2.1% of baseline sales). Weight concentration gives  $N_{\text{eff}} = 4.2$ , indicating that the counterfactual is not dominated by a single donor. Leave-one-out analysis shows that no single DMA drives the estimate.



**Fig. 18.3** Synthetic-control estimation results. Top panel: gap plot showing the difference in sales between the aggregated treated unit and the synthetic counterfactual. Bottom panel: distribution of donor weights, indicating which control DMAs contribute most to the counterfactual.

In the in-space placebo distribution, the aggregated treated unit ranks 4th out of 46 units, suggesting that the observed gap is somewhat unusual relative to placebo gaps, though not decisive on its own.

**Conclusion.** The campaign generated a point estimate of 8.2% lift with  $\widehat{i\text{ROAS}} = 2.4$ . Diagnostics are generally supportive: pre-period fit is strong and the estimate is stable to leave-one-out perturbations. The placebo distribution provides only moderate separation between treated and placebo gaps. We conclude that the evidence is suggestive rather than definitive. A larger-scale replication with more treated DMAs would provide sharper inference. The business decision to continue investment should weigh this uncertainty against the opportunity cost of inaction.

## 18.4 Digital Attribution and Multi-Touch

While geo-experiments measure the total lift, digital attribution attempts to assign credit to individual touchpoints. The industry standard of “last-click” attribution is purely descriptive and ignores the causal reality that many customers would have converted without the final click.

### Estimand: Marginal Channel Contribution

The phrase “marginal channel contribution” is ambiguous in multi-channel systems with interactions. We distinguish two causal questions. First, a *controlled direct effect* of channel  $\ell$  holds other channels fixed at a specified exposure pattern. Second, a *total effect* of a channel policy allows exposure to channel  $\ell$  to change exposures to other channels through substitution, mediation, or platform optimisation. Most business questions are about the total effect. The controlled direct effect can still be operationally useful when you can intervene on one channel while holding others approximately fixed.

Let  $Y_i$  denote the outcome for user  $i$  over a fixed campaign window (for example, a conversion indicator or incremental CLV aggregated over that window). Let  $D_{i\ell} \in \{0, 1\}$  indicate whether user  $i$  was exposed to channel  $\ell$  during the window, and let  $D_{i,-\ell}$  denote the vector of exposures to all other channels.

**Definition 18.3 (Marginal Channel Effect)** For channel  $\ell$ , a controlled direct effect is

$$\Delta_\ell^{\text{CDE}}(d_{-\ell}) = \mathbb{E}[Y_i(d_{-\ell}, 1) - Y_i(d_{-\ell}, 0)],$$

for a specified policy-relevant exposure pattern  $d_{-\ell}$ . This is not the same object as the total effect when channels interact.

This estimand differs from last-click attribution, which assigns full credit to the final touchpoint regardless of the causal contribution of earlier exposures.

**Assumption 91 (No Cross-Channel Interference)** Exposure to channel  $\ell$  does not causally alter exposure to other channels in a way that changes the exposure pattern  $d_{-\ell}$  relevant for the controlled direct effect. In many digital ad stacks this restriction is mechanically false.

This assumption fails when channels interact causally. For example, a display ad can increase the probability that a user searches and thus sees a search ad. In such cases,  $\Delta_\ell^{\text{CDE}}(d_{-\ell})$  is a controlled direct effect that does not match “incremental channel value”. You should either redefine the estimand as a total effect of a channel policy, or explicitly model mediation across channels, which is outside this chapter’s scope.

*Remark 18.5 (Static vs. Path-Dependent Attribution)* The binary indicator  $D_{i\ell} \in \{0, 1\}$  treats channel exposure as static: either the user saw the channel or not. In practice, user journeys are sequences such as display → email → search → conversion. Two users with identical exposure sets {display, search} may convert at different rates depending on the order and timing of exposures. The path-dependent framework  $Y_{it}(d_i^t)$  from

Chapter 10 applies here, where  $\underline{d}_i^t$  records the ordered sequence of exposures for user  $i$  up to time  $t$ . Full attribution requires modelling the entire path, not just the exposure set. Shapley-value methods attempt this decomposition but impose strong additivity assumptions. See Chapter 10 for the causal notation for dynamic interventions.

## Identification Strategy

Identifying a channel effect requires variation in channel  $\ell$  exposure that is independent of potential outcomes. In observational data, this assumption is implausible: users who see more ads differ systematically from those who see fewer. Algorithms target high-value users. Engaged users seek out brand content. Intent drives both ad exposure and conversion.

**Assumption 92 (Conditional Unconfoundedness for Channel  $\ell$ )** Conditional on other channel exposures  $D_{i,-\ell}$  and covariates  $X_i$ , exposure to channel  $\ell$  is independent of potential outcomes:

$$Y_i(D_{i,-\ell}, d_\ell) \perp\!\!\!\perp D_{i\ell} \mid D_{i,-\ell}, X_i, \quad \text{for } d_\ell \in \{0, 1\}.$$

This is a channel-specific version of the conditional-unconfoundedness assumption (see Assumption 5 for the general formulation). The assumption is rarely credible in observational digital data. Instead, credible identification usually requires quasi-experimental or experimental designs that generate exogenous variation in exposure.

Table 18.5 summarises quasi-experimental strategies that exploit features of digital ad delivery to approximate exogenous variation.

**Table 18.5** Quasi-experimental identification strategies for digital attribution

Strategy	Source of Variation	Design
Budget exhaustion	Daily caps cause ads to go dark for late-in-day users	RDD at cap threshold
Platform outages	Unplanned outages create exogenous exposure gaps	DiD or event study
Auction bid variation	Randomness in real-time bidding	IV or natural experiment
Geo-randomisation	Randomised ad delivery at geo level	Cluster RCT, geo-experiment

## Estimation

Given a valid identification strategy, estimation proceeds according to the design. For budget exhaustion, a regression discontinuity design compares users just above and below the daily cap threshold, using standard

RDD methods. This design targets a local effect for users whose exposure is marginal at the cap. For geo-randomisation, difference-in-differences or synthetic control applies at the geo level, with user-level outcomes aggregated to a geo-time outcome  $Y_{ct}$  (Section 18.3).

Inference should follow the sampling unit implied by the assignment mechanism. If the platform randomises or perturbs delivery at the geo level, treat geographies as the independent clusters  $c = 1, \dots, G$  and use cluster-robust or randomisation-based inference. If the design creates plausibly independent user-level assignment, user-level inference may be appropriate. In many digital systems, interference through auction spillovers, frequency caps, and algorithmic re-optimisation makes user-level independence doubtful. State the sampling unit explicitly.

The key challenge is that digital data are high dimensional: many channels, many touchpoints, and complex user journeys. Double machine learning (Chapter 12) can reduce functional-form bias in nuisance models and support valid inference *conditional on* unconfoundedness and overlap. It does not address selection on unobservables induced by targeting or platform optimisation.

## Measurement Under Privacy and Platform Constraints

The methods above assume access to user-level exposure and outcome data. This assumption is increasingly unrealistic.

*Remark 18.6 (The Privacy and Platform Opacity Problem)* Three forces are fragmenting digital measurement. First, privacy regulation and platform restrictions—iOS App Tracking Transparency, GDPR, and third-party cookie deprecation—limit cross-site and cross-device tracking, so advertisers increasingly receive aggregate-only reports rather than user-level logs. Second, platforms impute conversions using machine learning when deterministic tracking fails, and these modelled conversions are themselves noisy proxies for the true outcomes  $Y_i$ , introducing measurement error of unknown magnitude. Third, platform algorithm opacity means that audience-finding algorithms, CPC-optimising bidders, and other proprietary systems introduce latent selection into who sees which ads; the advertiser observes outcomes but not the algorithm’s internal targeting logic, creating a suspected but unquantifiable source of confounding.

These constraints do not eliminate the need for causal measurement. They elevate it. When platform metrics are unreliable, geo-experiments and MMM calibration (Section 18.5) become essential anchors. The methods in this chapter should be viewed as a hierarchy. Use experiments where possible. Use quasi-experiments where data permit. Use calibrated models where neither is feasible, acknowledging the uncertainty that platform opacity introduces.

## The Digital Measurement Paradoxes

Digital advertising presents unique measurement challenges that arise from the tension between granular data and causal identification.

### Box 18.3: The Digital Ad Paradox

**Setting:** Digital advertising promises precise, real-time measurement: impressions, clicks, viewability, conversions, and path data. Yet attribution remains fragile. Last-click rules ignore the contribution of earlier touchpoints. Viewability and fraud issues mean many recorded impressions are never seen by humans. Retargeting may chase consumers who would have converted anyway.

From a causal perspective, the key estimands are channel effects defined under an explicit intervention and dose-response functions  $\mu(a)$  for intensity, not raw platform metrics. The *digital ad paradox* is that as measurement becomes more granular, naive interpretation becomes more misleading. Finely measured but confounded metrics can distort budgets more than coarse but causally anchored measures.

**Measurement implications:** Reconciling rich digital data with credible causal inference requires combining three elements. These are experiments and quasi-experiments to identify causal effects for key channels and formats. They are structural or MMM-style models that aggregate these effects across time and touchpoints. They are disciplined attribution rules that inherit constraints from the first two elements. Rather than optimising on last-click ROAS or click-through rates, firms should calibrate attribution models to experimental lift and report uncertainty around channel contributions.

### Box 18.4: The Attention Paradox in Digital Advertising

**Setting:** A display campaign runs on a social platform. For each user  $i$ , let  $A_i$  denote attention intensity over the campaign window (for example, total viewable impressions or total seconds in view) and let  $Y_i$  be an outcome such as conversions or incremental CLV. Ideal causal dose-response curves compare  $Y_i(a)$  across different levels of attention  $a$ .

**Dose-response pattern:** Using an identification strategy (for example, budget caps that shut off impressions at different times), we estimate the causal response function  $\mu(a) = \mathbb{E}[Y_i(a)]$ . In some settings, estimates suggest diminishing marginal returns and possible wear-out at high frequency. This pattern is design- and context-dependent, and it requires enough identifying variation at high doses. When  $\mu(a)$  is smooth, the marginal effect  $\partial\mu(a)/\partial a$  can become small or negative at high  $a$ . The *attention paradox* arises because platform metrics (impressions, view time) keep rising with  $a$ , while the incremental causal effect  $\mu(a + \Delta a) - \mu(a)$  can flatten.

**Measurement implication.** Treating attention as a causal dose rather than as a simple KPI forces analysts to estimate  $\mu(a)$  non-parametrically or with flexible splines, report the range where marginal returns are positive, and avoid optimising on raw attention metrics. Integrating this dose-response perspective with geo-experiments and MMM (Section 18.5) aligns attention optimisation with true incremental outcomes rather than engagement for its own sake.

## Evidence from Large-Scale Studies

Several landmark studies have used experiments and credible quasi-experiments to benchmark advertising effectiveness.

### Box 18.5: What Do Large Studies Say About Advertising ROI?

**Television:** Shapiro et al. [2021] analyse TV advertising for hundreds of brands across many categories. Using observational panel data with a transparent product-selection protocol and extensive robustness checks, they find shorter-run TV elasticities that are substantially smaller than those in much of the earlier MMM literature and a sizeable share of statistically insignificant or even negative estimates. Many brands appear better off with their observed TV spend than with no advertising at all, but the evidence points to considerable over-investment at the margin.

**Search:** Blake et al. [2015] run large-scale field experiments at eBay that randomly shut off paid search in parts of the business. They show that naive observational estimates vastly overstate returns because search and purchase are jointly driven by underlying intent. Brand-keyword ads show no measurable short-run benefit, while non-brand keywords influence new and infrequent users but not frequent buyers. Since frequent buyers account for most of the advertising expense, average marginal ROI is often negative.

**Social and display:** Gordon et al. [2019] use large Facebook experiments to benchmark observational methods. Simple exposed-versus-unexposed comparisons, and even some sophisticated models, can be badly biased relative to experimental lift. The results underscore that online ad data are highly confounded: without randomisation or strong design structure, log-level ROAS and platform metrics are unreliable guides to causal impact.

**Measurement economics and market valuation:** Lewis and Rao [2015] document that the economics of measuring ad returns are often unfavourable: true effects may be modest and noisy, requiring very large samples to detect reliably. Erickson and Jacobson [1992] take a complementary perspective, relating discretionary spending on advertising and R&D to stock-market returns. They find that, on average and under their design and modelling assumptions, markets value advertising as a long-run intangible asset that contributes to firm value, even if marginal per-campaign lift is hard to pin down.

Taken together, these studies suggest that marginal short-run advertising effects are typically smaller and more heterogeneous than naive metrics imply, that observational methods can seriously mismeasure ROI, and that long-run brand and firm value can still be sensitive to advertising investments. The framework in this chapter—combining experiments, quasi-experiments, and dynamic MMM or brand-stock models—is designed to reconcile these facts.

## 18.5 Media Mix Modelling with Causal Foundations

Media Mix Modelling (MMM) uses time-series econometrics to estimate the effectiveness of various marketing channels. Traditional MMM often suffers from spurious correlation due to omitted seasonality or promotions. A causal MMM approach integrates structural assumptions and experimental priors.

*Remark 18.7 (MMM in the Two-Dimensional Taxonomy)* MMM is the canonical case that spans both dimensions of the taxonomy in Section 18.1. On the identification axis, MMM faces *endogeneity*: budgets respond to anticipated sales, creating simultaneity bias. On the temporal axis, MMM involves *dynamic effects*: advertising builds brand equity stocks that decay slowly, requiring adstock and saturation modelling. The identification strategy must address endogeneity (via experimental calibration or instruments). The estimator must accommodate dynamics (via distributed lags and state-space structures). Conflating these dimensions—using a dynamic model without addressing endogeneity, or addressing endogeneity with a static estimator—yields unreliable results.

### Adstock and Saturation

We model the delayed effect of advertising using adstock transformations, following the dynamic-treatment framework in Chapter 10.

**Definition 18.4 (Geometric Adstock)** The adstock for channel  $j$  at time  $t$  is given by

$$A_{jt} = D_{jt} + \lambda_j A_{j,t-1}, \quad \lambda_j \in [0, 1),$$

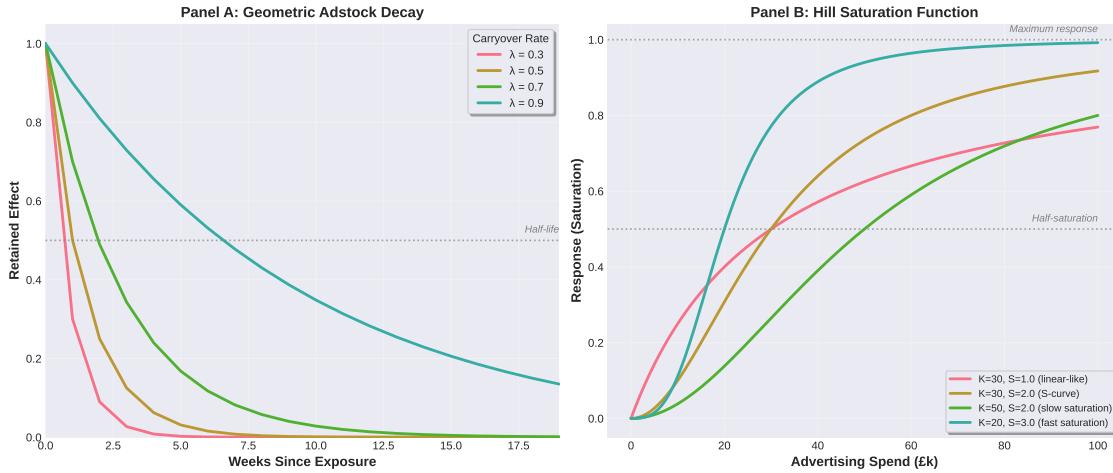
where  $D_{jt}$  is spend or exposure and  $\lambda_j$  is the carryover rate. The half-life of the adstock is  $h_j = -\log(2)/\log(\lambda_j)$  periods.

We also model diminishing returns using saturation functions.

**Definition 18.5 (Hill Saturation Function)** The Hill function maps adstock  $A$  to response  $f(A)$ :

$$f(A) = \frac{A^\alpha}{K^\alpha + A^\alpha},$$

where  $K > 0$  is the half-saturation point (the adstock level at which response reaches 50% of maximum) and  $\alpha > 0$  controls the steepness. When  $\alpha > 1$ , the curve is S-shaped with an inflection point. When  $\alpha = 1$ , it reduces to the Michaelis–Menten form.



**Fig. 18.4** Media transformation functions. Left panel: geometric adstock decay showing how an impulse of spend dissipates over time. Right panel: Hill saturation curve mapping adstock to incremental sales, illustrating diminishing returns.

## Identification and Calibration

MMM coefficients have causal interpretation only under strong assumptions about the data-generating process.

**Assumption 93 (Exogeneity of Media Spend)** A causal interpretation of MMM coefficients requires a credible source of exogenous variation in spend after conditioning on controls. A strong working version of this requirement is that, conditional on controls (seasonality, promotions, competitor activity), media spend  $D_{jt}$  is uncorrelated with the error term:

$$\mathbb{E}[\varepsilon_t | D_{1t}, \dots, D_{Jt}, X_t] = 0,$$

where  $X_t$  collects control variables.

This assumption fails when spend responds to anticipated demand, for example increasing television advertising before Christmas because sales are expected to rise anyway. In that case,  $D_{jt}$  and  $\varepsilon_t$  share common shocks and the regression coefficient reflects correlation rather than causal effect.

Two strategies can mitigate endogeneity. First, *experimental calibration* uses lift estimates from geo-experiments (Section 18.3) to inform Bayesian priors for MMM parameters, anchoring the model to externally identified causal effects. Second, *instrumental variables* may be feasible in rare settings with defensible exclusion restrictions, using exogenous variation in media costs or delivery as instruments for spend (see Remark 18.2 for scope limitations on IV in marketing).

Calibration is the preferred approach when experimental estimates are available. The geo-experiment provides a point estimate  $\hat{\Delta}_j$  for the incremental lift of channel  $j$  over a defined post-period window. The MMM prior is centred on the implied cumulative effect for that window, with variance reflecting experimental uncertainty. Because geo-experiment estimates include carryover effects if the post-period is sufficiently long

(Remark 18.3), the post-period window in the experiment should align with the adstock decay assumptions in the MMM to keep the calibration coherent.

## Bayesian Estimation and the Small-Sample Problem

MMM operates in a statistical regime fundamentally different from panel econometrics. A typical MMM dataset comprises weekly observations over two to five years, yielding  $T \approx 100\text{--}250$  periods. With ten media channels, each requiring adstock ( $\lambda_j$ ), saturation ( $K_j, \alpha_j$ ), and coefficient ( $\beta_j$ ) parameters, plus seasonality, trend, and controls, the analyst faces 40–60 parameters. Degrees of freedom are thin, multicollinearity between channels is endemic, and frequentist asymptotics provide poor guidance.

Dependence is central in this setting. The effective sample size is driven by the number of time periods  $T$  and by the strength of serial correlation in outcomes and marketing inputs. Any uncertainty statements must respect time-series dependence. At minimum, use HAC-style uncertainty for frequentist summaries. In Bayesian workflows, check that posterior uncertainty is not spuriously tight because the likelihood treats residuals as independent when the data are not.

This small- $T$  problem has three consequences. First, maximum-likelihood estimates of adstock decay and saturation curvature are weakly identified. The likelihood surface is flat, producing point estimates that are unstable across specifications. Second, standard errors derived from the Fisher information matrix understate true uncertainty. Third, overfitting is the default: high in-sample  $R^2$  coexists with poor out-of-sample prediction.

Bayesian methods address these pathologies not as a philosophical preference but as a practical necessity. Priors regularise weakly identified parameters, posterior distributions propagate uncertainty to downstream decisions, and hierarchical structures allow partial pooling across channels.

**Definition 18.6 (Bayesian MMM Specification)** Let  $Y_t$  denote the outcome (sales, conversions) in period  $t$ , and let  $A_{jt}$  be the adstocked media input for channel  $j = 1, \dots, J$  (Definition 18.4), where  $J$  is the number of channels. The observation equation is

$$Y_t = \alpha_0 + \sum_{j=1}^J \beta_j f_j(A_{jt}, K_j, \alpha_j) + X_t' \gamma + \varepsilon_t,$$

where  $\alpha_0$  is an intercept,  $f_j(\cdot)$  is the Hill saturation function (Definition 18.5), and  $X_t$  collects controls. As a working likelihood one may take  $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ . In MMM, residuals are often serially dependent, so uncertainty statements should allow for time-series dependence.

The prior structure is

$$\begin{aligned}\lambda_j &\sim \text{Beta}(a_\lambda, b_\lambda) \\ \log K_j &\sim \mathcal{N}(\mu_K, \sigma_K^2) \\ \alpha_j &\sim \text{Gamma}(a_\alpha, b_\alpha) \\ \beta_j &\sim \mathcal{N}(\mu_{\beta,j}, \sigma_{\beta,j}^2)\end{aligned}$$

Hyperparameters encode domain knowledge. For television,  $\lambda_j$  priors centre on half-lives of two to six weeks. For paid search, half-lives centre near one week. The saturation half-point  $K_j$  is scaled relative to observed spend ranges.

The prior on  $\beta_j$  deserves special attention. When experimental estimates are available, they provide a principled basis for prior specification.

**Definition 18.7 (Experimental Calibration as Prior Elicitation)** Suppose a geo-experiment (Section 18.3) yields an estimate  $\widehat{\Delta}_j$  for the incremental effect of channel  $j$  over a stated horizon and aggregation, with standard error  $\text{SE}(\widehat{\Delta}_j)$ . A coherent calibration defines an MMM-implied lift for that same experiment window. Let  $L_j(\beta_j, \lambda_j, K_j, \alpha_j)$  denote the incremental lift implied by the MMM for channel  $j$  over the experiment's horizon under the experiment's spend perturbation, computed using the same adstock and saturation transformations.

Calibration then anchors the MMM at the level of lift, rather than equating a regression coefficient to an experimental estimand:

$$L_j(\beta_j, \lambda_j, K_j, \alpha_j) \sim \mathcal{N}(\widehat{\Delta}_j, \text{SE}(\widehat{\Delta}_j)^2 + \sigma_{\text{drift}}^2),$$

where  $\sigma_{\text{drift}}^2$  accounts for drift between the experimental period and the MMM estimation window. In practice, you can implement this by choosing priors for  $\beta_j$  (and, if needed,  $\lambda_j, K_j, \alpha_j$ ) so that the induced prior over  $L_j(\cdot)$  matches the experimental evidence.

Calibration converts MMM from an exercise in curve-fitting to a synthesis of experimental and observational evidence. Without calibration, MMM coefficients primarily reflect correlation patterns that may be confounded. With calibration, the posterior is anchored to causal estimates and deviations from the anchor are penalised.

Posterior inference typically proceeds via Markov chain Monte Carlo or variational methods. These models are implementable in probabilistic-programming frameworks (Stan, PyMC, NumPyro), but they can be computationally demanding and sensitive to priors and parameterisations. Convergence diagnostics and prior predictive checks are not optional. The output is a joint posterior over all parameters, from which the analyst extracts marginal response curves with uncertainty bands that show how response varies with spend for each channel, return-on-investment distributions that enable probabilistic statements such as “there is a 90% probability that TV ROI exceeds 1.5,” and budget recommendations that integrate over posterior uncertainty so allocations are less sensitive to parameter risk.

**What Bayesian methods do not solve.** Regularisation and uncertainty quantification do not remedy identification failures. If media spend is endogenous—planned to coincide with anticipated demand—the

posterior will be centred on a biased value. No amount of computation fixes this unless the priors are anchored to credible external variation through experimental calibration, or the design supplies defensible instruments. The “causal” in causal MMM derives from the identification strategy, not from the estimation framework.

*Remark 18.8 (Hierarchical Priors Across Channels)* When the number of channels is large, hierarchical priors improve estimation for channels with limited temporal variation. Let  $\beta_j \sim \mathcal{N}(\bar{\beta}, \sigma_\beta^2)$ , where  $\bar{\beta}$  and  $\sigma_\beta^2$  are themselves given hyperpriors. Partial pooling shrinks noisy channel estimates toward the population mean, reducing mean squared error at the cost of introducing controlled bias. This is particularly valuable for long-tail digital channels with sporadic spend.

### Diagnostic Checklist for MMM

MMM requires careful validation before coefficients can be trusted for budget allocation.

*Remark 18.9 (Input Data Quality and Platform Opacity)* MMM reliability depends on input data quality. Platform-reported spend may not reflect true exposure: viewability issues, modelled reach, and opaque allocation algorithms (Remark 18.6) introduce measurement error. When digital channels dominate the media mix, it is important to use first-party spend data where available, treat platform-reported metrics explicitly as noisy proxies with appropriate uncertainty, and anchor digital-channel coefficients more heavily to experimental calibration. Measurement error and opacity in the inputs propagate directly into biased MMM coefficients if left unmodelled.

## Box 18.6: MMM Diagnostic Checklist

- 1. In-sample fit:** Report  $R^2$  and MAPE. Poor fit suggests missing variables or misspecified dynamics.
- 2. Out-of-sample validation:** Hold out the final 10–20% of periods. Compare predicted and actual sales. Large errors indicate overfitting or structural breaks.
- 3. Coefficient plausibility:** Check that elasticities are in plausible ranges informed by external evidence and calibration. Treat any numeric cut-offs as heuristics rather than rules.
- 4. Experimental calibration:** Compare MMM-implied lift to geo-experiment lift for channels with experimental estimates. Large discrepancies suggest confounding, misspecified dynamics, or a mismatch in horizon and aggregation.
- 5. Adstock and saturation stability:** Re-estimate with alternative  $\lambda_j$  and  $K_j$  values. Large sensitivity indicates weak identification of dynamics.
- 6. Residual diagnostics:** Use autocorrelation diagnostics (for example, Durbin–Watson and Ljung–Box statistics) and check for heteroskedasticity. Serial dependence distorts standard errors and undermines inference.
- 7. Counterfactual reasonableness:** Simulate a 50% budget cut for each channel. Assess whether the implied sales declines are consistent with business intuition and experimental evidence.

**Case Study: Brand Equity as a Dynamic Stock**

Brand equity can be formalised as a stock variable that accumulates past marketing investment and decays slowly over time. Returning to a commodity category such as baked beans with multiple brands  $b = 1, \dots, B$ , stores  $s = 1, \dots, S$ , and weeks  $t = 1, \dots, T$ , let  $B_{bt}$  denote an (observed or latent) brand-equity index for brand  $b$  at time  $t$ . This index may be proxied by a brand-tracking score (awareness or consideration) or treated as an unobserved state in a state-space model.

Advertising contributes to  $B_{bt}$  through an adstocked input. Let  $D_{bt}$  be weekly advertising spend for brand  $b$  and  $A_{bt}$  its geometric adstock, as in the previous subsection:

$$A_{bt} = D_{bt} + \lambda_b A_{b,t-1}, \quad \lambda_b \in [0, 1].$$

We then specify a simple brand-stock evolution equation

$$B_{bt} = \rho_b B_{b,t-1} + \kappa_b A_{bt} + u_{bt}, \quad |\rho_b| < 1,$$

where  $\rho_b$  governs the persistence of brand equity,  $\kappa_b$  translates adstock into brand-stock increments, and  $u_{bt}$  captures other shocks (earned media, competitor actions).

Brand equity in turn shifts demand and willingness to pay. Let  $Q_{bst}$  denote quantity sold and  $P_{bst}$  price for brand  $b$  in store  $s$  and week  $t$ . A demand equation that links sales to price, brand equity, and controls is

$$\log Q_{bst} = \alpha_s + \gamma_t + \delta_b + \beta_p \log P_{bst} + \theta B_{bt} + X'_{bst} \eta + \varepsilon_{bst},$$

where  $\alpha_s$  and  $\gamma_t$  are store and week fixed effects,  $\delta_b$  is a time-invariant brand intercept, and  $X_{bst}$  collects short-run drivers such as promotions and displays. The coefficient  $\theta$  measures how a unit increase in brand equity shifts log sales holding price fixed. Combining the stock and demand equations, a one-period pulse in adstock  $A_{bt}$  has a long-run effect on brand equity of  $\kappa_b/(1 - \rho_b)$  and a corresponding long-run effect on log sales of  $\theta\kappa_b/(1 - \rho_b)$ .

In some discrete-choice demand models, the brand equity index can be interpreted as a component of mean utility. This is a modelling interpretation, not an identification result. Holding covariates fixed, the implied long-run brand premium—the price increase that leaves expected demand unchanged after a permanent increase in  $B_{bt}$ —solves approximately

$$\beta_p \log \left( 1 + \frac{\Delta p_b}{p_{bst}} \right) \approx -\theta \frac{\kappa_b}{1 - \rho_b},$$

linking dynamic advertising investment to a steady-state willingness-to-pay premium. In practice, parameters  $(\lambda_b, \rho_b, \kappa_b, \theta, \beta_p)$  can be estimated using panel data on prices, quantities, and advertising, with experimental or quasi-experimental variation in spend anchoring  $\kappa_b$  and  $\theta$ . This case study treats brand equity not as a vague intangible but as a measurable stock that mediates between marketing actions and long-run pricing power. The static demand-based brand premium model in Chapter 13, Section 13.11, can be viewed as taking  $B_{bt}$  as given and inferring the implied willingness-to-pay premium at a point in time.

## Box 18.7: The Brand vs Performance Paradox in Budget Allocation

**Setting.** Extend the brand-stock model to include a short-lived performance channel (for example, paid search or retargeting). Let  $A_t^B$  and  $A_t^{\text{perf}}$  denote adstocked brand and performance spend for a focal brand, and let  $B_t$  be the brand equity stock evolving as

$$B_t = \rho B_{t-1} + \kappa A_t^B + u_t,$$

while log demand follows

$$\log Q_t = \alpha + \beta_{\text{perf}} A_t^{\text{perf}} + \theta B_t + X'_t \eta + \varepsilon_t.$$

The performance channel  $A_t^{\text{perf}}$  generates strong contemporaneous response  $\beta_{\text{perf}}$  but little persistence, whereas brand spend  $A_t^B$  feeds into the stock  $B_t$  and, via  $\theta$ , into long-run sales and willingness to pay.

**Estimated dynamics.** Suppose MMM and dynamic-panel estimates imply a short-run elasticity of performance spend  $\beta_{\text{perf}} = 0.12$  with a long-run multiplier  $\text{LRM}_{\text{perf}} \approx 1.1$  reflecting little carryover, and a short-run elasticity of brand spend via  $B_t$  that is small on impact but with a long-run multiplier  $\text{LRM}_B \approx 3.5$  once  $B_t$  reaches steady state.

**Budget experiment.** With a fixed budget  $M$  per period, compare two steady-state policies. A performance-heavy allocation directs  $0.8M$  to  $A_t^{\text{perf}}$  and  $0.2M$  to  $A_t^B$ , yielding high immediate sales response but modest contribution to  $B_t$  and to the long-run brand premium  $\Delta p_b$ . A brand-heavy allocation directs  $0.2M$  to  $A_t^{\text{perf}}$  and  $0.8M$  to  $A_t^B$ , yielding lower short-run incremental sales but much higher steady-state  $B_t$  and thus higher long-run revenue and pricing power. Evaluated over a one-week or one-month horizon, the performance-heavy mix can exhibit superior ROI because  $\beta_{\text{perf}}$  dominates. Evaluated over a one-year horizon, the brand-heavy mix can dominate once the contribution of  $B_t$  to both quantity and the brand premium  $\Delta p_b$  is accounted for.

**Measurement implication.** The *brand vs performance paradox* arises when MMM or attribution focuses on short-run conversions or revenue and ignores the stock  $B_t$  and induced price premiums. Embedding brand equity as a state variable in MMM, and reporting both short-run and long-run elasticities and multipliers by channel, allows budget decisions to balance immediate ROI against the slower accumulation of brand value.

**Box 18.8: Joe Sixpack and Lifetime Brand Exposure**

**Setting.** Consider Joe Sixpack choosing beer in a supermarket. Any given Budweiser ad this week has a small immediate effect on his purchase probability, but the *cumulative* effect of years of Budweiser advertising makes him more likely to choose Bud over store brands. In the brand-stock model, Joe's latent predisposition is summarised by a brand equity stock  $B_t$  that evolves as

$$B_t = \rho B_{t-1} + \kappa A_t + u_t,$$

where  $A_t$  is adstocked Budweiser exposure in Joe's market,  $\rho$  is the persistence of brand equity, and  $\kappa$  translates advertising into increments in  $B_t$ . Joe's purchase probability in week  $t$  depends on  $B_t$  through the demand equation

$$\log s_t = \alpha + \beta_p \log P_t + \theta B_t + X'_t \eta + \varepsilon_t,$$

where  $s_t$  is Budweiser's share,  $P_t$  is price, and  $X_t$  collects promotions and competition.

**Lifetime effects in the model.** Under a roughly constant advertising schedule  $A_t = \bar{A}$ , the brand stock converges to a steady state  $B^* = \kappa \bar{A} / (1 - \rho)$ . A brand that has advertised heavily for decades maintains a high  $\bar{A}$  and thus a high steady-state  $B^*$ . A brand that has advertised sparingly has a lower  $\bar{A}$  and  $B^*$ . The *lifetime* effect of Budweiser advertising on Joe's choice is captured by the difference in steady states:

$$\Delta B^* = \frac{\kappa}{1 - \rho} (\bar{A}^{\text{Bud}} - \bar{A}^{\text{baseline}}),$$

and the corresponding long-run shift in log choice probability is  $\theta \Delta B^*$ . Even if any single week's ad has a tiny impact, the discounted sum of exposures over Joe's life is encoded in  $B_t$ .

**Measurement strategies.** In practice we do not observe individual lifetime exposure, but we can estimate  $(\rho, \kappa, \theta)$  using panel data on advertising, brand tracking, and sales at the market or cohort level. Long panels or historical variation in advertising intensity across regions allow us to approximate how markets with decades of heavy Budweiser advertising differ in baseline share from markets with weaker histories, holding current spend fixed. Combined with the steady-state formula, this provides an operational measure of Joe Sixpack's long-run advertising exposure and its contribution to brand choice.

## 18.6 Loyalty Programme Valuation

Loyalty programmes create a classic selection problem. Members typically spend more than non-members, but this difference combines the treatment effect of the programme with the selection effect that high spenders are more likely to join. Naive comparisons of member versus non-member spending confound these two effects.

*Remark 18.10 (Loyalty Programmes in the Taxonomy)* In the taxonomy of Section 18.1, loyalty programme valuation is primarily a *confounding* problem: high-value customers self-select into programmes and higher tiers. The identification challenge is severing the link between potential outcomes and enrolment. Secondary considerations include dynamics—programme effects may grow with tenure as habits form, or decay as initial excitement fades—and interference if member benefits affect non-member behaviour (for example, crowding at redemption counters).

### Estimand: Programme Incremental Value

We seek to estimate the causal effect of programme participation on customer spending, as a concrete instance of the ATT framework developed in Chapters 2 and 4. Let  $Y_{it}$  be spending by customer  $i$  in period  $t$ , and let  $D_{it} \in \{0, 1\}$  indicate whether customer  $i$  is a member in period  $t$ , with  $D_{it} = 1$  for members and  $D_{it} = 0$  for non-members. Let  $Y_{it}(d)$  be the potential spending in period  $t$  if membership status were set to  $d \in \{0, 1\}$ .

**Definition 18.8 (Programme Incremental Value)** The time-specific programme effect for members is

$$\text{ATT}_t = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1, t],$$

which captures the net effect of enrolment, behaviour change, and any discounts or rewards received for customers who are members at time  $t$ .

For tier upgrades, we need a distinct treatment definition to avoid overloading  $d \in \{0, 1\}$  with two meanings. Let  $Y_i^{\text{tier}}(d)$  denote next-period spending for customer  $i$  if tier status were set to  $d \in \{0, 1\}$ , where  $d = 1$  denotes Gold and  $d = 0$  denotes Silver. In a threshold-based RDD, the causal estimand is a local effect at the upgrade threshold, which we discuss below.

These estimands are static in the sense that they compare spending in a given period under alternative membership states. In practice, programme effects often vary with tenure. Early-period effects may reflect novelty and experimentation. Later-period effects may reflect habit formation or fatigue. When tenure effects are important, treat enrolment as a staggered-adoption event with join time  $G_i$  and use the event-time estimands and tools in Chapter 10.

### Identification Strategy 1: Staggered Rollout

If the programme is introduced in waves across stores or regions, we can use staggered difference-in-differences (Chapter 4). The treatment is at the store level (programme availability), but outcomes are measured at the customer level. This multi-level structure requires care: customer composition may differ across stores, and customers may shop at multiple locations. In practice, analysts often assign customers to a “home store” based on historical shopping patterns and restrict attention to transactions at that store to reduce cross-store interference.

**Assumption 94 (Parallel Trends for Programme Rollout)** In the absence of programme introduction, customer spending trends would be parallel across store cohorts. Let  $Y_{ist}(0)$  denote potential spending by customer  $i$  at store  $s$  in period  $t$  when store  $s$  has not launched the programme, holding the rest of the system fixed. If customers shop across stores and programme availability elsewhere affects behaviour, this restriction can fail via spillovers. Let  $H_s$  be the rollout cohort for store  $s$ . Then

$$\mathbb{E}[Y_{ist}(0) - Y_{is,t-1}(0) \mid H_s = g] = \mathbb{E}[Y_{ist}(0) - Y_{is,t-1}(0) \mid H_s = g'], \quad \forall g, g'.$$

For inference, it is helpful to index stores as clusters  $c = 1, \dots, G$ , with customers  $i$  nested within store  $c$ .

Under this assumption, customers at not-yet-treated stores serve as valid controls for customers at newly treated stores. Use heterogeneity-robust estimators such as Callaway and Sant’Anna [2021] or Sun and Abraham [2021] to avoid negative weighting when treatment effects vary across cohorts. Because rollout is at the store level, treat stores as the independent assignment units for inference. Cluster uncertainty at the store level, and interpret effective sample size in terms of the number of independent stores  $G$  rather than the number of customers. If customers shop across stores, state and diagnose the interference or contamination risk explicitly.

### Identification Strategy 2: Tier Threshold RDD

For tier upgrades, regression discontinuity exploits the sharp threshold rule.

**Assumption 95 (Continuity of Potential Outcomes at Threshold)** Let  $X_i$  be qualifying spend in the current year and let  $Y_i^{\text{tier}}(0)$  denote the potential outcome (for example next-year spend) in the absence of upgrade (Silver status). Potential outcomes are continuous functions of qualifying spend at the threshold  $x_0$ :

$$\lim_{x \downarrow x_0} \mathbb{E}[Y_i^{\text{tier}}(0) \mid X_i = x] = \lim_{x \uparrow x_0} \mathbb{E}[Y_i^{\text{tier}}(0) \mid X_i = x].$$

Since treatment (Gold status) jumps discontinuously at  $x_0$  (Gold for  $X_i \geq x_0$ , Silver otherwise), any discontinuity in observed outcomes  $\mathbb{E}[Y_i^{\text{tier}} \mid X_i = x]$  at  $x_0$  is attributable to the treatment.

This assumption fails if customers manipulate their spending to just exceed the threshold in ways correlated with potential outcomes.

*Remark 18.11 (Manipulation vs. Rational Bunching)* In loyalty programmes, customers can observe their progress toward thresholds and may rationally time purchases to qualify. This *bunching* is not manipulation in the sense of fraud. It is the intended programme design. Bunching invalidates RDD only if it is *selective*, meaning that customers who bunch have different potential outcomes than those who do not. A McCrary density diagnostic can detect bunching but cannot distinguish selective from non-selective bunching. Covariate balance at the threshold provides additional evidence. If observable characteristics are continuous at  $x_0$  despite bunching, selective manipulation is less likely.



**Fig. 18.5** Regression discontinuity design for loyalty tier valuation  
(Illustrative.) Average future spend plotted against qualifying spend, with a discontinuity at the \$500 threshold for Gold status. The vertical jump at the threshold represents the local causal effect of Gold status on subsequent spending. Bandwidth selection and density diagnostics are design choices and should be reported transparently. A density diagnostic can reveal bunching near the threshold. Bunching alone is not dispositive without evidence of selective manipulation.

## Estimation

For staggered rollout, apply the estimators from Chapter 4 and aggregate cohort–time effects into an overall programme ATT using appropriate weights and clear reporting of heterogeneity. For RDD, estimate the local average treatment effect at the threshold using local polynomial regression:

$$\widehat{\Delta}_{\text{RDD}} = \lim_{x \downarrow x_0} \hat{\mathbb{E}}[Y_i^{\text{tier}} | X_i = x] - \lim_{x \uparrow x_0} \hat{\mathbb{E}}[Y_i^{\text{tier}} | X_i = x].$$

Use Imbens–Kalyanaraman [Imbens and Kalyanaraman, 2012] or Calonico–Cattaneo–Titiunik [Calonico et al., 2014] bandwidth selectors, and report estimates for multiple bandwidths as a sensitivity check to demonstrate stability of the local effect.

## Diagnostic Checklist

### Box 18.9: Loyalty Programme Diagnostic Checklist

- For staggered rollout:** Run pre-trend diagnostics and placebos. If you use event-time leads, report the lead coefficients and interpret them cautiously in light of low power. Compare store characteristics across rollout waves to assess balance. Verify adequate observations in each cohort–time cell for support. Report cohort-specific effects and check for effect dynamics to assess heterogeneity and to detect early-adopter bias.
- For tier threshold RDD:** Run a McCrary density diagnostic to assess bunching at the threshold and interpret it in light of rational bunching versus selective manipulation. Compare demographics and prior behaviour at the threshold to assess covariate balance. Report estimates for multiple bandwidths to assess sensitivity. Use discontinuities at non-threshold values as falsification checks.

## Case Study: Retail Loyalty Tier Upgrade

We illustrate the RDD approach with a hypothetical retail loyalty programme. The numbers are illustrative and do not represent real data.

**Setting.** A retailer's loyalty programme awards Gold status to customers spending \$500 or more in a calendar year. Gold members receive 5% cashback (versus 2% for Silver). We estimate the causal effect of Gold status on next-year spending for customers near the qualifying threshold.

**Data.** Fifty thousand customers with qualifying spend between \$400 and \$600. The outcome is total spend in the following year.

**Results.** The estimated discontinuity at the \$500 threshold is \$142 (SE = \$38). Customers who just qualified for Gold spent about \$142 more in the following year than customers who just missed the threshold.

**Diagnostics.** The McCrary density diagnostic shows little bunching at \$500 in this illustration, which reduces concern about selective manipulation but does not eliminate it. Covariate balance at the threshold

looks stable: age, tenure, and prior-year spend show no visible discontinuities. The estimate is stable across bandwidth choices: \$128 (half bandwidth), \$142 (optimal), \$151 (double).

**Interpretation.** The \$142 incremental spend exceeds the incremental cashback cost (approximately  $\$15 = 3\% \times \$500$ ), yielding a positive programme ROI at the margin. This is a *local* effect for marginal customers close to the threshold. Extrapolating to all Gold members requires additional assumptions about effect heterogeneity and how far behaviour away from the cutoff resembles behaviour at the margin.

## 18.7 Promotion and Price Elasticity

Pricing is an equilibrium phenomenon. Prices respond to demand, and demand responds to prices. Naive regression of quantity on price yields biased estimates because high demand causes high prices (simultaneity) and unobserved quality drives both price and sales (omitted variables). Estimating causal price elasticity requires isolating exogenous variation in prices.

*Remark 18.12 (Price Elasticity in the Taxonomy)* In the taxonomy of Section 18.1, price-elasticity estimation is primarily an *endogeneity* problem: prices are set in response to anticipated demand, creating simultaneity bias. This is one domain where instrumental variables are relatively well established in marketing. Cost shifters can provide instruments with defensible exclusion restrictions (see Remark 18.2 for general IV scope limitations). Secondary considerations include dynamics—short-run and long-run elasticities differ as consumers adjust behaviour—and interference when products are substitutes and cross-elasticity estimation is required.

### Estimand: Price Elasticity

We define the causal effect of price on demand.

**Definition 18.9 (Price Elasticity)** Let  $Q_{jst}(d)$  denote the potential quantity that would be sold if the price of product  $j$  in store  $s$  at time  $t$  were set to  $d$ . The own-price elasticity at price  $d$  is

$$\eta_{jj}(d) = \frac{\partial \log \mathbb{E}[Q_{jst}(d)]}{\partial \log d},$$

when this derivative exists. When substitution across products matters, define a demand-system potential outcome  $Q_{jst}(d_j, d_k)$  under an intervention that sets  $(P_{jst}, P_{kst}) = (d_j, d_k)$  and define cross-price elasticities as derivatives of  $\log \mathbb{E}[Q_{jst}(d_j, d_k)]$  with respect to  $\log d_k$ . Elasticity is typically negative for own price ( $\eta_{jj} < 0$ ) and positive for substitutes ( $\eta_{jk} > 0$ ).

For temporary promotions, it is often more convenient to work with a discrete treatment indicator. Let  $D_{jst} = \mathbf{1}\{\text{product } j \text{ in store } s \text{ is promoted in week } t\}$ , and let  $Q_{jst}(d)$  be the potential quantity with  $d \in \{0, 1\}$ . The promotion ATT is

$$\text{ATT}_{\text{promo}} = \mathbb{E}[Q_{jst}(1) - Q_{jst}(0) | D_{jst} = 1],$$

which measures the causal lift from a temporary price reduction for promoted store–product–week cells.

*Remark 18.13 (Short-Run vs. Long-Run Elasticity)* The static elasticity  $\eta_{jj}$  captures the contemporaneous response of quantity to price. In practice, consumer response has temporal structure. First, adjustment lags mean consumers may not immediately notice price changes or may wait for confirmation that changes are

permanent, so long-run elasticity typically exceeds short-run elasticity in absolute value. Second, reference-price effects mean consumers evaluate prices relative to an internal reference formed from past prices, with response to price increases (losses) often exceeding response to equivalent decreases (gains) due to loss aversion. Third, stockpiling and intertemporal substitution mean that for storable goods, temporary promotions shift purchases across time, inflating short-run response but reducing post-promotion sales.

When dynamics matter, the analyst should estimate distributed-lag models or error-correction specifications that distinguish short-run from long-run effects, linking to Chapter 10.

### Identification Strategy 1: Instrumental Variables

The standard approach uses cost shifters as instruments for price.

**Assumption 96 (IV Exclusion Restriction)** Let  $Z_{jst}$  be an instrument (for example, commodity cost, transport cost, exchange rate). Conditional on fixed effects and controls, the exclusion restriction requires that  $Z_{jst}$  affects demand only through its impact on price.

**Assumption 97 (IV Relevance)** The instrument shifts prices after conditioning on fixed effects and controls, so the first stage has non-trivial variation.

Table 18.6 summarises common instruments for price.

**Table 18.6** Instruments for price-elasticity estimation

Instrument	Mechanism	Exclusion Restriction
Commodity costs	Oil, wheat, sugar prices shift production costs	Input costs affect demand only via retail price
Exchange rates	Currency fluctuations shift import costs	Requires arguing away channels through income, tourism, or macro conditions that also move demand
Hausman instruments	Prices of same product in other markets [Hausman, 1996]	Cost shocks common, demand shocks local. Vulnerable if common demand shocks exist
Regulatory shocks	Tax changes, tariffs, minimum prices	Policy changes exogenous to local demand

Hausman instruments in particular require caution: if common demand shocks or brand-wide promotions affect prices across markets, the exclusion restriction can fail and instruments can be weak.

### Identification Strategy 2: Promotion Timing

For promotion effects, we exploit the timing of promotional events.

**Assumption 98 (Exogenous Promotion Timing)** Conditional on store–product fixed effects and time effects, promotion timing is as-good-as-random with respect to demand shocks:

$$\mathbb{E}[\varepsilon_{jst} | D_{jst}, \alpha_{js}, \gamma_t] = 0.$$

This assumption is more credible for manufacturer-driven promotions planned months in advance than for retailer-initiated promotions, which may respond to inventory levels or competitive dynamics. It fails if retailers promote products that are selling poorly (endogenous promotion) or if promotions coincide systematically with demand peaks (for example, holiday displays). Use event-time leads and lags as diagnostics for anticipation and stockpiling. Treat these as diagnostics for the plausibility of the identifying assumption, not as proof.

## Estimation

For IV estimation, use two-stage least squares (2SLS):

$$\text{First stage: } \log P_{jst} = \pi_0 + \pi_1 Z_{jst} + \alpha_{js} + \gamma_t + u_{jst}, \quad (18.1)$$

$$\text{Second stage: } \log Q_{jst} = \beta_0 + \eta_{jj} \widehat{\log P_{jst}} + \alpha_{js} + \gamma_t + \varepsilon_{jst}, \quad (18.2)$$

where  $\alpha_{js}$  are store–product fixed effects and  $\gamma_t$  are time fixed effects.

For promotion effects, use a panel specification that exploits within store–product variation and includes appropriate fixed effects:

$$\log Q_{jst} = \alpha_{js} + \gamma_t + \beta_{\text{promo}} D_{jst} + X'_{jst} \beta + \varepsilon_{jst}.$$

With scanner-panel data, state the independent sampling unit for inference and cluster accordingly. If pricing and promotions are set at the store level, cluster at the store level. If shocks are correlated within store–product series, consider clustering at the store level and reporting sensitivity to alternative clustering schemes.

When store-level policies drive identifying variation, treat stores as the independent assignment units for inference. In that case, index stores as clusters  $c = 1, \dots, G$  and interpret precision in terms of the number of independent stores  $G$ . Do not treat the number of store–product–week observations as the effective sample size.

*Remark 18.14 (Cross-Elasticity and Interference)* Estimating cross-price elasticities raises an interference concern. If products  $j$  and  $k$  are substitutes, promoting  $j$  shifts demand from  $k$ . From the perspective of product  $k$ , this is a SUTVA violation:  $k$ 's outcome depends on  $j$ 's treatment. In these settings, the estimand is a demand-system object, not a single-equation slope. Credible estimation therefore requires either explicitly modelling the substitution structure via demand systems (for example, logit or AIDS models) or randomising promotions at the product–store level and measuring effects on the full product set. Simply including cross-

price terms in a regression without addressing the joint determination of prices across substitutes yields biased estimates. This connects directly to the interference issues in Chapter 11.

### Diagnostic Checklist

#### Box 18.10: Price Elasticity Diagnostic Checklist

These diagnostics map to the identifying assumptions. IV diagnostics probe instrument relevance and the plausibility of the exclusion restriction. Promotion-timing diagnostics probe the exogeneity of promotion timing. Spillover diagnostics probe the interference and substitution structure that makes single-product analyses misleading.

**For IV estimation:** Report first-stage strength using statistics appropriate for the error structure, such as the Kleibergen–Paap rk statistic under clustering. When instruments are weak, use weak-IV procedures designed for weak instruments rather than relying on conventional t-statistics. Argue why the instrument affects demand only through price and check plausible failure modes in context. If multiple instruments are used, report overidentification diagnostics and interpret them cautiously.

**For promotion analysis:** Run pre-trend diagnostics to check for demand changes before promotion. Diagnose stockpiling by examining whether a post-promotion dip occurs, indicating that consumers bought forward. Check for spillovers by examining substitution effects on competing products. Report heterogeneity in effects by product category, store type, and promotion depth.

### Case Study: Scanner Panel Price Elasticity

We illustrate IV estimation using a hypothetical scanner panel. The numbers are illustrative and do not represent real data.

**Setting.** A retailer estimates own-price elasticity for a national beverage brand across 500 stores over 104 weeks. The instrument is the wholesale cost of sugar, a key input.

**Data.** Store-week level data on quantity sold, shelf price, and wholesale sugar cost. Controls include store-week fixed effects, display indicators, and feature advertising.

**First stage.** Sugar cost strongly predicts shelf price: a 10% increase in sugar cost raises shelf price by 2.3% ( $SE = 0.4\%$ ). In this illustration, the first stage is strong. In applications, report first-stage strength using statistics appropriate for the error structure and interpret any rule-of-thumb thresholds cautiously.

**Results.** The IV estimate of own-price elasticity is  $\hat{\eta}_{jj} = -2.1$  (SE = 0.4). A 10% price increase reduces quantity sold by 21%. The OLS estimate is  $-0.8$ , biased toward zero due to simultaneity (high-demand periods have both higher prices and higher sales).

**Diagnostics.** The exclusion restriction is plausible: sugar cost affects consumer demand only through its effect on price, not directly. The coefficient is in a plausible range for beverages. Residual diagnostics show no clear autocorrelation after clustering at the store level.

**Promotion analysis.** Separately, we estimate promotion effects. A 20% temporary price reduction increases quantity by 85% during the promotion week (SE = 12%). Quantity in the two weeks following the promotion is 15% below baseline, consistent with stockpiling. The net effect over a four-week window is a 40% quantity increase.

**Interpretation.** The IV elasticity of  $-2.1$  implies that a permanent 5% price increase would reduce sales by approximately 10.5%. The promotion analysis shows strong short-run response but significant stockpiling, suggesting that promotion ROI depends critically on the time horizon of evaluation and on how intertemporal substitution is valued.

## 18.8 Customer Lifetime Value and Acquisition

Customer Lifetime Value (CLV) is a long-term outcome that summarises the total value a customer generates over their relationship with a firm. Marketing decisions—acquisition channel, onboarding experience, early promotions—may have persistent effects on CLV that short-term metrics miss. Estimating the causal effect of these decisions on CLV requires addressing both selection and measurement challenges.

*Remark 18.15 (CLV Attribution in the Taxonomy)* In the taxonomy of Section 18.1, CLV attribution is primarily a *confounding* problem: customers self-select into acquisition channels based on characteristics correlated with their potential lifetime value. Paid search attracts high-intent customers; referrals attract customers similar to existing high-value customers. The temporal dimension is also central: CLV is a cumulative outcome over time, and channel effects may operate through retention (how long customers stay), intensity (how much they spend per period), or both. Early-tenure interventions—onboarding experiences, welcome promotions—create path-dependent effects where initial treatment shapes the entire margin trajectory. See Section 18.6 for related analysis of loyalty programme effects on spending, and Section 18.4 for multi-touch attribution when customers interact with multiple channels.

### Estimand: CLV and Channel Effects

We first define CLV, then the causal effect of acquisition channel.

**Definition 18.10 (Customer Lifetime Value)** For customer  $i$  acquired at time  $t_0$ , the Customer Lifetime Value over a horizon of  $H$  periods is:

$$\text{CLV}_i(H) = \sum_{h=0}^H \delta^h \cdot m_{i,t_0+h},$$

where  $m_{it}$  is the margin (revenue minus variable cost) from customer  $i$  in period  $t$ , and  $\delta \in (0, 1]$  is the discount factor. For  $\delta = 1$ , this is undiscounted cumulative margin. In practice, CLV is often only partially observed because the horizon extends beyond the data window. We then work with a model-based estimate  $\hat{\text{CLV}}_i(H)$ . The causal estimands below are defined in terms of the underlying true  $\text{CLV}_i(H)$ .

The causal question is whether the acquisition channel affects CLV beyond its effect on who is acquired.

**Definition 18.11 (Channel Causal Effect on CLV)** Let  $C_i \in \{1, \dots, K\}$  denote the observed acquisition channel for customer  $i$ . To define a causal effect, we must specify an intervention. Let  $Y_i(\ell)$  denote the potential outcome under a channel policy  $\ell$ , where the outcome is CLV over horizon  $H$ , so  $Y_i(\ell) := \text{CLV}_i^{(\ell)}(H)$ .

The channel policy effect comparing  $\ell$  to  $\ell'$  is

$$\text{ATE}_{\ell, \ell'} = \mathbb{E}[Y_i(\ell) - Y_i(\ell')].$$

This object is policy-relevant only after you state the target population. If changing channel policy changes who enters the customer base, then  $\text{ATE}_{\ell,\ell'}$  includes selection into acquisition, not only post-acquisition spending.

This estimand requires the *overlap* (positivity) assumption for the target population. In practice, overlap often fails. Customers without social media accounts cannot be acquired via social. Customers who never search for the product cannot be acquired via paid search. When overlap fails, options include restricting comparison to an overlap population and reporting the overlap-restricted estimand, trimming observations with extreme propensity scores and reporting the trimmed estimand, or focusing on an ATT-type estimand that changes the target population.

A more practical estimand is the Average Treatment Effect on the Treated (ATT) for customers acquired via channel  $\ell$ , relative to a chosen reference channel  $\ell_0$ :

$$\text{ATT}_{\ell|\ell_0} = \mathbb{E}[Y_i(\ell) - Y_i(\ell_0) \mid C_i = \ell].$$

This answers the question: for customers we currently acquire via channel  $\ell$ , what would their CLV have been had we instead acquired them via channel  $\ell_0$ ?

### Identification Challenge: Selection into Channels

Customers who respond to different channels differ systematically. Paid search attracts high-intent customers. Social media attracts younger demographics. Referrals attract customers similar to existing customers. Naive comparison of CLV across channels confounds channel effects with selection effects.

**Assumption 99 (Unconfoundedness for Channel Assignment)** Conditional on observed pre-channel covariates  $X_i$  (demographics and acquisition context measured before channel assignment), channel assignment is independent of potential CLV:

$$Y_i(\ell) \perp\!\!\!\perp C_i \mid X_i, \quad \forall \ell.$$

This is a channel-specific version of Assumption 5. It is strong and often implausible—customers who click on paid ads may differ from organic arrivals in unobservable ways (e.g., price sensitivity, brand awareness).

### Identification Strategy 1: Propensity Score Methods

When unconfoundedness is plausible, propensity score methods adjust for selection.

For multi-valued channel  $C_i \in \{1, \dots, K\}$ , we use propensity scores  $e_\ell(X_i) = \mathbb{P}(C_i = \ell \mid X_i)$  from Chapter 12. Inverse probability weighting reweights observations to balance covariates across channels. Doubly robust estimators combine this propensity weighting with an outcome regression  $m_\ell(X_i) = \mathbb{E}[\text{CLV}_i(H) \mid X_i, C_i = \ell]$ . Rather than re-deriving the influence-function formula here, we rely on the estimators developed in Chapter 12, typically implemented via double machine learning to handle high-dimensional  $X_i$ .

In practice, analysts often work with an estimated outcome  $\widehat{CLV}_i(H)$ . If you do this, state the prediction model and ensure the features used are not themselves downstream of channel assignment unless you are willing to redefine the estimand accordingly. This is a design decision, not a modelling detail.

## Identification Strategy 2: Instrumental Variables

When unconfoundedness fails, we need exogenous variation in channel exposure.

**Assumption 100 (Channel Instrument)** Let  $Z_i$  be an instrument that shifts channel assignment through a defensible mechanism. Conditional on controls,  $Z_i$  affects  $CLV_i(H)$  only through the induced change in channel assignment.

Table 18.7 summarises potential instruments. In practice, IV is rarely used for CLV attribution because valid instruments are scarce and the exclusion restriction is difficult to defend—most factors that affect channel exposure also plausibly affect customer quality.

**Table 18.7** Potential instruments for channel effects on CLV

Instrument	Mechanism	Exclusion Concern
Ad auction randomness	Real-time bidding variation	Auction winners may differ in quality
Geographic variation	Differential channel availability	Regions differ in customer characteristics
Temporal variation	Channel-specific promotions/outages	Timing correlates with demand shocks

## Estimation

For CLV prediction, common models include the BG/NBD + Gamma-Gamma probabilistic framework for transaction frequency and monetary value [Fader et al., 2005], survival models (Cox proportional hazards for churn, with CLV as expected discounted margin conditional on survival), and direct regression approaches that predict CLV from early-tenure features using gradient boosting or neural networks. Keep the conceptual distinction clear: prediction models estimate  $\widehat{CLV}_i(H)$ , while the causal design determines whether differences in  $\widehat{CLV}$  across channels can be interpreted as causal effects. Predictive accuracy is not evidence of unconfoundedness.

*Remark 18.16 (CLV Trajectory and Path Dependence)* The CLV definition aggregates margins over time, but channel effects may operate differently across the customer lifecycle. First, retention effects mean the channel affects how long customers remain active, with referral customers potentially having higher retention because social ties create switching costs. Second, intensity effects mean the channel affects per-period spending

conditional on remaining active, with paid search customers potentially having higher order values because they arrived with purchase intent. Third, trajectory effects mean the channel affects the shape of the margin path—whether spending grows, stays flat, or declines over tenure.

Decomposing CLV into these components clarifies the mechanism and informs intervention design. If the channel effect operates through retention, onboarding and engagement programmes may amplify it. If it operates through initial order value, the effect may be harder to replicate.

## Diagnostic Checklist

### Box 18.11: CLV Attribution Diagnostic Checklist

**For Propensity Score Methods:** Check overlap by verifying that all channels have positive probability for all covariate values and trim extreme propensity scores. Report standardised mean differences (SMD) for key covariates before and after weighting, separately for each channel pair being compared, and interpret magnitudes in context. Use sensitivity analyses appropriate to the identification strategy. Interpret any sensitivity parameter as a reporting device, not as proof of robustness.

**For CLV Measurement:** Report CLV at multiple horizons (6 months, 1 year, 2 years) since short horizons may miss channel effects while long horizons have more noise. Account for customers who are still active (right-censoring) using survival-based CLV models. Report sensitivity to the discount rate since high discount rates favour channels with early revenue.

**For Causal Interpretation:** Run placebo and falsification checks where feasible (for example, two organic search variants). Decompose CLV into frequency, monetary value, and tenure to identify which component drives the channel effect.

## Case Study: Acquisition Channel CLV Comparison

We illustrate propensity score methods with a hypothetical e-commerce retailer. The numbers are illustrative and do not represent real data.

**Setting.** An online retailer acquires customers through three channels: paid search (40%), social media ads (35%), and organic/referral (25%). The question is whether paid search customers have higher CLV than social media customers, or whether the observed difference reflects selection.

**Data and Notation.** 100,000 customers acquired over 12 months, observed for 24 months post-acquisition. Covariates include age, gender, device type, and acquisition month. The outcome  $Y_i$  is 24-month CLV, i.e.,  $CLV_i(H)$  with  $H = 24$  months. Contrasts are pairwise (paid search vs. social media), with organic analysed separately as a reference.

**Naive Comparison.** Mean CLV by channel: paid search = \$285, social media = \$195, organic = \$340. Paid search appears \$90 higher than social media.

**Propensity Score Analysis.** We estimate the propensity score model using gradient boosting on the covariate set. After IPW reweighting, covariate balance improves: all SMDs fall below 0.05.

**Results.** The doubly robust estimate of the paid search vs. social media effect is \$42 (SE = \$18). This is less than half the naive difference of \$90, indicating substantial selection bias. Paid search customers are older and use desktop devices more often, both predictors of higher CLV.

**Sensitivity Analysis.** We assess sensitivity to unobserved confounding using Rosenbaum bounds [Rosenbaum, 2002], which ask: how strong would an unobserved confounder need to be to explain away the estimated effect? The bounds show that a confounder would need to increase the odds of paid search assignment by a factor of  $\Gamma = 1.8$  (holding covariates fixed) to reduce the effect to zero. Since plausible confounders—such as prior brand awareness or purchase intent—could reasonably have this magnitude, the causal interpretation is tentative. For a formal treatment of sensitivity analysis, see Section 17.8.

**Mechanism.** Decomposing CLV: paid search customers have similar purchase frequency but 15% higher average order value. The channel effect operates through basket size, not retention.

**Interpretation.** Paid search generates higher CLV than social media, but the effect (\$42) is smaller than the naive comparison (\$90) suggests. The difference is driven by order value, not frequency. Given sensitivity to unobserved confounding, we recommend a randomised channel experiment to confirm the causal effect before reallocating budget.

## 18.9 Dynamic Pricing in Transport Networks

Transport operators—airlines, railways, ride-hailing platforms—use dynamic pricing algorithms that adjust prices in real time based on demand forecasts. Estimating the causal effect of price on demand is challenging because prices respond to demand. High prices occur precisely when demand is high, which creates simultaneity bias. Naive regression of bookings on price yields biased estimates, typically finding elasticities that are too small or even positive.

The independent sampling unit in these settings is rarely an individual booking. Designs and inference are usually coherent at an aggregated level such as route–day, route–departure-time, or zone–time blocks. Cross-route substitution and network shocks create dependence across these units, so uncertainty should be expressed using clustering or common-shock corrections rather than treating each transaction as independent.

*Remark 18.17 (Dynamic Pricing in the Taxonomy)* In the taxonomy of Section 18.1, dynamic pricing estimation is primarily an *endogeneity* problem: prices are set algorithmically in response to demand signals, creating simultaneity bias. The term “dynamic” here refers to real-time algorithmic pricing, not dynamic treatment effects in the temporal sense of Chapter 10. However, temporal considerations do arise: customers may anticipate price changes and shift purchases intertemporally, and past prices affect reference price expectations. For retail price elasticity using similar identification strategies, see Section 18.7.

### Estimand: Dynamic Price Elasticity

We define the causal effect of price on demand in the transport context.

**Definition 18.12 (Route-Time Price Elasticity)** Let  $Q_{rt\ell}(p)$  denote the potential number of bookings on route  $r$  at departure time  $t$  and lead time  $\ell$  (days before departure) if the posted price were set to  $p$ , holding the rest of the environment fixed as specified by the design. The price elasticity at price  $p$  is

$$\eta_{rt\ell}(p) = \frac{\partial \log \mathbb{E}[Q_{rt\ell}(p)]}{\partial \log p},$$

when this derivative exists. Dynamic pricing implies that  $\eta$  may vary with  $\ell$ : customers booking far in advance may be more price-sensitive than last-minute travellers.

For ride-hailing, we often focus on the effect of surge pricing on completed rides. Completed rides are an equilibrium outcome affected by both rider demand and driver supply. Let  $D_{zt} \in \{0, 1\}$  indicate whether zone  $z$  at time  $t$  is in surge ( $D_{zt} = 1$ ) or not ( $D_{zt} = 0$ ), and let  $Q_{zt}(d)$  be the potential number of completed rides under surge state  $d$ , defined as the equilibrium outcome under policy state  $d$ . The surge effect for surge periods is

$$\text{ATT}_{\text{surge}} = \mathbb{E}[Q_{zt}(1) - Q_{zt}(0) \mid D_{zt} = 1],$$

which measures the causal effect of surge on completed rides during surge periods in a given zone. Interpreting this object as a rider-demand elasticity requires separate outcomes or assumptions, for example using ride requests rather than completed rides.

### Identification Challenge: Simultaneity

Dynamic pricing algorithms set prices based on predicted demand. If the algorithm observes signals that predict high demand (e.g., a concert ending, bad weather), it raises prices. This creates positive correlation between price and demand even if the causal effect is negative.

To highlight why naive OLS fails, consider unobserved demand shocks  $U_{rt\ell}$ . Under algorithmic pricing we typically have:

$$\text{Cov}(P_{rt\ell}, U_{rt\ell}) > 0,$$

because the algorithm sets  $P_{rt\ell}$  as a function of demand forecasts that incorporate  $U_{rt\ell}$ . Any regression of  $\log Q_{rt\ell}$  on  $\log P_{rt\ell}$  that treats  $P_{rt\ell}$  as exogenous is biased toward zero or even positive elasticities. Identification therefore requires either discontinuities in the pricing rule or instruments that move price but not  $U_{rt\ell}$ .

### Identification Strategy 1: Algorithmic Discontinuities (RDD)

Many pricing algorithms exhibit discrete jumps at thresholds, creating regression discontinuity designs.

**Definition 18.13 (Pricing Algorithm RDD)** If the pricing algorithm increases price discretely when a threshold is crossed (e.g., days to departure, capacity utilisation), we estimate the demand response at the discontinuity:

$$\hat{\eta}_{\text{Wald}} = \frac{\lim_{v \downarrow c} \mathbb{E}[\log Q_{rt\ell} | V = v] - \lim_{v \uparrow c} \mathbb{E}[\log Q_{rt\ell} | V = v]}{\lim_{v \downarrow c} \mathbb{E}[\log P_{rt\ell} | V = v] - \lim_{v \uparrow c} \mathbb{E}[\log P_{rt\ell} | V = v]},$$

where  $V$  is the running variable (e.g., days to departure) and  $c$  is the threshold. This is a fuzzy RDD where the first stage is the price jump. The ratio is a local Wald (local IV) estimand for the causal effect of  $\log P$  on  $\log Q$  induced by the algorithmic threshold. It is well defined under continuity of counterfactual outcomes in  $V$  at  $c$  and no manipulation of the running variable at the cutoff. The resulting  $\hat{\eta}_{\text{Wald}}$  is local to the threshold and to the price variation induced by the algorithmic rule.

Table 18.8 summarises common algorithmic discontinuities that enable RDD estimation.

*Remark 18.18 (Strategic Anticipation and RDD Validity)* Algorithmic pricing rules are increasingly transparent. Sophisticated travellers know that airline prices rise as departure approaches. Apps like Hopper advise when to buy. If customers strategically time purchases to avoid price jumps, several issues arise. First, bunching occurs when customers cluster bookings just before price thresholds, violating the continuity assumption. Second, the population becomes selected because those who book after a price jump may be

**Table 18.8** Algorithmic discontinuities for pricing RDD

Discontinuity	Running Variable	Application
Days to departure	Calendar time before flight	Airlines: tiers at 21, 14, 7, 3 days
Capacity thresholds	Load factor (% seats sold)	Airlines/hotels: jumps at 70%, 85%, 95%
Surge multipliers	Demand/supply ratio	Ride-hailing: discrete tiers (1.0×, 1.5×, 2.0×)

systematically different (less price-sensitive, higher willingness to pay), so the RDD identifies a local effect for a non-representative population. Third, algorithm response may occur if the platform observes strategic behaviour and adjusts thresholds, creating a moving target.

Density diagnostics such as the McCrary diagnostic can reveal bunching, but absence of bunching does not guarantee that anticipation is absent—customers may shift purchases smoothly rather than bunch. RDD-based local Wald estimates are local to the threshold and to the subpopulation that books near it. They may not generalise to customers who book earlier or later in the booking window.

### Identification Strategy 2: Instrumental Variables

Cost shifters and algorithmic features provide instruments for price.

**Assumption 101 (Transport Pricing Instruments)** Let  $Z_{rtl}$  be an instrument. Conditional on fixed effects and controls, the instrument affects demand only through price, and it shifts prices with non-trivial first-stage variation.

Table 18.9 summarises potential instruments. However, several of these are more fragile than they may appear, and the exclusion restriction warrants careful scrutiny.

**Fuel costs** are the cleanest example. Jet-fuel price shocks shift airline operating costs and therefore fares. The main exclusion concern is whether fuel shocks coincide with macro conditions that also shift demand on the same routes. This is usually modest for short-run variation but should still be discussed.

**Competitor capacity** is much riskier. Additional competitor seats alter the choice set and perceived service quality for travellers, so they can affect demand directly as well as via the focal airline's price. Unless capacity variation is clearly cost-driven and route-independent in its demand impact, this instrument is hard to defend.

**Surge in adjacent zones** can help in ride-hailing by shifting driver supply, but nearby zones often share the same demand shocks—concerts, sports events, weather. The exclusion restriction that demand shocks are spatially local therefore needs explicit argument and diagnostics, not just an assumption.

These caveats reinforce the identification discipline emphasised in Remark 18.2: instruments in transport pricing are context-dependent and require case-by-case justification.

**Table 18.9** Instruments for transport pricing

Instrument	Mechanism	Exclusion Restriction
Fuel costs	Jet fuel prices shift airline operating costs	Requires arguing away channels through macro conditions that also move demand
Competitor capacity	Competitor seats can shift focal carrier pricing	Plausible only in special settings. Capacity can change the choice set and quality directly
Algorithmic lags	Delayed price updates create timing variation	Requires arguing lag structure is orthogonal to current demand shocks
Surge in adjacent zones	Nearby surge affects driver supply, thus local price	Requires careful argument. Nearby zones can share demand shocks

## Network Structure and Spatial Dependence

Transport networks exhibit spatial dependence: demand shocks on one route affect prices and demand on substitute routes. This has two implications.

*Remark 18.19 (Two-Sided Markets in Ride-Hailing)* Ride-hailing platforms operate two-sided markets where surge pricing affects both sides. On the rider side, higher prices reduce rider demand (the elasticity we seek to estimate). On the driver side, higher prices attract driver supply, reducing wait times and improving service quality.

The equilibrium effect on completed rides depends on both responses. A price increase that reduces rider requests may be partially offset by improved driver availability. Estimating rider-side elasticity alone understates the platform's pricing power. Conversely, ignoring supply-side effects overstates the revenue loss from price increases. Full analysis requires modelling both sides of the market, which is beyond our scope but important for platform pricing optimisation.

**Definition 18.14 (Spatial Dependence in Transport)** Let  $r$  and  $r'$  be routes sharing an origin or destination. Suppressing the lead-time index  $\ell$  for notational simplicity, demand shocks are spatially correlated:

$$\text{Cov}(U_{rt}, U_{r't}) \neq 0.$$

Standard errors that ignore this correlation are too small, leading to uncertainty statements that are too narrow.

State the independent sampling unit and the dependence structure explicitly. If you work with route panels and expect common shocks across routes, cluster at the route level and consider additional adjustments for cross-sectional dependence when  $T$  is moderately large. Driscoll–Kraay standard errors are appropriate for panels with moderately large  $T$  and cross-sectional dependence, not as a default for short panels. For ride-hailing, clustering at the city-hour level is often a more coherent starting point than clustering at the observation level.

## Estimation

For RDD, use local polynomial regression with the Calonico-Cattaneo-Titiunik bandwidth selector. The first stage is the price jump, the reduced form is the demand response, and the ratio is a local Wald estimand for the causal effect of log price on log demand induced by the algorithmic threshold.

For IV, use 2SLS with route-time fixed effects:

$$\text{First stage: } \log P_{rt\ell} = \pi_0 + \pi_1 Z_{rt\ell} + \alpha_r + \gamma_t + u_{rt\ell}, \quad (18.3)$$

$$\text{Second stage: } \log Q_{rt\ell} = \beta_0 + \eta \widehat{\log P_{rt\ell}} + \alpha_r + \gamma_t + \varepsilon_{rt\ell}. \quad (18.4)$$

For ride-hailing with high-frequency data, include hour-of-day and day-of-week fixed effects to absorb predictable demand patterns.

## Diagnostic Checklist

### Box 18.12: Dynamic Pricing Diagnostic Checklist

**For Algorithmic RDD:** Verify that price actually jumps at the threshold by plotting price against the running variable. Use density diagnostics for bunching just below the threshold, which can indicate strategic timing. Check that route characteristics are smooth through the threshold for covariate balance. Report estimates at multiple bandwidths for sensitivity. For inference, treat the route-time (or zone-time) unit implied by the design as the independent sampling unit rather than individual transactions.

**For IV:** Report first-stage strength using statistics appropriate for the error structure (for example, clustered or heteroskedasticity-robust diagnostics such as Kleibergen–Paap-type measures). When instruments are weak, use weak-IV robust procedures rather than relying on conventional  $t$ -statistics. Argue why the instrument affects demand only through price. If multiple instruments are used, report overidentification diagnostics and interpret them cautiously.

**For Spatial Dependence:** Cluster at a level consistent with the assignment and sampling structure, such as the route (origin-destination) level, and report robustness to common-shock corrections when  $T$  is moderately large. Use spillover diagnostics to assess whether price changes on route  $r$  affect demand on substitute routes  $r'$ .

## Case Study: Airline Pricing with Days-to-Departure RDD

We illustrate the algorithmic RDD approach with a hypothetical airline. The numbers are illustrative and do not represent real data.

**Setting.** A low-cost carrier uses a pricing algorithm that increases fares by approximately 15% when bookings occur within 7 days of departure. We estimate the price elasticity at this threshold.

**Data.** 500,000 bookings across 200 routes over 12 months. Running variable is days to departure. Outcome is log bookings per route-day. Treatment is log price.

**First Stage.** At the 7-day threshold, log price jumps by 0.14 (approximately 15%). The discontinuity is sharp and precisely estimated (SE = 0.02).

**Reduced Form.** Log bookings drop by 0.21 at the threshold (SE = 0.05). Customers booking within 7 days face higher prices and book less.

**Elasticity Estimate.** The fuzzy RDD elasticity is  $\hat{\eta} = -0.21/0.14 = -1.5$  (SE = 0.4). A 10% price increase reduces bookings by 15%.

**Diagnostics.** Density diagnostics suggest limited bunching at the 7-day threshold in this illustration. Covariate balance looks stable: route distance, competitor presence, and day-of-week are smooth through the threshold. The estimate is stable across bandwidths: -1.3 (half), -1.5 (optimal), -1.6 (double).

**Heterogeneity.** Elasticity varies by customer segment: leisure routes show  $\hat{\eta} = -1.9$ , while business routes show  $\hat{\eta} = -0.8$ . Business travellers are less price-sensitive, consistent with lower flexibility.

**Interpretation.** The elasticity of -1.5 implies that demand is elastic at this threshold: a 10% price increase reduces bookings by 15%, and since  $|\eta| > 1$ , the percentage quantity loss exceeds the percentage price gain, so *revenue falls*. This suggests the airline may be over-pricing at the 7-day threshold. A lower price jump would increase revenue. The heterogeneity analysis reinforces this. Business routes with  $|\eta| = 0.8 < 1$  (inelastic) can sustain price increases, while leisure routes with  $|\eta| = 1.9$  (elastic) should see smaller increases. Differentiated pricing by route type could improve overall revenue. These are local elasticities at the 7-day threshold. Elasticities earlier in the booking window may differ.

## 18.10 Subscription and Paywall Effects

For media companies, the paywall is a critical lever: stricter paywalls increase subscription revenue but reduce reach, advertising impressions, and brand awareness. Estimating the causal effect of paywall strictness on subscriptions—and the trade-off with advertising revenue—requires careful design because users who hit the paywall differ systematically from those who do not.

*Remark 18.20 (Paywall Effects in the Taxonomy)* In the taxonomy of Section 18.1, paywall optimisation involves both *confounding* and *dynamics*. On the identification axis, high-engagement users self-select into paywall exposure, creating positive selection bias. On the temporal axis, subscription is an inherently dynamic outcome: it creates recurring revenue over months or years, and early-tenure experiences shape long-run retention. The short-run conversion effect understates lifetime value; see Section 18.8 for CLV perspectives on customer acquisition. Subscription also parallels loyalty programme membership (Section 18.6): both involve self-selection, ongoing engagement, and habit formation.

### Estimand: Paywall Conversion and Revenue Trade-off

Encountering the paywall is not a primitive intervention independent of user behaviour. Users hit the paywall because they read more, so an exposure indicator is endogenous by construction. We therefore frame the causal question as a *policy effect*: change the paywall strictness and measure subscription and engagement outcomes.

Let  $D_i \in \{0, 1\}$  denote assignment to a strict policy ( $D_i = 1$ ) versus a lenient policy ( $D_i = 0$ ), for example three free articles versus ten, over a stated horizon. Let  $S_i(d)$  denote the potential subscription indicator under policy  $d$ , and let  $Y_i(d)$  denote a measure of engagement such as total page views or ad impressions over the same horizon. The policy effects are

$$\text{ATE}^S = \mathbb{E}[S_i(1) - S_i(0)], \quad \text{ATE}^Y = \mathbb{E}[Y_i(1) - Y_i(0)].$$

**Definition 18.15 (Paywall Reach Effect)** The paywall reach effect is the average change in engagement when moving from a lenient to a strict policy:

$$\text{ATE}^Y = \mathbb{E}[Y_i(1) - Y_i(0)].$$

The business question is whether the subscription-value gain from a stricter policy, which depends on  $\text{ATE}^S$  and subscriber value, exceeds the advertising-value loss from reduced reach, which depends on  $\text{ATE}^Y$ . Throughout, evaluate these effects over a consistent horizon so that subscription and advertising components are comparable.

## Identification Challenge: Selection into Paywall Exposure

Users who hit the paywall are high-engagement users who differ from casual visitors. This makes “exposure effects” hard to interpret causally without design-based variation. Policy experiments avoid this post-behaviour selection by varying strictness directly.

To understand why naive comparisons fail, consider user characteristics  $X_i$  (visit frequency, content preferences, device). In many settings, users with higher  $X_i$  are more likely to reach the paywall threshold *and* more likely to subscribe even in the absence of the paywall:

$$P(A_i \geq \bar{A} | X_i) \text{ increasing in } X_i, \quad P(S_i(0) = 1 | X_i) \text{ increasing in } X_i.$$

This describes the selection problem: without design-based variation, naive estimates overstate the causal impact of the paywall because engaged users both hit the paywall and subscribe at higher rates. Identification therefore comes from designs that create local randomisation around thresholds, explicit randomisation of policy, or staggered rollouts.

## Identification Strategy 1: Article Limit RDD

If the paywall triggers at a fixed article count, regression discontinuity compares users just above and below the threshold.

**Definition 18.16 (Article Limit RDD)** Let  $A_i$  be the number of articles read in a billing period and let  $\bar{A}$  be the free-article limit. An RDD for subscription uses:

$$\hat{\Delta}_{\text{RDD}} = \lim_{a \downarrow \bar{A}} \mathbb{E}[S_i | A_i = a] - \lim_{a \uparrow \bar{A}} \mathbb{E}[S_i | A_i = a],$$

which estimates the *local* effect of crossing the paywall threshold on subscription for users at the margin—those with  $A_i$  near  $\bar{A}$ .

**Assumption 102 (Continuity of Potential Outcomes at Article Threshold)** Potential subscription outcomes are continuous functions of article count at the threshold:

$$\lim_{a \downarrow \bar{A}} \mathbb{E}[S_i(0) | A_i = a] = \lim_{a \uparrow \bar{A}} \mathbb{E}[S_i(0) | A_i = a].$$

This is a discrete RDD at a behavioural threshold. Identification requires that users just below and just above the limit are comparable in their latent subscription propensity absent treatment. Treat this as a diagnostic-driven quasi-experiment rather than a default design. Continuity of covariates and density diagnostics for bunching are central.

This assumption fails if users strategically stop reading to avoid the paywall, creating selective bunching below the threshold.

*Remark 18.21 (Rational Avoidance and RDD Validity)* Users who stop at  $\bar{A} - 1$  articles to avoid the paywall are engaging in rational avoidance, not manipulation in the sense of fraud. However, this behaviour affects RDD validity. First, missing “would-be treated” occurs because users who would have read  $\bar{A}$  articles but stopped short are absent from the treatment group, so the RDD identifies the effect for users who could not or would not avoid the threshold. Second, the population becomes selected because those who cross the threshold despite awareness may be less price-sensitive or more committed readers, and the local effect may not generalise to the broader population. Depending on how avoidance relates to latent subscription propensity, the RDD estimate may be biased upward or downward. In some monotone-selection stories it can be interpreted as an upper bound.

Density diagnostics can detect bunching, but modest bunching warrants caution rather than a mechanical decision rule. Report the RDD estimate with appropriate caveats about external validity.

### Identification Strategy 2: Randomised Paywall Experiments

The preferred approach is randomising paywall strictness across users or time periods.

**Definition 18.17 (Paywall A/B Experiment)** Randomly assign users to paywall policies. For example, assign strict versus lenient thresholds. The average treatment effect of a stricter policy on subscription is:

$$\text{ATE}^S = \mathbb{E}[S_i(1) - S_i(0)].$$

Randomisation eliminates selection bias within the experimental sample. External validity still depends on whether the experiment population (e.g., new visitors) reflects the broader user base and on the horizon used to measure subscription and retention.

### Identification Strategy 3: Staggered Paywall Rollout

If paywall strictness changes over time or across regions, staggered difference-in-differences applies.

**Assumption 103 (Parallel Trends for Paywall Rollout)** In the absence of the paywall policy change, subscription trends would be parallel across cohorts. Let  $S_{it}(0)$  be the subscription indicator for user  $i$  under the baseline policy and  $G_i$  the rollout cohort (wave). Then:

$$\mathbb{E}[S_{it}(0) - S_{i,t-1}(0) | G_i = g] = \mathbb{E}[S_{it}(0) - S_{i,t-1}(0) | G_i = g'], \quad \forall g, g'.$$

See Assumption 6 for the general formulation.

Use heterogeneity-robust estimators (Chapter 4) to avoid negative weighting when effects vary across cohorts.

## Estimation

For RDD, use local polynomial regression with the Calonico-Cattaneo-Titiunik bandwidth selector. The running variable is article count. The outcome is the subscription indicator.

For A/B experiments, simple difference-in-means with robust standard errors suffices when each  $i$  is an independent user and outcomes are aggregated at the user level over the analysis horizon. If interference is plausible (for example, password sharing or article sharing across users), interpret estimates as effects under the platform's partial-treatment regime and consider clustering at a higher level of assignment when feasible. For long-horizon outcomes (e.g., 12-month retention), account for attrition.

For staggered rollout, apply Callaway–Sant'Anna or Sun–Abraham estimators with clustering at the independent rollout unit. If rollout occurs at a higher level (region, market, content category, or time block), treat that rollout level as the independent sampling unit for inference and interpret effective sample size in terms of the number of independent rollout units  $G$ .

## Revenue Trade-off Analysis

The optimal paywall balances subscription and advertising revenue.

**Definition 18.18 (Paywall Revenue Trade-off)** Let  $R^S$  be subscription revenue and  $R^A$  be advertising revenue over a horizon of  $H$  months. Let  $\Delta p_{\text{sub}}$  denote the increase in subscription probability under a strict versus lenient policy (estimated via RCT, RDD, or DID), and let  $\Delta y_{\text{reach}}$  denote the change in average page views or impressions per user. The net revenue effect of moving from a lenient to a strict paywall is:

$$\Delta R = \underbrace{\Delta p_{\text{sub}} \times \text{CLV}_{\text{sub}}(H) \times N_{\text{users}}}_{\text{Subscription gain over horizon } H} - \underbrace{\Delta y_{\text{reach}} \times (\text{CPM}/1000) \times N_{\text{users}}}_{\text{Advertising loss over horizon } H},$$

where  $\text{CLV}_{\text{sub}}(H)$  is the expected discounted revenue per marginal subscriber over horizon  $H$  and  $\Delta y_{\text{reach}}$  is incremental impressions per user over the same horizon.

Report both components separately, as they accrue to different business units. Diagnostics should use the same horizon and discount rate for both components so that subscription gains and advertising losses are commensurate.

*Remark 18.22 (Habit Formation and Lock-In)* Subscription creates lock-in and habit formation that extend beyond the initial conversion. First, reading habits develop when subscribers establish daily reading routines, leading to higher retention than those who subscribe impulsively. Second, sunk cost effects mean that having paid for a subscription, users may consume more content to “justify” the cost, reinforcing engagement. Third, switching costs accumulate as subscribers build up saved articles, personalised recommendations, and familiarity with the interface, raising barriers to cancellation.

These dynamics imply that the long-run effect of paywall-induced subscription may exceed the short-run conversion effect. Marginal subscribers acquired via aggressive paywalls may, however, have lower retention

than subscribers acquired organically. Tracking cohort-level retention by acquisition source informs long-run paywall optimisation.

### Diagnostic Checklist

Box 18.13: Paywall Diagnostic Checklist

**For Article Limit RDD:** Use density diagnostics for bunching just below the threshold. Use survey or behavioural data to assess threshold awareness. Check that user characteristics are smooth through the threshold for covariate balance. Report estimates at multiple bandwidths for sensitivity.

**For A/B Experiments:** Verify that randomisation succeeded by checking covariate balance across arms. Check for differential dropout across paywall conditions to assess attrition. Consider that users may share articles, contaminating the control group through spillovers. Note that short experiments may miss long-run subscription effects.

**For Revenue Trade-off:** Ensure that subscription gains and advertising losses are evaluated over the same horizon  $H$  and discount rate. High-value users may respond differently to paywall than casual visitors, creating heterogeneity. Stricter paywall may shift users to competitor sites, creating cannibalisation.

### Case Study: News Publisher Paywall Optimisation

We illustrate the RDD approach with a hypothetical news publisher. The numbers are illustrative and do not represent real data.

**Setting.** A digital news publisher operates a metered paywall: users can read 5 free articles per month before encountering a subscription prompt. The subscription price is \$10/month. We estimate the local discontinuity in subscription conversion at the 5-article threshold.

**Data.** 2 million unique visitors over 6 months. The running variable is monthly article count. The outcome is subscription within 30 days of crossing the threshold.

**RDD Estimate.** At the 5-article threshold, subscription probability jumps from 0.8% (just below) to 2.1% (just above). The estimated discontinuity is  $\hat{\Delta}_{RDD} = 1.3$  percentage points ( $SE = 0.2$  pp).

**Diagnostics.** Density diagnostics suggest modest bunching at 5 articles in this illustration, suggesting some users may stop reading to avoid the paywall. Covariate balance is acceptable: age, device type, and referral source are smooth through the threshold. The estimate is stable across bandwidths: 1.1 pp (half), 1.3 pp (optimal), 1.4 pp (double).

**Revenue Trade-off.** Of 2 million visitors, 400,000 hit the paywall (20%). The subscription gain is:

$$0.013 \times \text{CLV}_{\text{sub}}(H) \times 400,000,$$

where  $\text{CLV}_{\text{sub}}(H)$  is the expected discounted revenue per marginal subscriber over the chosen horizon  $H$ , accounting for churn. However, users who hit the paywall and do not subscribe reduce their visits. Estimated reach loss is 15% of impressions among paywall-exposed non-subscribers, translating to an advertising loss of approximately  $(\text{CPM}/1000) \times \Delta y_{\text{reach}} \times N_{\text{users}}$  over horizon  $H$ , with units made explicit.

**Net Effect.** The net revenue effect depends on the horizon- $H$  subscriber value  $\text{CLV}_{\text{sub}}(H)$  and the reach-loss term. A stricter 3-article paywall would increase conversions but also increase reach loss. Choosing the optimal threshold requires estimating the full trade-off curve.

**Interpretation.** The paywall generates substantial subscription revenue that exceeds the advertising loss. The 1.3-percentage-point jump is a *local* effect for users near the five-article threshold; the annual-revenue calculation extrapolates this to all threshold-crossers, implicitly assuming similar responsiveness across that group. The modest bunching suggests some users are gaming the system. The publisher might consider: (1) making the threshold less salient; (2) offering a “hard” paywall for high-value content; or (3) personalising the threshold based on predicted user value.

## 18.11 Platforms and Two-Sided Markets

In platforms—ride-hailing, e-commerce marketplaces, food delivery—interference is the norm, not the exception. Treating drivers affects riders. Treating sellers affects buyers. Treating one side of the market equilibrates through prices, wait times, and matching quality to affect the other side. Standard experimental designs that assume no interference (SUTVA) yield biased estimates. This section builds on the interference foundations from Chapter 11. It applies them to platform settings.

*Remark 18.23 (Platform Experiments in the Taxonomy)* In the taxonomy of Section 18.1, platform experiments are the canonical *interference* problem: a unit's outcome depends on others' treatment status, violating SUTVA. Chapter 11 develops the theoretical foundations for causal inference under interference. This section applies those foundations to platform settings. Related applications include ranking algorithm effects (Section 18.12) and seller interventions (Section 18.13). Secondary considerations include dynamics: platform interventions may have carryover effects, and long-run network effects differ from short-run equilibrium responses.

### The Interference Problem

**Definition 18.19 (Platform Interference)** In a two-sided market, unit  $i$ 's period- $t$  potential outcome can depend on its own treatment and on others' treatments:

$$Y_{it}(d, D_{-i,t}),$$

where  $d$  is unit  $i$ 's treatment status and  $D_{-i,t}$  denotes the collection of treatments for all other units in period  $t$ . SUTVA fails if there exist two treatment allocations for others,  $D_{-i,t}$  and  $D'_{-i,t}$ , such that  $Y_{it}(1, D_{-i,t}) \neq Y_{it}(1, D'_{-i,t})$ .

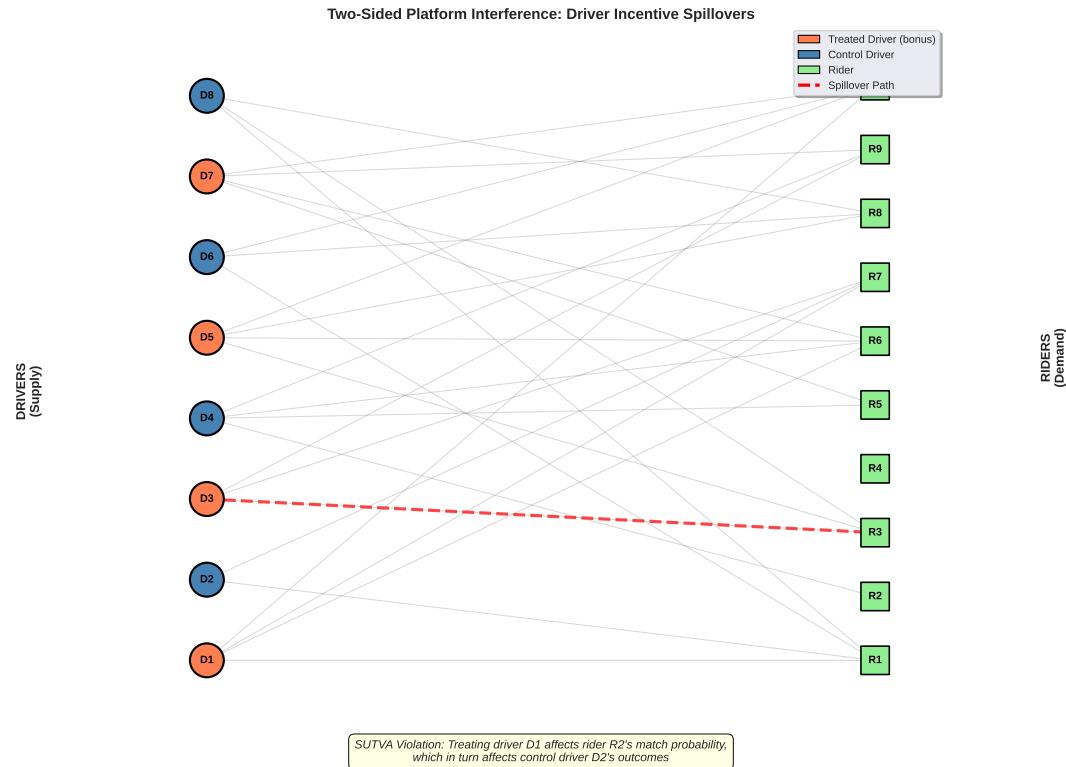
In applications, it is often useful to summarise spillovers via an exposure mapping. Writing outcomes as  $Y_{it}(d, h)$  with  $h = h_i(D_{-i,t})$ , the interference restriction becomes dependence on  $(d, h)$  rather than the full vector of others' treatments.

Table 18.10 summarises the channels through which interference operates in platforms.

### Estimand: Global Average Treatment Effect

Under interference, the standard ATE is not well-defined because potential outcomes depend on others' treatments. We instead target the Global Average Treatment Effect.

**Definition 18.20 (Global Average Treatment Effect (GATE))** The GATE compares outcomes when everyone is treated to outcomes when no one is treated:



**Fig. 18.6** Network interference in two-sided markets

The diagram illustrates how treating a subset of supply-side units (drivers/sellers) spills over to demand-side units (riders/buyers) and subsequently affects untreated supply units through market equilibration. A driver incentive reduces wait times for all riders, not just those matched with treated drivers.

**Table 18.10** Interference channels in platform markets

Channel	Mechanism	Example
Supply-side spillovers	Incentivising some suppliers increases total supply	Driver bonuses reduce wait times for all riders
Demand-side spillovers	Promotions to some buyers increase total demand	Rider discounts raise prices for all riders
Matching externalities	Treating one unit affects matches for others	Prioritising one rider delays others' matches
Price equilibration	Interventions shift market-clearing prices	Seller promotions affect marketplace price levels

$$\text{GATE} = \mathbb{E}[Y_i(\mathbf{1}) - Y_i(\mathbf{0})],$$

where **1** denotes the all-treated assignment and **0** denotes the all-control assignment. The expectation averages over the population indexed by  $i$ , which must be stated in each application (for example, riders, drivers, sellers, or market-level aggregates). In terms of Definition 18.19,  $Y_i(\mathbf{1})$  corresponds to the potential outcome under full rollout and  $Y_i(\mathbf{0})$  corresponds to the potential outcome under full hold-out. This captures the total market effect of a policy, including all spillovers.

The GATE answers the business question: “What happens to total platform outcomes if we roll out this policy to everyone?” A distinct object is the direct treatment effect, which holds others at control:

$$\text{DTE} = \mathbb{E}[Y_i(1, \mathbf{0}_{-i}) - Y_i(0, \mathbf{0}_{-i})],$$

where  $\mathbf{0}_{-i}$  denotes the all-control assignment for everyone other than unit  $i$ . Large gaps between DTE and GATE indicate economically meaningful interference.

### Identification Strategy 1: Switchback Experiments

Switchback designs randomise treatment across time windows within a market, exploiting temporal variation.

**Definition 18.21 (Switchback Design)** Divide time into windows  $t = 1, \dots, T$ . Randomly assign each window to treatment ( $D_t = 1$ ) or control ( $D_t = 0$ ). All units in the market receive the same treatment during window  $t$ . The GATE estimator is:

$$\hat{\Delta}_{\text{SB}} = \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \bar{Y}_t - \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \bar{Y}_t,$$

where  $\bar{Y}_t$  is the average outcome in window  $t$  and  $\mathcal{T}_1$  and  $\mathcal{T}_0$  are the sets of treatment and control windows.

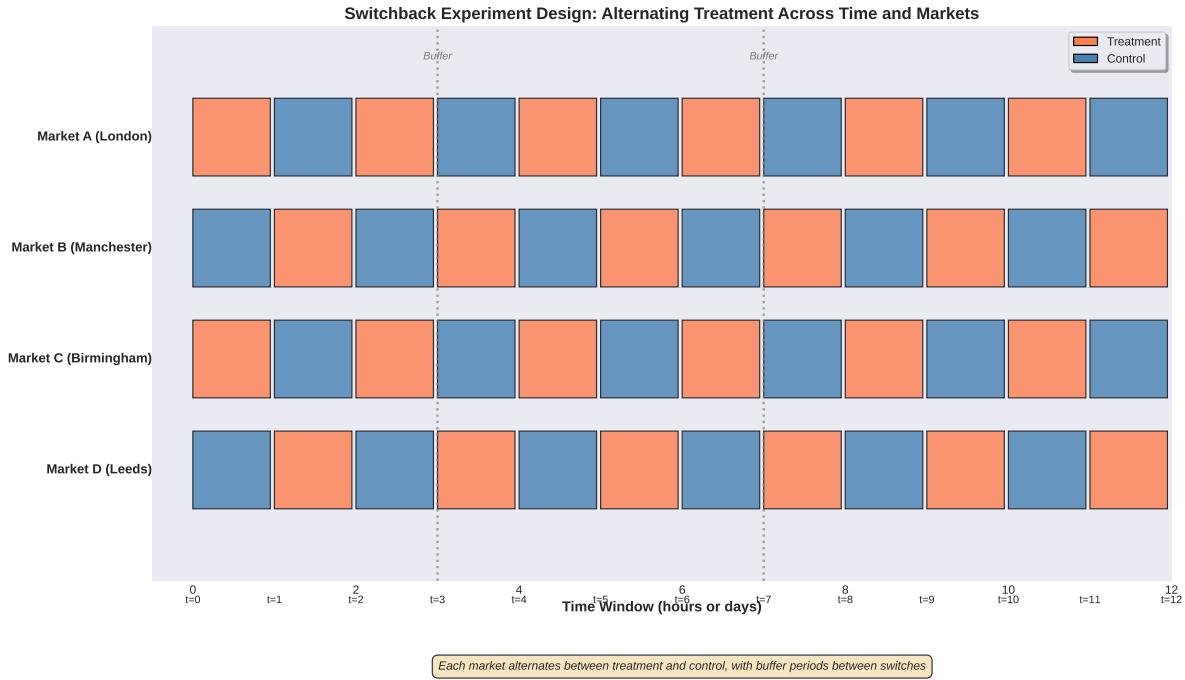
Identification is by randomisation over windows. Inference should treat windows as the randomisation units. Use randomisation-based inference where feasible, or cluster uncertainty at the window level because within-window observations are not independent.

**Assumption 104 (No Carryover Effects)** Let  $\underline{d}^{t'} = (D_1, \dots, D_{t'})$  denote the treatment path across windows up to window  $t'$ . No carryover requires that outcomes in window  $t'$  depend only on the current window's treatment, so that for any two paths  $\underline{d}^{t'}$  and  $\underline{d}'^{t'}$  with  $D_{t'} = D'_{t'}$ ,

$$Y_{it'}(\underline{d}^{t'}) = Y_{it'}(\underline{d}'^{t'}).$$

This assumption fails if treatment effects persist (e.g., driver incentives change behaviour in subsequent periods). Mitigation strategies include buffer periods (excluding observations immediately after switches), longer windows (4-hour or daily instead of hourly), and carryover modelling (including lagged treatment indicators in the regression). See Chapter 11 for formal treatment of temporal spillovers.

*Remark 18.24 (Effective Sample Size Heuristic)* Switchbacks generate identifying variation at the window level. A practical way to think about information is therefore to work with window averages  $\bar{Y}_t$  and treat them as a time series. If  $\bar{Y}_t$  is serially correlated across windows, the effective number of independent windows is smaller than  $T$ . Under an AR(1)-style heuristic for window means, an effective-window count scales like  $T \times (1 - \phi)/(1 + \phi)$ , where  $\phi$  is the autocorrelation of  $\bar{Y}_t$  across adjacent windows. This is only a heuristic. Its purpose is to reinforce that inference and power should be driven by the number of independent windows, not the raw number of user-level observations.



**Fig. 18.7** Switchback experiment timeline

Treatment allocation alternates between Treatment (blue) and Control (orange) across time windows (e.g., hourly) within a market. Grey buffer periods between switches allow carryover effects to dissipate. Outcomes during buffer periods are excluded from analysis.

## Identification Strategy 2: Graph Cluster Randomisation

When markets can be partitioned into weakly connected clusters, randomise at the cluster level.

**Definition 18.22 (Graph Cluster Randomisation)** Partition the market graph into clusters  $\mathcal{C}_1, \dots, \mathcal{C}_G$  such that interference within clusters is strong but interference across clusters is weak. Randomly assign clusters to treatment or control. The GATE estimator is:

$$\widehat{\Delta}_{\text{GCR}} = \frac{1}{|\mathcal{G}_1|} \sum_{c \in \mathcal{G}_1} \bar{Y}_c - \frac{1}{|\mathcal{G}_0|} \sum_{c \in \mathcal{G}_0} \bar{Y}_c,$$

where  $\mathcal{G}_1$  and  $\mathcal{G}_0$  are the sets of treated and control cluster indices and  $\bar{Y}_c$  is the average outcome in cluster  $c$ .

**Assumption 105 (Limited Cross-Cluster Interference)** Cluster-level outcomes in cluster  $c$  depend only on treatments within cluster  $c$ :

$$Y_{ct}(\mathbf{D}_c, \mathbf{D}_{-c}) = Y_{ct}(\mathbf{D}_c, \mathbf{D}'_{-c}), \quad \forall \mathbf{D}_{-c}, \mathbf{D}'_{-c}.$$

In practice,  $Y_{ct}$  is an aggregate outcome for cluster  $c$  in period  $t$  (for example, a mean over  $i \in \mathcal{C}_c$ ).

This assumption is plausible when clusters are geographically separated (e.g., different cities) or temporally separated (e.g., different weeks). It fails when users travel across clusters or when platform-wide effects (e.g., reputation, learning) spill across clusters.

*Remark 18.25 (Marketplace Applications)* While the case study below focuses on ride-hailing, the same methods apply to e-commerce marketplaces and other multi-sided platforms. First, seller promotions affect the broader market: subsidising some sellers affects buyer traffic and prices for all sellers, and cluster randomisation at the product category or geographic level can isolate effects. Second, search ranking changes create zero-sum competition: changing ranking for some products affects visibility and sales of competing products (see Section 18.12). Third, commission structure changes spill over through equilibrium: changing fees for some sellers affects their pricing, which spills over to buyer choices and competitor responses (see Section 18.13). The key insight is that marketplace interventions operate through equilibrium effects: any treatment that shifts supply, demand, or prices for treated units will spill over to untreated units through the market mechanism.

## Estimation

For switchback experiments, use regression with calendar controls and cluster standard errors at the window level:

$$Y_{it} = \alpha + \delta D_t + f(\text{calendar}_t) + X'_{it}\beta + \varepsilon_{it},$$

where  $f(\text{calendar}_t)$  captures time-of-day and day-of-week patterns (e.g., hour-of-day dummies and weekday indicators), but is *not* a full set of window fixed effects—including window fixed effects would be collinear with  $D_t$ , which is constant within each window. Cluster standard errors at the window level to account for within-window correlation and interference.

For graph cluster randomisation, use difference-in-means at the cluster level with cluster-robust standard errors. With few clusters, use randomisation inference over the set of possible cluster assignments.

## Variance and Power

Platform experiments face a trade-off between variance and bias from interference or carryover. Shorter time windows reduce carryover bias but increase variance (fewer independent windows). Fewer, larger clusters reduce cross-cluster spillover bias but also increase variance (fewer clusters to compare). Power calculations must account not just for sample size, but for the number of independent windows or clusters and for serial and spatial correlation.

*Remark 18.26 (Long-Run Platform Dynamics)* Switchback experiments and short-run cluster trials capture immediate equilibrium effects but may miss long-run dynamics. First, behavioural adaptation occurs when drivers and sellers learn to game incentive structures over time, so short-run effects may overstate (novelty)

or underestimate (learning curve) long-run effects. Second, network effects mean platform value grows with user base, so experiments in new markets may not generalise to mature markets with established network effects. Third, reputation accumulation creates path dependence as driver and seller ratings accumulate over time, which short-run experiments cannot capture.

For long-run dynamics, consider staggered market rollouts (Chapter 4) or structural models that extrapolate short-run estimates to long-run equilibria.

*Remark 18.27 (Diagnostic Checklist)*

Box 18.14: Platform Experiment Diagnostic Checklist

**For Switchback Experiments:** Diagnose carryover by comparing outcomes in early versus late portions of each window, where persistent differences suggest carryover. Assess buffer adequacy by varying buffer length and checking stability of estimates. Verify time-of-day balance to ensure treatment and control windows are balanced across hours and days. Estimate autocorrelation in window means to assess effective information.

**For Graph Cluster Randomisation:** Diagnose cluster independence using outcomes in control clusters adjacent to treated clusters. Verify cluster balance by comparing cluster-level covariates across treatment arms. With fewer than 20 clusters, use randomisation inference rather than asymptotic standard errors for few-cluster inference.

**For Both Designs:** Report both GATE and DTE if possible, as large differences indicate strong interference. Check whether effects differ by market size, time of day, or user segment to assess heterogeneity. Consider that platform experiments in one city may not generalise to others, limiting external validity.

### Case Study: Ride-Hailing Driver Incentive

We illustrate the switchback approach with a hypothetical ride-hailing platform. The numbers are illustrative and do not represent real data.

**Setting.** A ride-hailing platform runs a driver incentive pilot: \$5 bonus per completed ride during peak hours. The goal is to estimate the effect on completed rides (a supply-and-demand equilibrium outcome).

**Design.** Switchback experiment in one city over 4 weeks. Time windows are 2 hours. Treatment and control alternate randomly, with 30-minute buffer periods between switches. Total: 336 windows (168 treatment, 168 control after excluding buffers).

**Data.** 1.2 million ride requests and 950,000 completed rides. Outcomes are measured at the window level: completed rides, average wait time, driver earnings.

**Results.** The switchback estimate of the GATE on completed rides is +8.2% (SE = 2.1%). Average wait time decreased by 1.4 minutes (SE = 0.3 min). Driver earnings per hour increased by \$4.20 (SE = \$0.80).

**Diagnostics.** Carryover diagnostics compare outcomes in the first 30 minutes of each window to outcomes in the last 30 minutes, suggesting the 30-minute buffer is adequate in this illustration. Time-of-day balance appears even across hours. Autocorrelation in window means is  $\rho = 0.35$ , suggesting fewer effectively independent windows than  $T$ .

**Cost-Benefit.** Compute cost and benefit on a consistent population and horizon. Let  $Q^{\text{treat}}$  denote the number of bonus-eligible completed rides in treated windows. The treated-window incentive cost is then  $\$5 \times Q^{\text{treat}}$ . If the +8.2% estimate applies to treated windows for the same bonus-eligible population, the implied incremental rides are approximately  $0.082 \times Q^{\text{treat}}$ . Valuing incremental rides requires a per-ride contribution margin, not gross revenue. Under a \$15 contribution margin, incremental contribution would be about  $15 \times 0.082 \times Q^{\text{treat}}$ , which is below the incentive cost in this illustration.

**Interpretation.** The driver incentive increases supply and reduces wait times, leading to more completed rides. However, the direct cost exceeds the incremental revenue. The platform might consider targeting incentives to high-demand periods only. It could also reduce the bonus amount. A longer-run evaluation could incorporate driver retention and rider satisfaction that the short-run analysis misses.

## 18.12 Ranking and Recommendation Algorithms

Ranking algorithms determine what users see first—in search results, product listings, content feeds, and recommendations. Items in top positions receive more clicks, but this conflates two effects: the item’s intrinsic relevance and the position’s prominence. Estimating the causal effect of position is essential for optimising ranking algorithms and for understanding how much of an item’s success is due to algorithmic placement rather than quality.

*Remark 18.28 (Ranking in the Taxonomy)* In the taxonomy of Section 18.1, ranking algorithm evaluation involves both *endogeneity* and *interference*. On the identification axis, position is endogenous: the algorithm assigns top positions to items with high predicted relevance, creating simultaneity bias. On the interference axis, ranking is a zero-sum system: moving one item up necessarily moves others down. The effect on the promoted item includes the displacement effect on demoted items. For the broader platform context, see Section 18.11. For seller-side interventions that affect rankings, see Section 18.13.

### Estimand: Position and Relevance Effects

We describe position and relevance effects in a potential-outcomes framework. Let  $Y_{ij} \in \{0, 1\}$  indicate whether user  $i$  clicks on item  $j$  in a given impression. Let  $S_{ij} \in \{1, 2, \dots, K\}$  denote the slot (position) of item  $j$  in user  $i$ ’s list, with 1 the top position.

**Definition 18.23 (Position Effect)** In a ranking, moving item  $j$  to position  $p$  necessarily moves other items. A feasible counterfactual is therefore swap-based. Define  $Y_{ij}(p)$  as the click indicator we would observe if item  $j$  were swapped into position  $p$  within the same impression, displacing the item currently shown at  $p$ . The causal position effect of moving an item from position  $p'$  to  $p$  is:

$$\text{ATE}_{p,p'} = \mathbb{E}[Y_{ij}(p) - Y_{ij}(p')],$$

where the expectation is taken over a suitable population of user–item pairs.

**Definition 18.24 (Relevance Stratification)** Let  $R_j$  denote a relevance proxy for item  $j$  (for example, historical conversion, ratings, or a model score). Relevance is not a manipulable treatment, so we use it as a descriptive stratification. Conditional on a fixed position  $p$ , compare click rates across relevance strata to understand how click propensity varies with relevance holding position fixed.

A commonly used structural assumption is the multiplicative click model:

$$\mathbb{P}(Y_{ij} = 1 | S_{ij}, R_j) = \theta_{S_{ij}} \times \gamma_{R_j},$$

where  $\theta_p$  is interpreted as an examination probability and  $\gamma$  as a relevance-dependent conditional click propensity. This separability is an *assumption*, not a law of nature. Interpreting  $\theta_p$  as an examination proba-

bility requires that the experimental randomisation breaks the correlation between position and unobserved relevance.

### Identification Challenge: Confounding by Relevance

Ranking algorithms place relevant items in top positions. Naive comparison of click rates across positions confounds position effects with relevance effects.

Under algorithmic ranking, let  $R_j$  be item relevance. Then:

$$\text{Cov}(S_{ij}, R_j) < 0,$$

because more relevant items receive better (lower-numbered) positions. In reality, the algorithm's position assignment responds to a high-dimensional relevance signal, which includes *unobserved* components beyond  $R_j$ . Naive comparison of click rates across positions therefore confounds position effects with both observed and unobserved relevance. To identify causal position effects, we need variation in position that is independent of these relevance components.

*Remark 18.29 (Zero-Sum Interference in Rankings)* Ranking is inherently a zero-sum system: there is exactly one position 1, one position 2, and so on. Improving one item's position necessarily worsens others'. This creates interference in several ways. First, displacement effects mean that moving item A from position 5 to position 1 displaces items B, C, D, E downward, so the platform-level effect includes both A's gain and the displaced items' losses. Second, competitive dynamics arise when sellers can influence rankings via advertising, pricing, or quality improvements, making one seller's gain come at competitors' expense. Third, GATE interpretation becomes complex because the Global Average Treatment Effect of a ranking policy change must account for winners and losers, and item-level position effects do not aggregate simply. For the formal treatment of GATE under interference, see Section 18.11.

Experiments that randomise position for a subset of items must consider how displacement affects the non-randomised items.

### Identification Strategy 1: Position Randomisation

The preferred approach is randomising position for a subset of traffic.

**Definition 18.25 (Position Randomisation Experiment)** For a random subset of users, shuffle the ranking of items (fully or partially). Compare click rates across positions in the randomised sample:

$$\hat{\theta}_p = \frac{\sum_i \sum_j Y_{ij} \cdot \mathbf{1}\{S_{ij} = p\}}{\sum_i \sum_j \mathbf{1}\{S_{ij} = p\}},$$

which estimates the average click probability at position  $p$  under the randomisation policy.

**Assumption 106 (Random Position Assignment)** In the experimental sample, position is independent of relevance:

$$S_{ij} \perp\!\!\!\perp R_j.$$

Full randomisation degrades user experience (irrelevant items appear at top). Partial randomisation—swapping adjacent pairs or randomising within quality tiers—balances identification with user experience.

Inference must respect the randomisation unit. If randomisation occurs at the session level, treat sessions as the independent units for uncertainty. If it occurs at the user level with repeated sessions per user, cluster at the user level. In either case, remember that rankings are zero-sum within an impression, so item-level outcomes within the same list are mechanically dependent.

### Identification Strategy 2: Instrumental Variables

When randomisation is infeasible, instruments exploit exogenous variation in position.

**Assumption 107 (Position Instrument)** Let  $Z_{ij}$  be an instrument that shifts rank position. Conditional on controls, IV validity requires relevance and an exclusion restriction:  $Z_{ij}$  affects clicks only through its effect on rank position.

Table 18.11 summarises potential instruments for position. However, several of these are more delicate than their brief descriptions suggest.

**Table 18.11** Instruments for position effects

Instrument	Mechanism	Exclusion Restriction
Algorithmic tie-breaking	Equal predicted relevance resolved by arbitrary rule (item ID, random seed)	Tie-breaker uncorrelated with true relevance. This can fail if the system uses logged historical performance
Display constraints	Screen size or pagination creates visibility discontinuities	Items near boundary have similar relevance
Temporal variation	Algorithm updates change rankings for non-quality reasons	Updates uncorrelated with item quality changes (high exclusion risk: many updates aim to improve relevance)

### Identification Strategy 3: Regression Discontinuity

Pagination and screen boundaries create sharp discontinuities in visibility.

**Definition 18.26 (Pagination RDD)** Consider items ranked around a page boundary at position  $L$ . Since position is discrete, we compare click outcomes at positions  $L$  and  $L+1$  (and possibly in small neighbourhoods on each side) to estimate the effect of being on the first page rather than the second. The basic contrast is:

$$\widehat{\Delta}_{\text{RDD}} = \mathbb{E}[Y_{ij} \mid S_{ij} = L] - \mathbb{E}[Y_{ij} \mid S_{ij} = L + 1].$$

This is a discrete RDD at the page boundary, with position treated as an ordered running variable.

**Assumption 108 (Continuity of Relevance at Page Boundary)** Expected item relevance is continuous at the page boundary:

$$\lim_{p \rightarrow L} \mathbb{E}[R_j \mid S_{ij} = p] = \lim_{p \rightarrow L+1} \mathbb{E}[R_j \mid S_{ij} = p],$$

so that items placed at positions  $L$  and  $L + 1$  have similar relevance. Since visibility jumps discretely at the boundary (first page vs. second page), any discontinuity in click rates at that boundary can be attributed to the visibility difference rather than to systematic relevance differences. This is the standard continuity-of-potential-outcomes assumption applied to a discrete running variable.

This assumption is most plausible when the algorithm's ranking is relatively noisy near the page boundary, or when items are pre-grouped into quality tiers and the page break falls within, rather than between, tiers.

## Estimation

For position randomisation, estimate position effects using the randomised sample only. Fit a click model with item fixed effects and position indicators:

$$\text{logit}(\mathbb{P}(Y_{ij} = 1)) = \alpha_j + \sum_p \delta_p \mathbf{1}\{S_{ij} = p\},$$

where  $\alpha_j$  are item fixed effects absorbing relevance and the position dummies  $\delta_p$  estimate log-odds differences by position.

For IV, use 2SLS with the instrument predicting position in the first stage.

For RDD, use local polynomial regression at the page boundary.

*Remark 18.30 (Dynamic Considerations in Ranking)* Position effects may vary over time and across user experience levels. First, user learning occurs as experienced users learn ranking patterns and may scroll further or trust algorithmic ordering less, so position effects may decay with user sophistication. Second, algorithm updates improve ranking quality over time, shrinking the relevance gap between positions and potentially reducing position effects. Third, item lifecycle matters because new items may benefit more from top positions (discovery) than established items (already known to users).

Reporting position effects by user tenure and item age, rather than as a single scalar, provides insight into these dynamics and helps avoid over-generalising from a single experiment.

## Diagnostic Checklist

### Box 18.15: Ranking Algorithm Diagnostic Checklist

**For Position Randomisation:** Verify that item quality is balanced across positions in the experimental sample. Monitor bounce rates and session duration in the randomised group to assess user experience impact. Ensure adequate power since position effects are often small.

**For Pagination RDD:** Check that item quality metrics are smooth through the page boundary to assess relevance balance. Use diagnostics for strategic behaviour around the boundary. Report estimates at multiple bandwidths for sensitivity.

**For Click Models:** Compare predicted versus observed click rates by position to assess model fit. Diagnose whether the multiplicative assumption is plausible in the experimental sample. Check whether position effects vary by user segment or device type to assess heterogeneity.

## Case Study: E-Commerce Search Ranking

We illustrate position randomisation with a hypothetical e-commerce platform. The numbers are illustrative and do not represent real data.

**Setting.** An e-commerce platform displays 20 products per search results page. The ranking algorithm orders products by predicted purchase probability. We estimate the position effect to understand how much of top-position success is due to visibility vs. relevance.

**Design.** For 5% of search sessions, we randomise the order of the top 10 products. The remaining 95% receive the algorithmic ranking. We compare click-through rates (CTR) by position in the randomised sample.

**Data.** 10 million search sessions over 2 weeks, with 500,000 in the randomised condition. Outcome is click on product listing.

**Results.** Position effects in the randomised sample:

Position	CTR (Randomised)	CTR (Algorithmic)
1	12.3%	18.5%
2	9.8%	14.2%
3	7.4%	11.1%
5	4.2%	7.8%
10	1.8%	4.1%

The randomised CTR provides a baseline click probability at each position under the randomisation policy. The gap between algorithmic and randomised CTR reflects sorting, not an identified causal decomposition.

**Sorting gap.** At position 1, the algorithmic CTR is 18.5% and the randomised CTR is 12.3%. The 6.2 percentage point gap quantifies how much the production algorithm concentrates higher-relevance items at the top, relative to the randomised baseline.

**Diagnostics.** Balance diagnostics suggest that product ratings and prices are similar across positions in the randomised sample (all SMDs  $< 0.05$ ). User experience diagnostics show a modest increase in bounce rate in the randomised group. Model fit diagnostics for the multiplicative click model compare predicted and observed CTR by position.

**Interpretation.** Position effects are substantial: even random items get 12% CTR at position 1 vs. 1.8% at position 10. This implies that ranking algorithm improvements have high leverage—moving a good item from position 10 to position 1 increases its CTR by approximately 10 percentage points from position alone. The platform should invest in ranking quality, as the position effect amplifies relevance differences.

## 18.13 Marketplace Seller Interventions

Marketplace platforms intervene on sellers through fee changes, promotional tools, quality badges, and policy enforcement. These interventions affect not only the treated sellers but also buyers (through prices and selection) and competing sellers (through market share shifts). Estimating the causal effect of seller interventions requires accounting for these equilibrium spillovers.

*Remark 18.31 (Seller Interventions in the Taxonomy)* In the taxonomy of Section 18.1, seller interventions are primarily an *interference* problem: treating one seller shifts demand away from competitors, violating SUTVA. The effect on the treated seller includes business stolen from untreated sellers. The market-level effect nets out these transfers. For the general platform experiment framework, see Section 18.11. Seller interventions may also affect rankings (Section 18.12) and interact with pricing decisions (Section 18.7).

### Estimand: Seller and Market Effects

We distinguish direct effects on treated sellers from spillover effects on the market.

**Definition 18.27 (Direct Seller Effect)** Let sellers be indexed by  $i$  and periods by  $t$ . Let  $D_{it} \in \{0, 1\}$  indicate whether seller  $i$  receives the intervention in period  $t$ . Under interference, seller outcomes can depend on competitors' treatments. Write potential outcomes as  $Y_{it}(d, D_{-i,t})$ , where  $D_{-i,t}$  denotes the collection of competitors' treatments in period  $t$ . The direct effect on treated seller-periods is:

$$\text{ATT}_{\text{direct}} = \mathbb{E}[Y_{it}(1, D_{-i,t}) - Y_{it}(0, D_{-i,t}) \mid D_{it} = 1],$$

which averages the contrast between treated and untreated potential outcomes over the experiment's assignment mechanism for competitors' treatments. This captures how much the intervention changes treated sellers' outcomes, given the actual mix of treated and untreated competitors in the experiment.

**Definition 18.28 (Spillover Effect on Competing Sellers)** The spillover effect on untreated seller  $k$  from treating seller  $i$  is:

$$\Delta_{\text{spillover}}(i \rightarrow k) = \mathbb{E}[Y_{kt}(0, D_{it} = 1, D_{-\{i,k\},t}) - Y_{kt}(0, D_{it} = 0, D_{-\{i,k\},t})],$$

holding  $k$ 's own treatment at zero and averaging over the experiment's assignment mechanism for other competitors  $D_{-\{i,k\},t}$ . In practice we seldom estimate pairwise spillovers. Instead we summarise how outcomes for untreated sellers vary with their exposure to treated competitors. In competitive markets, spillovers are typically negative (business stealing). For quality improvements that expand the category, they may be positive.

**Definition 18.29 (Buyer Welfare Effect)** Let  $V_{bt}(\mathbf{D}_t)$  be buyer  $b$ 's surplus (e.g., consumer surplus, transaction value) in period  $t$  under a full seller-treatment assignment  $\mathbf{D}_t$ . The buyer welfare effect of rolling out the intervention to all sellers is:

$$\text{GATE}_{\text{buyer}} = \mathbb{E}[V_{bt}(\mathbf{1}) - V_{bt}(\mathbf{0})],$$

where  $\mathbf{1}$  denotes the all-treated assignment (all sellers treated) and  $\mathbf{0}$  denotes the all-control assignment (no sellers treated). This is a buyer-side GATE: it captures the total change in buyer outcomes when all sellers are treated versus none.

The platform's objective typically combines seller revenue, buyer welfare, and platform fee revenue. These components may move in different directions, so experiments should report them separately.

### Identification Challenge: Competitive Spillovers

Seller interventions create competitive externalities. Treating one seller shifts demand away from competitors.

*Remark 18.32 (Competitive Interference)* In a marketplace with  $J$  sellers, treating seller  $j$  can affect seller  $k$ 's outcomes for  $k \neq j$ . The sign depends on the intervention: fee reductions or promotions often create negative spillovers (business stealing), while quality improvements can create positive spillovers (market expansion).

Naive randomisation at the seller level estimates the direct effect but misses spillovers. If spillovers are negative, the direct effect overstates the market-level impact.

### Identification Strategy 1: Market-Level Randomisation

Randomise the intervention at the market level to capture full equilibrium effects.

**Definition 18.30 (Market-Level Experiment)** Partition the platform into markets indexed by  $c = 1, \dots, G$  (for example, cities or product categories). Randomly assign markets to treatment (all sellers receive intervention) or control (no sellers receive intervention). Let  $\bar{Y}_c(\mathbf{D}_c)$  be the average outcome in market  $c$  under a within-market assignment  $\mathbf{D}_c$ . Under the no-cross-market-spillovers assumption (Assumption 109), the difference in average  $\bar{Y}_c$  between treated and control markets identifies a market-level GATE:

$$\text{GATE}_{\text{market}} = \mathbb{E}[\bar{Y}_c(\mathbf{1}) - \bar{Y}_c(\mathbf{0})],$$

where  $\mathbf{1}$  and  $\mathbf{0}$  denote “all sellers in the market treated” and “no sellers treated” respectively. This estimand nets out business stealing within the market and measures how the intervention changes total market outcomes. The natural estimator is:

$$\hat{\Delta}_{\text{market}} = \frac{1}{|\mathcal{G}_1|} \sum_{c \in \mathcal{G}_1} \bar{Y}_c - \frac{1}{|\mathcal{G}_0|} \sum_{c \in \mathcal{G}_0} \bar{Y}_c,$$

where  $\mathcal{G}_1$  and  $\mathcal{G}_0$  are the sets of treated and control market indices.

**Assumption 109 (No Cross-Market Spillovers)** Outcomes in market  $c$  depend only on treatments within market  $c$ :

$$Y_{ct}(\mathbf{D}_c, \mathbf{D}_{-c}) = Y_{ct}(\mathbf{D}_c, \mathbf{D}'_{-c}), \quad \forall \mathbf{D}_{-c}, \mathbf{D}'_{-c}.$$

This assumption holds when markets are geographically or categorically separated. It fails when sellers operate across markets or when buyers substitute across markets.

### Identification Strategy 2: Seller Cluster Randomisation

When market-level randomisation is infeasible, cluster sellers into groups with strong within-cluster competition.

**Definition 18.31 (Seller Cluster Randomisation)** Partition sellers into clusters  $\mathcal{C}_1, \dots, \mathcal{C}_G$  based on competitive proximity (e.g., same subcategory, same price tier). Randomly assign clusters to treatment or control. The cluster-level effect captures within-cluster spillovers:

$$\widehat{\Delta}_{\text{cluster}} = \frac{1}{|\mathcal{H}_1|} \sum_{c \in \mathcal{H}_1} \bar{Y}_c - \frac{1}{|\mathcal{H}_0|} \sum_{c \in \mathcal{H}_0} \bar{Y}_c,$$

where  $\mathcal{H}_1$  and  $\mathcal{H}_0$  are the sets of treated and control seller-cluster indices.

Cluster definition is critical: clusters should contain close competitors. Use product similarity, price overlap, or buyer co-consideration data to define clusters.

### Identification Strategy 3: Difference-in-Differences with Staggered Rollout

If the intervention rolls out to sellers over time, staggered DiD exploits timing variation.

**Assumption 110 (Parallel Trends for Seller Rollout)** In the absence of the intervention, seller outcomes would follow parallel trends:

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid G_i = g'], \quad \forall g, g',$$

where  $G_i$  is the cohort (rollout wave) for seller  $i$ .

Use heterogeneity-robust estimators (Chapter 4). Note that staggered DiD captures the direct effect plus spillovers to not-yet-treated sellers, but not the full equilibrium effect.

## Estimation

For market-level experiments, treat markets as the independent assignment units for inference. Inference is therefore driven by the number of independent markets  $G$ , not by the number of seller transactions. Use difference-in-means with uncertainty clustered at the market level. With few markets, use randomisation inference.

For seller cluster randomisation, treat seller clusters as the independent assignment units for inference. Aggregate outcomes to the cluster level and use cluster-robust standard errors, with effective sample size driven by the number of independent clusters.

For staggered rollout, apply Callaway and Sant'Anna [2021] or Sun and Abraham [2021] estimators. Include seller fixed effects and time fixed effects:

$$Y_{it} = \alpha_i + \gamma_t + \delta D_{it} + X'_{it}\beta + \varepsilon_{it}.$$

## Decomposing Direct and Spillover Effects

When both treated and control sellers are observed within the same market, decompose the total effect.

**Definition 18.32 (Effect Decomposition (Approximation))** In a partial-treatment design where fraction  $p$  of sellers are treated, we can *approximate* direct and spillover components using exposure to treated competitors. Let  $E_k$  measure untreated seller  $k$ 's exposure to treated competitors (e.g., the share of its closest rivals that are treated). Comparing outcomes for untreated sellers with high versus low  $E_k$  provides an empirical spillover gradient:

$$\text{Direct contrast: } \hat{\Delta}_{\text{direct}} = \bar{Y}_{\text{treated}} - \bar{Y}_{\text{control, low exposure}}, \quad (18.5)$$

$$\text{Spillover contrast: } \hat{\Delta}_{\text{spillover}} = \bar{Y}_{\text{control, high exposure}} - \bar{Y}_{\text{control, low exposure}}. \quad (18.6)$$

The total market effect when fraction  $p$  of sellers are treated is then summarised heuristically as  $p \cdot \hat{\Delta}_{\text{direct}} + (1 - p) \cdot \hat{\Delta}_{\text{spillover}}$ .

This decomposition is an *approximation*: it requires defining an exposure threshold to classify “low exposure” as a proxy for the idealised “no spillover” group, which may not exist in a dense competitive network. The linear aggregation ignores non-linearities in competitive response and assumes spillovers scale linearly with treatment intensity. In practice, construct  $\hat{\Delta}_{\text{spillover}}$  by contrasting untreated sellers in markets or clusters with many treated competitors against those with few treated competitors, using a clearly defined exposure-threshold rule.

*Remark 18.33 (Dynamic Effects of Seller Interventions)* Seller interventions may have temporal dynamics beyond immediate effects. Sellers adjust pricing, inventory, and listing strategies over time in response to changed incentives, so short-run effects may differ from long-run equilibrium effects. Fee changes can also induce entry and exit, so the seller composition in treated markets may diverge from control over time. Improved

selection or lower prices may attract buyers who develop purchasing habits, creating persistent demand even after the intervention ends. Short-run experiments capture immediate responses. Longer observation windows or staggered rollouts (Chapter 4) are needed to assess persistence and equilibrium effects.

## Diagnostic Checklist

Box 18.16: Seller Intervention Diagnostic Checklist

**For Market-Level Experiments:** Diagnose market independence by checking outcomes in control markets adjacent to treated markets. Compare market-level covariates (size, seller count, category mix) across arms. With fewer than 20 markets, use randomisation inference rather than asymptotic standard errors.

**For Seller Cluster Randomisation:** Validate that clusters contain close competitors (for example, high cross-elasticity or high buyer co-consideration). Check seller characteristics are balanced across treated and control clusters. Diagnose cross-cluster spillovers by examining outcomes in control clusters near treated clusters.

**For Effect Decomposition:** Define the competitive network (same category, price tier, geography). Compute each control seller's exposure to treated competitors. Verify that spillover gradients have the expected sign (negative for business stealing, positive for market expansion).

## Case Study: Marketplace Fee Reduction

We illustrate market-level randomisation with a hypothetical e-commerce marketplace. The numbers are illustrative and do not represent real data.

**Setting.** A marketplace runs a 2 percentage point reduction in seller commission (from 15% to 13%) to stimulate seller activity. The goal is to estimate the effect on total marketplace GMV (gross merchandise value), accounting for competitive dynamics.

**Design.** Market-level experiment across 40 cities. 20 cities receive the fee reduction for all sellers. 20 cities serve as control. Duration: 8 weeks.

**Data.** 500,000 sellers across 40 cities and 50 million transactions. Outcomes: seller GMV, number of listings, average price, buyer transactions.

**Results.** Market-level effects:

Outcome	Treatment Effect	SE
Seller GMV	+8.4%	2.1%
Number of listings	+12.1%	3.2%
Average price	-1.8%	0.6%
Buyer transactions	+5.2%	1.4%

The fee reduction increased seller activity (more listings) and was partially passed through to buyers (lower prices), leading to more transactions.

**Decomposition.** Within treated markets, we compare sellers by exposure to competing treated sellers. Sellers with few treated competitors show GMV +11.2%, while sellers with many treated competitors show GMV +6.1%. The difference (5.1 percentage points) reflects competitive spillovers: when all competitors also receive the fee reduction, the relative advantage is smaller.

**Revenue Impact.** The 8.4% GMV increase generates additional platform revenue of approximately \$4.2 million (at 13% take rate). However, the 2 pp fee reduction on baseline GMV costs approximately \$6.8 million. Net platform revenue decreases by \$2.6 million.

**Interpretation.** The fee reduction successfully stimulates marketplace activity: more listings, lower prices, more transactions. However, the direct revenue loss exceeds the indirect gain from higher GMV. The intervention is profitable only if the GMV increase persists after the fee reduction ends (habit formation), if new sellers acquired during the promotion have high lifetime value, or if the platform values buyer welfare beyond direct revenue. A longer-term analysis with seller and buyer retention outcomes would clarify the full ROI.

## 18.14 Method Selection Summary

This section synthesises the application-specific guidance from the preceding sections into a unified framework for method selection. The choice of causal method depends on three factors: the source of identifying variation, the data structure, and the plausibility of required assumptions. No single method dominates. Credibility comes from matching the method to the context.

*Remark 18.34 (Method Selection and the Two-Dimensional Taxonomy)* The taxonomy in Section 18.1 organises marketing problems along two dimensions: the *identification threat* (confounding, endogeneity, interference) and the *temporal structure* (static or dynamic effects). Method selection requires addressing both. The identification strategy responds to the main threat. Confounding calls for selection-on-observables or parallel-trends designs. Endogeneity calls for instruments or experimental calibration. Interference calls for cluster or switchback designs. The estimation approach responds to the temporal structure. Static effects permit standard panel estimators. Dynamic effects require distributed lags, adstock models, or impulse-response estimation.

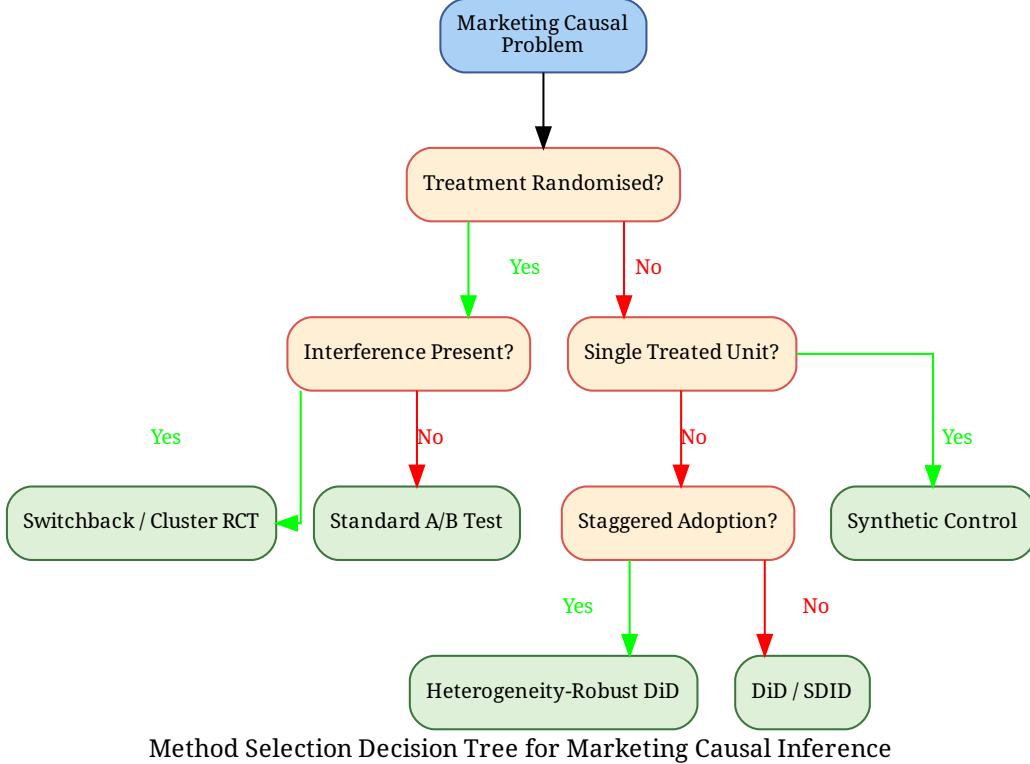
The table below maps each application to both dimensions, linking conceptual classification to operational method choice. Abbreviations such as “Confound.”, “Endog.”, and “Interf.” follow Section 18.1. Method acronyms (SC, SDID, MMM, IV-FE, PSM, DR) follow the earlier method chapters.

### The Decision Process

Method selection proceeds in three steps. First, define the estimand: what causal quantity do you need—ATT, ATE, GATE (when interference matters), an elasticity, or a dose–response function? The estimand determines which methods are even applicable (Section 18.2). Second, identify the source of variation: is treatment randomised, quasi-randomly assigned, or observational? Randomisation permits simple comparisons. Quasi-experiments require design-based methods. Purely observational data require strong modelling and identification assumptions. Third, assess assumption plausibility. Every method requires assumptions—parallel trends, exclusion restrictions, no carryover, limited spillovers—and you should defend them in context rather than assert them by default. Figure 18.8 provides a visual guide to this process.

### Problem-Method Mapping

Table 18.12 maps the marketing problems covered in this chapter to recommended methods. The “Key Assumption” column highlights the primary identification condition associated with the primary method in each row. It is shorthand rather than an exhaustive list.



**Fig. 18.8** Method selection decision tree

*The flowchart guides the practitioner from key problem characteristics—such as the presence of randomisation, interference, or time-series data—to the most appropriate causal estimator. Start at the top and follow branches based on your data and design.*

## When Assumptions Fail

Every method in Table 18.12 requires assumptions that may be violated. The following guidance addresses common failure modes.

**Parallel trends diagnostics indicate departures.** If pre-trend diagnostics suggest departures from parallel trends, consider alternative methods that change the identifying assumptions, such as SC/SDID/factor approaches that rely on counterfactual fit and factor stability. You can also use sensitivity analysis to bound effects under plausible trend deviations (Section 17.8). If neither route is credible, report that the design does not support a causal claim.

**Table 18.12** Marketing problems and recommended causal methods

Problem	Taxonomy	Data	Primary Method	Alternative	Key Assumption (primary)
Ad incrementality (randomised geo hold-out)	Interf. + Dyn.	Geo experiment	Cluster RCT	—	Random assignment and limited spillovers
Ad incrementality (observational geo panel)	Confound. + Dyn.	Geo-panel	SC/SDID	DiD	Valid synthetic/factor counterfactual
Digital attribution	Endog. + Interf.	User-session	Natural expt	IV	Local randomisation / exogenous variation
Media mix	Endog. + Dyn.	Time-series	Bayesian MMM	Calibration	Experimental calibration informs coefficients
Loyalty programme	Confound. + Dyn.	Customer panel	Staggered DiD	RDD	Parallel trends with limited anticipation
Price elasticity	Endog. + Static	Store-week	IV-FE	DiD	Exclusion restriction
CLV attribution	Confound. + Dyn.	Customer cohort	PSM/DR	DiD	Selection on observables
Dynamic pricing	Endog. + Static	Route-day	IV-FE	RDD	Instrument validity
Subscription	Confound. + Dyn.	User-session	A/B experiment	RDD	Random assignment of paywall policy
Platform expt	Interf. + Static	User-time	Switchback	Cluster RCT	Limited carryover and limited cross-cluster spillover
Ranking effects	Interf. + Endog.	Session	Position RCT	IV	Random position assignment (independent of relevance) and a defined randomisation policy respecting the zero-sum constraint
Seller intervention	Interf. + Static	Market-level	Market/cluster RCT	DiD	Limited cross-market/cross-cluster spillover

**Weak instruments.** If the first-stage F-statistic is below conventional thresholds, IV estimates are unreliable. Consider finding stronger instruments or using weak-instrument-robust inference such as Anderson–Rubin-type procedures. If needed, report reduced-form effects only. See Remark 18.2 for the scope of IV coverage in this book.

**Carryover effects detected.** If switchback experiments show persistent effects across windows, the no-carryover assumption fails. Consider lengthening time windows or extending buffer periods. If carryover is part of the product, model it explicitly with lagged treatment indicators.

**Cross-cluster spillovers.** If cluster randomisation shows effects in control clusters adjacent to treated clusters, the limited-spillover assumption fails. Consider using larger clusters or excluding boundary observations. If spillovers are central to the mechanism, model them using exposure measures.

**Selection on unobservables.** If propensity score methods show sensitivity to unobserved confounding (low Rosenbaum bounds), the unconfoundedness assumption is suspect. Consider finding an instrument or natural experiment. You can also use bounds under partial identification. Otherwise, report that the causal interpretation is tentative.

**Manipulation at thresholds.** If RDD shows bunching at the threshold (for example, via density diagnostics), the local continuity restriction is doubtful. Consider treating a "donut" RDD excluding observations at the threshold as a sensitivity check rather than a repair. You can also switch to a different identification strategy. If neither approach is credible, report that the design is compromised.

## Triangulation and Robustness

When multiple methods are applicable, triangulation strengthens conclusions. If geo-experiments, MMM, and digital attribution all point to similar advertising effects, confidence increases. If they disagree, investigate why: different estimands, different time horizons, or assumption violations.

Report results from multiple methods when feasible. A specification curve (Section 17.7) across methods, not just within a method, provides the most honest assessment of what the data support.

### Box 18.17: Method Selection Checklist

Before committing to a method, verify:

- **Estimand clarity:** Is the causal quantity precisely defined in potential outcomes notation?
- **Identification source:** What is the source of exogenous variation? Can you defend it?
- **Assumption plausibility:** Are the required assumptions credible in this context?
- **Diagnostic plan:** What diagnostics will you run? What would cause you to abandon the method?
- **Sensitivity analysis:** How will you assess robustness to assumption violations?
- **Alternative methods:** What other methods could address the same question? Do they agree?

If you cannot answer these questions, return to the application protocol (Section 18.2) before proceeding.

The taxonomy in Section 18.1 provides the conceptual foundation for classifying marketing problems. This section provides the operational guidance for selecting and validating methods. Together, they form the practitioner's handbook for credible causal inference in marketing.

## 18.15 Synthesis and Reporting

Rigorous causal analysis is wasted if findings are not communicated effectively. This section provides guidance on reporting standards, stakeholder communication, and common pitfalls. The goal is to translate statistical estimates into actionable business insights while maintaining intellectual honesty about uncertainty and assumptions.

This synthesis section completes the workflow established in this chapter: the taxonomy (Section 18.1) classifies the problem, the application protocol (Section 18.2) structures the analysis, the method selection summary (Section 18.14) guides design choices, and this section ensures findings are communicated credibly.

### Reporting Standards

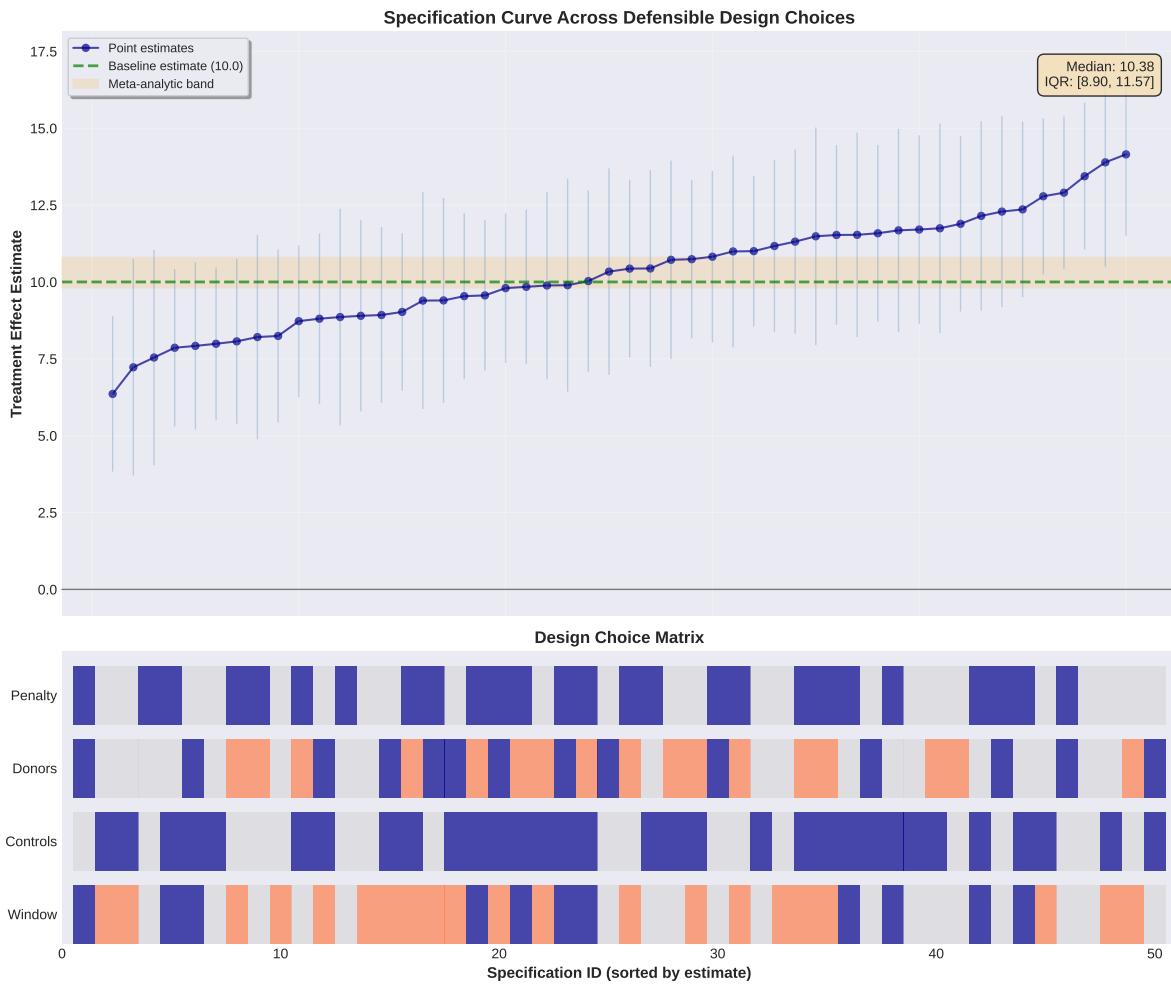
Every causal analysis should document the estimand, the assumptions under which it is identified, the estimator aligned with those assumptions, the diagnostics used to probe plausibility, and the inference procedure used to quantify uncertainty. Pre-registration—documenting the design before seeing results—reduces discretion and makes specification searching visible. It ties directly back to the design and diagnostics principles in Chapter 17 and the sensitivity frameworks in Section 17.8.

#### Box 18.18: Reporting Standards Checklist

- 1. Pre-registration:** Document the estimand, design, and primary specification before analysis. Register with a timestamp (internal wiki, OSF, or email to stakeholders).
- 2. Estimand clarity:** State the causal quantity (e.g., ATT, ATE, GATE, an elasticity, or incremental CLV) and its business interpretation. Avoid vague language like “the effect of X on Y.”
- 3. Assumption disclosure:** List the identification assumptions and discuss their plausibility in context. State assumptions in the main text rather than only in footnotes.
- 4. Diagnostic reporting:** Include pre-trend diagnostics, balance summaries, fit statistics, and placebo or falsification checks (Chapter 17). Report p-values and effect sizes, not just pass/fail.
- 5. Sensitivity analysis:** Report specification curves or robustness to alternative designs (Section 17.8). Show what would change the conclusion.
- 6. Confidence intervals:** Report uncertainty using appropriate inference (cluster-robust, bootstrap, permutation). Never report point estimates without uncertainty.
- 7. Business translation:** Convert the ATT to actionable metrics (iROAS, incremental CLV, elasticity). Include confidence intervals on business metrics.

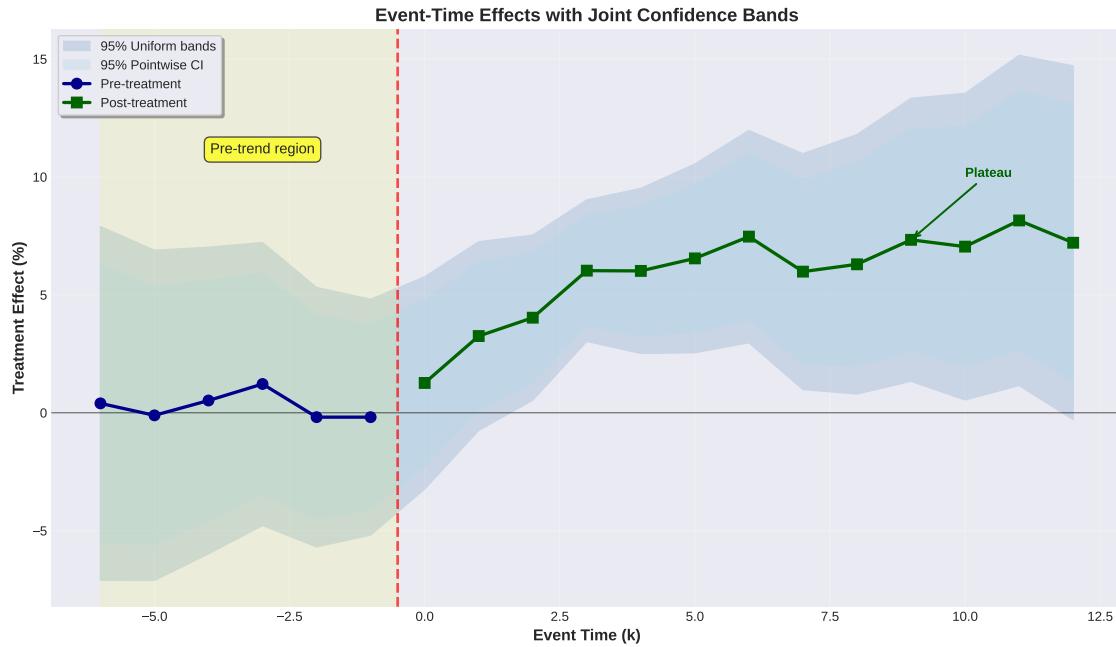
## Visualising Causal Evidence

Two visualisations are essential for communicating causal findings: specification curves and event-study plots with joint confidence bands.



**Fig. 18.9** Specification curve analysis

*Top panel:* Point estimate and 95% confidence interval for each specification, sorted by effect size. *Bottom panel:* Analytical choices (covariates, bandwidths, estimators) corresponding to each estimate. The curve reveals stability. If most specifications yield similar estimates, the finding is less sensitive to analyst discretion. If estimates vary widely, the conclusion depends on modelling choices.



**Fig. 18.10** Event-study estimates with joint confidence bands

*Dynamic treatment effects relative to event time ( $t = 0$ ). The shaded region illustrates a joint confidence band constructed to control family-wise coverage across event times under the dependence assumptions used for inference (see Chapter 16). Pre-treatment coefficients near zero are consistent with parallel trends. Post-treatment coefficients describe the estimated effect dynamics.*

## Communicating to Stakeholders

Technical audiences (data scientists, economists) and business audiences (executives, marketers) require different communication strategies.

### For technical audiences:

- Lead with the identification strategy: What is the source of exogenous variation?
- Report diagnostics in detail: pre-trends, balance, placebos, sensitivity.
- Discuss limitations honestly: What assumptions are most vulnerable?
- Provide a reproducible internal analysis package (code, parameter settings, and input extracts or synthetic validation fixtures) and a clear audit trail so colleagues can rerun the pipeline.
- Use appendices for full formal statements of assumptions (such as parallel trends or SUTVA) so that technical readers can audit the logic without overloading the main narrative.

### For business audiences:

- Lead with the business question and answer: "The campaign generated \$2.4M incremental revenue."
- Translate to familiar metrics: iROAS, lift percentage, payback period.

- Communicate uncertainty in business terms: "We are 95% confident the true lift is between 3% and 13%."
- Use decision framing: "If the true effect is at the lower bound, the campaign still breaks even."
- Avoid jargon: Say "comparison group" not "counterfactual". Say "similar trends before the campaign" not "parallel trends assumption."

### The one-page executive summary:

1. **Question:** What business question did we answer?
2. **Answer:** What is the causal effect in business terms?
3. **Confidence:** How certain are we? What is the range of plausible effects?
4. **Caveats:** What assumptions did we make? What could change the conclusion?
5. **Recommendation:** What action should the business take?

### Common Reporting Pitfalls

Practitioners frequently make errors that undermine the credibility of causal claims.

1. **Reporting only significant results.** If you ran multiple analyses and report only the one that "worked," you are p-hacking. Report all pre-registered analyses, including nulls.
2. **Ignoring uncertainty.** A point estimate of 8% lift with a 95% CI of [-2%, 18%] is not "an 8% effect"—it is "an effect that could plausibly be zero or as high as 18%." Communicate the full range.
3. **Conflating statistical and practical significance.** A statistically significant effect of 0.1% lift is not actionable. Report effect sizes in business terms and assess whether they matter.
4. **Hiding failed diagnostics.** If pre-trends are rejected, do not proceed as if they passed. Report the failure and either switch methods or acknowledge the limitation.
5. **Overstating external validity.** A geo-experiment in 5 DMAs does not generalise to all markets. Be explicit about the population to which results apply.
6. **Confusing correlation with causation in language.** Avoid phrases like "X is associated with Y" when you mean "X causes Y." If your design supports causation, say so. If not, do not imply it.
7. **Presenting sensitivity analysis as robustness.** If the specification curve shows sign changes or wide swings, the result is not robust—it is fragile. The role of the specification curve is to reveal this fragility, not to certify robustness. Do not claim robustness when the evidence does not support it.

### **Example Report Structure**

A complete causal analysis report should include the following sections:

**Box 18.19: Causal Analysis Report Template****1. Executive Summary** (1 page)

- Business question and answer
- Key finding with confidence interval
- Recommendation

**2. Background and Motivation** (1–2 pages)

- Business context and decision at stake
- Prior evidence and expectations

**3. Data and Design** (2–3 pages)

- Data sources, sample, and time period
- Treatment definition and timing
- Identification strategy and key assumptions

**4. Results** (2–3 pages)

- Main estimate with confidence interval
- Event-study or dynamic effects plot
- Heterogeneity analysis (if applicable)

**5. Diagnostics** (2–3 pages)

- Pre-trend diagnostics and balance tables
- Placebo or falsification checks and fit statistics
- Specification curve or robustness checks

**6. Sensitivity Analysis** (1–2 pages)

- What would change the conclusion?
- Bounds under alternative assumptions

**7. Limitations and Caveats** (1 page)

- Assumptions that may be violated
- External validity concerns
- What we cannot conclude

**8. Business Implications** (1 page)

- Translation to business metrics
- Decision recommendation
- Next steps and follow-up analyses

**Appendix:** Technical details, reproducible code (with package versions), data dictionary, and pre-registration documentation

### The Honest Conclusion

The goal of causal inference in marketing is not to produce impressive-looking results but to inform better decisions. An honest null result—"we could not detect an effect, and the confidence interval rules out effects larger than X"—is more valuable than a spurious positive. A finding that depends on fragile assumptions should be reported as tentative, not definitive.

The credibility of marketing analytics depends on the integrity of its practitioners. Report what the data support, acknowledge what they do not, and resist pressure to overstate findings. The methods in this book provide the tools for rigorous causal inference. The responsibility for honest reporting lies with the analyst.

## 18.16 Common Pitfalls and Anti-Patterns

This section catalogues common errors that undermine causal credibility in marketing analytics. These pitfalls recur across organisations and domains. Recognising them is the first step to avoiding them.

*Remark 18.35 (Pitfalls and the Taxonomy)* The pitfalls below map to the identification threats in Section 18.1:

1. **Confounding pitfalls:** Selection into treatment, survivorship bias, unobserved heterogeneity.
2. **Endogeneity pitfalls:** Simultaneity and algorithmic targeting (see Sections 18.5 and 18.4), budget responding to sales.
3. **Interference pitfalls:** Ignoring spillovers, cannibalisation, competitive externalities (see Section 18.11).
4. **Temporal pitfalls:** Short-run optimisation ignoring long-run brand effects, carryover, habit formation.

These pitfalls are not isolated mistakes but organised manifestations of the identification threats in Section 18.1. Recognising which threat applies guides both diagnosis and remedy.

### Conceptual Pitfalls

These errors reflect fundamental misunderstandings about causality.

**Attribution is not causality.** Last-click attribution, multi-touch models, and data-driven attribution are all descriptive allocations of credit. They do not answer the counterfactual question: what would have happened without this touchpoint? Only designs with a defensible counterfactual—randomised experiments or credible quasi-experiments—support causal claims. Purely descriptive attribution rules do not. See Section 18.4 for causal approaches to attribution.

**ROAS is not a causal metric.** Return on Ad Spend divides revenue by spend but ignores the baseline. If a customer would have purchased anyway, the "return" is illusory. Incremental ROAS (iROAS) corrects this by subtracting the counterfactual revenue (Definition 18.2).

**Platform metrics without counterfactuals.** Engagement metrics such as clicks, views, and impressions measure activity but not effect. A campaign with high click-through rate may simply be well-targeted to users who would have converted regardless. Without a control group, these metrics are not causal.

**Confusing prediction with identification.** A machine learning model that predicts sales well does not identify the causal effect of marketing. Predictive accuracy on held-out data is not the same as unbiased estimation of treatment effects. Causal inference requires assumptions about the data-generating process, not just out-of-sample fit. See Chapter 12 for how to integrate ML into causal designs without conflating prediction with identification.

**Platform opacity as latent confounding.** Many advertisers now see only aggregate data from platforms that has been processed by opaque algorithms. Audience-finding algorithms, automated bidding, and modelled conversions create unquantifiable confounding: the platform’s black-box optimisation affects both treatment assignment and measured outcomes. When the data you receive has already been filtered through an algorithm you cannot inspect, standard identification strategies may fail silently. When the platform’s internal model both selects treatment and imputes outcomes, standard identification assumptions—including unconfoundedness and IV exclusion—may fail even if you observe rich covariates. See Remark 18.6 for implications and mitigations.

## Design Pitfalls

These errors arise from flawed experimental or quasi-experimental designs.

**Ignoring interference.** In platform settings, treating one user affects others. A user-level A/B experiment may estimate a direct effect under the experiment’s partial-treatment regime, which can differ sharply from the rollout (all-treated vs all-control) effect. If the lift comes from cannibalising the control group, the total platform effect can be small or even negative. Switchback and cluster designs are essential (Section 18.11).

**Selection into treatment.** Comparing loyalty members to non-members, or ad-exposed users to unexposed users, confounds the treatment effect with the selection effect. High-value customers select into programmes and are targeted by algorithms. Design-based identification is required (Section 18.6, Section 18.8).

**Survivorship bias.** Analysing only customers who remained active ignores those who churned. If treatment affects retention, conditioning on survival biases the estimated effect on spending. Use intention-to-treat analysis or model attrition explicitly.

**Short-run optimisation ignoring long-run effects.** Optimising for immediate conversions may cannibalise brand equity. Performance marketing shows strong short-run ROI. Brand marketing shows weak short-run ROI but builds the stock that sustains long-run demand. This reflects the *temporal structure* dimension of the taxonomy: static estimators miss dynamic effects. Evaluate over appropriate time horizons (Section 18.5).

## Analysis Pitfalls

These errors occur during estimation and inference.

**p-hacking and the garden of forking paths.** Running many specifications and reporting only the significant ones inflates false positive rates. Pre-registration, specification curves, and honest reporting of all analyses are essential safeguards (Section 18.15).

**Simpson's paradox.** An effect that appears positive in aggregate may be negative within every segment (or vice versa). Always check whether aggregate effects mask heterogeneity. Report segment-level effects when relevant.

**Ignoring multiple testing.** Running 20 hypothesis tests at  $\alpha = 0.05$  yields one false positive on average. Define outcome families and pre-specify primary estimands. Use FWER control for a small number of primary hypotheses and FDR for exploratory screens. Report both adjusted and unadjusted results with clear labels (Chapter 16).

**Treating point estimates as truth.** A point estimate of 8% lift with a 95% CI of [-2%, 18%] is not "an 8% effect"—it is "an effect that could plausibly be zero or as high as 18%." Always communicate uncertainty.

## Assumption Violations by Domain

Table 18.13 summarises the key assumption violations and their mitigations across marketing domains. Each row links to the relevant section for detailed guidance.

## The Anti-Pattern Checklist

Before finalising any causal analysis, verify that you have avoided these anti-patterns.

**Table 18.13** Assumption violations and mitigations by marketing domain

Domain	Key Assumption	Common Violation	Mitigation
Geo-experiment (Section 18.3)	Random assignment and limited spillovers	Regional shocks, leakage	Buffer zones, spillover diagnostics, randomisation inference
Digital attribution (Section 18.4)	Exogenous variation	Algorithmic targeting	Natural experiments, IV
Media mix (Section 18.5)	Exogeneity of media spend after controls	Budget responds to sales	Experimental calibration, IV, sensitivity
Loyalty programme (Section 18.6)	No anticipation	Customer gaming thresholds	Donut RDD, staggered DiD
Price elasticity (Section 18.7)	Exclusion restriction	Demand-correlated costs	Alternative IVs, bounds
Dynamic pricing (Section 18.9)	Instrument validity	Algorithmic endogeneity	RDD at thresholds (when IV implausible)
Subscription/paywall (Section 18.10)	Local continuity at the threshold (RDD) or random assignment (A/B)	Avoidance and bunching	Density diagnostics, sensitivity (donut RDD), A/B experiments
Platform experiment (Section 18.11)	Limited carryover	Habit formation	Longer washout, buffers
Platform experiment (Section 18.11)	Restricted interference (limited cross-cluster spillover)	Interference, cannibalisation	Switchback, cluster RCT
CLV attribution (Section 18.8)	Selection on observables	Unobserved heterogeneity	Sensitivity (essential), IV
Ranking (Section 18.12)	Random position	Relevance confounding	Position randomisation
Seller intervention (Section 18.13)	Limited spillover	Competitive externalities	Market-level RCT

#### Box 18.20: Anti-Pattern Checklist

##### Conceptual:

- Did you define a causal estimand, not just a descriptive metric?
- Did you specify the counterfactual explicitly?
- Did you distinguish prediction from causal identification?

##### Design:

- Did you account for interference if units interact?
- Did you address selection into treatment?
- Did you consider survivorship bias?
- Did you evaluate over an appropriate time horizon?

##### Analysis:

- Did you pre-register the primary specification?
- Did you correct for multiple testing if applicable?
- Did you report uncertainty, not just point estimates?
- Did you check for Simpson's paradox in subgroups?

##### Reporting:

- Did you report all pre-registered analyses, including nulls?
- Did you disclose failed diagnostics?
- Did you acknowledge limitations and external validity concerns?

By avoiding these pitfalls and adhering to the protocols in this chapter, marketing analysts can produce evidence that meets the standards of credible causal inference. Combined with the method-selection and reporting checklists in Sections 18.14 and 18.15, this anti-pattern list provides a practical guardrail for everyday analysis. The goal is not to claim effects that do not exist but to identify effects that do—and to be honest when the evidence is inconclusive.

## Chapter 19

# Data, Measurement, and Platforms

This chapter defines a measurement workflow. We first fix the target estimand. We then define  $D_{it}$ ,  $Y_{it}$ , and  $X_{it}$  from raw sources. Finally, we state which identification assumptions—parallel trends, exclusion, and overlap—remain defensible after those transformations. Measurement error is not merely a nuisance that inflates standard errors. Depending on where it enters (outcome, treatment, timing, or sample inclusion) and whether it is differential, it can also invalidate the identification assumptions that underpin causal designs. For instrumental variables, mismeasurement can weaken first-stage relevance by attenuating  $\text{Cov}(Z_{it}, D_{it})$ . It can also undermine exclusion when the measurement process itself depends on unobservables that affect outcomes. For difference-in-differences, parallel trends is a restriction on the outcome actually used in estimation. If outcome mismeasurement evolves differently by group or over time, it can mimic or mask treatment effects even when latent outcomes would have satisfied parallel trends [Bound et al., 2001]. We outline design-relevant practices for identity linking, keys, joins, deduplication, and time or calendar alignment under privacy and policy constraints. We diagnose and mitigate attribution bias, exposure mismeasurement, missingness, and measurement error, with sensitivity and partial-identification framing. We reconcile data sources and method outputs across designs and document divergences. Finally, we describe reproducible, governed pipelines that reduce leakage risk and make design-faithful diagnostics and inference auditable.

## 19.1 Motivation and Scope

Credible causal inference depends on data that support a well-defined estimand and make the identification assumptions for that estimand defensible. Platform metrics (impressions, clicks, attributed conversions) are not, by themselves, causal estimands such as ATT on sales or the dose–response function  $\mu(d)$  for viewable exposure. They are often assignment-dependent (defined using the platform’s own delivery and attribution rules), mechanically correlated with treatment, and subject to change when platforms update policies. A defensible design requires that you document how raw data map into  $D_{it}$ ,  $Y_{it}$ , and the target estimand, because that mapping determines what identification assumptions could plausibly hold.

Mismeasured treatments and outcomes do more than inflate standard errors. In a simple linear regression with a continuous regressor measured with classical additive error, OLS slopes attenuate towards zero. This benchmark does not cover binary misclassification, timing mismeasurement, or differential measurement processes, which are common in platform data. Once measurement error correlates with unobserved confounders, or when treatments are misclassified, the bias can go in either direction. Worse, measurement error can violate identification assumptions outright. Mismeasurement can weaken instrument relevance, and it can undermine exclusion if the measurement process itself depends on unobservables that affect outcomes. A mismeasured outcome may break parallel trends if the error process differs across groups [Bound et al., 2001]. In the panel designs developed in Chapters 4–9 and the DML methods in Chapter 12, mismeasured treatment timing or exposure can misclassify cohorts, event times, or propensity scores, so that even formally robust estimators target the wrong potential-outcome contrasts.

This chapter outlines data sources in marketing panels, identity linking and join strategies under privacy constraints, transformations and aggregation choices that preserve identification, alignment of platform metrics with econometric estimands, governance and reproducibility standards, and validation strategies. We use the potential-outcomes framework developed earlier in the book to state what platform data can and cannot identify. For the underlying identification framework and inferential machinery, see Chapters 2 and 16.

## 19.2 Panel Data Sources in Marketing

Marketing causal inference relies on diverse data sources, each with distinct strengths and limitations. The data source fixes how we can define  $D_{it}$  and  $Y_{it}$ , which in turn determines the estimand and the credibility of any identification assumptions. This section surveys the major data types and their implications for causal analysis.

### Retail Scanner Data

Retail scanner data, provided by syndicated data vendors (Nielsen, IRI, Circana) or directly by retailers, record sales, prices, and promotions at a granular level—typically store-week-UPC. These data can be used for price-elasticity estimation, but credible causal interpretation requires an assignment mechanism (e.g., instruments or exogenous price shocks) beyond the scanner panel itself (Section 18.7). They also support promotion analysis and category management studies.

Scanner data offer precise measurement of transactions, consistent unit definitions, and long time series. Store-level panels make DiD and synthetic-control designs feasible, but validity still hinges on design assumptions such as parallel trends or factor stability.

Coverage can be incomplete because not all retailers participate and online sales are often excluded. Panel churn—stores entering and exiting the sample—creates unbalanced panels. Aggregation choices, such as daily to weekly or SKU to brand, trade noise against bias and can change the practical meaning of the estimand.

Price endogeneity is pervasive because retailers set prices based on expected demand. Instrumental variables (Section 18.7) or natural experiments are required for causal elasticity estimates. Classical measurement error in prices attenuates elasticity estimates, but scanner pricing errors can be non-classical (e.g., promotion mis-coding), so bias direction can be ambiguous. Aggregation can also induce composition bias if the mix of products sold shifts with price changes. When treatments are assigned at a higher level than the scanner unit (for example, national campaigns or DMA-level tests), the independent sampling unit is the treatment cluster  $c = 1, \dots, G$ . In that setting, the effective sample size is  $G$  and inference should use cluster-robust variance estimators at the cluster level (Chapter 16).

### Platform Event Logs

Digital platforms generate massive event logs recording impressions, clicks, conversions, and user journeys. These data support digital attribution (Section 18.4), ranking experiments (Section 18.12), and platform experiments (Section 18.11).

Event logs offer granular, real-time data at massive scale. When you can randomise exposure, user-level tracking supports clean causal inference.

What is observed versus what is inferred is critical. An impression may be logged but never seen (viewability). A click may not reflect genuine interest (bot traffic, accidental clicks). A conversion is an attributed outcome, not a direct consequence, and attribution rules can change without notice.

Privacy-enhancing technologies—including Apple’s App Tracking Transparency, cookie deprecation, and differential privacy—reduce our ability to link user activity across devices and time. Cross-device identity resolution becomes probabilistic rather than deterministic, introducing measurement error in treatment assignment and outcomes. Browser vendors are replacing third-party cookies with privacy-preserving alternatives. Google’s Privacy Sandbox, for example, includes the Topics API (interest-based targeting without user-level tracking) and the Attribution Reporting API (aggregate conversion measurement with noise injection). These APIs shift the data available for causal inference from user-level panels to aggregate cohort-level signals with built-in differential privacy noise.

Selection into ad exposure is algorithmic and endogenous. Naive exposed-versus-unexposed comparisons are badly biased (Section 18.4). Randomisation or quasi-experimental designs remain essential. Because ranking algorithms allocate impressions via auctions and relevance scores, one user’s exposure can affect others’ exposures through budget constraints and auction competition. If outcomes depend on these cross-unit allocations, SUTVA fails and you need an explicit exposure mapping (Chapter 11). From an identification perspective, privacy APIs implement a known measurement-error mechanism in which conversions and exposures are reported with censored and noisy counts governed by the API documentation. A known reporting mechanism lets you write down the likelihood of the reported outcome conditional on latent outcomes. Recovering a latent causal estimand still requires additional assumptions. Absent those, it is more honest to treat the latent estimand as partially identified.

## Geo and Mobility Data

Geo data define markets—designated market areas (DMAs), postcodes, census tracts, custom polygons—for geo-experiments (Section 18.3) and spillover analysis. Mobility data from mobile devices or transportation records delineate catchment areas and cross-shopping patterns.

Geographic units are stable over time and enable market-level experiments that shift most interference to within-market interactions rather than between treated and control markets. Mobility data can reveal consumer movement patterns that administrative boundaries miss.

Definitions are often administratively convenient (DMA boundaries) rather than economically motivated (commuting zones, trade areas). Misalignment between treatment units and economic catchments introduces measurement error in exposure and spillovers, which often forces you to use buffer zones or explicit exposure mappings (Chapter 11). The Modifiable Areal Unit Problem (MAUP) compounds this issue because results can change substantially depending on how boundaries are drawn, and there is no theory-free way to choose among alternatives [Openshaw, 1984]. From an identification standpoint, MAUP means that the estimand itself—for example, ATT defined on cluster-level outcomes  $Y_{ct}$ —depends on the chosen partition. Sensitivity checks over plausible boundary definitions are therefore part of the design.

With few geographic units, inference must respect the fact that markets are the independent sampling units. If markets are indexed by  $c = 1, \dots, G$ , then  $G$  is the effective sample size, and inference should use randomisation inference (when available) or cluster-robust variance estimators at the market level (Chapter 16). Spillovers across geographic boundaries must be modelled or excluded. The ecological inference problem poses a deeper threat because relationships observed at the aggregate level need not hold at the individual level [Robinson, 1950]. A positive correlation between regional advertising intensity and regional sales does not imply that exposed individuals bought more. Individual-level experiments or designs that exploit within-region variation are needed to rule out compositional confounding.

## CRM and Transaction Data

Customer relationship management (CRM) systems and transaction databases record individual purchase histories, loyalty programme activity, and customer service interactions. These data support CLV analysis (Section 18.8), loyalty programme evaluation (Section 18.6), and customer-level experiments.

CRM data provide longitudinal customer histories with precise transaction timing, and loyalty programmes can create natural variation in treatment exposure. The same systems also create predictable data problems. CRM data cover only existing customers, missing prospects and churned customers. Selection into loyalty programmes is endogenous, and data quality depends on identity resolution across channels.

Survivorship bias is a major concern because analysing only active customers ignores those who churned. Intention-to-treat analysis or explicit attrition modelling is therefore essential. Inverse Probability of Censoring Weights (IPCW) provide a formal correction. They weight each observation by the inverse probability of remaining in the sample, estimated from a model of attrition on pre-treatment covariates. In the notation of Chapter 16, this amounts to choosing observation weights  $w_{it}$  proportional to the inverse of the estimated survival probability. Under censoring at random given observed history (and adequate overlap for the censoring model), IPCW can identify the causal estimand for the full target population that would be observed under complete follow-up. Without these assumptions, IPCW targets a different estimand, often closer to a survivor effect. Extreme weights can inflate variance and make estimates unstable (Chapter 17). When attrition depends on unobserved factors correlated with outcomes, IPCW is biased and sensitivity analysis is required. Measurement error in customer identity—from duplicate accounts, merged households, or probabilistic matching—attenuates treatment effect estimates and can bias subgroup analyses if matching quality varies with customer characteristics. In potential-outcomes terms, mis-resolved identities reassign outcomes to the wrong units, so that  $Y_{it}(d)$  is matched with the wrong treatment history. This is a non-classical error that can distort both level and heterogeneity estimates.

## Survey and Tracking Data

Brand tracking surveys measure awareness, consideration, and preference at regular intervals. Customer satisfaction surveys (NPS, CSAT) capture attitudinal outcomes. These data support brand equity analysis and long-run advertising effects (Section 18.5).

Surveys measure constructs (brand perception, satisfaction) that transaction data cannot capture, and tracking studies can provide consistent time series for trend analysis. They also suffer from response bias, social desirability, and low response rates. Sample sizes are typically small, limiting statistical power, and linking survey responses to individual transactions is often infeasible.

Survey outcomes define different causal targets (attitudes rather than purchases). Use them as primary outcomes only when the estimand is attitudinal, and treat behavioural claims as requiring behavioural outcomes.

## Data Source Summary

Table 19.1 summarises the major data sources, their typical applications, and key causal considerations.

**Table 19.1** Marketing data sources and causal considerations

Data Source	Typical Unit	Primary Applications	Key Causal Issue
Scanner data	Store-week-UPC	Price elasticity, promotions	Price endogeneity
Platform logs	User-session	Attribution, ranking, A/B tests	Algorithmic selection
Geo/mobility	DMA, postcode	Geo-experiments, spillovers	Few clusters, boundary effects
CRM/transaction	Customer-time	CLV, loyalty, retention	Survivorship bias
Survey/tracking	Respondent-wave	Brand equity, satisfaction	Response bias, low power

## 19.3 Identity, Linking, and Keys

Data integration is foundational to marketing causal inference. We rarely analyse a single data source in isolation. Instead, we join treatment data, outcome data, and covariates from multiple systems. Joins implicitly impose a measurement model (who gets linked, with what error rates, and at what timestamps). That model can redefine the estimand or break identification assumptions.

### Keys and Record Linkage

We rely on primary and foreign keys to link records across tables and systems. A customer ID links transactions to CRM records. A cookie links impressions to conversions. A store code links scanner data to geographic covariates.

**Key stability.** Keys can be unstable. Customer IDs change when accounts are merged or migrated. Cookies expire or are cleared. Device IDs reset. Late-arriving data and platform schema changes require careful versioning to ensure reproducibility.

**Probabilistic matching.** When exact keys are not available, we turn to probabilistic matching—linking records based on name, address, or behavioural similarity. This introduces measurement error: false positives (linking distinct individuals) and false negatives (failing to link the same individual). Bias depends on whether linkage error is (i) independent of treatment assignment and potential outcomes, and (ii) symmetric across treated and control groups. Errors in treatment assignment and errors in outcomes have different consequences.

**Causal implications.** Linkage errors create measurement error in treatment assignment. Under non-differential treatment misclassification (misclassification rates independent of potential outcomes, conditional on covariates) many estimators are biased towards zero in magnitude, but the direction and size depend on the misclassification matrix and the estimand. Differential misclassification can generate bias of either sign. In many settings linkage quality varies systematically with observed covariates (e.g., device type, region, app/web usage). If those covariates also predict outcomes, linkage error becomes differential. Sensitivity analysis should assess how estimates change under plausible linkage error rates, reporting bounds for a range of false positive and false negative rate pairs.

### Cross-Device and Cross-Platform Linking

Modern consumers interact with brands across mobile, desktop, tablet, and in-store environments. Cross-device linking attempts to stitch together this fragmented activity into a unified user journey.

**Deterministic linking.** Relies on login events, hashed emails, or authenticated sessions. Offers high precision (low false positive rate) but limited coverage (only logged-in users).

**Probabilistic linking.** Expands coverage by using behavioural signals—device fingerprints, IP addresses, browsing patterns—to infer identity. Precision is lower, and the linkage is inherently noisy.

**Identity graphs.** Third-party identity providers (LiveRamp, Experian, Oracle) maintain identity graphs that map devices to individuals. For causal analysis, you must validate whether coverage and error rates differ by treatment status, outcomes, and key covariates. Identity graphs that systematically miss or mis-link particular segments (for example, low-value customers or privacy-conscious users) can distort both average and heterogeneous treatment effect estimates (Chapter 15).

**Privacy constraints.** Privacy regimes have degraded cross-device linkage. GDPR and CCPA require explicit consent for tracking, reducing opt-in rates. Apple’s App Tracking Transparency requires per-app consent, and opt-in rates are often below 30%. Third-party cookies are being phased out, eliminating a primary linkage mechanism. Platforms increasingly add noise to user-level data via differential privacy, degrading linkage precision. These regimes often induce partially documented censoring and noise processes. From an identification perspective, they turn full user-level panels into partially observed, noisy cohorts. Even with random assignment, you identify effects for the consented or observable subset and for the platform’s reported outcome definition. Extending conclusions to the full population or to latent outcomes requires additional assumptions.

**Causal implications.** When linkage fails, we lose sight of treated users’ post-treatment activity. If treatment is assigned on one device but outcomes are only observed when cross-device linkage succeeds, ‘being observed’ becomes an outcome-dependent selection variable. Analyses restricted to linked conversions can be biased even under random assignment. Sensitivity analyses should simulate linkage failures under worst-case scenarios.

## Leakage Prevention

Leakage occurs when information from the future contaminates the past, biasing estimates and invalidating inference.

**Temporal leakage.** Temporal leakage occurs when features incorporate information from periods after treatment but are treated as if they were pre-treatment. For example, if a covariate is updated after treatment at time  $t$  but the join uses the updated value in pre-treatment rows, then post-treatment information contaminates the pre-treatment record. This is distinct from a simple mis-specified join and can be subtle when aggregation windows cross treatment boundaries.

**Cross-validation leakage.** Fold structures in cross-validation must respect temporal order. Training on future data to predict past outcomes yields optimistic performance estimates that do not generalise. Block on time to avoid this, as emphasised in Chapter 12 and Chapter 13.

**Feature leakage.** Features that are consequences of treatment (mediators) should not be used as controls. Including post-treatment variables as controls changes the estimand: you no longer target the total effect of  $D_{it}$  on  $Y_{it}$ , but a controlled direct effect under additional assumptions (Chapter 2).

## Handling Missing Identifiers

In practice, identifiers are often missing or incomplete. A fraction of transactions lack customer IDs. Some impressions have no cookie. Some survey respondents cannot be linked to CRM records.

**Missing completely at random (MCAR).** If missingness is unrelated to treatment or outcomes, complete-case analysis is unbiased but loses power.

**Missing at random (MAR).** Under a missing-at-random assumption for identifiers (conditional on observed covariates) and adequate overlap for the missingness model, weighting/imputation can identify the estimand for the target population that would be observed under complete linkage. Without overlap or with misspecified models, the procedure targets a different estimand.

**Missing not at random (MNAR).** If missingness depends on unobserved factors (e.g., privacy-conscious users who block tracking are also less likely to convert), bias is unavoidable without strong assumptions. Partial identification methods provide bounds on causal effects without requiring the missing-at-random assumption [Manski, 2003]. The width of these bounds depends on what we are willing to assume about the relationship between missingness and outcomes. When bounds are wide, they honestly reflect our uncertainty. When they are narrow, they indicate that the data are informative despite missingness. Sensitivity analysis is essential: report point estimates under MAR alongside Manski-style bounds under MNAR to show how conclusions depend on assumptions.

## Identity and Linkage Checklist

Box 19.1: Identity and Linkage Checklist

### Before joining data:

- Are keys stable over the analysis period? Document any migrations or resets.
- What is the linkage rate? Report the fraction of records successfully linked.
- What is the false positive/negative rate for probabilistic matches?

### For cross-device analysis:

- What is the opt-in rate for tracking consent?
- How does the identity graph handle unlinked devices?
- What fraction of treatment-to-outcome paths are observable?

### For leakage prevention:

- Are all covariates measured before treatment assignment?
- Does cross-validation respect temporal order?
- Are any features post-treatment mediators?

### For missing identifiers:

- What is the missing rate for key identifiers?
- Is missingness plausibly MCAR, MAR, or MNAR?
- Have you conducted sensitivity analysis for missing data?

## 19.4 Transformations and Aggregation

Raw data rarely arrive in a form suitable for causal analysis. This section covers the transformations required to align time, aggregate appropriately, handle outliers, and construct exposure variables. Each transformation involves trade-offs that affect the validity and interpretation of causal estimates.

### Time and Calendar Alignment

Time and calendar alignment reduces mechanical confounding that arises when treatment timing and outcome measurement are recorded on incompatible calendars. ISO weeks start on Monday and span seven days, while fiscal weeks may start on different days or vary in length. Misalignment between treatment timing (which follows business calendars) and outcome measurement (which follows ISO weeks) can contaminate pre- and post-treatment periods when weeks straddle treatment dates. When aggregation windows straddle the rollout date, the measured 'pre' outcome can include post-treatment exposure (and vice versa), so the estimand becomes a mixture of untreated and treated potential outcomes rather than  $Y_{it}(0)$  versus  $Y_{it}(1)$ . In event-time or difference-in-differences designs, this misalignment effectively mismeasures treatment timing and outcomes, risking violations of parallel trends and attenuation of estimated effects.

**Seasonality controls.** Week-of-year or month fixed effects absorb regular patterns. Flexible splines or interactions better capture non-linear trends. Event calendars document holidays, product launches, and competitor actions, allowing pre-specified exclusions or robustness checks around confounding events.

**Look-ahead bias.** Arises when aggregating forward (for example, using week  $t + 1$  outcomes in week  $t$  predictors). This must be prevented via strict temporal ordering in ETL pipelines.

### Temporal Aggregation and Data Intervals

The choice of time interval for aggregation involves a trade-off. Tellis and Franses [2006] discuss 'unit exposure time' as a practical modelling choice in advertising response settings. In causal designs, the aggregation interval should instead be chosen to match the treatment assignment and the dynamic response horizon, with explicit sensitivity checks.

Very high-frequency outcomes amplify noise and serial dependence. If you estimate models that assume a single-period treatment effect while the true causal response unfolds over several periods, short-interval outcomes conflate immediate responses with dynamic adjustment, leading to mis-specified distributed lags rather than classical measurement error (see Chapter 10). Overly aggregate data obscure short-run dynamics. Long aggregation windows can obscure dynamic responses. Whether this attenuates or amplifies summary effects depends on the response shape and the estimand (e.g., whether you target  $\theta_k$  or a cumulative effect).

These aggregation effects are relative to the underlying dynamic process. A subtler problem arises when aggregation changes the composition of units: parallel trends may hold at the individual level but fail at the aggregate level if the mix of individuals in treatment and control groups evolves differently over time. For example, store-level sales may satisfy parallel trends, but regional aggregates may not if new stores enter treatment regions at different rates than control regions. Always verify that identification assumptions hold at the level of aggregation used for estimation, not just at finer granularities. The choice depends on the substantive question, the treatment schedule, and the outcome's temporal structure.

Marketing panels often span multiple time scales. Weekly sales data may be appropriate for television advertising campaigns that air on consistent schedules. Daily data may be required for promotions that vary day-to-day. Hourly data may be necessary for digital advertising with real-time bidding and dynamic creative optimisation. The analyst must balance the desire for granularity against the reality that more frequent data amplify measurement error, increase computational burden, and complicate inference under serial dependence.

## Quality Signals from the Crowd

User-generated content (UGC) provides granular signals about perceived product and brand quality at scale. Text and ratings can be mapped to features, but these features inherit (i) timing risks (post-treatment measurement) and (ii) analyst-choice risks (pipeline degrees of freedom). Treat the text pipeline as part of the design, not a neutral preprocessing step. These measures inform market response models, dynamic panels, and financial-market links. Quality is multi-faceted. Golder et al. [2012] propose an integrative framework that clarifies the processes and states underpinning measurable quality constructs. Tirumillai and Tellis [2014] show how to map UGC chatter into actionable metrics for strategic brand analysis. Borah and Tellis [2016] document how social media events can spill over across brands, amplifying or dampening signals and requiring designs that account for interference (see Chapter 11). When UGC indices enter as controls, you must justify that they are pre-treatment and not mediators. Otherwise they can induce post-treatment bias ('bad control'). When they are used as instruments, you must defend an exclusion restriction that rules out direct effects of quality shocks on outcomes beyond the specified channel.

## Measurement Paradoxes in Digital Marketing

The following boxes formalise three paradoxes that arise when platform metrics diverge from causal effects.

**Box 19.2: The Content Paradox**

**Setting.** Brands are urged to “act like media companies” and publish ever more content across social platforms. Let the unit index be  $i$  (brand) for this box. Write  $C_{it}$  for the number of organic posts (or content units) published by brand  $i$  in period  $t$ ,  $A_{it}$  for the resulting effective exposure (for example, viewable impressions across users) after algorithmic filtering, and  $Y_{it}$  for outcomes including visits, sign-ups, and sales.

Organic reach has declined even as  $C_{it}$  has risen: most posts receive few or no impressions, and engagement is highly skewed. Consumers say they value authenticity and relevance, but ranking algorithms may amplify sensational or polarising material. The raw volume of content  $C_{it}$  is therefore a poor proxy for causal impact.

**Paradox and dose–response.** From a causal perspective, content volume or exposure is a continuous treatment. The relevant estimand is a dose–response curve  $\mu(a) = \mathbb{E}[Y_{it}(a)]$  mapping effective exposure  $a$  to expected outcomes. The *content paradox* is that increasing  $C_{it}$  beyond a modest level often leaves  $A_{it}$  and  $Y_{it}$  unchanged: additional posts are unseen, ignored, or shown to users who were already at saturation. Incremental effects  $\mu(a + \Delta a) - \mu(a)$  are frequently near zero.

**Measurement implications.** Estimating  $\mu(a)$  requires designs that separate content decisions from distribution and demand. Randomising posting volume, timing, or format within brands, exploiting exogenous shocks to ranking algorithms, or using geo-experiments on content-heavy vs content-light strategies all create variation in  $A_{it}$  that can be linked to  $Y_{it}$ . These variations can identify causal effects of exposure only if the variation shifts effective exposure  $A_{it}$  in a way that is as-good-as-random (conditional on the design), and if the relevant interference and measurement assumptions hold. Otherwise they remain informative for prediction or descriptive calibration, not identification. Rather than counting posts or headline engagement metrics, brands should report the range of exposure where marginal causal returns are positive and recognise that most incremental content lies beyond that range.

## Box 19.3: The Trust Paradox

**Setting.** Surveys routinely report that only a small minority of consumers say they trust advertising or influencers, yet ad markets and influencer campaigns move billions in sales. Let  $D_{it}$  denote exposure of consumer (or segment)  $i$  to an ad or influencer campaign at time  $t$ ,  $S_{it}$  a stated trust score from surveys or brand trackers, and  $Y_{it}$  a revealed-preference outcome such as clicks, purchases, or CLV. Stated trust and behaviour often diverge. Survey responses reflect norms, self-presentation, and coarse beliefs (“I don’t trust ads”), while behaviour reflects marginal trade-offs in context (prices, availability, attention). A credible measurement strategy must treat  $S_{it}$  and  $Y_{it}$  as distinct outcomes rather than assuming that stated trust is a sufficient statistic for how persuasion works.

**Paradox and latent trust.** One way to formalise the trust paradox is via a latent trust stock  $U_{it}$  that evolves with exposure,

$$U_{it} = \phi U_{i,t-1} + \kappa D_{it} + v_{it},$$

and affects both stated and revealed outcomes:

$$S_{it} = g(U_{it}) + \xi_{it}, \quad Y_{it} = h(U_{it}, D_{it}, X_{it}) + \varepsilon_{it}.$$

Here  $S_{it}$  is a noisy, distorted measurement of  $U_{it}$  (subject to social-desirability and survey biases), while  $Y_{it}$  captures how  $U_{it}$  and current exposure  $D_{it}$  translate into behaviour. The trust paradox arises when experiments or quasi-experiments show large causal effects of  $D_{it}$  on  $Y_{it}$  but small or even negative effects on  $S_{it}$ .

**Measurement implications.** Randomised ad or influencer campaigns that collect both survey trust and behavioural outcomes allow separate estimation of  $D \rightarrow S$  and  $D \rightarrow Y$ . Mediation analysis can be used only under additional assumptions about mediator assignment and mediator–outcome confounding. Without those assumptions, report separate causal effects of  $D$  on survey trust and of  $D$  on behavioural outcomes, and treat mediation as sensitivity analysis rather than a default. In practice, revealed-preference outcomes  $Y_{it}$  anchor ROI and CLV, while stated trust  $S_{it}$  is best treated as a complementary diagnostic. Design-faithful measurement acknowledges that consumers may say one thing and do another, and structures empirical work to learn from both rather than privileging surveys by default.

Box 19.4: The Personalisation Paradox

**Setting.** Platforms encourage ever finer targeting: lookalike audiences, propensity scores, dynamic creative, and per-user bids. Let  $S_{it}$  be a targeting score or predicted conversion probability for user  $i$  at time  $t$ , and let  $D_{it}$  denote personalised treatment intensity (for example, number of tailored ads or discount depth). Outcomes  $Y_{it}$  include conversions, revenue, or CLV.

Consumers report that they value relevance but dislike feeling “tracked”. At low levels of  $D_{it}$ , personalisation can be helpful. At high levels or with sensitive features, it can feel creepy, trigger avoidance, or invite regulatory risk. The personalisation paradox is that the very data and models that enable high relevance also increase the risk that marginal treatment becomes intrusive or legally constrained.

**Dose-response and constraints.** From Chapter 14, we can treat  $D_{it}$  as a continuous dose and estimate  $\mu(d) = \mathbb{E}[Y_{it}(d)]$  over the support of observed intensities, subject to privacy and fairness constraints. Empirically,  $\mu(d)$  may be increasing and concave over low doses, then flatten or decline when users are over-targeted or when discounts train deal-seeking.

**Measurement implications.** Randomised experiments on targeting intensity (for example, high vs medium personalisation) and policy changes to data access (for example, removal of third-party cookies) provide variation to recover  $\mu(d)$  and to quantify how privacy constraints shift optimal intensity. Under the designs in Chapter 14, these variations identify the dose-response function  $\mu(d)$  rather than relying on observational correlations. Rather than maximising predicted response at the individual level, firms should report ranges of  $d$  with positive marginal causal returns that also satisfy regulatory and reputational constraints, recognising that the apparent gains from ever-finer personalisation in observational data may not survive causal and policy-aware analysis.

## UGC Quality Signal Validity

Design-faithful measurement aligns UGC-derived quality indices with econometric estimands. Sampling and selection affect representativeness as vocal users differ from the broader customer base. Platform policy changes (filtering, de-duplication, bot mitigation) induce breaks. Manipulation and gaming risk bias.

**Validation requirements.** Validity requires external benchmarks (audits, satisfaction indices), pre-specified dictionaries or models, and sensitivity to alternative text pipelines. Pre-registration of text analysis choices is essential to avoid researcher degrees of freedom: specify the sentiment lexicon, topic model parameters, and embedding method before seeing outcomes. When quality signals form part of identification (as instruments or controls), document exclusion restrictions and run placebos. Treat text-derived features as pre-treatment covariates only if they are measured before treatment assignment. Using post-treatment text (such as reviews written after a campaign) as a control variable introduces post-treatment bias. In potential-outcomes terms, using reviews written after a campaign as covariates conditions on a mediator and blocks part of the causal path from  $D_{it}$  to  $Y_{it}$ . Event-study designs can diagnose market reactions to discrete quality

shocks. Vector autoregressions (VARs) can trace dynamic feedback from quality to sales and stock returns, but without credible shocks or instruments they remain descriptive, not identifying causal effects.

## Robustness Transformations

Robustness checks often require transforming raw signals to handle outliers, zero-inflation, and sparse cells.

**Winsorisation.** Replaces extreme values with percentile thresholds (such as the 1st and 99th percentiles) to dampen sensitivity to outliers, trading a small amount of bias for significant gains in stability. Winsorisation changes the target quantity to a trimmed/winsorised estimand. Report thresholds and interpret results as applying to that transformed outcome distribution.

**Two-part models.** The zero-inflation common in conversion data—where most users do nothing—often motivates two-part models that separate the likelihood of an event from its magnitude. Two-part models impose separate modelling and identification conditions for the extensive and intensive margins. State which causal estimand each part targets and whether the same design assumptions plausibly hold for both margins.

**Minimum cell thresholds.** To protect privacy and ensure stability, apply minimum cell thresholds, pooling or excluding sparse segments such as DMAs with few treated stores. Minimum cell thresholds exclude or pool sparse units, which changes the effective population and can remove influential treated or control clusters. Treat these thresholds as part of the design: document them and report how estimates change under alternative cut-offs.

## Exposure Construction for Continuous Treatments

Exposure construction maps raw impressions, reach, and frequency to dose variables aligned with estimands in Chapter 14.

**Key metrics.** Reach counts unique users exposed at least once during a campaign window. Frequency averages impressions per exposed user and captures intensity of exposure conditional on any exposure. Viewability adjusts raw impression counts for whether ads were actually displayed in the viewable area of the screen for a minimum duration, addressing the gap between served and seen. Frequency caps, imposed by platforms or advertisers, truncate dose at pre-specified thresholds, creating bunching in the exposure distribution that must be accounted for in dose-response estimation.

**Dose variable choices.** Mapping these to dose requires choices: use raw impressions, capped impressions, or viewability-adjusted impressions? Each definition corresponds to a different causal estimand: the effect of served impressions, of viewable impressions, or of capped exposures. Be explicit about which you target. Each

choice affects overlap, support, and the interpretation of marginal effects. Document the choice and report sensitivity to alternatives.

**Attribution model transformations.** When users are exposed to multiple touchpoints before conversion, raw exposure data must be transformed into attributed credit. Last-touch attribution assigns all credit to the final touchpoint. First-touch assigns all credit to the initial exposure. Neither reflects causal contribution. More sophisticated approaches include: (i) Shapley value attribution, which allocates credit based on each touchpoint's marginal contribution across all possible orderings (computationally expensive but axiomatically justified), (ii) Markov chain attribution, which models the conversion path as a state transition process and allocates credit based on removal effects (how much would conversion probability drop if this touchpoint were removed?), and (iii) data-driven attribution from machine learning models that predict conversion probability from path features. Each transformation embeds assumptions about how touchpoints interact. Shapley assumes separable contributions. Markov assumes memoryless transitions. ML models assume the training data distribution reflects causal structure. These attribution schemes are typically fitted to observational paths and are not identification strategies by themselves. For causal claims, treat them as pre-processing choices that define candidate treatment variables, then apply the experimental or quasi-experimental designs from earlier chapters to those variables. If findings are not robust across attribution models, you are learning more about attribution assumptions than about treatment effects.

## 19.5 Platform Metrics vs Econometric Estimands

Platform-reported metrics and econometric estimands often measure different quantities. This section clarifies the distinction, identifies common pitfalls, and provides strategies for reconciliation.

### The Platform Lift Problem

Platform-reported lift metrics often do not correspond to causal estimands. These often report descriptive contrasts between exposed and unexposed users. Without a design that makes exposure as-good-as-random, the contrast does not identify ATE or ATT.

**Why platform lift is biased.** Platform algorithms target ads to users with high predicted conversion probability. Users who see ads differ systematically from those who do not—they were selected precisely because they were likely to convert anyway. Comparing conversion rates between exposed and unexposed groups captures both the causal effect and the selection effect.

**Example.** A platform reports 15% lift: exposed users converted at 3.0% vs. 2.6% for unexposed. Suppose the targeted group would have converted at 3.0% even without ads, while the untargeted group would have remained at 2.6%. Then the observed 0.4 percentage point gap reflects pure selection, and the causal effect is zero. The platform metric is not wrong—it accurately describes the data—but it does not answer the counterfactual question: what would have happened without the ad? Geo-experiments (Section 18.3) provide the randomised evidence needed to ground such calibration.

**Econometric estimands.** Our estimands—ATE and ATT—are defined on potential outcomes and counterfactual contrasts [Pearl, 2009]. They require identification strategies that sever the link between treatment assignment and potential outcomes (Section 18.4).

Delivery metrics report what the platform did, not what the user experienced. Four concepts matter in practice. Coverage describes the share of the target audience that receives at least one impression over the analysis window. Pacing describes how spend unfolds over time, for example whether it is front-loaded at launch, roughly even across the period, or back-loaded towards the end. Auction diagnostics summarise how often the advertiser wins auctions, the bid landscape it faces, and the intensity of competition. Frequency captures how many impressions the average reached user sees over the window.

Measurement gaps open up between these delivery metrics and the actual exposure consumers receive.

Viewability is the first gap: served impressions and viewable impressions define different treatments  $D_{it}$  and therefore different causal estimands. A platform logs an impression when it serves an ad, not when the ad appears on screen for a viewable duration, so the share of served impressions that are truly viewable can vary substantially by placement.

A second gap comes from bot traffic: non-human activity generates fraudulent impressions and clicks, inflating counts. Industry studies report non-trivial levels of fraud in digital advertising [Gordon et al., 2019].

A third gap arises from post-hoc filters: fraud and quality filters applied after delivery can remove impressions retroactively, creating discrepancies between real-time logs and reconciled datasets.

From the standpoint of our panel notation, these gaps mean that treatment  $D_{it}$ , outcomes  $Y_{it}$ , and sometimes the observation process itself (which  $i, t$  cells are observed) are measured with error, often in ways that vary over time and across placements. This is non-classical measurement error that can both attenuate and distort estimated treatment effects unless designs and estimators explicitly adjust for it.

**The prediction–identification tension.** Delivery optimisers allocate exposure using predicted conversion probability. This induces selection on (often unobserved) predictors of outcomes, so exposed users differ systematically from unexposed users. The same model that improves targeting efficiency creates selection bias for causal estimation. Orthogonalised machine learning (Chapter 12) can reduce sensitivity to nuisance-function estimation error, but it does not fix selection on unobservables. It still requires a credible identification condition (e.g., unconfoundedness with overlap, or a valid instrument), and careful cross-fitting to avoid overfitting-induced bias.

## Design-Faithful Measurement

Design-faithful measurement means that metric definitions align with identification assumptions. When definitions change mid-study, causal estimates are compromised.

**Method-specific requirements.** Difference-in-differences (Chapter 4) and event studies (Chapter 5) require stable outcome definitions across pre and post periods. If outcome definitions change, the parallel-trends assumption applies to different variables, so the contrast no longer identifies the same potential-outcome difference. Synthetic control (Chapter 6) and SDID (Chapter 7) require donor outcomes measured identically to treated outcomes. Any change in measurement between treated and donor units violates the assumption that untreated potential outcomes share a common factor structure. Factor models (Chapter 8) require missing data patterns that are stable or explicitly modelled. Time-varying missingness patterns can break low-rank structure and identification of factors.

**Common violations.** Attribution windows can change mid-campaign, for example from 28-day to 7-day click attribution. Viewability standards can shift, for example when the MRC definition changes (from 50% pixels for 1 second to 100% pixels for 2 seconds). Cookie policy updates, including iOS App Tracking Transparency and browser cookie deprecation, reduce trackable conversions. Platforms may also begin imputing conversions for untracked users, which changes the outcome definition.

These violations break comparability of  $Y_{it}$  across time and groups. At minimum you should restrict analysis to windows with stable definitions and report sensitivity to alternative windows. In many cases you must treat estimates outside stable windows as descriptive checks rather than as causal estimates of a stable estimand. In each case, definition drift changes the estimand itself: your pre-period outcome no longer represents the same underlying  $Y_{it}(0)$  as your post-period outcome.

## Reconciliation Strategies

When platform metrics diverge from external data, reconciliation is required.

**External validation.** Use independent outcome data—retail scanner sales, CRM transactions, financial results—to validate platform-reported conversions. Discrepancies reveal measurement error or attribution inflation.

**Incrementality calibration.** Run periodic geo-experiments or randomised holdouts to estimate true incrementality. Use these estimates to calibrate platform lift metrics (Section 18.3, Section 18.5). Remember that geo-experiment estimates themselves carry uncertainty, both from sampling variation and from the small number of geographic units typically available. This uncertainty should propagate into calibrated platform metrics: if the geo-experiment confidence interval spans (0.5, 1.5) for the calibration multiplier, report calibrated lift ranges rather than point estimates. Calibration also assumes that the relationship between platform-reported lift and true incrementality is stable across time, audiences, and creatives. If that link drifts, a single geo-experiment multiplier cannot repair future platform metrics, and you must update calibration regularly.

In most geo-experiments, the independent sampling unit is the geographic cluster. The effective sample size is therefore the number of clusters  $G$ , not the number of users. Inference should use randomisation inference (when available) or cluster-robust variance estimators at the cluster level.

**Multi-source triangulation.** Compare estimates from multiple methods: platform lift, MMM, geo-experiments, and attribution models. Convergence across methods increases confidence that you are learning about a real effect rather than artefacts of a single design. Divergence signals measurement or identification problems and should trigger further diagnostics rather than selective reporting.

**Audit trails.** Document all metric definitions, platform versions, and policy changes. Maintain a changelog that maps dates to definition changes for sensitivity analysis. These changelogs should be under version control and tied to the analysis repository to reinforce reproducibility.

## Platform Metric Validation Checklist

Box 19.5: Platform Metric Validation Checklist

**Before using platform metrics:**

- What is the platform's definition of "conversion"? Does it match your business outcome? If not, your estimand is "effect on platform-defined conversions", not "effect on sales".
- What attribution window is used? Has it changed during the analysis period?
- What fraction of conversions are modelled (imputed) vs. observed?
- What is the viewability rate? Are non-viewable impressions included?

**For lift metrics:**

- How was the control group constructed? Is it a true holdout or algorithmic?
- What is the selection mechanism into exposure? Is it endogenous? If exposure is algorithmic, treat reported lift as descriptive and rely on experimental or quasi-experimental designs for causal claims (see Section 18.4).
- Has the lift been validated against external incrementality tests?

**For reconciliation:**

- Do platform conversions match CRM or sales data? What is the discrepancy rate?
- Have you run geo-experiments to calibrate platform estimates? Remember to carry the geo-experiment uncertainty into any calibrated figures.
- Are metric definitions stable across the pre and post periods?

## 19.6 Privacy, Policy, and Governance

Privacy regulations, platform policies, and data governance practices shape what data are available for causal analysis. These constraints define which estimands are feasible and which identification strategies remain credible.

### Privacy Regulations and Consent

Privacy regulations fundamentally constrain our ability to collect and link data.

**Key regulations.** GDPR (EU) requires explicit consent for personal data processing and grants data subjects rights to access, rectification, and erasure. CCPA and CPRA (California) require disclosure of data collection practices and grant opt-out rights for data sales. Apple’s App Tracking Transparency requires per-app consent for cross-app tracking, and opt-in rates are typically 20–30%. Third-party cookies are being phased out in major browsers, eliminating a primary tracking mechanism.

**Causal implications of consent.** Consent requirements shrink samples. They threaten external validity when consenters differ from non-consenters, and they can threaten identification when consent is affected by treatment and the analysis conditions on consent. Users who consent differ systematically from those who do not—they may be more engaged, more trusting, or less privacy-conscious. If consent correlates with potential outcomes, estimates from consented samples do not generalise to the full population. Selection bias grows with both the gap in consent rates between treatment and control and the outcome gap between consenters and non-consenters. When both are non-zero, estimates from the consented sample will generally not identify effects for the full population. When treatment itself affects consent—for example, if an experiment changes prompts or messaging about tracking—consent becomes a post-treatment collider. Conditioning on consent can then induce spurious associations even under random assignment. Avoid this restriction where possible, or state clearly that the estimand becomes an effect among consenters under additional assumptions.

**Sample size impact.** With opt-in rates around 25%—common under App Tracking Transparency—effective sample size drops by roughly 75%. Power calculations must account for expected consent rates. For rare outcomes, user-level experiments may become infeasible, requiring a shift to geo-level designs (Section 18.3).

### Privacy-Preserving Techniques

When individual-level data are unavailable, privacy-preserving techniques enable aggregate analysis.

**Differential privacy.** Adds calibrated noise to query results, providing formal privacy guarantees. Platforms (Google, Apple, Meta) increasingly report differentially private aggregates. The privacy-utility trade-off means that stronger privacy (smaller  $\epsilon$ ) yields noisier estimates. When the noise mechanism is fully documented and independent across released statistics, you can sometimes adjust variance estimates. In practice, DP systems may include clipping, thresholding, and correlated noise, so you often need conservative inference or partial-identification framing. For small cells or very strong privacy parameters, the noise can dominate the signal so that only bounds or very wide intervals are defensible [Manski, 2003].

**Secure aggregation.** Computes aggregates across users without exposing individual records. Useful for federated learning and cross-platform measurement.

**Data clean rooms.** Secure environments where multiple parties can join data without exposing raw records. Enable advertiser-publisher matching while preserving privacy. Examples include Google Ads Data Hub, Amazon Marketing Cloud, and LiveRamp Safe Haven. Clean rooms impose query restrictions that can prevent certain causal designs and force the unit of analysis towards cohorts or geographies, which changes the estimand and the appropriate inference unit. Minimum aggregation thresholds (e.g., cells must contain at least 50 users) preclude user-level difference-in-differences, and pre-specified query templates limit ad hoc sensitivity analyses. Differential privacy noise injected within clean rooms can invalidate inference for small treatment cells, where the noise magnitude exceeds the signal. These constraints redefine the estimand to effects on high-volume cohorts. Because volume correlates with user behaviour and platform targeting, treat the resulting population restriction as substantively meaningful, not innocuous. They often make user-level estimands unidentifiable, so designs must be formulated at the cohort or geo level from the outset.

**Aggregated conversion APIs.** Platform-provided APIs (Meta Aggregated Event Measurement, Google Privacy Sandbox) report conversions at aggregate levels with noise injection. These replace user-level conversion tracking but introduce measurement error. You are estimating effects on the API-defined aggregates, not on latent user-level outcomes, so the estimand is implicitly redefined by the API's aggregation and noise rules.

**Causal implications.** Aggregation reduces statistical power and may introduce bias if aggregation boundaries (e.g., cohorts, time windows) do not align with treatment variation. Sensitivity analysis should assess how privacy-induced noise affects confidence intervals.

## Platform Policy Changes

Platform policy changes create regime breaks that threaten identification.

### Types of policy changes:

- **Auction mechanism updates:** Changes to bidding algorithms, reserve prices, or quality scores alter who sees ads conditional on observables and unobservables, introducing algorithmic confounding (Chapter 15).
- **Ranking algorithm revisions:** Updates to search or social feed algorithms change organic versus paid exposure correlation.
- **Conversion definition shifts:** Platforms redefine attribution windows (7-day to 28-day), add post-view conversions, or change deduplication rules.
- **Targeting restrictions:** Removal of sensitive targeting categories (e.g., political, health) changes the feasible treatment space.

From an identification standpoint, regime breaks change the assignment mechanism and can act as time-varying confounders if they coincide with treatment adoption. A DiD or event-study design that treats post-break periods as “post-treatment” for some campaign can inadvertently pick up effects of the policy change instead, unless break dates are explicitly modelled or excluded.

**Monitoring for breaks.** Track changelogs, release notes, and A/B test schedules to flag break dates. Subscribe to platform developer blogs and API update notifications. Maintain an internal calendar of known policy changes.

#### Guarding designs against regime breaks:

- Restrict analysis windows to stable periods before and after breaks. This may sacrifice power and should be documented as a design choice.
- Model breaks explicitly via time-varying parameters or structural break tests.
- Report before-and-after estimates separately with transparency about comparability.
- Use falsification checks (placebos) around break dates to assess impact on control outcomes. These checks should focus on untreated or baseline outcomes. Significant jumps at break dates indicate that policy changes, not treatments, are driving apparent effects.

## Data Governance for Reproducibility

Data governance ensures that causal analyses are reproducible and auditable.

**Dataset versioning.** Snapshot dates and schema hashes prevent silent updates from changing results. Tag datasets with version identifiers and link these identifiers in your analysis repository (for example, Git tags) so that code and data versions move together. Never overwrite historical extracts.

**Schema migration logs.** Document field additions, deletions, or redefinitions. A field renamed from “clicks” to “valid\_clicks” with a new definition invalidates historical comparisons.

**Reproducible extracts.** Use version-controlled queries (SQL in Git) and deterministic sampling (fixed random seeds). Document all filters, joins, and transformations. Treat SQL and ETL code as part of the analysis: keep them in the same version-control system as your modelling code so that any change to data extraction is reviewable and reproducible.

**Access controls and audit trails.** Track who accessed data when, supporting compliance and preventing leakage. Role-based access limits exposure to sensitive fields.

**Retention policies.** Balance replication (keep data long enough for independent verification) with privacy (delete data after study completion). Document retention periods in study protocols. Where regulations require deletion of identifiers, consider retaining de-identified aggregates that are sufficient for planned designs, and document which future re-analyses will no longer be possible once raw data are purged.

## Governance Checklist

Box 19.6: Privacy and Governance Checklist

### Privacy compliance:

- What consent is required for data collection? What is the expected opt-in rate?
- Does the analysis require individual-level data, or can aggregates suffice? If you design at the geo or cohort level from the start, you avoid promising user-level estimands that clean rooms and privacy constraints cannot support.
- Are privacy-preserving techniques (differential privacy, clean rooms) available?
- Have you assessed selection bias from consent-based sampling?

### Policy monitoring:

- Have you documented all platform policy changes during the analysis period?
- Are there known regime breaks that require window restrictions?
- Have you run falsification checks (placebos) around break dates? These should focus on control or baseline outcomes.

### Data governance:

- Are datasets versioned with snapshot dates and schema hashes?
- Are extraction queries version-controlled and reproducible, in the same repository as the analysis code?
- Are access controls and audit trails in place?
- Is the retention policy documented and compliant with regulations?

## 19.7 Missing Data and Measurement Error

Missing data and measurement error are pervasive in marketing panels. Both threaten the validity of causal estimates, but through different mechanisms. Missing data reduce sample size and can redefine the target population. When missingness depends on potential outcomes, it can also break identification. Measurement error can attenuate estimated effects, increase variance, or introduce spurious correlations, depending on where it enters the model and how it relates to other variables. This section provides a framework for diagnosing and addressing both problems.

### Missing Data Mechanisms

The appropriate remedy for missing data depends on the mechanism generating missingness.

**Missing completely at random (MCAR).** Missingness is unrelated to observed or unobserved variables. Complete-case analysis is unbiased but loses precision. Example: random server failures that drop some impressions.

**Missing at random (MAR).** Missingness depends on observed covariates but not on the missing values themselves, conditional on observables. Multiple imputation or inverse-probability weighting can mitigate bias provided the model for missingness is correctly specified. Example: users with older devices are less likely to have tracking enabled, but device type is observed.

**Missing not at random (MNAR).** Missingness depends on unobserved factors or on the missing values themselves. Bias is unavoidable without strong assumptions such as exclusion restrictions or instruments. Example: privacy-conscious users who block tracking are also less likely to convert—both the outcome and the missingness are driven by an unobserved privacy preference.

**Diagnosing the mechanism.** Compare observed characteristics of complete vs incomplete cases. Large differences make MCAR implausible. You can use diagnostics to check whether missingness correlates with observed outcomes after conditioning on covariates (when outcomes are observed for a subset). These diagnostics cannot distinguish MAR from MNAR: data alone cannot identify the missingness mechanism. Treat such diagnostics as ways to rule out MCAR and to stress-test MAR assumptions, not as proof of mechanism.

### Panel-Specific Missing Data Patterns

Panels exhibit structured missingness patterns that require tailored approaches.

**Attrition.** Units drop out of the panel over time. If attrition correlates with treatment or outcomes, estimates are biased. ITT targets the effect of assignment, not compliance. If outcomes are missing due to attrition, ITT still requires assumptions or design features (e.g., follow-up) to recover the ITT estimand. Per-protocol analysis conditions on completion and may be biased. Per-protocol analyses implicitly target effects among survivors rather than the full randomised population. They are only causal under strong assumptions about attrition, typically stronger than those required for ITT.

**Staggered entry.** Units enter the panel at different times, creating unbalanced panels. Early entrants have longer pre-treatment histories. Late entrants may differ systematically.

**Intermittent missingness.** Units appear, disappear, and reappear. Common in platform data where users are active sporadically.

**Structured missingness for causal inference.** In DiD and synthetic control, the counterfactual outcomes are unobserved by definition. Do not conflate this with data-missingness from attrition or logging gaps. Matrix completion and factor models (Chapter 8) impute missing entries via low-rank approximations, exploiting the panel structure.

**Design-consistent imputation.** Most designs in this book estimate effects without explicitly filling in missing potential outcomes. When you do choose to impute for reporting or secondary analyses, the imputation scheme must mirror the identification strategy: For DiD, this means imputing treated outcomes by extending pre-treatment trends under the maintained parallel-trends assumption. For synthetic control, it means imputing using weighted donor outcomes. For factor models, it means imputing using estimated latent factors. These imputations are for interpretation and diagnostics, not a substitute for a valid design. Imputing with post-treatment information or treatment-correlated covariates violates identification assumptions.

## Measurement Error Types

Different types of measurement error have different consequences. Classical error in outcomes inflates variance without bias under standard linear models. Classical error in treatments or regressors attenuates estimated effects toward zero. Non-classical or differential error can bias estimates in either direction.

**Classical measurement error.** Error is independent of true values and other variables. In a simple linear regression with a single continuous regressor  $X^*$  measured with classical additive error  $X = X^* + u$ , where  $u$  is mean-zero, independent of  $X^*$  and the structural error, the probability limit of the OLS slope equals  $\beta \cdot \frac{\sigma_{X^*}^2}{\sigma_{X^*}^2 + \sigma_u^2}$ . This attenuation factor, sometimes called the reliability ratio, is always less than one, so estimates are biased toward zero. Note that this result applies to continuous regressors under specific assumptions and does not describe binary misclassification in treatment indicators  $D_{it}$ , which behaves differently.

**Non-classical measurement error.** Error is correlated with true values, treatment, or outcomes. Bias direction depends on the correlation structure. Example: viewability is lower for ads shown to less engaged users, who also have lower conversion rates—viewability error is correlated with outcomes.

**Differential measurement error.** Error magnitudes differ between treated and control groups. Example: attribution windows capture more conversions for treated users if treatment increases purchase timing, creating spurious lift.

## Measurement Error by Data Type

Different marketing data sources exhibit characteristic measurement errors.

**Impression data.** Logged impressions may be served but not viewed. A substantial fraction of logged impressions are non-viewable in many placements [Gordon et al., 2019]. Non-human activity also inflates counts. Industry studies report non-trivial shares of non-human or fraudulent traffic in some channels [Gordon et al., 2019]. In addition, invalid clicks and impressions vary by channel and can generate spurious exposure.

**Conversion data.** Attribution windows can misalign with true causal lags. Last-click attribution ignores earlier exposures in the conversion path. Cross-device gaps also matter because conversions on unlinked devices are missed.

**Sales data.** Scanner coverage is incomplete because not all retailers report. Stockouts produce zero recorded sales when the product is unavailable, which is not the same as zero demand. Data are also often aggregated to protect retailer identity, which can change the unit of analysis.

## Remedies for Measurement Error

**Validation studies.** Use audits, surveys, or external benchmarks to estimate error magnitudes. Compare platform-reported conversions to CRM data. Compare impression counts to viewability audits.

**Instrumental variables.** If an instrument is available that is correlated with the true treatment but not with the measurement error, IV estimation recovers consistent estimates (see the instrumental-variables chapter). The instrument effectively proxies for the true treatment, purging the measurement error that contaminates OLS. This requires stronger conditions than standard IV for endogeneity: the instrument must be independent of both the structural error and the measurement error in the observed regressor. If the instrument itself is mismeasured, or if measurement error depends on the instrument, IV will generally not fully correct the bias.

**SIMEX.** When measurement error variance is known or can be estimated from validation data, Simulation Extrapolation (SIMEX) provides a practical correction [Carroll et al., 2006]. The method adds increasing amounts of artificial measurement error, estimates the relationship between error variance and bias, then extrapolates back to zero error. SIMEX is particularly useful when IV is unavailable or when multiple variables are measured with error.

**Partial identification.** When error magnitudes are unknown, accept that the estimand is set-valued and report bounds rather than point estimates. Manski bounds provide worst-case intervals under minimal assumptions about the error process [Manski, 2003]. If prior information constrains the error distribution (e.g., known sign, bounded magnitude), tighter bounds are achievable. The width of the bounds reflects genuine uncertainty about the estimand. Narrow bounds indicate that the data are informative despite measurement error, while wide bounds signal that conclusions are sensitive to assumptions.

**Negative controls.** Negative control outcomes and negative control exposures provide falsification checks. Interpret failures cautiously because a detected 'effect' on a negative control can reflect misspecification, multiple testing, or measurement issues rather than a single identifiable bias source. A negative control outcome is a variable that should be unaffected by treatment if the causal model is correct—for example, sales of an unrelated product category, or conversions that occur before the ad could plausibly have an effect. Finding a treatment effect on a negative control outcome signals bias from confounding or measurement error. A negative control exposure is a placebo treatment that should have no effect on outcomes—for example, ads shown but not viewed (due to viewability failure), or campaigns in markets where the product is unavailable. Finding an effect of negative control exposure indicates confounding or spurious correlation. Use negative controls as model diagnostics: passing negative-control checks can increase confidence at the margin, but it does not validate identification. Low power can also yield false reassurance. If it fails, investigate the source of bias before interpreting the main results. Negative controls rarely fix bias on their own. They are diagnostics that should be integrated with the sensitivity analyses in Chapters 15 and 17.

**Sensitivity analysis.** Specify the range of error magnitudes consistent with substantive conclusions (Chapter 16, Chapter 17). Report how large measurement error would need to be to overturn the finding.

## Missing Data and Measurement Error Checklist

Box 19.7: Missing Data and Measurement Error Checklist

### For missing data:

- What is the missing rate for key variables? Is it differential by treatment status?
- Is missingness plausibly MCAR, MAR, or MNAR? What evidence supports this? Remember that data can rarely distinguish MAR from MNAR. Justify your choice conceptually, not just statistically.
- Have you compared characteristics of complete vs. incomplete cases?
- Does your imputation method respect the identification strategy?
- Have you reported ITT estimates alongside per-protocol estimates?

### For measurement error:

- What are the known sources of measurement error in your data?
- Is error likely classical, non-classical, or differential?
- Have you validated platform metrics against external data?
- Have you conducted sensitivity analysis for plausible error magnitudes?
- If error is substantial, have you reported bounds rather than point estimates, or at least complemented point estimates with Manski-style bounds?

## 19.8 Validation and Reconciliation

Credible causal inference requires validation at two levels: data source validation (do different data sources agree on basic facts?) and method reconciliation (do different estimation methods yield consistent causal estimates?). Discrepancies at either level can indicate measurement problems or estimand mismatches. Treat them as signals to diagnose, not as noise to ignore.

### Data Source Reconciliation

Platform-reported metrics rarely match external data sources perfectly. Reconciliation quantifies and explains these discrepancies.

**Common discrepancy sources.** Discrepancies arise from coverage differences (platform data may include or exclude channels, devices, or segments that external data capture differently), attribution differences (for example, platforms use 7-day click attribution while CRM uses first-touch and finance uses cash basis), timing differences (platforms report in UTC while retailers report in local time and finance uses fiscal periods), definition differences (“conversion” may mean leads, sales, or revenue), and deduplication rules (platforms may count unique users differently than CRM systems).

#### Reconciliation workflow:

1. **Identify overlap period:** Find a time window where all data sources are available and definitions are stable.
2. **Align definitions:** Map each source’s metrics to a common definition (e.g., “completed purchase within 7 days of click”).
3. **Compute discrepancy:** Calculate the ratio or difference between sources for the aligned metric.
4. **Check by treatment status:** Compute discrepancies separately by treatment and control groups (or exposed vs unexposed) because differential discrepancies by treatment status can bias contrasts even when overall reconciliation looks acceptable.
5. **Examine time variation:** Examine discrepancies over time. Stable discrepancies may be less damaging for DiD if they are time-invariant and do not change differentially with treatment. Time-varying or treatment-interacting discrepancies threaten identification.
6. **Diagnose root cause:** Decompose discrepancy into coverage, timing, and definition components.
7. **Document and adjust:** Record the discrepancy magnitude and apply calibration factors if justified.

**Acceptable discrepancy thresholds.** There is no universal threshold. If you want a rule of thumb, scale discrepancy tolerances to the smallest effect size that would change a decision. A discrepancy larger than the expected causal effect is, by definition, design-threatening. If the expected treatment effect is 2%, a 5%

discrepancy in baseline metrics is catastrophic. If the expected effect is 50%, the same 5% discrepancy is noise. Scale thresholds relative to the magnitude of effects you seek to detect.

**Persistent gaps.** If discrepancies persist after alignment, they may signal measurement drift (gradual change in definitions) or structural breaks (sudden policy changes). Monitor discrepancies over time. Sudden jumps indicate breaks that require window restrictions or explicit modelling. If drift or breaks differ between treated and control units or coincide with treatment adoption, they can mimic treatment effects in DiD and related designs. Treat discrepancy series as an additional diagnostic alongside pre-trend diagnostics.

## Method Reconciliation

Different causal methods impose different identification assumptions. Comparing estimates across methods is informative only after you harmonise the estimand (outcome definition, population, exposure) as far as possible.

**Methods to compare.** Common comparisons include difference-in-differences (Chapter 4), synthetic control (Chapter 6), SDID (Chapter 7), factor models (Chapter 8), double machine learning (Chapter 12), and continuous-dose estimators (Chapter 14). The comparison is only meaningful after you align the estimand and then state the identifying assumptions each method relies on.

**Interpreting agreement and divergence.** When multiple methods with different assumptions yield similar estimates, confidence in the causal effect increases through triangulation. Triangulation is most informative when methods rest on genuinely different designs or identifying restrictions. Using several variations of the same DiD specification with shared preprocessing rarely delivers independent evidence. When methods disagree, treat that disagreement as information and diagnose the source. Design diagnostics (Chapter 17) can indicate whether overlap, pre-trends, weights, or functional form drive differences. Triangulation is most compelling when methods have \*different\* biases that would push estimates in opposite directions. If DiD is biased upward by anticipation effects but SC is biased downward by poor pre-treatment fit, agreement between them is stronger evidence than agreement between two methods with similar bias structures. Specification curve analysis (add citation) provides a formal framework for this: enumerate all defensible specifications, estimate effects under each, and report the distribution of estimates rather than a single point.

### Root cause analysis for divergent estimates:

1. **Check sample differences:** Do methods use the same sample? Exclusions for overlap or balance may differ.
2. **Check pre-trend diagnostics:** Do event-study lead coefficients suggest differential pre-trends for DiD while synthetic control achieves strong pre-treatment fit?
3. **Check weight distributions:** Are SC or IPW weights concentrated on a few units?

4. **Check functional form:** Does the dose-response curve assumption matter?
5. **Check time horizon:** Are methods estimating effects at different post-treatment windows?

**Reporting standards.** Report all estimates with sample sizes and exclusion criteria, identification assumptions stated explicitly, diagnostic results (pre-trends, balance, fit), and confidence intervals with appropriate inference. This allows readers to assess robustness and prefer estimates supported by multiple methods.

## Triangulation Strategy

Triangulation combines evidence from multiple data sources and methods to strengthen conclusions.

**Data triangulation.** Compare platform conversions to CRM sales to financial revenue. If all three show similar lift, confidence increases.

**Method triangulation.** Compare DiD, SC, and MMM (Section 18.5) estimates. If all three point to similar advertising effects, and diagnostics show that they fail for different reasons when stressed, the finding becomes more credible. Shared data problems (for example, mismeasured outcomes) can still invalidate all three.

**Design triangulation.** Compare geo-experiments to user-level A/B tests to natural experiments. Different designs have different threats. Agreement across designs is strong evidence.

**When triangulation fails.** If sources or methods disagree, do not average or cherry-pick. Instead: report all estimates transparently, diagnose the source of disagreement, acknowledge uncertainty in conclusions, and recommend follow-up studies to resolve discrepancies. Avoid informal “vote-counting” rules (for example, declaring an effect positive because most methods are). Focus on diagnosing why methods differ instead.

**Formal estimate combination.** When multiple estimates exist and triangulation neither clearly succeeds nor fails, formal combination methods can produce principled summaries. Bayesian Model Averaging (BMA) treats each estimation method as a model, assigns prior probabilities to each based on credibility of assumptions, updates weights based on model fit, and produces a posterior distribution over causal effects that reflects uncertainty across methods. In practice, priors over identification assumptions are rarely specified transparently, and ‘model fit’ is not a proxy for causal validity. BMA should be treated as a thought experiment more than a default production tool. Meta-analytic approaches pool estimates by inverse-variance weighting, though this assumes estimates are independent draws from a common effect distribution—often violated when methods use overlapping data. When methods share data or preprocessing, inverse-variance weights can overstate effective information. In such cases, ranges or partial pooling may be more honest than a single pooled estimate. A simpler alternative is to report the range of estimates (minimum to maximum) as an informal bound, acknowledging that the true effect lies somewhere within if at least one method’s

assumptions are satisfied. The key principle is that combination should reflect genuine uncertainty, not hide disagreement: a wide range across methods is informative, signalling that conclusions depend on modelling assumptions.

## Validation and Reconciliation Checklist

Box 19.8: Validation and Reconciliation Checklist

### Data source reconciliation:

- Have you compared platform metrics to external data (CRM, scanner, financial)?
- What is the discrepancy magnitude? Is it within acceptable thresholds?
- Have you diagnosed root causes (coverage, timing, definitions)?
- Are discrepancies stable over time, or do they show drift or breaks?

### Method reconciliation:

- Have you estimated effects using multiple methods with different assumptions?
- Do estimates agree or diverge? If divergent, have you diagnosed the source?
- Have you reported all estimates with sample sizes, assumptions, and diagnostics?

### Triangulation:

- Does evidence from multiple data sources point in the same direction?
- Does evidence from multiple methods point in the same direction?
- If triangulation fails, have you acknowledged uncertainty and recommended follow-up?

## 19.9 Pipelines and Reproducibility

A reproducible data pipeline is necessary for credible causal inference, but it does not by itself validate identification. This section provides practical guidance on building pipelines that are transparent, auditable, and reproducible—from raw data extraction to final estimates.

### Extract-Transform-Load (ETL) Pipelines

The ETL process documents the full lineage of data, from raw logs to analysis-ready panels.

**Extract.** Pull data from source systems (platform APIs, databases, data warehouses) with explicit timestamps and version identifiers. Document the extraction query, date range, and any filters applied.

**Transform.** Apply cleaning, joining, deduplication, aggregation, and feature engineering. Each transformation should be:

- **Documented.** What does this step do and why?
- **Version-controlled.** Code in Git with meaningful commit messages.
- **Verified.** Unit tests verify expected behaviour (e.g., no duplicate keys after deduplication). These checks should include design-specific invariants, for example no treatment assignments before rollout dates, a balanced panel structure where required, stable definitions of key outcomes across pre and post periods, and no post-treatment information in ‘pre-treatment’ covariates.
- **Idempotent.** Rerunning the pipeline on the same inputs yields identical outputs. Where transformations rely on random sampling or stochastic algorithms, fix random seeds and document them so that idempotence holds in practice.

**Load.** Write the analysis-ready panel to a versioned location with schema documentation. Tag the output with a version identifier (date, hash, or semantic version).

**Tools.** Examples of tools include:

- **dbt.** SQL-based transformations with testing and documentation.
- **Airflow, Prefect, or Dagster.** Workflow orchestration for scheduled pipelines.
- **Great Expectations.** Data quality checks and validation.
- **Git.** Version control for code and Git LFS for large files.

These tools do not substitute for design thinking: they enforce what you specify. A flawed identification strategy implemented in dbt or Airflow remains flawed, just better automated.

## Pre-Registration for Observational Studies

Pre-registration—specifying the analysis plan before seeing outcome data—reduces specification searching and p-hacking risk in observational studies.

### What to pre-register:

- **Estimand.** The causal quantity of interest, for example ATE, ATT, or dose-response functions  $\mu(d)$ .
- **Design.** The identification strategy (DiD, SC, IV, RDD).
- **Sample.** Inclusion and exclusion criteria, time windows, and geographic scope.
- **Treatment definition.** How treatment is operationalised from raw data.
- **Outcome definition.** Primary and secondary outcomes with measurement details.
- **Controls and donors.** Covariates for adjustment or donor pool for synthetic control.
- **Diagnostics.** Pre-trend diagnostics, balance diagnostics, and placebo falsification checks.
- **Inference.** Standard error method, confidence level, and multiple-testing correction.

**When to pre-register.** Before accessing outcome data. It is acceptable to explore treatment and covariate data to refine the design, but outcome data should remain blinded until the plan is locked. Even when outcomes are blinded, document any sample or variable decisions made after exploratory work, because they can introduce implicit selection. This applies to confirmatory analysis—analysis intended to test a specific hypothesis. Exploratory analysis, where researchers search for patterns without pre-specified hypotheses, is legitimate and valuable but should be clearly labelled as such. The distinction matters: confirmatory findings carry more weight because they cannot be explained by specification searching. Pre-registration does not by itself solve identification problems. A pre-registered but misspecified DiD or SC design still fails if its assumptions (parallel trends, factor structure, overlap) do not hold. Registration protects against selective reporting, not against bad designs.

**Pre-registration deviations.** Pre-registration is not a straitjacket. Deviations from the pre-registered plan are acceptable if disclosed and justified. Unexpected data quality issues, unforeseen confounding events, or discovery of coding errors may require plan modifications. Document all deviations, explain why they were necessary, and report both pre-registered and modified analyses when feasible. The goal is transparency, not rigidity.

**Where to register.** Options include OSF, AsPredicted, internal wikis with timestamps, or email to stakeholders with a fixed date. For internal wikis or email, ensure that timestamps and contents are immutable (for example, by archiving PDFs or using systems with version history) so that pre-registration is auditable.

## Design Timelines and Policy Documentation

Design timelines document the temporal structure of the study and anchor comparability claims.

**Key dates to document:**

- Treatment rollout dates (by unit, cohort, or geography).
- Platform policy changes (attribution windows, auction updates, privacy changes).
- Data collection start and end dates.
- Known confounding events (holidays, competitor actions, economic shocks).

These dates anchor pre- and post-treatment windows and define where parallel-trends, stability, and no-interference assumptions are meant to hold. If rollout or policy-change dates are misrecorded, downstream causal claims rest on misaligned periods (see also Section 19.6).

**Policy changelogs.** Maintain a changelog that maps dates to definition changes. This supports sensitivity analysis around break dates and justifies window restrictions.

## Snapshotting and Version Control

Snapshotting inputs at a fixed date prevents silent updates from altering results.

**Data snapshots.** Extract and freeze input data at a specific date. Store snapshots with version identifiers. Never overwrite historical snapshots.

**Code version control.** All analysis code should be in Git. Tag code releases corresponding to specific data snapshots and published results, and record snapshot identifiers in the analysis repo so that anyone can recover the exact data–code pairing used for a result. Document dependencies (package versions) in requirements files or lock files.

**Environment reproducibility.** Use Docker containers or virtual environments to freeze the computational environment. This ensures that code runs identically on different machines. When environments include access credentials or keys, store them outside the image or environment definition (for example, via secrets management), so that reproducibility does not compromise security.

## Audit Trails and Checksums

Audit trails support accountability and debugging.

**What to log:**

- Dataset versions accessed (with hashes or version IDs).
- Code commits used for each analysis run.
- Analyst identity and timestamp for each action.

- Key intermediate outputs (row counts, summary statistics).

Audit logs should reference dataset versions and record-level aggregates, not raw personally identifiable information, to respect privacy constraints while still enabling reproducibility.

**Checksums.** Compute checksums (e.g., SHA-256) on key outputs. An independent researcher with access to the same snapshots and code should regenerate identical checksums, validating reproducibility.

**Levels of reproducibility.** Distinguish three levels of evidence strength:

- **Computational reproducibility.** Same data and same code produce identical results. This is the minimum standard.
- **Robustness.** Different reasonable specifications of the same analysis produce qualitatively similar conclusions. This probes sensitivity to researcher degrees of freedom.
- **Replicability.** New data collected under similar conditions produce consistent findings. This probes external validity and generalisability.

Computational reproducibility is necessary but not sufficient. A finding that is computationally reproducible but not robust to specification changes provides weak evidence. A finding that is reproducible and robust but fails to replicate in new contexts has limited external validity.

## Common Reproducibility Failures

**Silent data updates.** Source data changes without notification, altering results. *Prevention:* Snapshot inputs and compare checksums across runs. A more systematic solution is **data contracts**—formal agreements between data producers and consumers that specify schema (column names, types, constraints), freshness requirements, and quality guarantees. Data contracts make implicit expectations explicit. If a platform changes an attribution window or renames a column, the contract is violated and the pipeline fails loudly rather than silently producing incorrect results. Tools like dbt contracts, Great Expectations, or custom schema validators can enforce contracts automatically at ingestion time. For causal analyses, data contracts should cover variables that encode treatments, outcomes, and key covariates so that any change to their definitions or timing triggers a visible contract violation rather than silently altering the estimand.

**Dependency drift.** Package updates change function behaviour. *Prevention:* Pin package versions and use lock files. Containerise environments.

**Undocumented manual steps.** Analyst performs ad-hoc fixes not captured in code. *Prevention:* Automate all steps and require code review before merge.

**Path dependencies.** Code assumes specific file paths that differ across machines. *Prevention:* Use relative paths and configuration files.

**Random seed omission.** Stochastic algorithms produce different results each run. *Prevention:* Set and document random seeds.

## Reproducibility Checklist

Box 19.9: Reproducibility Checklist

### Pipeline:

- Is the full ETL pipeline documented and version-controlled?
- Are transformations tested with unit tests that include design-specific invariants (for example, no treatment leakage into pre-periods)?
- Are input data snapshots versioned and immutable?

### Pre-registration:

- Was the analysis plan registered before accessing outcome data?
- Are estimand, design, sample, and diagnostics pre-specified?
- Are deviations from the pre-registered plan documented and justified?

### Environment:

- Are package versions pinned in requirements or lock files?
- Is the computational environment containerised or documented?
- Are random seeds set and recorded?

### Validation:

- Can an independent researcher regenerate results from snapshots and code without access to undocumented manual steps?
- Do checksums on key outputs match across runs?
- Is there an audit trail of dataset versions, code commits, and analyst actions?

## 19.10 Assumptions and Guardrails

This section states the core assumptions underpinning data and measurement workflows. These assumptions are implicit in most causal analyses using marketing panel data. Some violations change the estimand (by redefining  $Y_{it}$ ,  $D_{it}$ , or the target population). Others break identification even for a fixed estimand. Each assumption links to diagnostic checks that can detect violations.

Some assumptions are partly diagnosable: stable keys can be verified by auditing migrations, temporal ordering can be checked via join logic, and support can be assessed via overlap plots. Other assumptions are fundamentally untestable: full SUTVA (no interference whatsoever) cannot be verified without knowing the true spillover structure, and correct functional form cannot be confirmed without observing counterfactuals. However, partial-interference assumptions can be probed: buffer zone tests can detect whether estimated effects attenuate with distance from treated units, and exposure-density designs can identify spillover gradients (Chapter 11). These diagnostics do not prove the assumed structure, but they can falsify extreme forms of interference and support more plausible mappings. For untestable assumptions, sensitivity analysis and domain knowledge must substitute for diagnostics. When multiple assumptions are violated simultaneously, biases compound and may reinforce or partially cancel. This interaction makes interpretation difficult and argues for conservative conclusions.

### Data Stability Assumptions

**Assumption 111 (Stable unit keys and measurement invariance)** Unit identifiers (store, SKU, DMA, user) remain stable across the study window, and outcome or exposure definitions do not change. Platform policy changes and schema migrations are documented and either excluded or modelled explicitly.

**Diagnostics.** Check for key migrations or resets (Section 19.3). Monitor metric definitions for breaks (Section 19.5). Compare pre/post summary statistics around suspected break dates. If key migrations or definition changes differ systematically by treatment status or timing, they induce differential measurement error in  $D_{it}$  or  $Y_{it}$  rather than simple attenuation, and can violate parallel trends or factor-structure assumptions.

**Assumption 112 (Documented and stable attribution rules)** Attribution windows, deduplication logic, and conversion definitions are documented and stable across the study window, or breakpoints are flagged and sensitivity is assessed.

**Diagnostics.** Maintain policy changelogs (Section 19.6). Run placebo-style diagnostics around break dates. Report sensitivity to alternative attribution windows. Any change in attribution windows or deduplication logic implicitly changes the estimand (for example, from “effect on 28-day click + view conversions” to “effect on 7-day click conversions”). Analyses that cross such breaks compare different outcomes.

## Temporal Ordering Assumptions

**Assumption 113 (No leakage from post-treatment information)** Joins, transformations, and feature engineering use only information available at or before the time of prediction or treatment assignment. Fold structures in cross-validation block on time to prevent training on future data.

**Diagnostics.** Audit join keys for temporal ordering (Section 19.3). Verify that covariates are measured before treatment. Use time-blocked cross-validation (Chapter 12). Violations here typically invalidate the entire analysis: once post-treatment outcomes leak into “pre-treatment” covariates or training folds, no amount of reweighting or robustness checks can recover the original estimand.

## Interference Assumptions

**Assumption 114 (Limited interference or explicit exposure mapping)** Either (i) treatment effects are effectively confined to own units over the study window (a SUTVA-style setting), or (ii) spillovers are modelled via an exposure mapping  $h_i(D_{-i,t})$  as in Chapter 11. In designs that rely on cross-device or cross-platform linkage to define exposure mappings, linkage rates must be high enough for these mappings to be credible, or sensitivity to linkage failure must be reported.

**Diagnostics.** Diagnose spillovers using buffer zones or exposure gradients (Chapter 11). Assess linkage rates and conduct sensitivity analysis for linkage failure (Section 19.3), because linkage failure can change the constructed exposure mapping and the implied treatment definition.

## Support and Overlap Assumptions

**Assumption 115 (Adequate support after privacy aggregation)** Privacy-driven aggregation, consent-based sample restrictions, and trimming preserve sufficient variation in exposure and outcomes for identification. Overlap and balance diagnostics confirm common support as in Chapter 17.

**Diagnostics.** Check overlap for the treatment definition you use (binary propensity score  $e(X_{it})$  or generalised propensity score for continuous dose), and report how trimming changes the target population. Verify that trimming does not eliminate key subgroups. Report effective sample size and weight dispersion (for example, variance of  $w_{it}$ ) after consent-based restrictions and trimming, so that readers can see how much information the design actually uses (Section 19.6). Trimming and consent-based restrictions redefine the target population. Under adequate support and correctly specified models, you identify effects for the trimmed/consented population, not necessarily for all units. State this target explicitly.

**Assumption hierarchy.** Not all assumptions are equally consequential. Violations of temporal ordering (leakage) typically invalidate the entire analysis and cannot be repaired post hoc. Violations of stable keys create measurement error that often attenuates estimates under simple models, but in practice can also induce differential mismeasurement by treatment status or outcome. Such errors may be partially assessable via sensitivity analysis, but they are harder to correct than pure noise. Violations of adequate support can often be addressed by trimming or restricting the target population. Prioritise diagnostic effort accordingly: verify temporal ordering first, then key stability, then support.

## Assumption Verification Summary

Table 19.2 maps each assumption to its diagnostic checks and relevant sections.

**Table 19.2** Data assumptions and diagnostic checks

Assumption	Diagnostic Check	Reference
Stable keys (Assump. 111)	Key migration audit, break detection	Section 19.3
Stable attribution (Assump. 112)	Policy changelog, placebo diagnostics	Section 19.5
No leakage (Assump. 113)	Temporal ordering audit, time-blocked CV	Sections 19.3 and 12
Limited interference (Assump. 114)	Buffer diagnostics, linkage sensitivity	Chapter 11
Adequate support (Assump. 115)	Overlap diagnostics, effective sample size	Chapter 17

## 19.11 Workflow Checklist

This section consolidates the data and measurement practices from this chapter into a unified workflow. The checklist, figures, and summary table provide a practical reference for implementing design-faithful measurement in marketing causal inference.

The workflow is iterative because diagnostics can force you to revise the treatment/outcome definitions and therefore the estimand. Diagnostics at later stages often reveal problems that require returning to earlier stages. For example, reconciliation failures at Stage 6 may indicate metric definition problems from Stage 3, and overlap violations at Stage 5 may require revising the sample definition from Stage 1. Budget time for multiple passes through the workflow.

## End-to-End Workflow

### Box 19.10: Data–Measurement–Platforms Workflow Checklist

#### Stage 1: Source Inventory

- List all data inputs (platform logs, scanner data, CRM, geo/mobility).
- Define primary and foreign keys that link sources (Section 19.3).
- Document coverage gaps and missingness patterns (Section 19.7).
- Specify the target population, estimand, and assignment/inference unit for each analysis (for example, ATT for existing CRM customers, or a dose–response  $\mu(d)$  over viewable impressions). Make clear which data sources contribute to treatment  $D_{it}$ , outcomes  $Y_{it}$ , and covariates  $X_{it}$  for that estimand.

#### Stage 2: Temporal Alignment

- Standardise on ISO weeks or fiscal periods (Section 19.4).
- Specify seasonality controls (week-of-year, holiday indicators).
- Verify temporal ordering to prevent leakage (Assumption 113).
- Confirm that pre- and post-treatment windows used in later DiD, SC, or event-study designs line up with these aligned periods, and that aggregation windows do not straddle rollout dates. Misaligned calendars can break parallel trends even if within-source timing is correct.

#### Stage 3: Metric-Estimand Alignment

- Map each platform metric (for example, 7-day click conversions) to a clearly defined econometric estimand (for example, the effect of treatment on that platform-defined outcome). State explicitly when the estimand is “effect on platform conversions” versus “effect on sales” or revenue (Section 19.5).
- Construct exposure/dose variables from raw impressions (Section 19.4).
- Validate platform metrics against external data (Section 19.8).

#### Stage 4: Policy Documentation

- Document attribution windows and conversion definitions.
- Maintain changelogs for platform policy changes (Section 19.6).
- Flag break dates and restrict analysis to stable windows.

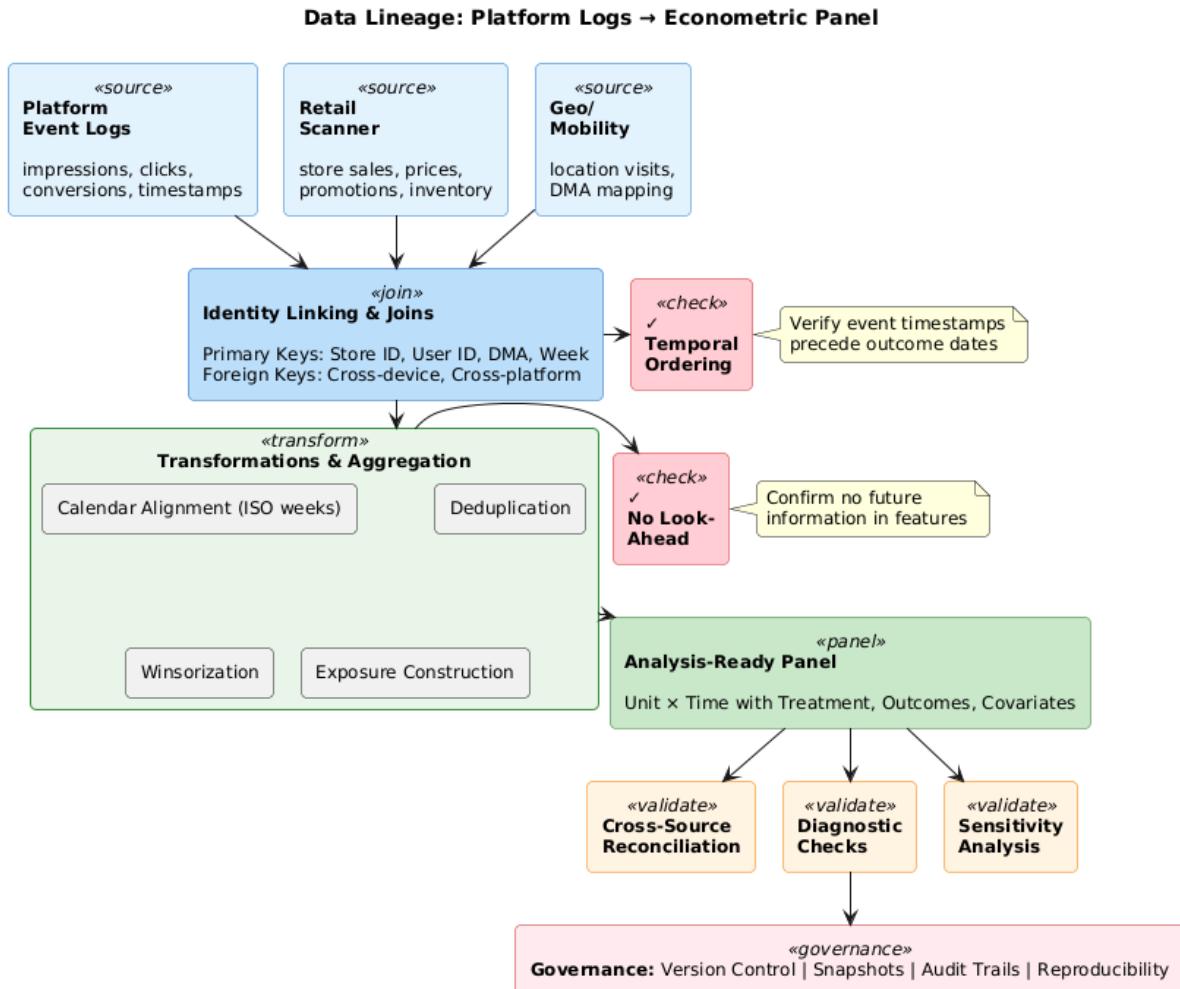
#### Stage 5: Imperfection Handling

- Decide, based on diagnostics and domain knowledge, whether to proceed under MCAR, MAR, or MNAR assumptions (Section 19.7), and document the justification.
- If you impute for reporting/diagnostics, use an imputation scheme that mirrors the identification strategy (e.g., donor-weight imputation for synthetic control) and avoid using post-treatment information.
- Where feasible, bound measurement error using validation studies or sensitivity analysis.

#### Stage 6: Reconciliation and Governance

- Compare estimates across methods (DiD, SC, DML) (Section 19.8). Only include methods in this comparison whose design diagnostics are acceptable on their own. Triangulation strengthens credible designs; it does not salvage those that fail basic checks.

## Visual Workflow Guides

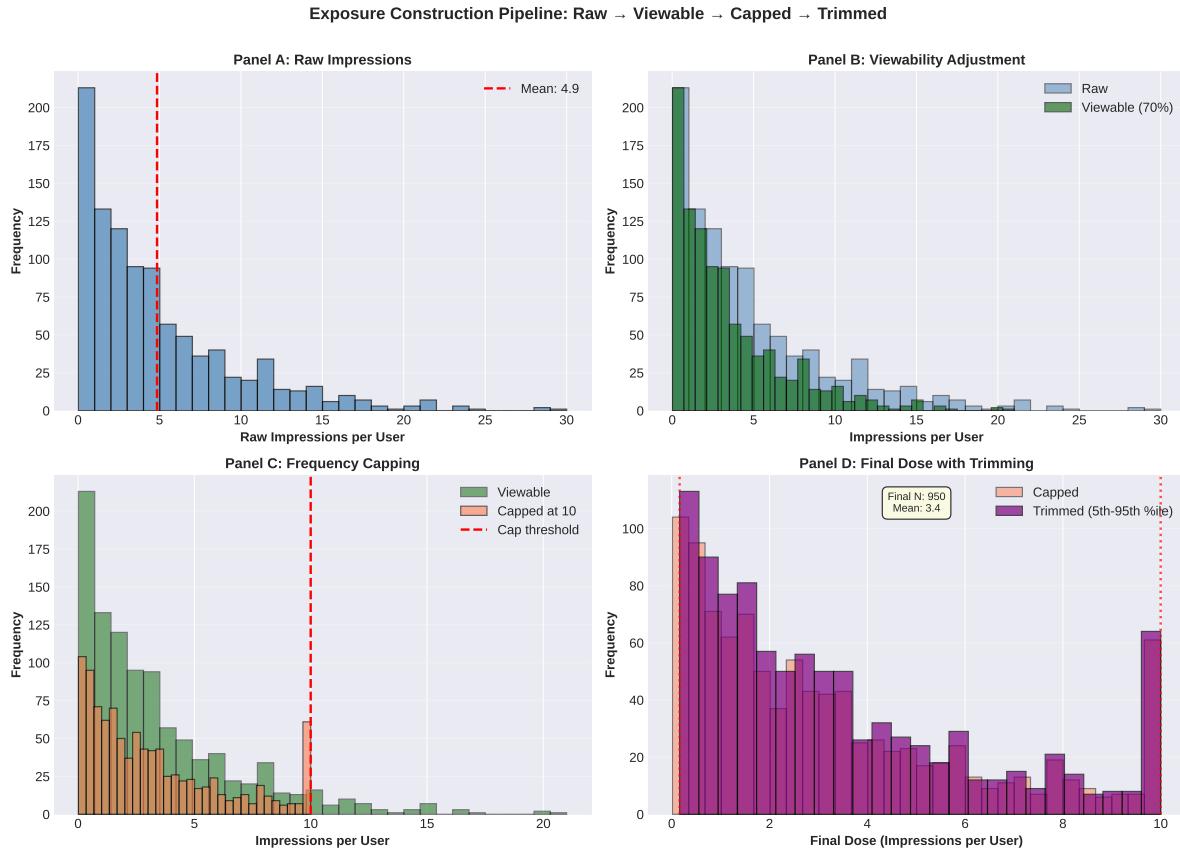


**Fig. 19.1** Data lineage from platform logs to econometric panel with leakage guardrails

*This flowchart depicts the complete data pipeline from raw sources to analysis-ready panel with temporal ordering verification at key stages (red checkmarks). The bottom governance layer ensures version control, snapshots, and audit trails.*

## Metric-Estimand Mapping

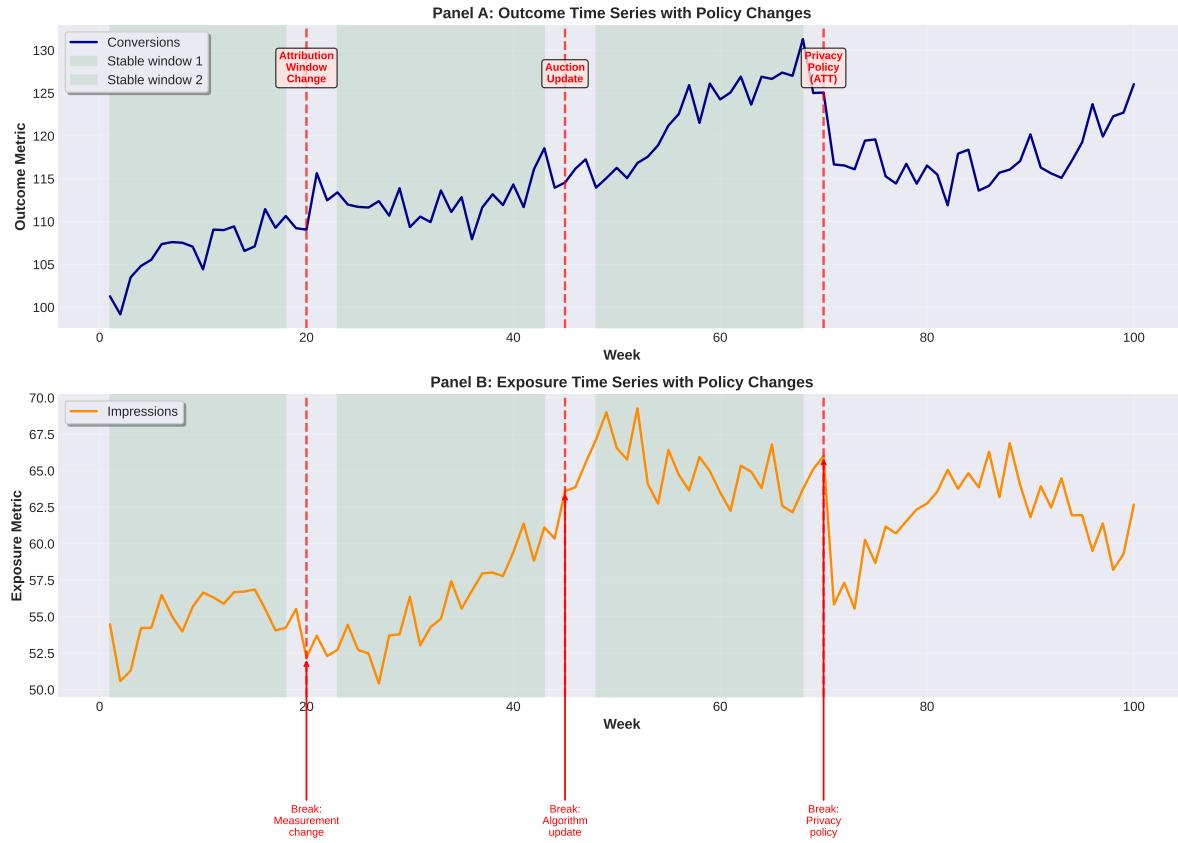
Table 19.3 provides a quick reference for mapping common marketing metrics to their econometric estimands, design constraints, and diagnostic checks. Note that multi-touch attribution and cross-channel effects complicate this mapping: when users are exposed to multiple channels (TV, digital, social) simultaneously, isolating



**Fig. 19.2** Exposure construction and mapping to dose with frequency caps and viewability

This illustrative example shows a four-panel pipeline for exposure construction. Each panel shows how transformations (viewability adjustment, frequency capping, trimming) affect the dose distribution and support for continuous-treatment analysis. All values are hypothetical and chosen to demonstrate how each transformation affects overlap and support rather than to describe a specific campaign.

the effect of any single channel requires additional design considerations such as partial randomisation, sequential experimentation, or structural modelling of the attribution problem.



**Fig. 19.3** Policy change timeline overlaid on outcomes and exposure metrics  
*This stylised example shows how platform policy changes (attribution window, auction update, privacy policy) create measurement breaks that can mimic or obscure treatment effects. Green shaded regions indicate stable windows suitable for analysis. Magnitudes are hypothetical and chosen to illustrate the importance of policy timelines for design-faithful measurement.*

**Table 19.3** Mapping from metric class to econometric estimand, design constraints, and diagnostic checks

Metric class	Econometric estimand	Design constraints	Diagnostic checks
Platform conversions	ATT, ATE on platform-defined conversions	Stable attribution, external validation	Reconcile with retail/finance, sensitivity to attribution windows
Impressions/reach	$D_{it}$ (exposure) and dose-response $\mu(d)$	Viewability, deduplication	Overlap, trimming, frequency distribution
Retail scanner sales	ATT, event-time effects $\theta_k$	Coverage, churn, calendar alignment	Balance, pre-trend diagnostics, support-by- $k$
Geo/mobility flows	Spillover exposure	Catchment definitions, buffers	Exposure maps, buffer sensitivity
Platform lift	Descriptive “lift” (generally biased for ATE)	Endogenous exposure, algorithmic targeting	Use external outcomes and, where possible, randomised or quasi-experimental designs (e.g., geo-experiments). Treat model-based adjustments (e.g., propensity-score methods or factor models) as secondary analyses that still require explicit identification assumptions. Treat platform-reported lift as descriptive unless backed by such designs

## **Chapter 20**

### **Outlook and Open Problems**

This chapter looks ahead. We organise the agenda around the same chain used throughout the book: estimand → identification assumptions → estimator → inference → diagnostics. We focus on failure modes common in platforms, including large-scale interference, nonstationarity, endogenous and adaptive assignment, support erosion, and privacy-induced selection in what we observe. We discuss uncertainty quantification under realistic dependence and we summarise design and reporting practices that practitioners can implement now, while being clear about what remains unsettled.

## 20.1 Motivation and Scope

Methods for panel causal inference have advanced rapidly, yet key challenges remain in marketing, where platforms, privacy rules, and real-time optimisation shape both treatment assignment and what is observed. In this chapter we set an agenda for these open problems. Our approach remains grounded in the design-based principles of this book, with a sharp focus on clear estimands and transparent identification assumptions. Predictive accuracy is useful for imputation and nuisance estimation, but it does not by itself justify causal claims [Angrist and Pischke, 2010, Pearl, 2009, Breiman, 2001].

### What This Book Has Covered

The preceding chapters have developed a toolkit for panel causal inference in marketing. Chapters 2–15 set out the design-based foundations. Chapters 4–8 develop the core estimators. Later chapters extend the core designs to staggered adoption, continuous treatments, high-dimensional nuisance estimation, and interference. Chapters 16–19 then cover inference, applications, and measurement.

These tools address many practical problems, but the frontier continues to advance. New challenges emerge as platforms evolve, privacy constraints tighten, and marketing systems become more adaptive.

### What Remains Open

This chapter isolates thirteen settings in which at least one link in the design chain fails: the estimand is unclear, identification assumptions are implausible, diagnostics are weak, or inference does not match the dependence structure. Our list shares themes with the open problems identified by Cinelli et al. [2025], particularly regarding interference and external validity, but focuses specifically on the panel data structures common in marketing. We begin with interference at scale (Section 20.2). We then consider structural instability and regime change (Section 20.3), adaptive experimentation (Section 20.4), and method selection when assumptions are uncertain (Section 20.5). Next we discuss inference and estimation in difficult settings: robust and distribution-free inference (Section 20.6), continuous treatments (Section 20.7), and partial identification (Section 20.8). Finally, we cover ML integration beyond nuisance (Section 20.10), the synthetic-data trap (Section 20.9), privacy-preserving measurement (Section 20.11), reproducibility standards (Section 20.12), a practitioner roadmap (Section 20.13), and assumptions for future practice (Section 20.14).

### Audience and Intent

This chapter is addressed to both researchers and practitioners. For researchers, we identify open problems that merit formal investigation, where new theory, methods, or computational tools are needed. For practi-

tioners, we highlight areas where current best practices are uncertain and where caution is warranted (see Section 20.13).

Our intent is not to provide turnkey estimators. Many of these problems remain unsolved. Instead, we aim to state the causal obstacles precisely and to describe what additional data, assumptions, or designs would be needed for credible inference. The field advances when researchers and practitioners engage with the hardest problems rather than avoiding them.

## 20.2 Interference at Scale

The partial-interference framework from Chapter 11 assumes that units can be partitioned into non-overlapping clusters with no between-cluster spillovers. This assumption fails in modern digital platforms where users belong to multiple social groups, geographic and online networks overlap, and treatment effects propagate through dense, interconnected systems. In platform settings, the no-interference component of SUTVA fails and the exposure mapping becomes high-dimensional. The resulting problem is not only statistical. Identification hinges on the assignment mechanism and on design choices that create usable, well-supported variation in exposure.

### The Challenge of Dense Networks

Marketing increasingly operates on platforms where interference is the rule, not the exception. In social networks, a user's response to an ad depends on whether their friends were also exposed. Viral effects, social proof, and word-of-mouth create spillovers that propagate through the network, and the exposure mapping  $h_i(D_{-i,t})$  from Chapter 11 becomes high-dimensional when each user has hundreds of connections.

Marketplace platforms exhibit similar complexity. Treating one seller affects competing sellers through price competition, search ranking displacement, and inventory effects. Treating one buyer affects sellers through demand shifts and affects other buyers through supply constraints. The two-sided nature of platforms means that interference flows in multiple directions simultaneously.

Consider a ride-sharing platform testing a new surge pricing algorithm. Treating drivers with higher surge multipliers affects rider wait times, which affects rider demand, which feeds back to driver earnings. A naive experiment that randomises drivers ignores this feedback loop. An experiment that randomises riders ignores the supply-side response. The platform must design experiments that account for both sides of the market simultaneously.

Geographic overlap compounds the problem. A geo-experiment in one DMA spills over to adjacent DMAs through commuting, media spillover, and supply chain effects. Buffer zones help but waste experimental power and may not fully eliminate contamination. Platform algorithms create yet another layer of indirect interference: treating one user changes the algorithm's predictions, which affects recommendations to other users. This algorithmic spillover is difficult to model because the algorithm itself is often a black box.

### Foundational Theory

The theoretical foundations for causal inference under interference were established by Hudgens and Halloran [2008], who introduced the partial-interference framework and defined direct and indirect (spillover) effects. Athey and Imbens [2017] provide a comprehensive review of this design-based perspective. Under partial interference, we can identify both a direct effect of own treatment and an indirect effect of exposure to

treated neighbours, provided the assignment mechanism is known and delivers overlap in the relevant within-cluster treatment patterns.

Under interference, an “effect” is not a single object. You must state whether you target a direct effect holding exposure fixed, a spillover effect of changing exposure, or a policy effect that includes equilibrium adjustment in the network or market.

This is one reason interference resists generic templates. The way a setting violates no-interference is typically idiosyncratic, which is why Cinelli et al. [2025] describe interference as an “Anna Karenina” problem. For practice, that means you should treat the exposure mapping  $h_i(D_{-i,t})$  and the assignment mechanism as part of the design, and report sensitivity to alternative exposure definitions rather than presenting a single preferred mapping as definitive.

Aronow and Samii [2017] extended this framework to general interference, where we do not assume any cluster structure. Their approach uses inverse probability weighting with exposure probabilities rather than treatment probabilities. These exposure probabilities are computable only under a known assignment mechanism and require positivity for the exposure levels induced by that mechanism. The estimand becomes the effect of a particular exposure mapping, a function that summarises how each unit’s outcome depends on the full vector of treatment assignments. This generality comes at a cost: we need to specify the exposure mapping, and inference requires knowledge of the joint distribution of exposures.

The practical implication is that we cannot escape modelling assumptions about how interference operates. We must specify which units affect which others and how. The challenge at platform scale is that these assumptions become high-dimensional and difficult to verify.

## Current Approaches and Their Limitations

Chapter 11 described several designs that partially address interference. At platform scale, we encounter binding limitations.

Cluster randomisation assigns treatment at the cluster level—graph clusters or geographic regions—to contain spillovers within clusters. When clusters  $c = 1, \dots, G$  are treated as independent, the effective sample size for inference is  $G$ , not  $N$ . In dense networks, clusters are rarely well-separated, so between-cluster spillovers remain while  $G$  is small, and precision deteriorates. We may partition a social network into 50 clusters, only to find that 30% of edges cross cluster boundaries.

Exposure mappings summarise neighbour treatments into low-dimensional variables, such as the fraction of friends treated. In dense graphs, the true exposure may depend on higher-order connections and on interaction between network position and treatment. Misspecification becomes almost inevitable. Identification requires variation in exposure conditional on own treatment, which may be scarce when assignment is correlated across neighbours.

Switchback experiments randomise treatment over time rather than across units, providing exogenous variation even when spatial interference is pervasive. They rely on limited carryover effects, however, and do not address spatial interference directly. Temporal dependence complicates inference.

Ego-network designs randomise treatment to focal users and measure outcomes on their neighbours. In dense networks, ego-networks overlap heavily, violating the non-overlap conditions that make these designs interpretable.

## Bipartite Experiments for Two-Sided Markets

Two-sided marketplaces—buyers and sellers, riders and drivers, guests and hosts—have a special structure we can exploit. The interaction graph is often approximately bipartite, though same-side competition can reintroduce within-side interference through congestion or rationing.

Bajari et al. [2023] develop experimental design principles for marketplaces that exploit this bipartite structure. The key insight is that we can randomise on one side of the market while measuring outcomes on both sides. If we randomise which sellers receive a promotional treatment, buyers provide the variation we need to measure seller-side effects, and we can measure buyer-side spillovers by comparing buyers who transacted with treated versus control sellers.

This design breaks the simultaneity problem that plagues naive marketplace experiments. Rather than trying to randomise the entire market, we fix one side and let the other side respond. The resulting estimand is a partial-equilibrium effect, the effect of changing assignment on one side while treating the other side’s response as part of the outcome process induced by the design. In applications you should state explicitly whether the target is, for example, a seller-side ATT<sub>t</sub>, a buyer-side ATT<sub>t</sub>, or the value of a marketplace policy rule.

Practical implementation requires care. We need sufficient within-market variation to identify effects. We need to track which buyers transacted with which sellers to construct the relevant exposure mappings. We need to account for selection into transactions, since buyers who transact with treated sellers may differ from those who transact with control sellers. Some platforms can support these requirements, but only when transaction links and assignment logs are retained and accessible for analysis.

## Platform-Facilitated Experiments

Platforms have unique advantages for addressing interference. They can partition their user graph into clusters using community detection algorithms, then randomise at the cluster level. This graph-cluster randomisation requires platform cooperation and access to the social graph, but it enables designs that external researchers cannot implement.

Market-level experiments offer another approach. Platforms can randomise entire markets—cities, product categories—to contain equilibrium effects within markets. This sacrifices granularity for cleaner identification. When only one market is treated, synthetic control methods from Chapter 6 can construct counterfactuals from untreated markets, provided spillovers across markets are limited. Zervas et al. [2018] use this approach to estimate Airbnb’s effect on hotel revenue by exploiting city-level variation in Airbnb penetration.

Platforms can also create exogenous variation in neighbour treatment by randomising the information users receive about their neighbours' behaviour. This instrumented interference allows identification of peer effects without network-level randomisation, but designs must separate the effect of information from the effect of peers' actual behaviour.

## Computational Scaling

Even when identification is in place, computation poses challenges. Computing exposure mappings for millions of users with thousands of connections each is expensive. Approximate methods—sampling, sketching, and sparse representations—trade accuracy for speed, but the bias–variance trade-offs are not well understood.

Network-robust variance estimators require summing over all pairs of connected units, which scales as  $O(m)$  where  $m$  is the number of edges. For dense networks with billions of edges, this is prohibitive. Sparse approximations and parallel algorithms are needed to make this feasible at scale.

Even when a variance formula is available, its validity depends on a dependence model, for example weak dependence along edges. Scaling computation is necessary but not sufficient because approximations must preserve the variance target to avoid under-coverage.

Design optimisation—choosing which clusters to treat to maximise power while limiting spillover contamination—is a combinatorial problem. Greedy heuristics and approximation algorithms help, but optimal designs remain computationally intractable for large networks.

## Practical Guidance

Given the challenges above, how should practitioners proceed? We offer the following workflow.

1. Map the interference structure. Before designing an experiment, understand how units affect each other. Is interference local (only immediate neighbours) or does it propagate? Is the network dense or sparse? Is it bipartite?
2. Choose the design to match the structure. For sparse networks with clear clusters, cluster randomisation can work. For two-sided marketplaces, bipartite designs exploit the market structure. For dense networks with no clear clusters, switchbacks may be the only option.
3. Specify the exposure mapping explicitly. Do not assume that only immediate neighbours matter. Check sensitivity to alternative specifications and report results under multiple exposure definitions.
4. Treat the dependence and variance model as part of the design. State the independent sampling unit used for inference, for example clusters  $c = 1, \dots, G$  or time blocks  $s = 1, \dots, S$ . When you treat clusters as independent, use a cluster-robust variance estimator built from cluster-level aggregates and interpret uncertainty at rate  $\sqrt{G}$ .

5. Bound the bias from interference. If you cannot eliminate interference, you can sometimes bound it. Sensitivity analysis that asks “how large would spillovers need to be to overturn the result?” can be more informative than a single preferred specification.

These steps do not solve the fundamental problem of interference at scale, but they make the problem transparent and provide readers with the information needed to assess the credibility of the results.

## Research Directions

Progress on interference at scale will require new identification frameworks that relax the non-overlapping cluster assumption while maintaining tractability. We need scalable algorithms for exposure calculation, variance estimation, and design optimisation on large networks. Platform partnerships that provide access to network structure and enable large-scale experimentation will be essential.

Structural models that link experimental estimates to equilibrium policy effects would bridge the gap between what experiments identify and what policy requires. Sensitivity analysis that bounds estimates under plausible violations of the partial-interference assumption would help practitioners assess robustness.

Key foundations exist, but they do not yet deliver a routine workflow for billion-user settings with overlapping interference and proprietary algorithms. Aronow and Samii [2017] provide the theoretical framework for general interference. Bajari et al. [2023] provide design principles for marketplaces. The challenge now is to scale these ideas to the billion-user networks where modern marketing operates.

## 20.3 Structural Instability and Regime Change

The marketing environment is not static. Platform algorithms update, privacy policies change, competitors enter and exit, and macroeconomic conditions shift. These changes violate stability conditions that many panel designs rely on, for example stable untreated potential outcomes  $Y_{it}(0)$  for controls, stable factor loadings in interactive fixed effects, and stable measurement of outcomes and treatment. Diagnostics can flag when instability is likely, but they do not restore identification without additional assumptions or design changes.

### A Note on Terminology

The term “nonstationarity” has a precise meaning in time-series econometrics: a process is nonstationary if its joint distribution—or, under weaker definitions, its mean, variance, or autocovariance structure—depends on time. Econometricians use this term to describe several distinct phenomena. Deterministic trends arise when the mean drifts over time, as in  $y_t = a + bt + \varepsilon_t$  with i.i.d. errors. Time-varying heteroskedasticity appears when  $\text{Var}(\varepsilon_t)$  depends on  $t$ . Structural breaks occur when parameters shift at known or unknown dates. Unit roots arise when  $y_t$  follows a random walk,  $y_t = y_{t-1} + \varepsilon_t$ , generating an I(1) process.

The first three can still admit asymptotic approximations under explicit regularity conditions, for example weak dependence and stable moments, but panels require care because neither serial nor cross-sectional dependence is negligible. Unit roots are different. They require functional limit theory and specialised inference procedures.

In many marketing applications,  $T$  is short enough that unit-root diagnostics have low power and are rarely decisive. With  $T = 12$  or  $T = 24$  months, we cannot reliably diagnose unit roots or estimate their properties. The practical concern is structural breaks and measurement regime changes.

Throughout this section, we use “structural instability” rather than “nonstationarity” to avoid conflating these distinct issues. When we do use “nonstationarity,” we mean it in the broad sense of “the DGP changes over time,” which is the usage common in applied panel causal inference, even if it lacks the precision of the time-series definition.

### Types of Structural Instability

Structural instability manifests in several forms, each with different implications for identification.

*Discrete structural breaks* are sudden, persistent shifts in the data-generating process. Apple’s App Tracking Transparency update in April 2021 is a canonical example: overnight, the relationship between digital advertising exposure and measured conversions changed dramatically because users could now opt out of cross-app tracking. Other examples include algorithm updates (Google’s search ranking changes), regulatory events (GDPR implementation), and platform policy shifts (cookie deprecation). These breaks are often ob-

servable but may not be announced in advance. They change the conditional mean function, the variance, or both.

*Continuous parameter drift* involves gradual evolution of parameters over time. Treatment effects may decay as novelty wears off. Factor loadings may drift as market structure evolves. Consumer preferences may shift slowly. Drift is harder to detect than discrete breaks because there is no single break date to identify. Unlike unit roots, which are stochastic and cumulative, parameter drift is typically modelled as deterministic or as a slow-moving random walk with small innovations.

*Regime switching* occurs when the system alternates between multiple states—high-demand and low-demand regimes, or competitive and collusive pricing. The current regime may be unobserved, and transitions may be stochastic. Even if the environment is stable within regimes, aggregating across regimes produces apparent instability.

*Time-varying treatment effects* represent a distinct form of instability. Even if the environment is stable, the treatment effect parameter itself may vary over time due to learning, saturation, or competitive response. The average treatment effect on the treated in period  $t$ ,  $\text{ATT}_t$ , may differ from  $\text{ATT}_{t+k}$ , not because the DGP changed, but because the treatment operates differently at different horizons. In the notation of Chapters 2 and 4, this means objects like  $\text{ATT}_t$  or  $\tau(g, t)$  vary over  $t$  for reasons beyond the deterministic dynamics we intend to capture, so aggregating across periods without accounting for instability can misstate the target estimand.

## Implications for Panel Methods

Different estimators are affected differently by structural instability, and we must understand these vulnerabilities to choose methods wisely.

Difference-in-differences identifies effects only under restrictions on untreated potential outcomes, typically stated as a parallel-trends condition for  $Y_{it}(0)$  across treated and control groups. A structural break in the pre-period undermines the extrapolation to the post-period, and breaks in the post-period confound treatment effects with underlying trend shifts. Note that time fixed effects do not solve this problem. They absorb common shocks but not differential responses to those shocks across treated and control groups. If a privacy policy change affects treated and control DMAs differently, a year dummy for 2021 does not help.

Synthetic control depends on pre-treatment fit quality, which in turn depends on stability. If the relationship between treated and donor units changes, because factor loadings drift or a structural break differentially affects the treated unit, the synthetic control constructed from pre-treatment data may not approximate the treated unit's counterfactual in the post-period. SDID from Chapter 7 can improve pre-treatment fit, but it remains vulnerable when the treated–donor relationship changes at the break.

Interactive fixed effects from Chapter 8 allow for time-varying factors but assume stable factor loadings. If loadings drift, the low-rank structure breaks down. Matrix completion methods are similarly affected.

Double machine learning from Chapter 12 relies on using training folds that represent the same regime as the evaluation fold. If the conditional outcome model changes over time, cross-fitting can still be biased because the nuisance estimates are trained under a different regime than the target estimand.

## A Concrete Example: iOS App Tracking Transparency and Panel Estimation

Consider a brand measuring the effect of a TV campaign on measured attributed online conversions using synthetic control. The treatment—a TV campaign launch—occurs in February 2021. We construct a synthetic control from untreated DMAs using January 2020 to January 2021 as the pre-period, and measure effects through June 2021.

In April 2021, Apple rolls out the App Tracking Transparency policy. Before the policy change, we observe 80% of online conversions attributed to digital touchpoints. After it, this drops to 40% because users opt out of tracking. The synthetic control, constructed from pre-policy data, predicts what measured attributed online conversions would have been absent the TV campaign, but this prediction assumes the measurement environment is stable (see Section 20.11).

The problem is that the privacy change affects both treated and control DMAs, but potentially differently. If the treated DMAs have higher iPhone penetration, the change affects them more, and we may attribute to the TV campaign what is actually a differential measurement shock. Even if iPhone penetration is similar, the timing creates confounding: we cannot separate the TV campaign effect from the privacy change in the February–June window.

What can we do? One option is to restrict the post-period to February–March 2021, before the policy change. This changes the estimand to a shorter-run effect over that window. It also sacrifices statistical power and prevents us from measuring longer-run effects. Another option is to include a policy-change indicator interacted with iPhone penetration as a covariate. This is a strong modelling move. It assumes the measurement shock is fully captured by the interaction and that no other contemporaneous shocks differentially hit treated and donor units. A third option is to use a different outcome, total sales rather than attributed online conversions, that is less affected by the measurement shock. This changes the estimand but may be more robust.

The broader lesson is that we must inventory potential structural breaks before analysis, not after. If we know a major policy change is coming, we can design around it. If we discover it post hoc, we are left with imperfect fixes.

## What Time Fixed Effects and First Differencing Do (and Don't) Solve

Two common tools—time fixed effects and first differencing—are sometimes proposed as solutions to instability. Understanding their limitations is important.

Time fixed effects absorb any time-varying shock that affects all units equally. They handle aggregate demand shocks, common seasonality, and platform-wide changes that affect everyone identically. They do *not* handle unit-specific trends or differential responses to aggregate shocks, structural breaks that affect units heterogeneously, or time-varying parameters in the treatment effect itself.

First differencing transforms  $Y_{it}$  to  $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$ . This eliminates unit fixed effects (since  $\alpha_i$  cancels) and removes deterministic trends (since a linear trend becomes a constant). It changes the object being estimated and typically increases serial dependence in errors, often inducing MA(1) structure, which affects inference. It also does *not* solve structural breaks (the break is still present in the differenced data), time-varying variance (differencing does not stabilise heteroskedasticity), or regime switching (regime differences persist in differences).

The upshot: neither time effects nor first differencing is a general solution to structural instability. They address specific forms of instability and leave others untouched.

## Detection Methods

When breaks are announced—platform policy changes, for instance—we can restrict analysis to stable windows or model the break explicitly. When breaks are unannounced, detection is harder.

For time series, Chow-type break diagnostics can be used when break dates are known [Chow, 1960], while Bai–Perron procedures address unknown break dates by searching over candidate dates and checking for parameter instability [Bai and Perron, 1998]. CUSUM-type diagnostics track cumulative sums of residuals to flag gradual departures from stability [Brown et al., 1975]. These methods are well-developed for single time series but less so for panels with cross-sectional dependence. With the short  $T$  typical in marketing panels, these diagnostics have limited power and should be treated as heuristics that can flag gross instability, not as definitive arbiters of stability.

Rolling estimation provides another approach: we estimate parameters on rolling windows and monitor for instability. Large changes in rolling estimates suggest breaks or drift. The challenge is distinguishing genuine breaks from sampling variation, especially with short windows.

Placebo checks around suspected break dates can also help. Running placebo DiD or SC analyses around suspected breaks can reveal whether the break has contaminated the control group. If an apparent “treatment effect” appears at a placebo date, instability is a plausible explanation.

## Recent Advances: Honest DiD and Synthetic Trends

Two methodological developments directly address structural instability in panel causal inference.

Rambachan and Roth [2023] introduce “Honest Difference-in-Differences,” a framework for partial identification when parallel trends may not hold exactly. Instead of assuming zero violation, they bound the maximum post-treatment violation by a multiple of the observed pre-treatment trend differences. This allows

analysts to report a “breakdown value”—how much trend instability would be required to overturn the result? This framework has emerged as a leading approach for handling potential trend breaks and is increasingly used in applied work.

Liu [2025] develops synthetic parallel trends, which use SC-like methods to construct counterfactual trends under drift, under maintained assumptions about stability of the weighted untreated outcomes. Rather than assuming parallel trends hold exactly, this approach constructs weights that minimise trend differences in the pre-period and extrapolates these adjusted trends to the post-period.

## Handling Regime Switching

Regime switching—alternating between high-demand and low-demand states—requires explicit modelling.

Change-point detection tools (CUSUM-type diagnostics and related algorithms) can suggest candidate breakpoints where the mean or variance shifts. Once we posit breakpoints, we can stratify the analysis by regime or include regime indicators as covariates. This is a modelling move and it does not by itself restore identification when regimes are correlated with treatment timing or when breaks affect treated and control units differently.

When regimes are latent, Markov switching models and Hidden Markov Models (HMMs) can infer the probability of being in a given state at each time  $t$  [Hamilton, 1989]. We can then estimate regime-specific treatment effects or marginalise over the regime distribution. Markov-switching and HMM approaches require relatively long time series and strong assumptions about transition dynamics and within-regime behaviour. In many marketing panels they are better viewed as exploratory tools than as a basis for hard causal claims.

## Practical Guidance

When structural instability is likely, we offer the following workflow.

1. Document known breaks. Maintain changelogs of platform policies, algorithm updates, and market events. When major breaks occur, record the date and expected mechanism of impact. This documentation should be contemporaneous, not reconstructed post hoc.
2. Restrict analysis to stable windows or model breaks explicitly. If a major break occurs mid-analysis, either truncate the sample or include the break as a covariate. Do not simply ignore it. Time fixed effects alone do not solve this problem.
3. Use detection methods proactively. Rolling estimates and placebo checks can flag unannounced breaks. If estimates are unstable across windows, investigate why before reporting results.
4. Conduct sensitivity analysis. Show how estimates change when the analysis window is varied or when different pre-treatment periods are used. If results are sensitive to window choice, acknowledge this limitation.

5. Triangulate across methods. Use multiple estimators with different stability assumptions. Agreement across methods is supportive evidence. Disagreement is a diagnostic signal that warrants investigation.
6. Report estimates as conditional on stability. Acknowledge that extrapolation to other time periods may be invalid. If the environment has changed since the analysis period, the estimated effects may no longer apply.

These steps do not eliminate the problem of structural instability, but they make it transparent and provide readers with the information needed to assess the credibility and applicability of the results.

## 20.4 Adaptive Experimentation and Learning

Modern platforms rarely assign treatments randomly and hold them fixed. Instead, they deploy adaptive systems—bandits, pacing algorithms, budget optimisers, and recommendation engines—that continuously adjust treatment assignment based on observed outcomes. This adaptivity creates fundamental challenges for causal inference. It violates the fixed-design condition that assignment probabilities are set *ex ante* and are independent of realised outcomes conditional on the information set available at assignment time.

### Types of Adaptive Systems

Adaptive treatment assignment takes many forms in marketing, and we must understand each to diagnose the identification challenges they create.

Multi-armed bandits allocate traffic to ads, creatives, or product variants based on observed performance. Thompson sampling draws from posterior distributions over arm rewards and selects the arm with the highest draw. Upper confidence bound (UCB) algorithms select arms based on optimistic reward estimates.  $\epsilon$ -greedy algorithms exploit the best-performing arm most of the time but explore randomly with probability  $\epsilon$ . In all cases, treatment assignment probabilities evolve as data accumulate, and units assigned later face different probabilities than units assigned earlier.

Pacing and budget algorithms present similar challenges. Ad delivery systems pace spend over time to meet budget constraints while maximising value. Assignment depends on predicted outcomes, bid landscapes, and remaining budget—all of which evolve endogenously. A unit arriving late in the budget cycle faces different assignment probabilities than a unit arriving early.

Recommendation engines add another layer: personalisation algorithms assign content, products, or ads based on predicted user preferences, and these predictions update continuously as user behaviour is observed. These systems optimise predictions, not causal estimands. Without additional design or assumptions, the data they generate need not identify causal effects.

Dynamic pricing creates simultaneity rather than just selection. Prices adjust in real time based on demand signals, inventory, and competitor behaviour. The treatment (price) is endogenous to the outcome (demand), and standard methods cannot disentangle cause from effect without further structure.

Reinforcement learning agents, increasingly deployed on platforms, learn policies over time. The policy itself is the object of interest, but it evolves during the observation period, making the target of inference a moving target.

### Contextual Bandits and Personalisation

While the previous subsection focused on simple multi-armed bandits, most marketing applications involve *contextual bandits*—algorithms that assign treatments based on features, not just past rewards. Personalisa-

tion engines are contextual bandits: they observe features  $X_{it}$  and assign exposure  $D_{it}$  to maximise predicted outcomes given those features.

Contextual bandits create additional challenges beyond simple bandits. Write the logging policy as  $\pi_{\log}(d | \mathcal{H}_{it})$ , where  $\mathcal{H}_{it}$  is the logged information set available at assignment time. In many personalisation systems,  $\mathcal{H}_{it}$  is summarised by a high-dimensional feature vector. As the algorithm learns which treatment is best for each user type,  $\pi_{\log}(d | \mathcal{H}_{it})$  can concentrate on a single action, creating the *vanishing propensity problem*. Users whose history strongly predicts response to treatment A receive treatment A with probability approaching 1.

When propensities vanish, inverse probability weighting fails. The IPW estimator reweights observations by  $1/\pi_{\log}(d | \mathcal{H}_{it})$ , but as  $\pi_{\log}(d | \mathcal{H}_{it}) \rightarrow 0$ , these weights explode. The variance of IPW estimators becomes unbounded, and estimates become unstable. This is not a minor technical issue. It is the central obstacle to inference with personalisation data.

Solutions exist but require different approaches. Stabilising weights can reduce variance, but it typically changes the estimand unless you state the reweighted target explicitly, for example an overlap-restricted policy value. *Policy learning* approaches, such as those developed by Athey and Wager [2021], directly estimate optimal policies rather than treatment effects, sidestepping the most extreme propensity regimes. *Doubly robust methods* combine outcome modelling with propensity weighting. When propensities are near-degenerate, the outcome model carries most of the inferential burden.

For practitioners, the implication is clear: if your data come from a mature personalisation system, propensities may be too extreme for standard IPW. Doubly robust estimators are often essential. Better still, design experiments with exploration phases where propensities are bounded away from zero.

## Why Adaptivity Breaks Standard Methods

In the design-based paradigm of Chapters 2–15, treatment assignment is either randomised with known probabilities or determined by observable covariates. Adaptive systems violate both conditions because assignment probabilities depend on past outcomes.

Consider a Thompson sampling bandit allocating traffic between two ad creatives. After 1000 impressions, Creative A has received 700 impressions (70%) and Creative B has received 300 (30%). Why the imbalance? Because Creative A was performing better, and the algorithm shifted traffic toward it.

A common misconception is that this creates estimation bias in the usual sense. Under a stationary reward model with i.i.d. outcomes within arm, the sample mean  $\bar{Y}_A$  is unbiased for its arm mean. In marketing panels, stationarity often fails, which is why adaptivity can induce time and context confounding. The problems are subtler but equally damaging:

*Winner's curse (selection bias).* If we select the arm with the highest sample mean and report that mean as our estimate, *that* estimate is biased upward. The winning arm is more likely to have benefited from positive sampling variation. This is selection bias, not estimation bias—we are reporting a selected quantity, not a random draw.

*Data-dependent sample size.* The sample size  $N_A$  is random and correlated with outcomes. Arms that perform well early receive more samples, and arms that perform poorly receive fewer. Fixed-sample CLTs do not apply directly. Valid inference typically relies on sequential (martingale) arguments and requires logging the assignment probabilities and specifying the information set that governs assignment.

*Temporal confounding.* The algorithm shifted traffic to A precisely when A was performing well, so A's sample is enriched with high-converting conditions (time of day, user segments). B's sample is enriched with periods when A was underperforming. This creates confounding by time and context that naive comparisons ignore.

In simulations, these effects combine to produce substantial errors. Naive estimates from bandit data can overstate the winner's advantage by 20–30%, making a 2 percentage point true difference look like 2.5–3 points.

The fundamental issues are:

*Endogenous assignment.* In a bandit, the probability that unit  $i$  receives treatment depends on the outcomes of all previous units. This probability is not fixed ex ante and may not be computable ex post without access to the algorithm's internal state.

*Feedback loops.* Units that would have good outcomes under treatment are more likely to be assigned treatment, confounding the treatment–outcome relationship. This is not standard confounding by observables—it is confounding by the algorithm's beliefs, which are updated by observed outcomes.

*Temporal dependence.* Each assignment depends on all previous assignments and outcomes, inducing dependence that breaks standard independence assumptions. The observations are not i.i.d., and standard central limit theorems do not apply directly.

*Optional stopping.* Adaptive experiments often stop when a “winner” is declared, creating additional selection bias. The stopping rule depends on the data, invalidating fixed-sample inference. If we stop an experiment because one arm looks better, we are more likely to stop when sampling variation favours that arm.

## Off-Policy Evaluation

A primary use case for adaptive data in marketing is *off-policy evaluation* (OPE): using data collected under one policy (the logging policy) to evaluate a different policy (the target policy). The estimand is a policy value, not a treatment effect. For example, letting  $\mathcal{D}$  denote the feasible action set, we might target

$$V(\pi_{\text{target}}) = \mathbb{E} \left[ \sum_{d \in \mathcal{D}} \pi_{\text{target}}(d | \mathcal{H}_{it}) Y_{it}(d) \right],$$

under a sequential consistency condition and a sequential unconfoundedness condition given the logged information set  $\mathcal{H}_{it}$ . We also need overlap for the target policy, for example  $\pi_{\log}(d | \mathcal{H}_{it}) \geq \epsilon$  whenever  $\pi_{\text{target}}(d | \mathcal{H}_{it}) > 0$ . When we enforce overlap by trimming, truncating weights, or restricting the policy class, we change the estimand to the corresponding overlap-restricted policy value.

In what follows,  $n$  denotes the number of logged decision points in the dataset, that is, the number of observed cells  $(i, t)$  at which an action was assigned and an outcome was measured for the corresponding evaluation window.

When decisions repeat within units, the independent sampling unit for inference is typically the unit  $i$ . In that case, treat  $n$  as a count of observations but cluster uncertainty at the unit level, with effective sample size  $N$  (or the number of independent higher-level clusters, when applicable).

Three main approaches exist, each with different bias–variance trade-offs.

*The direct method (DM)* builds a model  $\hat{\mu}(d, \mathcal{H}_{it})$  of expected outcomes given action  $d$  and information set  $\mathcal{H}_{it}$ , then evaluates the target policy by computing  $\hat{V}^{\text{DM}} = \frac{1}{n} \sum_{i,t} \sum_{d \in \mathcal{D}} \pi_{\text{target}}(d | \mathcal{H}_{it}) \hat{\mu}(d, \mathcal{H}_{it})$ . The direct method ignores propensities entirely, relying solely on the outcome model. This gives low variance but high bias if the outcome model is misspecified.

*Inverse probability weighting (IPW)* reweights observations by the ratio of target to logging policy probabilities:

$$\hat{V}^{\text{IPW}} = \frac{1}{n} \sum_{i,t} \frac{\pi_{\text{target}}(D_{it} | \mathcal{H}_{it})}{\pi_{\log}(D_{it} | \mathcal{H}_{it})} Y_{it}.$$

IPW is unbiased only when the logged assignment probabilities are correct and overlap holds for the actions used by the target policy. The challenge is variance: when target and logging policies differ substantially, importance weights become extreme.

*Doubly robust (DR) estimators* combine the direct method and IPW:

$$\hat{V}^{\text{DR}} = \frac{1}{n} \sum_{i,t} \left[ \sum_{d \in \mathcal{D}} \pi_{\text{target}}(d | \mathcal{H}_{it}) \hat{\mu}(d, \mathcal{H}_{it}) + \frac{\pi_{\text{target}}(D_{it} | \mathcal{H}_{it})}{\pi_{\log}(D_{it} | \mathcal{H}_{it})} (Y_{it} - \hat{\mu}(D_{it}, \mathcal{H}_{it})) \right].$$

DR estimators are consistent when overlap holds and either the outcome regression is correctly specified for the relevant conditional mean or the logged propensities are correct for the actions used by the target policy [Dudík et al., 2011, Hadad et al., 2021]. They typically have lower variance than IPW while being more robust than DM.

For practitioners, doubly robust estimators are often the safest default for off-policy evaluation when overlap is plausible and propensities are logged. They are widely implemented in software packages.

## Anytime-Valid Inference

When experiments are monitored continuously and stopped based on results, standard  $p$ -values and confidence intervals are invalid. A 95% confidence interval computed at a data-dependent stopping time does not have 95% coverage. The more we peek at the data, the more likely we are to declare a winner spuriously.

The solution is *anytime-valid inference*—methods that provide valid error guarantees at any stopping time, not just at fixed sample sizes. Three related frameworks address this problem.

*Confidence sequences* are sequences of confidence intervals  $(C_n)_{n \geq 1}$  such that the probability that the true parameter lies in all of them simultaneously is at least  $1 - \alpha$ :

$$\Pr(\vartheta \in C_n \text{ for all } n \geq 1) \geq 1 - \alpha.$$

Unlike fixed-sample confidence intervals, confidence sequences remain valid if we stop early, stop late, or peek repeatedly. The cost is that confidence sequences are wider than fixed-sample intervals at any given sample size—we pay a price for the flexibility to stop whenever we want. Howard et al. [2021] provide a comprehensive treatment with tight bounds.

*Sequential tests* control type I error under optional stopping using alpha-spending functions that allocate error probability across interim analyses. Group sequential methods partition the experiment into stages. Alpha-spending allows continuous monitoring. The Pocock and O'Brien-Fleming boundaries are classic examples. These methods are standard in clinical trials and increasingly adopted in tech industry A/B testing.

*E-values and e-processes* provide a unified framework for anytime-valid inference, developed rigorously by Grünwald et al. [2020]. An e-value  $E$  is a non-negative random variable with  $\mathbb{E}[E] \leq 1$  under the null. Large e-values are evidence against the null. Ville-type inequalities yield valid tests from e-processes, but you should not treat  $p = 1/E$  as a general identity without stating the specific calibration.

For marketing practitioners, the practical implication is clear: if your experiment involves adaptive assignment or optional stopping, standard  $p$ -values and confidence intervals are unreliable. Use confidence sequences or sequential testing methods instead. Software implementations are increasingly available, including in major A/B testing platforms.

## A Concrete Example: Inference Failure in Bandit Data

To make the identification problem concrete, consider a Thompson sampling bandit choosing between two creatives with true conversion rates  $p_A = 0.10$  and  $p_B = 0.08$ . Creative A is genuinely better by 2 percentage points. We write  $D_{it} \in \{0, 1\}$  for assignment, where  $D_{it} = 1$  denotes exposure to Creative A and  $D_{it} = 0$  denotes exposure to Creative B.

In a standard A/B test with 50/50 allocation, we would estimate  $\hat{p}_A - \hat{p}_B \approx 0.02$  with minimal bias, and a standard confidence interval would have correct coverage. In Thompson sampling, the algorithm learns that A is better and shifts traffic toward it. After 10,000 impressions, suppose A has received 7,500 and B has received 2,500.

The sample means  $\bar{Y}_A$  and  $\bar{Y}_B$  are individually unbiased for  $p_A$  and  $p_B$ . But reporting the difference  $\bar{Y}_A - \bar{Y}_B$  as our estimate of  $p_A - p_B$  encounters three problems:

First, Winner's curse: we selected A as the winner based on observed performance, so we are more likely to report an estimate that is biased upward.

Second, Temporal confounding: A's sample is enriched with high-converting periods (when the algorithm was shifting traffic toward it), while B's sample is enriched with low-converting periods.

Third, Invalid inference: the CLT-based confidence interval assumes fixed sample sizes, but  $N_A$  and  $N_B$  are random and correlated with outcomes. The true coverage of a nominal 95% interval may be 80% or less.

Weighting can address time and context confounding only if assignment is ignorable given the logged history and the assignment probabilities are recorded correctly. At each decision point  $t$ , the Thompson

sampling algorithm induces a probability of assigning treatment A. If we recorded these probabilities, we can use IPW:

$$\hat{p}_A^{\text{IPW}} = \frac{1}{N_A} \sum_{i,t:D_{it}=1} \frac{Y_{it}}{\pi_{\log}(1 \mid \mathcal{H}_{it})}.$$

For valid inference, we need methods designed for adaptive data, such as the doubly robust estimators with variance estimation from Hadad et al. [2021].

The challenge is that  $\pi_{\log}(1 \mid \mathcal{H}_{it})$  depends on the algorithm’s internal state, which may not be logged. Without these probabilities, identification fails. This is why logging propensities is essential for any system that may later be used for causal inference.

## Open Problems

These difficulties mirror the broader agenda on adaptive and sequential experimentation highlighted by Cinelli et al. [2025].

When both own treatment and neighbour treatments evolve adaptively, identification becomes extremely difficult. We lack established results for settings with simultaneous adaptivity and interference.

Bandits optimise for reward, not causal effect estimation in the sense of this book. Designing algorithms that balance reward maximisation with causal learning—so-called causal bandits—is an active research area. The goal is to learn which arm is best while also estimating by how much, with valid confidence intervals.

How do we detect when adaptivity has compromised identification? The design diagnostics from Chapter 17 assume fixed designs. We need new diagnostics that can flag when adaptive assignment has induced confounding—for example, by checking whether treatment assignment is predictable from lagged outcomes.

Adaptive systems optimise for short-run outcomes, but we often care about long-run effects—brand equity, customer lifetime value, habit formation. Estimating long-run effects when assignment is adapted to short-run signals remains largely unexplored.

Most work focuses on evaluating a fixed policy. Learning optimal policies from adaptive data—and quantifying uncertainty about the learned policy—remains challenging. The policy that appears optimal in sample may not be optimal out of sample, and standard confidence intervals do not account for the search over policy space.

## Practical Guidance

Given the challenges above, how should practitioners proceed? We offer the following workflow.

1. Document the adaptive systems in use. Before analysing data from an adaptive experiment, understand how the algorithm assigns treatments. Is it Thompson sampling? UCB? A contextual bandit with vanishing propensities? A proprietary pacing algorithm? The identification strategy depends on the answer.

2. Log assignment probabilities. If you control the adaptive system, ensure it logs the assignment probability at each decision point. Without these probabilities, causal inference is severely limited.
3. Reconstruct assignment probabilities when possible. If you have access to algorithm logs, extract the probabilities at each decision point. With these probabilities, doubly robust estimators become feasible.
4. Use designs that preserve overlap when propensities become extreme. Switchbacks can provide exogenous variation when within-period adaptivity is hard to model, but they still require limited carryover and stable conditions within switch windows. Phased rollouts provide variation in early stages before the algorithm has fully optimised. Holdout groups that receive fixed assignment provide a benchmark.
5. Use doubly robust estimation for off-policy evaluation. Prefer DR estimators to pure IPW or pure outcome modelling when overlap is plausible and propensities are logged.
6. Use anytime-valid inference under optional stopping. If the experiment involved continuous monitoring, report confidence sequences or sequential-testing results rather than fixed-sample confidence intervals.
7. Conduct sensitivity analysis. If propensity reconstruction is imperfect, vary assumptions about the assignment mechanism and report how estimates change. If estimates are sensitive, acknowledge the uncertainty.
8. Separate exploration from exploitation. Design experiments with a dedicated exploration phase where assignment is random or near-random. Use this phase for causal inference. Use the exploitation phase for reward maximisation.

These steps do not fully solve the problem of causal inference under adaptivity, but they make the challenges transparent and provide practitioners with the tools to assess the credibility of their conclusions.

## 20.5 Method Selection and Design Guidance

Practitioners now face a large menu of estimators: difference-in-differences, synthetic control, SDID, factor models, matrix completion, double machine learning, and their many variants. How should we choose? The method selection guidance in Chapter 18 (Section 18.14) provides practical heuristics, but a formal decision framework remains elusive. This section articulates open problems in method selection and proposes directions for progress.

### A Decision Heuristic

While no formal decision framework exists, we can offer a practical heuristic based on problem characteristics. Start with the estimand and the identification threats, then choose an estimator whose diagnostics are informative about those threats and whose inference matches the dependence structure.

*How many treated units do you have?* If you have one or a few treated units, synthetic control methods are the natural choice—they were designed precisely for this setting. If you have many treated units (say, more than 20), difference-in-differences and its heterogeneity-robust variants become feasible and often preferable.

*How long is your pre-treatment period?* Synthetic control typically benefits from a long pre-treatment period. Whether 10–20 periods suffice depends on the complexity of  $Y_{it}(0)$  and the stability of the treated-donor relationship. With shorter pre-periods, DiD may be more robust because it relies on restrictions on  $Y_{it}(0)$  rather than tight pre-treatment fit. Factor models and matrix completion also benefit from longer panels.

*Is treatment timing staggered or simultaneous?* If treatment occurs simultaneously for all treated units, classic DiD and SC are straightforward. If timing is staggered, you must use heterogeneity-robust DiD estimators (Chapter 4) or stacked approaches that avoid the negative weighting problem.

*What are your main threats to identification?* If parallel trends is the primary concern, synthetic control's pre-treatment fit provides a check. If time-varying confounders are the concern, factor models that allow for latent structure may be more appropriate. If you suspect interference, the methods from Chapter 11 become necessary.

*Do you have high-dimensional covariates?* If covariates are important for selection or for improving precision, augmented methods (ASDID, augmented SC) or double machine learning become attractive.

These questions do not yield a unique answer, and the boundaries between methods blur. In practice, we recommend running multiple methods and comparing results, using diagnostics to assess which assumptions are most plausible.

## SDID as Principled Unification

Before presenting worked examples, we should highlight that synthetic difference-in-differences (SDID) can be a useful default when both DiD and SC appear plausible, but it does not remove the need to justify identification. SDID combines unit weights (like SC) with time weights (like DiD), trading off between unit reweighting and time reweighting.

The intuition is straightforward. SC constructs a weighted average of control units to match the treated unit's pre-treatment trajectory. DiD uses all control units equally but relies on restrictions on  $Y_{it}(0)$ . SDID weights both units (to improve fit) and time periods (to down-weight periods where fit is poor). In common implementations, the resulting estimator can be written as a weighted least squares problem with unit and time fixed effects, which helps interpret the mechanics.

For practitioners, the implication is practical: when both DiD and SC are plausible, SDID can provide a useful reference estimate, and its weights can serve as diagnostics about which units and periods drive the estimate.

## A Worked Example

Consider a retailer testing a new loyalty programme. The programme launches in 5 DMAs in January 2023, with 45 DMAs as potential controls. We have monthly sales data from January 2020 through December 2023—36 pre-treatment periods and 12 post-treatment periods.

How should we approach method selection?

*Number of treated units:* Five is too few for DiD to be well-powered but enough for synthetic control to construct separate controls for each treated DMA or to pool treated units.

*Pre-treatment length:* 36 months is ample for SC to achieve good pre-treatment fit. Factor models are also feasible.

*Treatment timing:* Simultaneous, so we do not face staggered adoption complications.

*Main threats:* The treated DMAs may have been selected because they were growing faster—a parallel trends concern. They may also be in different regions with different economic conditions—a time-varying confounder concern.

Given these features, our approach would be:

1. Construct synthetic controls for each treated DMA and assess pre-treatment fit.
2. Run two-way fixed effects DiD as a benchmark, using event-study diagnostics to examine pre-trends.
3. Run SDID as the principled combination of both approaches.
4. If neither DiD nor SC achieves good diagnostics, consider a factor model that allows for DMA-specific loadings on regional economic factors.
5. Compare estimates across methods. If they agree, we gain confidence. If they disagree, investigate which assumptions are likely violated.

## When Diagnostics Conflict

The worked example above assumes a clean resolution: methods either agree or we can diagnose why they disagree. Reality is messier. What do we do when SC achieves excellent pre-treatment fit but DiD has tighter confidence intervals? Or when an event-study diagnostic shows no large pre-trend deviations for DiD, but SC weights concentrate on a single donor that seems substantively implausible?

Consider a scenario where our loyalty programme analysis yields: DiD yields  $\widehat{ATT} = 0.08$  (8% lift) with  $SE = 0.02$ , but an event-study diagnostic shows a slight upward pre-trend (the lead coefficient on  $D_{it}^{-1}$  equals 0.015). SC yields  $\widehat{ATT} = 0.05$  (5% lift) and excellent pre-treatment RMSPE, but 70% of weight falls on a single DMA that is geographically distant. SDID yields  $\widehat{ATT} = 0.06$  (6% lift) with an intermediate weight structure.

The methods disagree. DiD's pre-trend suggests its estimate may be biased upward. SC's weight concentration raises concerns about unobserved differences between the treated DMAs and the dominant donor. What should we report?

First, investigate the pre-trend. Interpreting a non-zero lead as bias requires an extrapolation assumption. For example, a lead coefficient of 0.015 when the estimated effect is 0.08 suggests the pre-period deviation could account for roughly 20% of the estimate under a linear continuation calculation. This is concerning but not decisive.

Second, investigate the donor. Why does SC concentrate weight on one DMA? If that DMA has similar retail characteristics (similar demographics, similar baseline sales levels), the concentration may be appropriate. If it is concentrated there simply to match a spurious pre-treatment pattern, the SC estimate is fragile.

Third, use SDID's intermediate estimate as a robustness check. The fact that SDID lies between DiD and SC (0.06 vs. 0.08 and 0.05) suggests neither extreme is obviously correct.

Fourth, report the range. Rather than reporting a single estimate, report "5–8% with central estimate 6%" and explain the disagreement. This is a robustness interval over a set of maintained assumptions. It is not an identification bound unless you also state a partial-identification argument that justifies it.

## Cross-Validation for Method Selection

The guidance above remains somewhat subjective: we investigate disagreement and make judgement calls. A more principled approach uses cross-validation on pre-treatment data to select methods based on predictive performance.

The procedure is as follows. Take the pre-treatment period and split it into a training period and a validation period. For instance, with 36 pre-treatment months, use months 1–30 for training and months 31–36 for validation. Fit each candidate method (DiD, SC, SDID, factor model) on the training period. Predict outcomes in the validation period. Select the method with the lowest mean squared prediction error (MSPE) in the validation period.

This approach has several advantages. It provides an objective criterion for method selection—the method that best predicts held-out pre-treatment data. It avoids cherry-picking based on which method gives the preferred answer. It naturally favours methods whose assumptions are most consistent with the data structure.

The approach also has limitations. Predictive performance in the pre-treatment period may not translate to causal validity in the post-treatment period. A method can predict  $Y_{it}(0)$  well in the pre-period and still fail causally if treatment timing is endogenous or if the treated–donor relationship breaks at treatment. The validation period may be too short to reliably estimate MSPE. Cross-validation also does not account for the uncertainty introduced by method selection itself.

Despite these limitations, pre-treatment holdout validation provides a useful diagnostic for counterfactual fit. Do not present it as a causal selection rule. Predictive performance in pre-treatment data does not identify post-treatment validity.

## Multiverse Analysis and Specification Curves

Rather than selecting a single method and reporting a single estimate, we can map the space of reasonable analytic choices and report how estimates vary across this space. This approach goes by several names. *Multiverse analysis* examines how results change across the space of data processing and analytic decisions. *Specification curve analysis* plots estimates from all reasonable specifications on a single figure.

The multiverse for a causal analysis might include which control units to include, how to construct the outcome variable (levels vs. logs, seasonally adjusted or not), which pre-treatment periods to use for fit, whether to include covariates, and which method to apply. Each combination yields an estimate, and the full set of estimates forms the specification curve.

A specification curve that is tightly clustered around a central estimate suggests robustness—conclusions do not depend on arbitrary analytic choices. A curve that spans zero or includes estimates of both signs suggests fragility—the conclusion depends critically on choices that the data cannot adjudicate.

Reporting specification curves is more honest than reporting a single estimate from a single method. It acknowledges that researchers make choices and that different choices can yield different conclusions. The challenge is defining the space of “reasonable” specifications—including unreasonable specifications inflates variance, while excluding reasonable ones hides fragility.

Specification curves increase transparency, but they also create a multiplicity problem. Pair them with pre-specified primary estimands and appropriate uncertainty quantification (see Chapter 16).

## Sensitivity Analysis and Austen Plots

Beyond specification curves, formal sensitivity analysis quantifies how much unobserved confounding would be required to overturn a conclusion. Section 20.8 introduced sensitivity frameworks. Here we emphasise their role in method selection.

*Omitted-variable-bias plots* (often called Austen plots) provide a visual summary of sensitivity to unobserved confounding [Cinelli and Hazlett, 2020]. The horizontal axis shows how strongly an unobserved confounder would need to correlate with treatment. The vertical axis shows how strongly it would need to correlate with outcomes. Contours show the resulting bias. The point where the contour crosses zero indicates the minimum confounding strength required to overturn the conclusion.

These plots are available through the `sensemakr` package in R and similar implementations elsewhere. They provide two benefits for method selection. First, they quantify fragility: an effect that would be overturned by weak confounding is less credible than one that would require implausibly strong confounding. Second, they allow comparison across methods. Omitted-variable-bias style sensitivity plots are most directly justified for regression-type estimators. For DiD or SC-style estimators they are at best heuristic unless you state the mapping clearly.

For practitioners, we recommend reporting Austen plots (or equivalent sensitivity metrics) alongside point estimates. This moves sensitivity analysis from a vague “we conducted robustness checks” to a quantitative statement about how much unobserved confounding would change the conclusion.

## Triangulating Panel Methods with MMM

A key open problem is reconciling panel causal methods with media mix modelling. The two approaches estimate different quantities: panel methods estimate ATT or ATE for specific interventions, while MMM estimates marginal effects of continuous spend variables. We must understand when they should agree and what disagreement implies.

*When should they agree?* Relating discrete-campaign effects to MMM marginal effects requires an explicit dose-response model and an aggregation rule over the treatment change. You must state whether you use a local-derivative approximation at a baseline spend level or a finite-change approximation over the relevant range. Agreement also requires that MMM’s functional form (adstock, saturation) correctly captures how effects aggregate.

*When should they disagree?* Disagreement arises when (1) the panel intervention is outside the range of MMM’s training data (extrapolation), (2) the panel measures short-run effects while MMM captures long-run equilibrium, (3) one or both identification strategies fail, or (4) the estimands genuinely differ because of heterogeneity or nonlinearity.

*What does disagreement tell us?* If panel methods suggest a large effect and MMM suggests a small effect, the MMM may be underestimating due to attenuation bias from measurement error or confounding from correlated channels. If panel methods suggest a small effect and MMM suggests a large effect, the panel may be measuring a local effect that does not generalise, or the MMM may be capturing spurious correlation.

*Calibration.* MMM estimates can be calibrated using experimental lift from geo-experiments or holdout tests, as discussed in Section 18.5. This bridges the two approaches but requires that experimental effects generalise to the MMM’s continuous-dose framework—an assumption that is rarely tested. When calibration changes MMM estimates substantially, it signals that the MMM’s identifying variation was problematic.

No formal framework exists for reconciling disagreement, but we can offer guidance: investigate the source of disagreement before choosing which estimate to report. If the disagreement can be traced to a specific assumption that one method makes and the other does not, that assumption becomes the focus of sensitivity analysis.

## Toward Formal Decision Frameworks

A formal method selection framework would map problem characteristics to recommended methods. The inputs would include the number of treated and control units, the length of pre-treatment and post-treatment periods, treatment timing (simultaneous, staggered, continuous), outcome characteristics (continuous, count, binary, zero-inflated), covariate availability and dimensionality, and suspected threats (parallel trends violations, interference, structural instability).

The output would be a ranked list of methods with diagnostics to run and sensitivity analyses to report. Such a framework requires understanding method performance across the input space—knowledge we do not yet have.

Building this knowledge requires systematic benchmarking. Realistic simulations would generate panels with known treatment effects under realistic violations—parallel trends deviations, factor structure, interference, measurement error—and compare methods on bias, RMSE, and coverage. Semi-synthetic benchmarks would use real marketing data with synthetic treatment effects, preserving realistic covariate structure while providing ground truth.

Held-out experiments provide the most direct benchmark. When experiments are available, they can serve as ground truth for observational methods applied to the same data. This is rare because it requires experimental access and willingness to share results. Community benchmarks—shared benchmark suites that allow method developers to compare performance—would accelerate progress.

## Close Comparison Groups: Strategy-Robust Inference

The methods discussed above—DiD, synthetic control, factor models—each rely on distinct identifying assumptions. Parallel trends differs from the monotonicity required by change-in-changes, which differs from lagged outcome unconfoundedness. In practice, we rarely know which assumption holds. Recent work by Callaway et al. [2025] proposes a different approach: instead of committing to an identification strategy, search for comparison groups where the choice does not matter.

The target is a time-specific treated-group effect (for example,  $\text{ATT}_{t^*}$ ) defined on untreated potential outcomes in a once-treated design. The core insight is elegantly simple. Define a *close comparison group* as one whose pre-treatment outcome distribution matches the treated group exactly. Let clusters be indexed by  $c = 1, \dots, G$ , where  $c = 1$  denotes the treated cluster and  $c \in \bar{\mathcal{C}} \subseteq \{2, \dots, G\}$  denotes candidate comparison clusters. Let  $\mathcal{C}_c$  denote the set of units in cluster  $c$ . Then

$$\mathcal{C}^* = \left\{ c \in \bar{\mathcal{C}} : (Y_{i,t^*-1} \mid i \in \mathcal{C}_1) \stackrel{d}{=} (Y_{i,t^*-1} \mid i \in \mathcal{C}_c) \right\},$$

where  $\bar{\mathcal{C}}$  denotes the set of never-treated or not-yet-treated comparison clusters and  $t^*$  is the treated period. If such clusters exist, define the cluster-period mean  $\bar{Y}_{ct} = |\mathcal{C}_c|^{-1} \sum_{i \in \mathcal{C}_c} Y_{it}$ . Then a simple level comparison

$$\widehat{\text{ATT}}_{t^*} = \bar{Y}_{1t^*} - \sum_{c \in \mathcal{C}^*} w_c \bar{Y}_{ct^*}$$

can align the identifying restrictions implied by several strategies when pre-treatment distributions match exactly. This is a knife-edge condition and practical implementation is approximate. The goal is to reduce sensitivity to which identifying strategy you adopt, not to avoid stating assumptions.

Why does this work? Each strategy can be written in the form  $\text{ATT}_{t^*} = \mathbb{E}[Y_{it^*} \mid i \in \mathcal{C}_1] - \mathbb{E}[Y_{it^*} \mid i \in \mathcal{C}_c]$  – adjustment term, where the adjustment term differs across strategies and is computed from pre-treatment data. DiD subtracts the difference in pre-treatment means. Change-in-changes applies a nonlinear transformation based on quantile functions. Lagged outcome unconfoundedness conditions on pre-treatment outcomes. When the pre-treatment distributions match exactly, all three adjustment terms equal zero simultaneously. The strategies agree because there is nothing to adjust.

The approach extends to interactive fixed effects and latent unconfoundedness models, but with a stricter requirement. Distributional matching over a single pre-treatment period suffices for DiD, change-in-changes, and lagged outcome unconfoundedness. For factor models, we need matching over multiple periods, at least  $R+1$  pre-treatment periods when there are  $R$  factors. Mean-matching (rather than distributional matching) suffices for DiD and factor models but not for the nonlinear strategies.

A complementary robustness path uses *time homogeneity*. If outcomes for comparison clusters do not change over time—that is,  $\mathbb{E}[Y_{it} \mid i \in \mathcal{C}_c] = \mathbb{E}[Y_{i,t^*-1} \mid i \in \mathcal{C}_c]$  for all  $c \in \bar{\mathcal{C}}$  and  $t \geq t^*$ —then a before–after comparison  $\mathbb{E}[Y_{it^*} - Y_{i,t^*-1} \mid i \in \mathcal{C}_1]$  identifies  $\text{ATT}_{t^*}$  under DiD and interactive fixed effects. Combining close comparison groups with time homogeneity provides “multiply robust” inference that remains valid if either condition holds.

The practical limitation is stringent data requirements. Close comparison groups may not exist. Distributional matching requires large within-group samples to assess reliably. When no comparison group satisfies the criteria, the method returns no estimate—a feature the authors argue is preferable to reporting a fragile point estimate. The approach asks more of the research design but delivers more credible conclusions when the design delivers.

How does this relate to synthetic control? Synthetic control constructs weighted averages of donor units to match pre-treatment characteristics. The weights can combine dissimilar units. The close comparison group approach requires finding actual groups that match—no weighting. This is more demanding but more transparent: either a matching group exists or it does not. The two approaches are complements, not substitutes. When close comparison groups exist, they provide stronger identification. When they do not, synthetic control remains the fallback.

### When to Use Close Comparison Groups

Consider this approach when you have multiple potential comparison clusters with microdata (states, DMAs, stores) and can check whether pre-treatment distributions match. Start by plotting outcome distributions for each candidate cluster against the treated cluster in the pre-treatment period. Clusters with overlapping distributions are candidates for  $\mathcal{C}^*$ . Formal two-sample checks (for example, Kolmogorov-Smirnov statistics or quantile comparisons) can assess distributional equivalence, though power may be limited with small samples.

The payoff is robustness: if matching groups exist, you avoid committing to parallel trends, change-in-changes, or lagged outcome assumptions. The cost is that the set  $\mathcal{C}^*$  may be empty, in which case you must fall back to standard methods with their attendant assumptions. In marketing applications with many DMAs or stores, distributional matching is often feasible. In applications with few comparison units, it rarely is.

## When to Combine Methods

Rather than selecting a single method, we may combine methods. Several approaches exist.

*Model averaging.* Weight estimates from multiple methods based on diagnostic performance or prior beliefs. Bayesian model averaging provides a formal framework but requires specifying priors over methods—rarely straightforward in causal inference where “methods” correspond to different identifying assumptions, not different models of the same DGP.

*Ensemble estimation.* Use multiple methods as inputs to a meta-estimator. Stacking, for example, learns weights that minimise cross-validated prediction error. This can reduce variance but may not reduce bias if all methods are biased in the same direction.

*Robustness intervals from disagreement.* If methods disagree, report the range of estimates as a robustness interval over a set of maintained assumptions. If DiD yields 5% and SC yields 8%, reporting “5–8%” acknowledges that method choice matters without arbitrarily preferring one.

The challenge is that combining methods does not resolve the fundamental question of which identifying assumptions are more plausible. Averaging across methods implicitly assumes that biases cancel, which may be false. Reporting robustness intervals acknowledges uncertainty but may be too conservative for decision-making.

## Practical Guidance

Given the open problems above, how should practitioners proceed? We offer the following workflow.

1. Characterise the problem. Before touching data, assess the number of treated units, pre-treatment length, treatment timing, main threats, and covariate availability. These features narrow the menu of appropriate methods.
2. Consider SDID as a default reference estimate when both DiD and SC are plausible candidates. It does not remove the need to justify identification.
3. Use held-out pre-treatment periods as a diagnostic for counterfactual fit. Do not treat MSPE as a causal selection rule.
4. Run diagnostics to assess assumption plausibility. Use pre-trend diagnostics for DiD, pre-treatment fit for SC, and placebo checks for both. Diagnostics are imperfect but informative.
5. Conduct multiverse analysis. Vary the control group, outcome definition, covariates, and method. Report the specification curve, not just the preferred estimate.
6. Quantify sensitivity. Use omitted-variable-bias plots or equivalent metrics to show how much unobserved confounding would be required to overturn the conclusion.
7. Investigate disagreement. If methods disagree, do not average or choose arbitrarily. Trace the disagreement to specific assumptions and focus sensitivity analysis there.
8. Report honestly. State which methods were run, what diagnostics suggest, how sensitive estimates are to specification, and what the range of estimates implies for the conclusion.
9. Pre-register when possible. If the analysis plan can be specified before seeing outcomes, pre-registration limits specification searching. Leave room for exploratory analysis and label it clearly.

These steps do not provide a definitive answer to “which method should I use?” but they make the method selection process transparent and defensible.

## 20.6 Robust and Distribution-Free Inference

Chapter 16 developed robust inference methods—randomisation inference, conformal prediction, and bootstrap procedures—that avoid strong distributional assumptions. These methods can deliver valid uncertainty quantification without strong parametric assumptions, but only when their design and exchangeability conditions match the panel’s dependence structure. Panels pose challenges that standard implementations do not address. This section maps the frontier where theory meets practice.

### Randomisation and Permutation Inference

Randomisation inference delivers exact finite-sample  $p$ -values under the sharp null that treatment has no effect on any unit–time potential outcome. It does so by comparing an observed statistic to its distribution over treatment assignments. With  $N$  units  $i = 1, \dots, N$  and  $T$  periods  $t = 1, \dots, T$ , the number of possible assignments grows rapidly. The appeal is conceptual purity: we make no distributional assumptions beyond the assignment mechanism itself. In panels, however, we face three obstacles that limit this purity.

The first is computational. Even under staggered adoption, exact enumeration becomes infeasible for panels larger than a few dozen units. Monte Carlo draws introduce simulation error, and unknown or mis-specified assignment mechanisms remove the finite-sample guarantee entirely.

The second is structural. When units cluster—stores within DMAs, users within social networks—permutations must respect cluster boundaries. Permuting at the cluster level reduces the number of *effective* assignments and limits  $p$ -value granularity.

The third is identification. Randomisation inference requires knowledge of the assignment mechanism. When treatment is observational or adaptive, as in Sections 20.4 and Chapter 17, we rarely know the true mechanism. The theoretical guarantee of exactness then rests on a foundation we cannot verify.

### Mosaic Inference: Local Exchangeability in Panels

Recent work by Spector et al. [2025] offers a way forward. The key insight is that panels have a *mosaic structure*: even when global exchangeability fails, local exchangeability may hold within certain blocks. Consider a staggered adoption design. Pre-treatment periods for eventually-treated units are exchangeable with pre-treatment periods for never-treated units, even though post-treatment periods are not. We can permute within these locally exchangeable blocks while respecting the non-exchangeability across blocks.

This kind of exchangeability is credible primarily under explicit randomisation (or a justified as-if random timing model). In observational panels, treat it as an assumption to be defended, not a default fact.

Mosaic inference can provide finite-sample guarantees only for the blocks that are truly exchangeable under the assignment mechanism. Incorrect block specification yields invalid  $p$ -values.

This mosaic approach can recover valid permutation inference for panels without requiring global exchangeability. The analyst identifies which observations are plausible candidates for permutation, typically based on treatment timing and covariate strata, and constructs tests that exploit this local structure.

The practical implication is significant. Mosaic methods can scale to large panels while preserving finite-sample validity where the structure permits. The cost is that we must specify which blocks are locally exchangeable, which is a design judgement that requires careful thought about the treatment assignment process.

## Conformal Prediction for Counterfactuals

Conformal prediction provides distribution-free prediction intervals with finite-sample coverage guarantees. Extending these ideas to causal inference is attractive but faces two obstacles: panel dependence and counterfactual estimation.

For synthetic controls, Chernozhukov et al. [2021] provide exact inference by permuting residuals, exploiting the estimator's structure rather than raw outcome exchangeability.

For general heteroskedastic panels, **Conformalized Quantile Regression (CQR)** [Romano et al., 2019] offers a robust alternative. CQR offers distribution-free coverage under an exchangeability condition. In panels, you need a splitting scheme that respects dependence, for example unit-level splits when units are independent clusters. CQR estimates conditional quantiles  $Q_\alpha(Y_{it} | X_{it})$  and then calibrates these estimates using a holdout set.

## Bootstrap Methods for Dependent Data

Bootstrap methods resample data to approximate sampling distributions. Their flexibility makes them the workhorse of applied inference, but panels require care in how we resample.

Block bootstrap preserves temporal dependence by resampling contiguous time blocks rather than individual observations. The block length trades off bias from short blocks (which break dependence) against variance from long blocks (which yield few resamples). Optimal block length depends on the dependence structure, which we rarely know. In practice, we try several lengths and check sensitivity.

Cluster bootstrap preserves within-cluster dependence by resampling entire clusters. With many clusters, this works well. With few clusters—common in marketing where we might have 20 DMAs or 15 retail chains—the bootstrap distribution is poorly approximated. The problem is not sample size per se but the number of independent units we resample from.

Wild cluster bootstrap addresses the few-clusters problem by multiplying residuals by random weights rather than resampling observations. This expands the effective permutation space while respecting cluster structure. Implementation requires specifying the weight distribution (Rademacher weights are standard) and the residual type (HC2 or HC3 adjustments help with leverage). Wild cluster bootstrap is often a sensible

default when clusters are limited (roughly 10–50), but its performance depends on the estimator and the assignment level. Treat it as a sensitivity check rather than a universal fix.

When data exhibit both temporal and cross-sectional dependence, no single bootstrap is fully satisfactory. Hybrid approaches that resample in both dimensions are computationally demanding and theoretically underdeveloped. Recent work by Almeida et al. [2025] provides variance estimators for causal panel estimators that can guide bootstrap implementation, but a unified framework remains elusive.

## Practical Guidance

Given the limitations above, how should practitioners proceed? We offer the following guidance, recognising that all choices involve trade-offs.

In all cases, state the independent sampling unit and the dependence structure explicitly. If clusters  $c = 1, \dots, G$  are the independent units, the effective sample size is typically  $G$ , and uncertainty quantification should use variance estimators consistent with cluster dependence.

For analytical standard errors, cluster-robust methods remain the baseline. Cluster at the level of treatment assignment—if stores are randomised within DMAs, cluster at the store level, not the DMA level. When in doubt, cluster conservatively (at a higher level), accepting wider confidence intervals as the price of validity.

For few-cluster settings (fewer than 20 clusters), use wild cluster bootstrap with Rademacher weights. Report both analytical cluster-robust SEs and bootstrap confidence intervals. If they diverge substantially, investigate why—the divergence itself is diagnostic.

For temporal dependence, consider block bootstrap with multiple block lengths. If results are sensitive to block length, report the range and acknowledge the uncertainty. HAC standard errors (Newey-West) address serial correlation in a single time series or aggregated series. For panels with cross-sectional dependence you must specify what is assumed independent and what is not.

For permutation tests, use mosaic inference when you can identify locally exchangeable blocks. Pre-treatment periods across treatment groups are typically a safe choice. With fewer than 50 effective permutations, report exact  $p$ -values to two decimal places and acknowledge the granularity.

For conformal inference, the Chernozhukov et al. [2021] approach works for synthetic control settings. For other panel designs, conformal methods remain experimental. Use them to supplement, not replace, other inference methods.

## Research Directions

Three developments would advance the field. First, we need implementations of mosaic inference that scale to large panels and provide guidance on block specification. The theory exists. The software does not. Second, we need conformal methods for heterogeneous treatment effects in panels—methods that provide conditional

coverage under dependence. Third, we need bootstrap procedures that handle both temporal and cross-sectional dependence with clear guidance on when each dominates.

Until these advances become standard, practitioners should use multiple inference methods and report sensitivity to assumptions. Agreement across methods is reassuring. Disagreement is informative—it tells us that conclusions depend on modelling choices we cannot resolve with the data at hand.

## 20.7 Continuous Treatments and Structural Response

Marketing treatments are often continuous: ad spend, price discounts, promotion intensity, and content frequency all vary in degree rather than presence or absence. Chapter 14 developed dose–response methods for these settings, integrating with double machine learning (Chapter 12) and regularisation (Chapter 13). Yet continuous treatments in panels raise challenges that go beyond the static case. This section maps the frontier.

### Recent Methodological Advances

Recent work has expanded panel designs to accommodate non-binary intensities, though identification remains driven by design assumptions rather than functional-form flexibility. Callaway and Sant'Anna [2021] develop DiD estimators for multiple time periods under parallel-trends style conditions. Adapting DiD logic to continuous doses requires defining dose–response objects under a conditional parallel-trends restriction [Callaway and Sant'Anna, 2021] and remains an active area.

For heterogeneous effects, generalised random forests (GRF) [Athey et al., 2019] are widely used in practice. By modifying the splitting rule to maximise heterogeneity in the treatment coefficient (rather than outcome variance), GRF recovers individual-level partial effects on the conditional dose–response function. With  $\mu(d, x) = \mathbb{E}[Y_{it}(d) | X_{it} = x]$ , GRF targets conditional marginal effects such as  $\partial\mu(d, x)/\partial d$ , evaluated at a stated reference dose, for example the observed  $D_{it}$  or a policy-relevant baseline. This allows marketers to target pricing or spend based on local elasticity.

### Stochastic Shift Interventions

While the academic literature focuses on the full dose–response curve  $\mu(d) = \mathbb{E}[Y_{it}(d)]$ , marketing decisions often involve stochastic shifts. Instead of asking "What if everyone set Price = \$10?" (which is unrealistic for premium and budget items alike), we ask "What if we increased every item's price by 10%?".

Díaz and Hejazi [2020] formalise this as a *modified treatment policy* or shift intervention. In panels, you should clarify whether the intervention shifts dose period-by-period or shifts an aggregated dose over a window. If dynamics matter, the primitive object is path-dependent potential outcomes  $Y_{it}(d_i^t)$ . A shift intervention then corresponds to a shifted path, for example  $d_{i,\delta}^t$  with  $d_{is,\delta} = d_{is} + \delta$  for each  $s \leq t$ , and the corresponding estimand is  $\mathbb{E}[Y_{it}(d_{i,\delta}^t)]$ .

## Dynamic Dosing and Anticipation

In practice, treatment intensity varies over time within units. Ad spend fluctuates weekly. Prices change daily. Promotion intensity varies by season. The dose at time  $t$  may affect outcomes at  $t + 1, t + 2, \dots$ , creating distributed lag structures that complicate identification.

This means the primitive object is path-dependent potential outcomes  $Y_{it}(\underline{d}_i^t)$ , not a static  $Y_{it}(d)$ .

When future doses are predictable, units may respond in advance. Consumers stockpile before announced price increases. Advertisers front-load spend before budget exhaustion. Such anticipation violates the no-anticipation assumptions used in event studies from Chapter 5 and complicates dose-response interpretation.

Leads can diagnose anticipation but do not by themselves repair identification. Restricting attention to variation that is plausibly unanticipated requires a defensible assignment-mechanism argument, for example quasi-random algorithmic updates.

The Average Causal Response framework from Chapter 14 provides a foundation for dynamic dosing by defining sensitivity over lags. The challenge is that estimation requires assumptions about the decay structure—how does the effect of a dose at  $t$  persist to  $t + k$ ?—and these assumptions are difficult to verify without experimental variation in timing.

## Endogenous Dosing Rules

Platforms and firms set doses based on predicted responses, creating endogeneity that threatens identification. Bid optimisers, pacing algorithms, and dynamic pricing systems adjust dose based on predicted outcomes. High-performing units receive higher doses, but this reflects selection rather than causation.

Consider ad spend allocation. A retailer allocates budget across stores to maximise total revenue. Stores with higher predicted conversion rates receive more spend. The observed correlation between spend and revenue reflects both the causal effect of spend and the selection of high-potential stores into high-spend buckets. Naive regression estimates the selection effect, not the causal effect.

Several strategies can restore identification. Instrumental variables that shift dose without affecting outcomes directly are the textbook solution, but valid instruments for continuous marketing treatments are rare. More promising are algorithmic discontinuities: bid thresholds where small changes in predicted value cause discrete jumps in spend, budget exhaustion points that create exogenous variation in dose timing, and A/B tests embedded within optimisation systems.

When instruments are unavailable, we can sometimes exploit the timing of dose changes. If the dosing algorithm updates predictions weekly but outcomes respond immediately, the lag between prediction and dose creates variation that is plausibly exogenous to short-run outcome fluctuations. This identification strategy is fragile. It requires a maintained assumption that the prediction error that drives dose changes is conditionally mean-independent of  $Y_{it}(\underline{d}_i^t)$  given the information set used by the algorithm.

## Equilibrium and Structural Response

Partial-equilibrium effects and general-equilibrium effects are different estimands. Reduced-form dose–response estimates generally target the former unless the design induces market-wide variation.

When one firm increases ad spend, competitors may respond by increasing their own spend. The full effect of the initial increase includes both the direct effect on the firm’s outcomes and the indirect effect through competitive response. If all firms increase spend together, the marginal effect may decline as competition intensifies—a congestion effect that partial-equilibrium estimates miss.

Large interventions may alter market structure itself. A platform-wide change to ad auction rules affects not just individual advertisers but the equilibrium bids, prices, and entry decisions of all participants. These general-equilibrium effects lie beyond the scope of reduced-form dose–response estimation.

Bridging partial-equilibrium estimates to policy-relevant equilibrium effects requires either structural models that specify how agents respond to changes, or natural experiments where multiple firms were treated simultaneously. The first approach is often infeasible in marketing settings where agent behaviour is complex and heterogeneous. The second approach is rare but invaluable when available—industry-wide regulatory changes or platform policy shifts provide the cleanest identification of equilibrium effects.

## Overlap and Support Diagnostics

Continuous treatments require overlap: for each covariate value, a range of doses must have positive probability. In practice, this often fails. Some units always receive high doses because they are valuable customers. Others always receive low doses because they are unprofitable segments. Extreme regions of the dose distribution may be essentially unobserved for large parts of the population.

We can diagnose overlap problems using the generalised propensity score

$$r(d | X_{it}, \alpha_i, \lambda_t) = f_{D|X,\alpha,\lambda}(d | X_{it}, \alpha_i, \lambda_t),$$

the conditional density of treatment given covariates. When this density is near zero for certain dose–covariate combinations, we lack support for estimating effects in those regions. Unlike binary treatments, where you can inspect estimated propensities for extreme values, continuous treatments require examining support across the conditional dose distribution.

Practical diagnostics include plotting the distribution of doses within covariate strata to identify gaps in support, computing the effective sample size at different dose levels to assess estimation precision, and using inverse propensity weights to identify regions where weights explode. When support is limited, we should either trim the dose range to regions with adequate overlap or acknowledge that our estimates apply only to the supported region.

## Functional Form and Sensitivity

Dose-response curves require functional form choices: linear, polynomial, spline, or fully nonparametric. These choices matter. A linear specification assumes constant marginal effects. A polynomial allows diminishing returns but imposes smoothness. Splines are more flexible but require choosing knot locations. Nonparametric methods avoid functional form assumptions but require more data and may have poor performance in the tails.

We should routinely report sensitivity to functional form. Does the conclusion that “more spend increases revenue” hold under both linear and spline specifications? Does the estimated optimal dose change substantially across specifications? If conclusions are robust to functional form, we have stronger evidence. If they are sensitive, we should report the range of estimates and acknowledge the uncertainty.

## Practical Guidance

For marketing practitioners working with continuous treatments in panels, we offer the following workflow. First, document the dose assignment mechanism. Is dose set by algorithm, by managers, or by policy? Understanding the mechanism guides the identification strategy.

Second, assess overlap using generalised propensity score diagnostics. Plot dose distributions within covariate strata. Identify regions where support is thin and either trim or acknowledge the limitation.

Third, consider instruments or quasi-experimental variation. Budget exhaustion, algorithmic discontinuities, and timing variation may provide identification where simple conditioning fails.

Fourth, report sensitivity to functional form. Show results under multiple specifications. If conclusions differ, explain why and report the range.

Fifth, acknowledge the partial-equilibrium nature of the estimates. If the policy intervention would affect many units simultaneously, note that equilibrium effects may differ from the estimated partial-equilibrium response.

This workflow does not solve the fundamental identification challenges of continuous treatments in panels, but it makes the challenges transparent and provides readers with the information needed to assess the credibility of the results.

## 20.8 Partial Identification and Sensitivity

Point identification requires assumptions that fix the estimand as a unique functional of the observed data distribution. Parallel trends may not hold exactly. SUTVA may be violated by spillovers. Outcomes may be measured with error. When these assumptions are uncertain, we should report identified sets and sensitivity intervals rather than only standard errors around a point estimate. This section develops practical tools for partial identification in marketing panels.

### Why Partial Identification Matters

Point estimates with narrow confidence intervals can create false confidence when identification assumptions are uncertain. We might report that a loyalty programme increased retention by 3.2 percentage points ( $p < 0.01$ ), but this precision rests on parallel trends holding exactly—an assumption we cannot verify. If trends deviated by even a small amount, the true effect could be 2% or 4% or zero.

Partial identification makes this uncertainty explicit. Rather than assuming parallel trends hold exactly, we assume they hold within  $\pm \delta$  per period on the chosen outcome scale. You should state explicitly whether  $Y_{it}$  is in levels, logs, or standard deviations. We then report the range of effects consistent with that assumption.

### Bounded Parallel Trends: The Rambachan-Roth Framework

The most important recent advance in partial identification for panels is the work of Rambachan and Roth [2023], who develop a practical framework for “honest” difference-in-differences inference under bounded parallel trends violations.

The core idea is simple. We assume that the deviation from parallel trends in post-treatment periods is no larger than some multiple  $\bar{M}$  of the maximum deviation observed in pre-treatment periods. Define event time  $k$  relative to treatment, where  $k < 0$  are pre-treatment periods. Let  $\delta_k$  denote the deviation from the parallel-trends restriction on untreated potential outcomes  $Y_{it}(0)$  at event time  $k$ . Then we assume:

$$|\delta_k| \leq \bar{M} \cdot \max_{k' < 0} |\delta_{k'}|$$

When  $\bar{M} = 0$ , we recover the standard parallel trends assumption. When  $\bar{M} > 0$ , we allow for violations whose magnitude is bounded by what we observe pre-treatment.

This approach has two virtues. First, the scale is anchored to observed pre-period deviations, but the sensitivity parameter  $\bar{M}$  remains a maintained assumption chosen by the analyst. Second, the resulting procedures deliver confidence sets that cover the identified set with the stated probability under the maintained violation class.

The `HonestDiD` R package implements this framework, providing ready-to-use tools for constructing bounds in DiD and event-study settings. Practitioners can specify  $\bar{M}$ , compute the implied confidence set, and report how conclusions change as  $\bar{M}$  varies. This is exactly the “breakdown analysis” we want: at what violation magnitude does the conclusion change?

## Calibrating Violation Magnitudes

The hardest part of partial identification is choosing the violation magnitude  $\delta$  or  $\bar{M}$ . Several strategies help.

Pre-treatment calibration uses observed pre-trends to bound plausible post-treatment deviations. If pre-treatment trends differ by at most 0.5% per period, it is plausible (though not guaranteed) that post-treatment violations are similar. The Rambachan-Roth  $\bar{M}$  parameter formalises this:  $\bar{M} = 1$  means we allow violations as large as the worst pre-treatment deviation,  $\bar{M} = 2$  allows violations twice as large, and so on.

Domain-knowledge calibration uses external information about the mechanism. If we know that regional economic shocks affect treated and control regions differently, we can estimate the magnitude of these shocks from auxiliary data and use this to bound parallel-trends violations.

Benchmark calibration compares the required violation to observable confounders. If we would need parallel trends to deviate by 10% per period to overturn the result, but all observable confounders produce deviations of less than 1%, the result is robust in a meaningful sense. This logic underlies sensitivity analysis frameworks like those in Chapter 15.

More generally, it helps to report robustness in units that can be compared across studies. Rather than stating that a result is “robust”, report the smallest violation magnitude that would overturn the sign of the estimate, and relate it to observable benchmarks or substantive mechanisms. Cinelli et al. [2025] argue that developing and standardising such robustness metrics is a key step for making sensitivity analysis routine rather than exceptional.

## Bounding Other Violations

Different assumption violations call for different bounding approaches. Spillovers typically change the estimand itself. If you want bounds, specify (i) the exposure mapping, (ii) what is being bounded, for example a direct effect holding exposure fixed versus a policy effect under a given saturation, and (iii) a bound on how outcomes change with exposure. Without these, a generic “ $\pm\gamma$ ” statement is not well-defined.

In simple linear models with classical measurement error in the regressor, attenuation can bias effects toward zero. For panel estimators and non-classical measurement error, the direction can differ and requires an explicit measurement model. If you introduce a latent true treatment  $D^*$ , state how it relates to  $D_{it}$  in panels and what external validation data identify or bound a reliability ratio.

For unmeasured confounding, Rosenbaum bounds (Definition 15.6 and Proposition 15.2 in Chapter 15) assess how strong a hidden confounder would need to be to overturn the result. The sensitivity parameter  $I$

quantifies the required confounding strength. When  $\Gamma$  must be large to change conclusions, we have evidence of robustness.

### Combining Multiple Violations

In practice, multiple assumptions may be violated simultaneously. Parallel trends and SUTVA may both be uncertain, measurement error may be present, and some confounding may remain. We must then account for several types of violations at once.

Joint bounds are necessarily wider than bounds for any single violation. Their exact form depends on whether you assume worst-case alignment of violations or impose structure on how violations interact. Optimising over multiple violation parameters to find the implied range of the estimand can be computationally intensive but feasible for moderate-dimensional problems.

The practical challenge is specifying plausible magnitudes for multiple violations. There is no consensus on how to elicit or report these magnitudes, but transparency helps. We should state each assumed bound explicitly (“we allow parallel trends to deviate by up to 1% per period and spillovers to affect outcomes by up to 0.5%”) rather than hiding assumptions in sensitivity tables.

### Reporting Standards

Transparent reporting of partial identification requires several elements. We should state the assumed bounds on parallel-trends deviations, spillover magnitudes, or confounding strength. Vague statements such as “robust to violations” should be avoided in favour of specific magnitudes.

We should report the point estimate under maintained assumptions alongside bounds under explicit violations, so readers see both the best guess and the range of uncertainty. A useful format is:

Under exact parallel trends, the estimated effect is 3.2% (95% CI: 2.1–4.3%). If parallel trends deviate by up to  $\bar{M} = 1$  times the maximum pre-treatment deviation, the 95% robust confidence set is 1.8–4.8%. The conclusion that the effect is positive is robust to  $\bar{M} \leq 2.3$ .

In partially identified settings, the appropriate inferential object is a confidence set that covers the identified set with a stated probability under the maintained violation class.

Breakdown analysis should report the violation magnitude at which conclusions change and relate this to observable features where possible. Calibrating allowable deviations from pre-treatment gaps or from observable confounder effects grounds the bounds in evidence.

## Current Tools

Partial identification is now practical for DiD and event-study settings. The `HonestDiD` R package implements Rambachan-Roth bounds with pre-trend calibration. The `sensemakr` package [Cinelli and Hazlett, 2020] provides sensitivity analysis for regression designs, generating "Austen plots" that visualise how much confounding (in terms of explained variance) would be required to overturn a result.

These tools are most directly justified for regression-style estimators. Applying them to DiD or synthetic control requires an explicit mapping from the panel estimator to an omitted-variable-bias model.

For synthetic control, conformal inference methods from Chernozhukov et al. [2021] provide exact p-values that implicitly bound the effect under the sharp null.

For more complex settings—factor models, continuous treatments, dynamic panels—partial identification methods are less developed. Software is sparse, and calibration strategies are unclear. These remain open problems.

## Open Problems

Several questions remain. How should we construct bounds for factor-based estimators like matrix completion and GSC? The parallel-trends intuition does not directly apply. How do we extend the Rambachan-Roth framework to staggered adoption with heterogeneous effects? How do we communicate bounds to business audiences without creating the impression that “we do not know anything”?

Progress will require templates that extend bounded-violation analysis to modern panel estimators, software that makes these templates easy to use, and reporting standards that normalise bounds alongside point estimates. The foundations exist. The challenge is diffusion into practice.

## Practical Guidance

Until partial identification becomes routine, practitioners should adopt the following workflow. First, report the point estimate under maintained assumptions. Second, use `HonestDiD` or similar tools to report bounds under plausible violation magnitudes, calibrating from pre-trends or domain knowledge. Third, report the breakdown point: the violation magnitude at which substantive conclusions would change. Fourth, relate this breakdown point to observable features so readers can assess plausibility.

This workflow makes assumptions explicit, provides actionable uncertainty quantification, and allows readers to form their own judgements about the credibility of results.

## 20.9 Generative Synthetic Data: Promise and Peril

A popular claim in marketing analytics holds that Generative AI can create “synthetic customers” or “digital twins” to replace surveys, experiments, and observational data. The operational appeal is speed and low marginal cost. That appeal becomes dangerous when analysts treat generated data as evidence about causal effects.

This excitement is mirrored in finance and medicine, where synthetic data is rapidly adopted for privacy preservation and fraud detection [Potluru et al., 2023, Meldrum et al., 2025]. Adoption elsewhere reflects privacy and engineering needs, not causal identification. In marketing, the use of synthetic data for *causal inference* poses specific risks that are often overlooked. A nuanced view requires distinguishing between valid operational uses and invalid causal discovery.

### Valid Applications: Operations and Privacy

Synthetic data has a legitimate and valuable role in the marketing measurement stack, particularly for tasks that do not involve discovering unknown causal mechanisms.

**Privacy and Data Sharing.** As discussed in Section 20.11, sharing user-level data between advertisers and platforms is increasingly restricted. Synthetic data generation can create datasets that approximate selected statistical properties (for example, marginal distributions and some correlations) of the original data without exposing individual records. This allows vendors to develop code and audit pipelines without accessing sensitive PII.

**Pipeline Stress-Testing.** Synthetic data with *known* ground truth is essential for validating econometric software. Before deploying a complex difference-in-differences estimator on real sales data, an analyst should test it on synthetic data where the treatment effect is known to be exactly 5%. This is a form of Monte Carlo simulation, a standard practice.

**Augmenting Rare Events.** In finance, synthetic data is widely used to upsample rare classes, such as fraud cases, to improve classifier performance [Meldrum et al., 2025]. Similarly, in marketing, synthetic data can augment sparse events like churn or conversion for *predictive* models. This improves the stability of training but does not necessarily yield unbiased causal estimates.

### The Causal Discovery Trap

The danger arises when practitioners attempt to use generative models to *discover* causal effects or predict reactions to novel interventions. This confuses the simulation of a distribution with the discovery of a mechanism.

Generative models (LLMs, GANs, VAEs) are trained to approximate a joint distribution over covariates, outcomes, and exposures, for example  $P(X_{it}, Y_{it}, D_{it})$  for unit-time cells. They learn correlations and confounding structures as they exist in the observed data. If high-income consumers in the training corpus tend to buy luxury goods *and* see more luxury ads, the model will reproduce this correlation.

When a marketer asks an LLM, “How would this customer segment react to a 10% price hike?”, the model produces a prediction implied by patterns in its training data. If the training data contain confounded associations—where price hikes correlate with higher quality signals, for instance—generated responses will tend to reproduce that confounding.

Formally, generative models approximate  $P(Y_{it} | X_{it}, D_{it})$ —the observational conditional. Causal inference targets  $P(Y_{it} | \text{do}(D_{it}), X_{it})$ , the distribution under intervention. Recovering this distribution demands *identification assumptions*: exclusion restrictions for instrumental variables, parallel trends for difference-in-differences, or a correctly specified causal graph for do-calculus (see [Pearl, 2012]), which is rarely available in platform marketing settings. What unites these approaches is the need for domain knowledge about why certain variation is exogenous. Generative models trained on observational data do not, by themselves, identify which variation is exogenous or which backdoor paths are blocked.

Running a “virtual experiment” on synthetic data is therefore circular. It breaks the design chain: the estimand is interventional, but the data-generating mechanism is trained to reproduce an observational regime. Any “treatment effect” we measure is an artefact of the training distribution, not evidence about the world.

## The Digital Twin Illusion

The concept of a “digital twin”—a virtual replica of a specific customer—fails for similar reasons in causal contexts. A predictive twin can be useful: it can forecast a customer’s likely next purchase under *current* conditions, assuming the structural environment remains stable (see Section 20.3).

But causal inference asks what happens under *intervention*—a change in the system. If we introduce a novel product, a new channel, or a price point never seen before, the generative model is forced to extrapolate off-support. Without a validated structural model, generated responses are not grounded in an identified behavioural mechanism and should not be treated as causal evidence.

## A Disciplined Approach

We distinguish sharply between Synthetic Control methods (Chapter 6) and generative synthetic data. Synthetic Control constructs a counterfactual using real outcomes from unaffected units. It is empirical. Generative synthetic data constructs outcomes using probabilistic guesses from a black-box model. It is simulation.

As marketing moves toward automated decision systems, the temptation to substitute expensive real data with cheap synthetic data will grow. The causal analyst should insist on this distinction: use synthetic data

to test your *pipelines* and protect *privacy*, but use real data to learn about the *world*. To learn causal effects, we must either intervene in reality or invoke identification assumptions that no generative model, however sophisticated, can supply.

## 20.10 ML Integration Beyond Nuisance

Chapter 12 treated machine learning as a black box for estimating nuisance functions. Modern ML, particularly foundation models, offers a practical way to featurise unstructured inputs (text, images, graphs). The causal question is when such features can be used without inducing post-treatment adjustment or leakage.

### Embeddings as Proximal Covariates

Marketing data brims with unstructured confounders. A product’s description affects both its pricing and its conversion rate. We might ignore the text entirely, or create manual dummies. Large language models offer an alternative: construct embeddings  $E_{it} \in \mathbb{R}^p$  from pre-treatment text (for example, product descriptions fixed before a pricing decision) and include them as covariates in  $X_{it}$ . Timing matters. Post-treatment embeddings can be mediators or colliders rather than confounders.

Conceptually, embeddings can act as proxies for unobserved confounders in the sense of proximal causal inference [Miao et al., 2018, Tchetgen Tchetgen et al., 2020]. Proximal identification typically requires structured proxy assumptions, including completeness-type conditions, and often multiple proxy variables. Embeddings can be candidates, but they do not automatically satisfy these conditions. Veitch et al. [2020] formalise this for text in specific settings.

The danger lies in how we learn  $E_{it}$ . Two strategies exist, and they differ sharply in their causal implications. Frozen embeddings—off-the-shelf models like BERT or ResNet trained externally—reduce the risk of outcome leakage because they are not trained on the study outcomes. They can still be strongly predictive of treatment assignment because they encode attributes that drive pricing or targeting. Fine-tuning can make the representation a function of  $Y$  (or of post-treatment signals correlated with  $Y$ ). Conditioning on such a representation can induce selection bias, what practitioners call “leakage.”

The engineering rule follows directly. It makes sense to treat representation learning as part of the nuisance estimation pipeline. If we must fine-tune, we must cross-fit the representation itself: train the encoder on Fold A, encode features for Fold B, and estimate effects on Fold B. Cross-fitting the representation prevents an observation’s realised outcome from influencing the features used to adjust for it in the causal stage.

### High-Dimensional Treatments and Generative AI

Generative AI introduces treatments that are inherently high-dimensional. An ad is not just “A vs B,” but a specific image generated from a latent vector. The unit-level effect of a unique piece of content is not identified because overlap fails at that granularity. Any causal estimand must be defined on a coarser treatment space, for example clusters, attributes, or policies. This is the core challenge of what Fong and Grimmer [2016] and Keith et al. [2020] call text-as-treatment estimation.

We can make progress by reframing the estimand. Instead of asking about the effect of a specific generated image (which may never appear again), define an intervention as a policy that maps prompts and context into a distribution over content. The estimand is then a policy value under that stochastic intervention, and the relevant exposure can be recorded as  $D_{it}$  or as a lower-dimensional summary of the content delivered.

To operationalise this, we must map the high-dimensional treatment space back to something we can estimate. One approach clusters the generative latent space into discrete concepts—“Minimalist” versus “Cluttered,” say—and estimates the average treatment effect of these clusters. Another projects the high-dimensional treatment onto interpretable basis functions (text embeddings of the prompt, for instance) and estimates a causal effect function over the prompt space. Both strategies sacrifice some fidelity to gain identification.

## Causal Representation Learning

A longer-run research direction is causal representation learning ([Schölkopf et al., 2021]): learning valid causal variables from low-level inputs (pixels, tokens) without supervision. Standard supervised learning disentangles factors based on correlation. Causal learning seeks to disentangle factors based on invariance across environments.

For marketing, this means moving beyond “predicting churn with embeddings” to “learning a representation of user loyalty that is invariant to discount depth.” We want features that capture stable structural relationships, not features that happen to correlate with outcomes in our particular dataset. This remains an open engineering challenge, but the direction is clear: representations must be judged not by their predictive accuracy, but by their stability under intervention.

## Open Problems

Three problems define this frontier. First, when is a pre-trained text embedding “sufficient” to block confounding? We lack credible diagnostics for proxy sufficiency in high dimensions. The identification results from proximal causal inference assume we know something about the relationship between proxies and latent confounders—assumptions that are hard to verify when  $E_{it}$  is a high-dimensional vector.

Second, adaptive content creates propensity nightmares. When generative AI produces content dynamically based on user history, the assignment mechanism becomes a complex function of that history. Recovering assignment probabilities requires access to the logging policy, for example probabilities of content variants conditional on the history used for assignment. Many teams do not have this access.

Third, controlling for a high-dimensional vector “blocks confounding” but offers no insight into what was confounded. Was it product quality? Sentiment? Topic? We need methods that project bias corrections back onto interpretable axes—methods that tell us not just that we adjusted for something, but what that something was.

## 20.11 Privacy-Preserving Measurement and Data Clean Rooms

Privacy regulations have not merely constrained how we analyse data—they have eliminated data that once existed. Apple’s App Tracking Transparency reduced cross-app tracking consent sharply [Kesler, 2022]. Third-party cookies are disappearing from major browsers. GDPR and CCPA impose consent requirements that shrink samples and introduce consent-based selection, so estimates may target effects for the observed (consenting) subpopulation rather than the full population. Chapter 19 discussed these constraints from a data-engineering perspective. Here we focus on the open problems they create for causal inference and the emerging tools we have to address them.

### The Post-Privacy Measurement Landscape

For many practitioners, the binding constraint is not how to analyse data across silos. It is that the data no longer exist.

Before ATT, a mobile advertiser could track users across apps, link ad exposure to conversions, and estimate user-level treatment effects. After ATT, this linkage is available only for the minority who opt in. The opted-in population differs systematically from the opted-out population—younger, more engaged, more tech-savvy—and estimates from the consented sample may not generalise to the full population.

Before cookie deprecation, advertisers could track users across websites, build cross-site conversion paths, and attribute sales to touchpoints. Without cookies, this tracking disappears for users who do not log in. First-party data become the only reliable source, but they capture only part of the customer journey.

We cannot develop fancier estimators for data we do not have. Instead, we must identify what can be learned from the data that remain and be explicit about what cannot. This is a fundamental shift in how we approach measurement.

### Data Clean Rooms in Practice

Data clean rooms have emerged as a primary tool for privacy-preserving measurement. Platforms including Google Ads Data Hub, Amazon Marketing Cloud, and Meta Advanced Analytics allow advertisers to run queries on joined first-party and platform data without exporting user-level records.

Clean rooms enable aggregate queries. An advertiser can ask: “What was the conversion rate among users who saw my ad versus those who did not?” The platform returns aggregate statistics—counts, means, sometimes regression coefficients—without revealing which users converted. The data never leave the platform. Only the query results are exported.

In practice, you should also state whether reported aggregates are user-weighted, impression-weighted, or event-weighted. That choice changes the estimand.

This architecture constrains analysis in ways that matter for causal inference:

*No user-level diagnostics.* We cannot inspect propensity score distributions, construct balance tables at the individual level, or examine residual plots. The researcher must trust that the platform’s aggregation is correct and that the query returns what was requested.

*Minimum cell sizes.* To prevent re-identification, platforms suppress results when cell counts fall below thresholds (often 50 or 100 users). This creates a trade-off: finer cuts provide more relevant estimates but risk suppression. Coarser cuts survive suppression but average over heterogeneity.

*Query limits.* Platforms often limit the number of queries per day or per dataset. This prevents exhaustive exploration and limits the multiverse analysis we recommended in Section 20.5.

*Black-box implementation.* We cannot inspect the platform’s code to verify that it computes what we requested. Regression coefficients may use different defaults for clustering or weighting than we would choose.

The causal implications of these constraints are underexplored. Can we run difference-in-differences in a clean room when we cannot inspect pre-trends at the user level? Can we construct synthetic controls when donor weights must be computed inside the platform? Can we assess overlap when propensity scores are not exportable?

## A Practical Example: DiD in a Clean Room

Consider running a difference-in-differences analysis in Google Ads Data Hub. We want to estimate the effect of a new ad campaign on conversions, comparing exposed users to unexposed users before and after the campaign launch.

In a traditional analysis, we would extract user-level data on exposure and conversions. We would construct a pre-period and post-period for each user, inspect pre-trends by plotting average outcomes over time for treated and control groups, run a regression with user and time fixed effects, and cluster standard errors appropriately.

In the clean room, we can request aggregate counts: conversions by exposure status by time period. From these aggregates, we can compute a simple  $2 \times 2$  DiD estimate. But we face limitations:

We cannot inspect pre-trends at the user level. We can request average outcomes by period, but if this crosses cell-size thresholds, some periods may be suppressed. We cannot examine whether pre-trends diverge for subgroups.

We cannot include user-level covariates flexibly. We can request averages conditional on covariate bins, but high-dimensional covariate adjustment is infeasible when each additional cut risks hitting cell-size limits.

This limitation also weakens core diagnostics. You often cannot inspect overlap or balance at the unit level, so identification claims must rely more heavily on design arguments.

We cannot cluster standard errors at the user level because we do not observe user-level data. When the unit of randomisation is aggregate (geo), cluster at that level and treat the geo cell as the unit. When treatment is user-level but only aggregates are returned, standard errors necessarily rely on additional assumptions about within-user dependence or require platform-provided variance primitives.

The resulting estimate is still a DiD estimate, but our ability to assess its credibility is limited. We recommend supplementing clean-room analysis with aggregate-level diagnostics where possible: pre-trend diagnostics using period-level aggregates, balance checks on observable group-level characteristics, and sensitivity analysis varying the aggregation structure.

## What Aggregated Data Can and Cannot Identify

Aggregation changes what is identified. Some estimands survive aggregation. Others do not.

*Group-level treatment effects.* When treatment varies at the group level, we need only aggregate outcomes by group. Let  $c = 1, \dots, G$  index geos and let  $Y_{ct}$  denote the geo-level outcome (for example, average conversions per user) in geo  $c$  at time  $t$ . Geo-experiments [Vaver and Koehler, 2011], where assignment is at the DMA or city level and spillovers across geos are limited, require only aggregate outcomes by geography. The unit of analysis matches the unit of aggregation, and no user-level data are needed. Inference then treats geos as the independent sampling units, so the effective sample size is  $G$ .

*User-level treatment effects.* These cannot be recovered from aggregated data without strong assumptions. If we observe only group means, we cannot distinguish a treatment that helps everyone a little from one that helps a few people a lot. Heterogeneity is hidden within the aggregation.

*Selection bias diagnosis.* With user-level data, we can compare treated and control units on observables and assess balance. With aggregated data, we see only group-level summaries. Imbalance within groups is invisible. A treated group may appear balanced on average age while having very different age distributions than the control group.

*Attrition and missing data.* Aggregated counts do not reveal whether missingness is random or systematic. The analyst must trust that the platform’s data pipeline handles missing values correctly.

The implication is that privacy-preserving measurement pushes us toward designs where treatment varies at aggregate levels—geo-experiments, market-level rollouts, time-based switchbacks—using the tools from Chapters 18 and 17. User-level quasi-experiments become infeasible when user-level data are unavailable.

## Differential Privacy: Formal Guarantees and Practical Trade-Offs

Differential privacy (DP) provides formal guarantees that individual records cannot be inferred from released statistics. Platforms increasingly apply DP to clean-room outputs.

The formal definition is as follows. A randomised mechanism  $\mathcal{M}$  satisfies  $\epsilon$ -differential privacy if, for any two datasets  $\mathcal{D}$  and  $\mathcal{D}'$  differing in one record, and any set of outputs  $S$ :

$$\Pr[\mathcal{M}(\mathcal{D}) \in S] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{D}') \in S].$$

The parameter  $\epsilon$  controls the privacy guarantee: smaller  $\epsilon$  means stronger privacy. In practice, platforms may not disclose the privacy parameters or the exact mechanism, which complicates uncertainty quantification.

The privacy–accuracy trade-off follows directly from the definition. Achieving smaller  $\epsilon$  requires adding more noise to outputs, which inflates variance. For a count query, the noise typically has standard deviation proportional to  $1/\epsilon$ . For a mean, the noise depends on the range of the variable divided by  $\epsilon$  and the sample size.

*Composition.* A crucial property of DP is composition: running multiple queries depletes the privacy budget. If we run  $k$  queries each with privacy parameter  $\epsilon$ , the combined privacy guarantee is approximately  $k\epsilon$  under basic composition (and  $\sqrt{k}\epsilon$  under advanced composition). This means we cannot explore data indefinitely. Each query “spends” some of the privacy budget, and when it is exhausted, no more queries are permitted.

This has direct implications for causal inference. Running pre-trend diagnostics, balance checks, and sensitivity analyses all consume privacy budget. An analyst who runs many exploratory queries may have little budget left for the primary analysis. Pre-specification of the analysis plan becomes not just good practice but a practical necessity.

*Bias from DP noise.* DP noise is calibrated for simple statistics—counts, sums, means. More complex quantities can inherit DP noise in non-linear ways:

Ratios and regression coefficients can inherit bias because they involve non-linear transformations of noised quantities. Bias and coverage depend on the mechanism and estimator. Avoid assuming that standard confidence intervals remain valid.

*Practical guidance.* When working with DP-protected data, we recommend four steps. First, pre-specify the analysis to minimise the number of queries. Second, request raw counts rather than computed statistics where possible, so you can compute ratios yourself with uncertainty quantification. Third, report that estimates are subject to DP noise and, if the privacy parameters are known, quantify the additional variance. Fourth, favour simple estimators (differences in means) over complex ones (regression coefficients) because the former have better-understood noise properties.

## Open Problems

Several problems remain unresolved.

How do we conduct design diagnostics—pre-trend diagnostics, balance checks, overlap assessment—when user-level data are inaccessible? Can we develop aggregate-level diagnostics that provide similar assurance? Initial work suggests that group-level pre-trends can be informative, but the power of such diagnostics is lower than their user-level counterparts.

How do we construct valid confidence intervals when estimates are subject to both sampling error and DP noise? The two sources of uncertainty must be combined. Recent work by Agarwal et al. [2021] provides solutions for linear regression. Comparable results for DiD, synthetic controls, and panel estimators under DP remain open.

How do we reconcile estimates across platforms when each platform’s clean room provides different aggregations and applies different privacy protections? Can we combine estimates from Google, Meta, and Amazon into a coherent picture when we do not know the overlap in their user populations?

How do we audit clean-room analyses for reproducibility when the underlying data cannot be exported and the platform’s code cannot be inspected? Can we develop audit protocols that verify correctness without requiring data access?

## Practical Guidance

Given the challenges above, how should practitioners proceed? We offer the following workflow.

First, favour aggregate-level designs. Geo-experiments, market-level rollouts, and switchback designs require only aggregate outcomes. They are naturally suited to the privacy-constrained environment and do not require user-level linkage.

Second, pre-specify clean-room analyses. Because queries consume privacy budget and cannot be undone, specify the analysis plan before accessing the clean room. Include primary estimand, diagnostic queries, and sensitivity analyses in the plan.

Third, request raw counts where possible. Compute statistics yourself rather than requesting computed statistics from the platform. This gives you control over how ratios and standard errors are calculated.

Fourth, report DP-related uncertainty. If estimates come from DP-protected queries, acknowledge that reported uncertainty may underestimate true uncertainty. If privacy parameters are known, quantify the additional variance from DP noise.

Fifth, triangulate across data sources. Clean-room estimates from one platform may be biased in ways that differ from another platform’s biases. Comparing estimates across platforms—where possible—provides a check on credibility.

Sixth, be honest about limitations. When user-level diagnostics are impossible, say so. Do not report user-level estimates when only group-level data are available. Honesty about what we can and cannot learn is more valuable than false precision.

These steps do not solve the fundamental problem of inference with missing data, but they make the limitations transparent and provide a framework for credible analysis in the post-privacy era.

## 20.12 Reproducibility, Benchmarking, and Standards

Credible science requires that others can verify our claims. In panel causal inference for marketing, this verification is hard. Data are proprietary and cannot be shared. Methods are complex and implementation details matter. Results depend on choices—sample restrictions, control variables, inference procedures—that are rarely fully documented. This section outlines what reproducibility can mean in this context and what standards we can adopt to improve practice.

### Pre-Registration for Observational Studies

Pre-registration—committing to an analysis plan before seeing outcome data—originated in clinical trials and has spread to experimental social science. Its application to observational studies is less developed but equally important.

The core problem is specification search. With panel data, we face many choices: which units to include, which time periods to analyse, which controls to add, which estimator to use, how to cluster standard errors. Each choice affects the result. Without pre-commitment, the temptation to search for specifications that yield desired results is strong, even unconsciously.

Pre-registration disciplines the design chain by fixing the estimand, identification strategy, diagnostics, and inferential procedure before outcomes are inspected. In practice, this means specifying the estimand, the identification strategy, the sample definition, the primary and secondary outcomes, the control variables, the estimator, and the inference procedure. Deviations from the plan are permitted but must be documented and justified.

A pre-registration should describe the research question and the estimand, the identification assumption and why it is plausible, the data sources and sample restrictions, the treatment and outcome definitions, the estimation method, and the inference procedure. It should also state the dependence structure assumed for inference (unit clustering, geo clustering, serial correlation handling) and the rationale, list the diagnostics to be run (pre-trend diagnostics, balance diagnostics, placebo checks), and specify the planned sensitivity analyses.

Pre-registration does not guarantee correct results. A flawed design remains flawed whether pre-registered or not. But it shifts the burden of proof. Deviations from the plan require explanation, and readers can assess whether the reported analysis matches the registered plan or whether post-hoc adjustments suggest fishing.

### Why Benchmarking Is Hard in Marketing

Benchmarking—comparing methods on common datasets with known ground truth—has driven progress in machine learning. ImageNet, for example, allowed researchers to compare vision models on a shared task

with clear performance metrics. Section 20.5 discussed the role of benchmarking for method selection. Here we focus on why marketing panels make benchmarking difficult.

The broader point is not just that benchmarks are scarce. It is that evaluation criteria in causal work must track whether methods approximate what a well-designed experiment would deliver, not only whether they fit observed outcomes. Cinelli et al. [2025] emphasise this distinction and argue for more creative validation strategies than single headline benchmarks.

The obstacles are substantial.

Ground truth requires knowing the true treatment effect. In observational data, we never know this. The whole point is that we are trying to estimate something we cannot directly observe. Experiments provide ground truth, but they are expensive and proprietary. Firms that run experiments rarely share the data because it reveals strategic information.

Simulations require assumptions about the data-generating process. If we simulate parallel-trends violations, we must specify how trends deviate. If we simulate interference, we must specify the spillover structure. Methods that perform well under our simulated violations may fail under violations we did not anticipate. Simulations tell us how methods perform under assumed conditions, not under real conditions.

*Benchmarks become targets.* Once a benchmark exists, researchers optimise for it. Methods may overfit to the benchmark's specific features rather than performing well in general. This is Goodhart's law applied to methodology.

Semi-synthetic benchmarks offer a middle path. We take real covariate data from marketing panels, simulate treatment assignment and potential outcomes, and inject known effects. This preserves realistic covariate structure while providing ground truth. But the simulated treatment mechanism may not match real-world selection, and the injected effects may not capture true heterogeneity.

Held-out experiments provide the most direct benchmark. When a firm runs an experiment, it can also apply observational methods to the same data and compare estimates to the experimental benchmark. This is the gold standard but rare, because it requires experimental access and willingness to share results. A few academic–industry partnerships have produced such comparisons, but they remain exceptions.

The implication is that benchmarking in marketing will remain fragmented. No single benchmark will achieve the status of ImageNet. Progress will come from accumulating evidence across many studies, each with its own limitations.

## Reporting Standards

Even without shared benchmarks, reporting standards can improve transparency. A credible analysis report should contain enough information for a reader to assess validity and, in principle, replicate the analysis.

*State the estimand precisely.* “The effect of advertising on sales” is insufficient. State the estimand using the book’s notation, for example  $\text{ATT}_t$ ,  $\theta_k$ , or  $\mu(d)$ , and the time horizon over which it is defined.

*Make the identification strategy explicit.* What assumption justifies a causal interpretation—parallel trends, unconfoundedness, exclusion restriction? Why is this assumption plausible in this context? What would violate it?

*Describe the data in detail.* What is the unit of observation? What is the time period? What sample restrictions were applied and why? What is the treatment definition? What are the outcome and control variables?

*Report diagnostics.* Pre-trend diagnostics for DiD. Pre-treatment fit for synthetic control. Balance tables for matching. Overlap diagnostics for propensity methods. First-stage F-statistics for IV. These diagnostics, discussed in Chapters 17 and 18, allow readers to assess whether identification assumptions are credible.

*Include sensitivity analyses.* How do results change under alternative specifications? What magnitude of assumption violation would overturn the conclusion? Bounds under explicit violations, as discussed in Section 20.8, are more informative than vague claims of robustness.

*Use appropriate inference.* Cluster standard errors at the level of treatment assignment. Apply corrections for multiple testing when examining many outcomes or subgroups, as discussed in Chapter 16. Acknowledge uncertainty from small samples or few clusters.

Adoption of these standards requires that journals, firms, and funders demand them. The *Journal of Marketing Research* and other leading outlets have moved toward requiring pre-registration and detailed reporting, but enforcement varies. Without consistent enforcement, standards remain aspirational.

## Replication Without Data Sharing

Most marketing data cannot be shared. Contracts prohibit it. Privacy regulations restrict it. Competitive concerns preclude it. Yet replication remains essential. How do we verify results when data are proprietary?

*Code sharing.* Even if data cannot be shared, analysis code can be. A reader who obtains similar data can run the same code and check whether the methodology is sound. Code also reveals implementation details—how missing data were handled, how standard errors were computed—that prose descriptions often omit.

*Synthetic data.* We can share synthetic data for code verification when it is explicitly *not* used for causal discovery. Use simulation data with known ground truth for pipeline testing, and state clearly that this does not validate identification on the real data (see Section 20.9). A reader can run the code on this synthetic data to verify that it executes correctly and produces sensible output.

*Audit protocols.* An independent auditor with data access—perhaps under NDA—re-runs the analysis and certifies that the reported results match. Audits can certify computational reproducibility, not causal validity if the design assumptions are wrong. The auditor’s report provides assurance without requiring public data release.

*Registered reports.* Some journals commit to publishing regardless of results. The analysis plan is reviewed and accepted before data are collected or analysed. This eliminates publication bias and ensures that null

results are reported. The *Journal of Marketing Research* and *Marketing Science* have experimented with registered reports, though uptake remains limited.

None of these solutions is perfect. Code without data cannot be fully verified. Synthetic data may not capture the features that matter. Audits are costly and rare. Registered reports require advance planning that observational studies often lack. But each contributes to a culture of transparency that, over time, improves credibility.

## Open Problems

Several questions remain open. How can we create benchmarks for marketing causal inference when data are proprietary and ground truth is unknown? Can academic–industry partnerships produce shared benchmarks without compromising competitive interests? Federated learning and secure computation offer technical possibilities, but the governance challenges are substantial.

How can we enforce reporting standards when journals, firms, and funders have weak incentives to demand them? What would change that equilibrium? Perhaps reputational concerns, as the credibility revolution in economics has raised expectations for identification.

How can we make pre-registration practical for observational studies where the design evolves as data are explored? Can we distinguish legitimate design refinement from specification search? One approach is to separate exploratory and confirmatory phases explicitly, with pre-registration applying only to the latter.

How can we verify results when data cannot be shared and audits are costly? Can secure computation or differential privacy enable verification without disclosure? The technical tools exist but are not yet practical for complex causal analyses.

These are institutional problems as much as methodological ones. Solving them requires coordination across researchers, journals, firms, and funders. The methods in this book are only as credible as the practices that surround their use.

## Practical Guidance

Given the challenges above, what should practitioners do? We offer the following recommendations.

First, pre-register when possible. Even if public registration is impossible, internal pre-registration—a timestamped analysis plan shared with stakeholders before seeing outcomes—disciplines the process and creates a record.

Second, share code. Even if data cannot be shared, code can be. Document the code well enough that someone with similar data could run it. Include the code in an appendix or supplementary materials.

Third, report comprehensively. State the estimand precisely. Explain the identification strategy. Describe the data. Report diagnostics. Include sensitivity analyses. Follow the reporting templates in Chapter 18.

Fourth, acknowledge limitations. No analysis is perfect. State what assumptions are required and why they might fail. Discuss what would change the conclusion. Readers value clear statements of uncertainty more than false confidence.

Fifth, advocate for standards. When reviewing papers, demand pre-registration and detailed reporting. When working with firms, push for audit protocols and code sharing. When publishing, choose outlets that enforce high standards. Individual choices aggregate into norms.

These recommendations do not resolve the fundamental tension between proprietary data and open science. But they move practice in the right direction and contribute to a culture where credibility is valued.

## 20.13 Practitioner Roadmap

This book has covered many methods, each with its own assumptions, diagnostics, and failure modes. Practitioners facing real problems need a workflow that turns this material into actionable steps. This section provides that workflow by synthesising guidance scattered across earlier chapters into a single reference.

### Before You Start

Begin by defining the estimand and the assignment mechanism story, then choose an estimator whose identification assumptions are defensible in that setting. What causal effect do you want to estimate? Be precise. “The effect of advertising on sales” is too vague. Specify the treatment (TV spend, digital impressions, a specific campaign), the outcome (weekly store sales, online conversions, brand awareness), the population (all customers, new customers, a specific segment), and the time horizon (immediate, cumulative over 12 weeks, long-run equilibrium).

Assess data availability and the unit of assignment (user, store, geo). What data do you have, and what data would your preferred method require? Many methods fail because the data are aggregated above the level at which treatment varies. If the data do not support the method, choose a different method or acknowledge that the question cannot be answered with the data at hand.

Identify the dominant threats. Every setting has threats to validity. Is selection the main concern—treated units differ systematically from controls? Is interference likely—treating one unit affects others? Is structural instability a risk—the environment is changing during the study period? Is measurement error substantial—outcomes or treatments are measured with noise? The dominant threat determines which methods are viable and which diagnostics are essential. Section 20.5 provides a decision heuristic for mapping threats to methods.

### Design Phase

Choose the method based on data features and threats. Chapter 18 maps problem characteristics to methods. Few treated units with good pre-treatment data suggest synthetic control. Many treated units with staggered adoption suggest modern DiD estimators. Continuous treatments suggest dose-response methods. Dense interference suggests cluster randomisation or switchback designs. No single method dominates. The choice is context-specific.

Pre-register the analysis plan. Before inspecting treated-control outcome contrasts, commit to the estimand, sample definition, method, diagnostics, and inference procedure. Section 20.12 details what to include. Pre-registration does not prevent learning from the data, but it separates confirmatory from exploratory findings.

Plan the diagnostics in advance. Decide which diagnostics you will run and specify what diagnostic patterns would materially weaken the identification argument. Treat diagnostics as graded evidence rather than mechanical pass/fail tests.

## Estimation Phase

Run diagnostics before estimating effects to avoid outcome-driven rationalisation. The temptation is to estimate the effect first and check diagnostics later. Resist this. Use pre-trend diagnostics, balance diagnostics, and overlap diagnostics before computing treatment effects.

Interpret diagnostic failures honestly. A diagnostic failure does not automatically invalidate the analysis, but it should change the claim you are willing to make and the sensitivity analysis you report.

Apply multiple methods when feasible. Agreement across methods is reassuring but not dispositive because different estimators can share the same failure mode. Use disagreement to locate assumptions that matter, and use agreement as supportive evidence rather than proof. Section 20.5 discusses triangulation and multiverse analysis in detail.

## Reporting Phase

Report what you did, not just what you found. A credible report includes the estimand, the identification assumption, the data and sample, the method, the diagnostics, and the sensitivity analyses. Section 20.12 provides a template. Readers should be able to assess validity without relying solely on the author's judgment.

Communicate uncertainty honestly. Report confidence intervals, not just point estimates. When assumptions are uncertain, report bounds or sensitivity analyses. When diagnostics are marginal, say so. When the sample is small or clusters are few, acknowledge that inference is imprecise. Overstating precision undermines credibility. Clear statements of uncertainty build trust.

Tailor communication to the audience. Technical audiences want diagnostics, robustness checks, and methodological details. Business audiences want the bottom line, the level of confidence, and the implications for decisions. Both need honesty about limitations, but the framing differs. Section 18.15 in Chapter 18 provides guidance on stakeholder communication.

Even in business reporting, keep the estimand and the key identification assumptions explicit. Otherwise decisions are made on an undefined causal object.

## When Things Go Wrong

Diagnostics will sometimes fail. Pre-trends will not be parallel. Overlap will be limited. Fit will be poor. These patterns are information about which identification arguments are not credible in this dataset.

Consider alternative methods. If DiD fails pre-trends, can synthetic control achieve better fit? If propensity methods lack overlap, can you restrict to a subsample with common support? If factor models require too many factors, is the low-rank assumption appropriate? Failure of one approach does not imply failure of all approaches.

Report bounds instead of point estimates. When identification is partial, report what can be learned rather than pretending to know more. Section 20.8 discusses bounding strategies. A wide bound that is credible is more valuable than a narrow estimate that is not.

Acknowledge when the question cannot be answered. Sometimes the data do not support causal inference. In these cases, the honest response is to say so. Descriptive analysis may still be valuable, but avoid re-labelling descriptive contrasts as causal effects when identification fails.

## A Condensed Workflow

For quick reference, here is the workflow in condensed form:

1. **Define the estimand.** What effect, on what outcome, for what population, over what horizon?
2. **State the assignment mechanism.** Why is treatment timing or variation plausibly exogenous (or conditionally exogenous)?
3. **Assess data.** Do you have the data the method requires?
4. **Identify threats.** Selection? Interference? Structural instability? Measurement error?
5. **Choose method.** Match data features and threats to method families (Section 20.5).
6. **Pre-register.** Commit to estimand, sample, method, diagnostics, inference.
7. **Run diagnostics first.** Pre-trends, balance, overlap, fit—before seeing estimates.
8. **Estimate.** Apply the method. Apply multiple methods if feasible.
9. **Report transparently.** Estimand, assumptions, data, diagnostics, sensitivity, uncertainty.
10. **If diagnostics fail:** Try alternatives, report bounds, or acknowledge the question cannot be answered.

## The Practitioner's Mindset

The methods in this book are tools, not answers. They encode assumptions that may or may not hold. Diagnostics provide evidence about assumptions but cannot prove them. Triangulation increases confidence but does not eliminate uncertainty. Our job is to use these tools thoughtfully, report results honestly, and acknowledge what remains unknown.

This is harder than it sounds. Stakeholders want certainty. Deadlines pressure shortcuts. Incentives reward positive findings. Resisting these pressures requires discipline and institutional support. But the alternative—false confidence in unreliable estimates—serves no one. The methods in this book are only as good as the judgement and integrity of those who use them.

## 20.14 Assumptions for Future Practice

The open problems discussed in this chapter will not be solved by methods alone. Progress requires new disciplines: documenting what was once implicit, monitoring what was once assumed stable, and embracing uncertainty where point identification fails. This section states operational requirements and maintained assumptions that future work should make credible through better logging, documentation, and governance.

### Stability and Adaptation

**Assumption 116 (Documented change points and adaptive stability)** Measurement and assignment rules are stable within defined windows, or change points are documented and modelled explicitly. Adaptive assignment mechanisms are logged with sufficient detail (for example, timestamped assignment probabilities and the information set used to compute them) to support sequential identification and inference.

Current practice often assumes stability rather than testing it. Platform policies change without announcement. Algorithm updates occur silently. Breaks are discovered only when estimates behave unexpectedly. Future practice should define explicit stability windows, maintain changelogs for policy and algorithm updates, and treat these as core data objects, as emphasised in Section 20.3. When assignment is adaptive, logging must be rich enough to support identification under sequential designs—a standard that few current systems yet meet.

### Interference and Exposure

**Assumption 117 (Explicit exposure mappings under complex interference)** When interference is dense or overlapping, exposure mappings  $h_i(D_{-i,t})$  summarise neighbour treatments into documented summaries. Identification conditions for spillover effects are supported by transparent diagnostics and falsification checks where available, and reported with sensitivity analysis when point identification is contestable.

The partial-interference framework assumes clean cluster boundaries. Real networks have overlapping communities, and users belong to multiple groups, as discussed in Section 20.2. Future practice should require explicit specification of the exposure mapping—how neighbour treatments are summarised into an exposure variable—and sensitivity analysis to alternative mappings. When the exposure structure is too complex for credible point identification, bounds such as those in Section 20.8 should be reported rather than forcing a precise but fragile estimate.

## Overlap and Support

**Assumption 118 (Evolving overlap and support monitoring)** As environments shift, overlap and support are monitored continuously. Donor pools, propensity models, and trimming rules adapt to maintain identification. Diagnostics flag when support erosion compromises inference.

Overlap is not a one-off design-time check. As platforms evolve and populations shift, the overlap that existed when a study was planned may have eroded by the time results are reported. Future practice should build monitoring into the analysis pipeline: track propensity score distributions over time, record changes in trimming thresholds, and alert when donor pools shrink or become unrepresentative. Changes in trimming rules should be documented because they change the effective estimand, for example an overlap-restricted target.

## Governance and Auditability

**Assumption 119 (Governance ensuring long-run auditability)** Data versioning, code repositories, and audit trails support reproducibility over time. Privacy-preserving workflows balance transparency with compliance. Analysis registries document pre-specified plans and deviations from those plans.

Reproducibility requires infrastructure, not just good intentions. Code must be versioned. Data snapshots must be preserved. Analysis plans should be registered before outcomes are observed, and deviations from those plans documented. Audit trails support computational reproducibility over time. They do not substitute for design-based identification.

## The Gap Between Aspiration and Reality

These assumptions describe where the field should go, not where it is. Most marketing analytics today lack documented change points, explicit exposure mappings, continuous overlap monitoring, and rigorous governance. Closing this gap requires investment in infrastructure, in training, and in institutional incentives that reward transparency and robustness rather than speed and volume.

The methods in this book provide the analytical tools. The assumptions in this section describe the conditions under which those tools can be trusted. Bridging the two is the work that remains.

## 20.15 Chapter Summary and Visual Guide

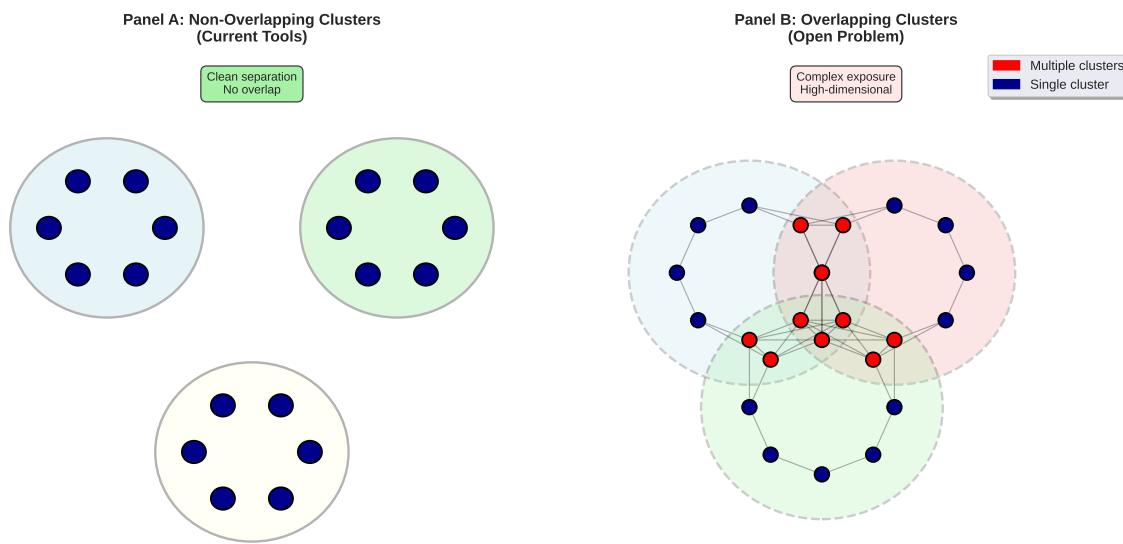
This chapter has mapped some of the open problems in panel causal inference for marketing. The obstacles are substantial: interference at scale, structural instability, adaptive experimentation, privacy constraints, and the difficulty of choosing methods when assumptions are uncertain. The figures and table below provide a visual summary of these challenges and the research agenda they suggest.

### Box 20.1: Key Takeaways from This Chapter

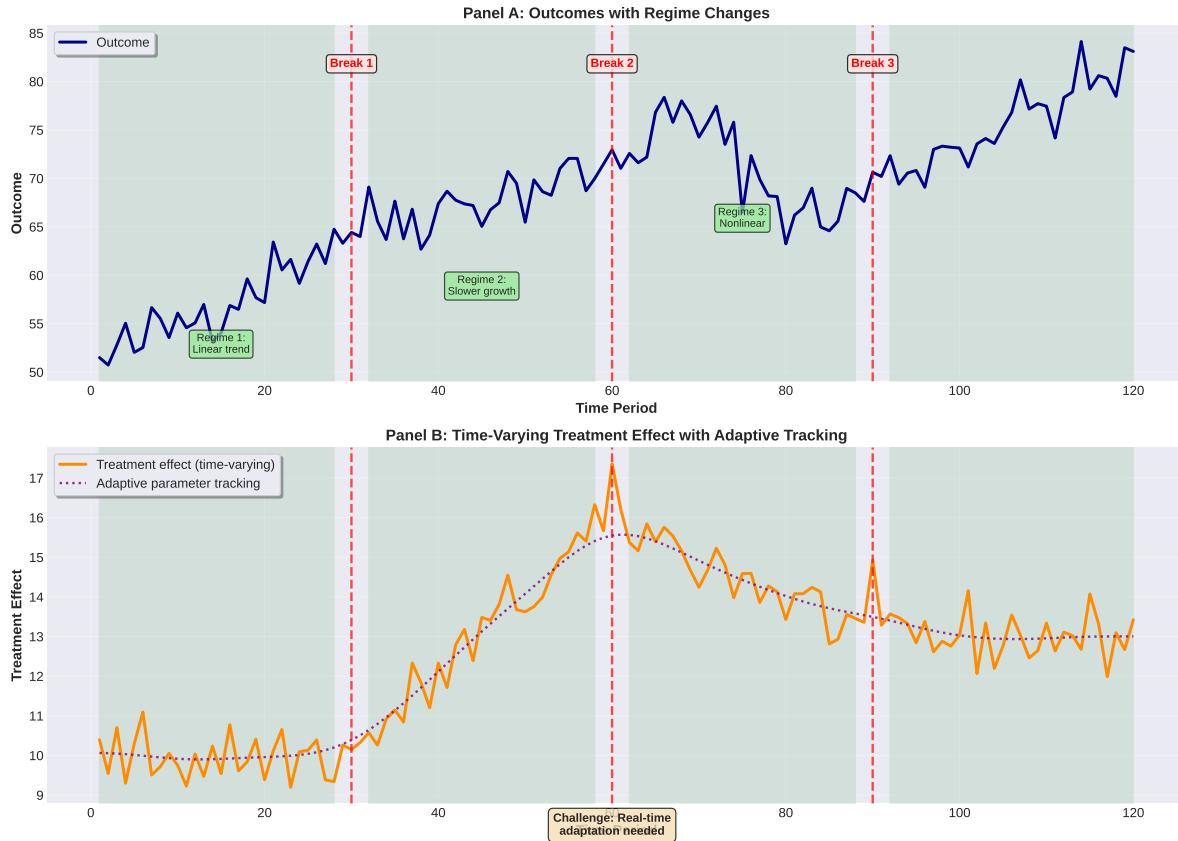
In settings with independent units, relatively stable environments, binary treatments, and accessible data, we can estimate well-defined panel causal estimands and interpret the usual diagnostics in a familiar way. The methods in Chapters 4–14 cover these settings.

The open problems arise when one of those conditions fails. Interference becomes dense or overlapping. Regimes shift without clear warning. Assignment adapts to outcomes. Privacy removes the data needed for diagnostics and inference. Method selection becomes difficult when assumptions are uncertain and diagnostics are imperfect.

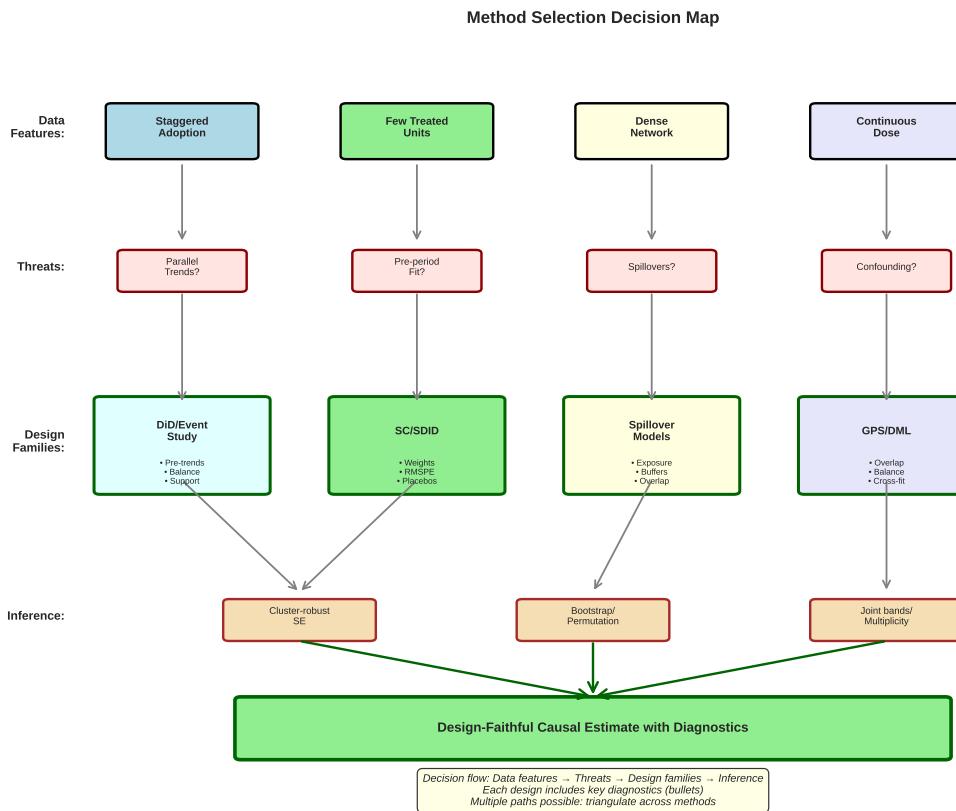
For now, practitioners should prioritise transparent design arguments, diagnostics, and sensitivity analysis, and be explicit when identification is partial. Progress will require new theory for interference and adaptivity, scalable algorithms for large and evolving networks, benchmarks with known ground truth, and reporting standards with real enforcement.



**Fig. 20.1** Interference-at-scale schematic with overlapping clusters and exposure mappings. *Panel A* illustrates non-overlapping clusters, which align with partial-interference assumptions. *Panel B* illustrates overlapping clusters, where exposure mappings are high-dimensional and inference must account for network-induced dependence.



**Fig. 20.2** Regime-change timeline with breakpoints overlaid on outcomes and exposures. *Illustrative schematic of regime breaks and drifting effects. Identification still requires design assumptions about assignment and about what is stable across regimes.*



**Fig. 20.3** Method-selection decision map linking threats to design families and diagnostics. A *decision flowchart linking data features and threats to design families, diagnostics, and inference choices. The output is a causal estimate conditional on the stated design assumptions, accompanied by diagnostics and sensitivity analysis.*

**Table 20.1** Open problems, current tools, research gaps, and diagnostics or evaluation criteria

Open problem	Current tools	Research gaps	Diagnostics / evaluation criteria
Interference at scale	Partial interference, buffer designs	Overlapping clusters, dense networks	Spillover sensitivity, scalability
Structural instability	Break diagnostics, time-varying factor models	Real-time adaptation, latent breaks	Post-break coverage, stability diagnostics
Adaptive experiments	Switchbacks, phased rollouts	Adaptivity + interference, sequential testing	Coverage under optional stopping, logging completeness
Method selection	Diagnostics, triangulation	Formal decision frameworks, benchmarks	Robustness to assumptions, transparency
Distribution-free inference	Permutation, conformal	Staggered adoption, few clusters	Finite-sample coverage, computational cost
Continuous treatments and dynamics	GPS, DML, adstock	Anticipation, endogenous rules	Overlap diagnostics, policy relevance of estimand
Partial identification	Sensitivity analysis, bounds	Defaults, joint violations	Bound width, informativeness
ML integration	Orthogonalisation, cross-fitting	Representation stability, leakage	Leakage risk, stability under shifts
Privacy and federated analysis	Aggregation, differential privacy	Federated DiD/SC, diagnostics	Privacy-accuracy trade-off, auditability

## Looking Forward

Looking forward, the open problems in Table 20.1 will not be solved quickly. Progress will be incremental. Until then, credible marketing evidence comes from designs that make assumptions explicit, diagnostics transparent, and uncertainty unavoidable rather than hidden.

We have written this book to give practitioners a pragmatic guide to the modern econometric and statistical toolkit: to sketch the broad landscape, to make the logic of a few core designs usable, and to be explicit about where the tools are reliable and where they are not. The methods we focus on work well in the settings they were designed for—*independent units, relatively stable environments, binary treatments, and accessible data*. As marketing moves toward *dense networks, adaptive platforms, and privacy-constrained measurement*, we will need new methods and new combinations of ideas.

Until those methods arrive, we must use what we have thoughtfully, report results honestly, and acknowledge what remains unknown. The credibility of causal inference in marketing depends not on having perfect methods—we never will—but on practitioners who understand the assumptions behind the tools they use,

apply them with care, and communicate their limitations. That is the mindset this book has sought to cultivate.



## **Part IX**

## **Appendices**



# Appendix A

## Time Series: Recap of Basic Principles

### Reader's Guide

This appendix collects probabilistic definitions and limit-theorem primitives used implicitly throughout the book, especially in Part V (Dynamics, Heterogeneity, and Spillovers) and Part VII (Inference). While the main text emphasises intuition and design-based identification, formal results for long panels ( $T \rightarrow \infty$ ) typically rest on rigorous time-series conditions stated here. This appendix is a technical reference rather than required reading. We write  $\mathcal{T}$  for the time index set to avoid confusion with  $T$ , which denotes the number of time periods in the main text. When we write a univariate process  $(X_t)_{t \in \mathcal{T}}$  here, you can read it as the time path of a fixed unit  $(Y_{it})_{t=1}^T$  or a fixed cluster  $(Y_{ct})_{t=1}^T$ .

- Refer to **Section A.3** for the ergodic properties that underpin distributed-lag estimators (Chapter 10).
- Refer to **Section A.3.1** for a compact summary of weak-dependence conditions and the long-run variance objects that appear in time-series robust inference.
- Refer to **Section A.3.2** for formal definitions of stationarity, which are essential for detecting structural breaks (Chapter 15) and specifying correct error structures (Chapter 16).

Source: Beran [2017].

### A.1 What Is a Time Series?

**Definition A.1** Let  $k \in \mathbb{N}$ ,  $\mathcal{T} \subseteq \mathbb{R}$ . A function

$$x : \mathcal{T} \rightarrow \mathbb{R}^k, \quad t \mapsto x_t$$

or, equivalently, a set of indexed elements of  $\mathbb{R}^k$ ,

$$\{x_t | x_t \in \mathbb{R}^k, t \in \mathcal{T}\}$$

is called an observed time series. We also write

$$x_t \ (t \in \mathcal{T}) \quad \text{or} \quad (x_t)_{t \in \mathcal{T}}.$$

**Definition A.2** Let  $k \in \mathbb{N}$ ,  $\mathcal{T} \subseteq \mathbb{R}$ ,

$$\Omega = (\mathbb{R}^k)^{\mathcal{T}} = \text{space of functions } X : \mathcal{T} \rightarrow \mathbb{R}^k,$$

$$\mathcal{F} = \sigma\text{-algebra on } \Omega,$$

$$P = \text{probability measure on } (\Omega, \mathcal{F}).$$

The probability space  $(\Omega, \mathcal{F}, P)$ , or equivalently the set of indexed random variables

$$\{X_t | X_t \in \mathbb{R}^k, t \in \mathcal{T}\}, \quad (X_t)_t \sim P$$

is called a ( $k$ -dimensional) time series (stochastic process) indexed by  $\mathcal{T}$ .

**Table A.1** Types of time series  $X_t \in \mathbb{R}^k$  ( $t \in \mathcal{T}$ )

Property	Terminology
$k = 1$	Univariate time series
$k \geq 2$	Multivariate time series
$\mathcal{T}$ countable, $\forall a < b \in \mathbb{R} : \mathcal{T} \cap [a, b]$ finite	Discrete time series
$\mathcal{T}$ discrete, $\exists u \in \mathbb{R}_+$ s.t. $t_{j+1} - t_j = u$	Equidistant time
$\mathcal{T} = [a, b]$ ( $a < b \in \mathbb{R}$ ), $\mathcal{T} = \mathbb{R}_+$ or $\mathcal{T} = \mathbb{R}$	Continuous time

A series  $(X_t)_{t \in \mathcal{T}}$  is termed a time series, or time series model. Instead of  $(\Omega, \mathcal{F}, P)$  we also write  $X_t$  ( $t \in \mathcal{T}$ ) or  $(X_t)_{t \in \mathcal{T}}$ .

Moreover, for a specific realization  $\omega \in \Omega$ , we write  $X_t(\omega)$  and

$$(x_t)_{t \in \mathcal{T}} = (X_t(\omega))_{t \in \mathcal{T}} = \text{sample path of } (X_t)_{t \in \mathcal{T}},$$

$$(x_t)_{t=1, \dots, n} = (X_t(\omega))_{t=1, \dots, n} = \text{finite sample path of } X_t.$$

*Remark A.1*  $\Omega$  may be more general than in Definition A.2. Similarly, the index set  $\mathcal{T}$  may be more general than a subset of  $\mathbb{R}$ , but it must be ordered and metric. Thus,  $(X_t)_{t \in \mathcal{T}}$  is a stochastic process with an ordered metric index set  $\mathcal{T}$ .

*Remark A.2* An overview of the most common types of time series  $X_t \in \mathbb{R}^k$  ( $t \in \mathcal{T}$ ,  $\mathcal{T} \neq \emptyset$ ) is given in Table A.1.

*Remark A.3* If  $X_t$  is an equidistant time series, then we may set without loss of generality  $\mathcal{T} \subseteq \mathbb{Z}$ .

## A.2 Time Series Versus iid Data

What distinguishes statistical analysis of iid data from time series analysis? We illustrate the question by considering the case of equidistant univariate real-valued time series  $X_t \in \mathbb{R}$  ( $t \in \mathcal{T} \subseteq \mathbb{Z}$ ).

Is consistent estimation of  $P$  possible? The answer depends on available a priori information and assumptions one is willing to make. This is illustrated in the following.

Let  $F_{X_t}(x)$  denote the marginal distribution function of  $X_t$  at time  $t$ ,

$$F_{X_t}(x) = P(X_t \leq x)$$

and let

$$F_n(x) = \frac{1}{n} \sum_{t=1}^n \mathbf{1}\{X_t \leq x\}$$

be the empirical marginal distribution function.

**Theorem:** *Assumption:*

$$X_t \in \mathbb{R} \ (t \in \mathbb{Z}) \text{ i.i.d.}$$

*Then*

$$\forall t \in \mathbb{Z} : F_{X_t} = F_{X_0}$$

and

$P$  is fully specified by  $F_{X_0}$ .

**Theorem:** (Glivenko-Cantelli) *Assumption:*

$$X_t \in \mathbb{R} \ (t \in \mathbb{Z}) \text{ i.i.d.}$$

*Then*

$$P \left( \lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F_{X_0}(x)| = 0 \right) = 1.$$

*Proof:* See e.g. van der Vaart [2000].

**Conclusion:** Given the i.i.d assumption,  $P$  can be estimated consistently. No additional assumptions are required.

This is not true in general under non-i.i.d assumptions.

**Example:**

$$X_t = \theta_t + Z_t \ (t \in \mathbb{Z}), Z_t \text{ i.i.d } \mathcal{N}(0, 1)$$

$$\theta_t \ (t \in \mathbb{Z}) \text{ unknown.}$$

Consistent estimation of  $\theta_t$  is not possible, unless additional assumptions on  $\theta_t$  are imposed.

**Example:**

$$X_t = U \quad (t \in \mathbb{Z}),$$

$$0 < p = P(U = 1) = 1 - P(U = 0) < 1.$$

Consistent estimation of  $p$  is not possible.  $\square$

**Conclusion:** In general, consistent estimation of  $P$  is not possible, unless additional assumptions are imposed. The problem is:

$$\begin{aligned} \text{observed time series} &= (x_1, \dots, x_n) \\ &= (X_1(\omega), \dots, X_n(\omega)) \\ &= \text{sample of size one from an } n\text{-dimensional distribution.} \\ &\neq \text{sample from the infinite dimensional distribution } P \text{ on } \mathbb{R}^{\mathbb{Z}}. \end{aligned}$$

In general, consistent estimation of  $P$  is not possible without additional assumptions. In this chapter, typical assumptions used in time series analysis are discussed. For simplicity we focus on equidistant univariate real valued time series  $X_t \in \mathbb{R}(t \in \mathbb{Z})$ .

### A.3 Fundamental Properties

The asymptotic distribution of many statistics follows—sometimes after a complicated proof or suitable transformations—from the asymptotic distribution of sums.

**Notation:**

$$\bar{x} = \bar{x}_n = n^{-1} \sum_{t=1}^n X_t$$

**Definition A.3**  $X_t \in \mathbb{R}(t \in \mathbb{Z})$  has the almost sure ergodic property with a constant limit, or a.s. EPCL, if

$$\exists \mu \in \mathbb{R} \text{ s.t. } P\left(\lim_{n \rightarrow \infty} \bar{x} = \mu\right) = 1.$$

Sometimes one also calls this the mean-ergodic property in the a.s. sense.

**Definition A.4**  $X_t \in \mathbb{R}(t \in \mathbb{Z})$  has the  $L^2$ -ergodic property with a constant limit, or  $L^2$ -EPCL, if

$$\exists \mu \in \mathbb{R} \text{ s.t. } \lim_{n \rightarrow \infty} \mathbb{E}[(\bar{x} - \mu)^2] = 0.$$

Sometimes one also calls this the mean-ergodic property in the  $L^2$ -sense.

The EPCL conditions above are law-of-large-numbers statements adapted to dependent data. Under which circumstances can it happen that the EPCL does not hold? Three main problems can occur, as outlined in the following.

**Problem:** Lack of stability: Distribution of  $X_{t+1}, X_{t+2}, \dots, X_{t+n}$  changes too much as a function of  $t$  so that  $(x_t)_{t=1, \dots, n}$  is not sufficiently representative for  $P$  on  $\mathbb{R}^{\mathbb{Z}}$ .

**Example:**  $\epsilon_t$  iid,  $E(\epsilon_t) = 0$ ,  $\sigma^2 = \text{var}(\epsilon_t) < \infty$ ,

$$X_t = \beta t + \epsilon_t.$$

Then

$$\bar{x} = \frac{\beta}{n} \left( \frac{n(n+1)}{2} \right) + \bar{\epsilon},$$

$$P(|\bar{x}| \rightarrow \infty) = 1.$$

**Example:**  $\epsilon_t$  iid,  $E(\epsilon_t) = 0$ ,  $\sigma^2 < \infty$ ,

$$X_s = 0(s \leq 0), X_t = \sum_{s=1}^t \epsilon_s (t \geq 1).$$

Then

$$\bar{x} = n^{-1} \sum_{t=1}^n \epsilon_{t-n+t},$$

$$\text{var}(\bar{x}) = \frac{\sigma^2}{n^2} \sum_{t=1}^n t^2 \rightarrow \infty.$$

**Problem:** High variability of the marginal distribution  $F_{X_t}$ .

**Example:** For Cauchy distributed iid  $X$ , we have

$$\bar{x} = X_1.$$

**Problem:** Absorbing states:

$$X_t (t \in \mathbb{Z}) \text{ is } \mathcal{F}\text{-measurable},$$

$$\exists t \in \mathbb{Z}, A_t \in \mathcal{F}, \text{s.t. } 0 < P(X_t = A_t) < 1 \text{ and } P(\forall s > t : X_s \in A_t | A_t) = 1.$$

Then

$$A_t \text{ is an absorbing state.}$$

□

**Example:**

$$X_t = U \ (t \in \mathbb{Z}),$$

$$0 < p = P(U = 1) = 1 - P(U = 0) < 1.$$

Then

$$A_t = \{X_t = 1\} \text{ is an absorbing state.}$$

### A.3.1 Weak Dependence and Long-Run Variance

Many asymptotic arguments in econometrics reduce to a central limit theorem for partial sums of dependent variables. Two widely used sufficient structures are (i) *martingale difference* conditions relative to a filtration  $(\mathcal{F}_t)$ , and (ii) *weak dependence* conditions (for example, strong mixing) for strictly stationary processes with suitable moment and summability requirements.

When a CLT holds for a mean-zero process  $(u_t)$ , one typically has

$$n^{-1/2} \sum_{t=1}^n u_t \Rightarrow \mathcal{N}(0, \Omega),$$

where the *long-run variance* is

$$\Omega = \sum_{k=-\infty}^{\infty} \text{cov}(u_t, u_{t-k}).$$

This object differs from  $\text{var}(u_t)$  when serial correlation is present and motivates time-series robust standard errors (HAC) and block-based resampling schemes.

### A.3.2 Strict Stationarity

**Definition A.5**  $X_t \in \mathbb{R} (t \in \mathbb{Z})$  is called strictly stationary or strongly stationary, if

$$\forall k \in \mathbb{Z}, \forall m \in \mathbb{N}, \forall t_1, \dots, t_m \in \mathbb{Z} : (X_{t_1}, \dots, X_{t_m}) \stackrel{d}{=} (X_{t_1+k}, \dots, X_{t_m+k}).$$

**Example:**  $X_t \in \mathbb{R} (t \in \mathbb{Z})$  iid is strictly stationary. □

**Example:**

$$X_t = \sum_{j=0}^q \psi_j \varepsilon_{t-j} \quad (t \in \mathbb{Z}),$$

$$\varepsilon_t \in \mathbb{R} (t \in \mathbb{Z}) \text{ iid, } \forall j \in \mathbb{R} (j = 0, \dots, q)$$

is strictly stationary.  $X_t$  is called a moving average process of order  $q$ , or MA( $q$ ) process. □

*Remark A.4* Strict stationarity solves Problem 6, but not Problems 9 and 11.

### A.3.3 Weak Stationarity

**Definition A.6** Let

$$X_t \in \mathbb{R} \quad (t \in \mathbb{Z}) \text{ s.t. } \forall t \in \mathbb{Z} : E(|X_t|) < \infty.$$

Then

$$\mu_t = E(X_t) \quad (t \in \mathbb{Z})$$

is called the expected value function, or mean function, of  $X$ . If

$$E(X_t^2) < \infty,$$

then

$$\gamma : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}$$

with

$$\gamma(s, t) = \text{cov}(X_s, X_t) = E[(X_s - \mu_s)(X_t - \mu_t)]$$

is called the autocovariance function of  $X$ , and

$$\rho(s, t) = \text{corr}(X_s, X_t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}$$

is called the autocorrelation function (ACF) of  $X$ .

*Remark A.5*

$$\gamma(t, t) = \text{var}(X_t), \quad \rho(t, t) = 1$$

*Remark A.6* For

$$X_t \in \mathbb{C} \quad (t \in \mathbb{Z}),$$

we define

$$\gamma(s, t) = \text{cov}(X_s, X_t) = E[(X_s - \mu_s)(\overline{X_t - \mu_t})].$$

**Lemma:**

$$\gamma(t, s) = \gamma(s, t)$$

**Proof**

$$\overline{\gamma(s, t)} = \overline{E[(X_s - \mu_s)(\overline{X_t - \mu_t})]} = E[\overline{(X_s - \mu_s)(X_t - \mu_t)}] = E[(X_t - \mu_t)(\overline{X_s - \mu_s})] = \gamma(t, s)$$

**Definition A.7**  $X_t \in \mathbb{R}$  ( $t \in \mathbb{Z}$ ) is called second order stationary, or weakly stationary, if

$$E(X_t^2) < \infty,$$

$$\exists \mu \in \mathbb{R} \text{ s.t. } \forall t \in \mathbb{Z} : E(X_t) = \mu,$$

$$\exists \gamma : \mathbb{Z} \rightarrow \mathbb{R} \text{ s.t. } \forall s, t \in \mathbb{Z} : \text{cov}(X_s, X_t) = \gamma(t - s).$$

**Lemma:**

$$\text{weak stationarity} \not\Rightarrow \text{strong stationarity}$$

**Proof** Counterexample for  $\not\Rightarrow$ :

$$X_{2i} = Z_i \quad (i \in \mathbb{Z}), \quad X_{2i+1} = \frac{Z_i^2 - 1}{\sqrt{2}} \quad (i \in \mathbb{Z})$$

where

$$Z_i \quad (i \in \mathbb{Z}) \text{ iid } \mathcal{N}(0, 1)\text{-variables}$$

Counterexample for  $\not\Leftarrow$ :

$$X_t \quad (t \in \mathbb{Z}) \text{ iid, Cauchy distributed}$$

**Lemma:** Assumptions:

$$X_t \in \mathbb{R} \quad (t \in \mathbb{Z}) \text{ strictly stationary, } E(X_t^2) < \infty$$

Then

$$X_t \quad (t \in \mathbb{Z}) \text{ weakly stationary}$$

**Proof** a)  $\mu$ : The Cauchy-Schwarz inequality implies

$$E^2(|X_t|) \leq E[X_t^2] < \infty,$$

and hence

$$\exists \mu_t = E(X_t) \in \mathbb{R}.$$

Thus, together with strong stationarity

$$\forall s, t \in \mathbb{Z} : \mu_s = \mu_t = \mu \in \mathbb{R}$$

b)  $\gamma$ : The Cauchy-Schwarz inequality implies

$$(E[X_s X_t])^2 \leq E(X_s^2) E(X_t^2) < \infty.$$

Together with strong stationarity we then have

$$\forall t, k \in \mathbb{Z} : E(X_t X_{t+k}) = E(X_0 X_k),$$

and

$$\text{cov}(X_t, X_{t+k}) = \text{cov}(X_0, X_k) = E(X_0 X_k) - \mu^2 = \gamma(k).$$

*Remark A.7* Weak stationarity solves Problem 6 w.r.t. first two moments of  $X_t$ , and Problem 9 in the sense that  $E(X_t^2) < \infty$ . It does not solve Problem 11.

**Example:**

$$X_t = U \quad (t \in \mathbb{Z})$$

where

$$0 < p = P(U = 1) = 1 - P(U = 0) < 1,$$

is weakly and strictly stationary, but

$A_t = \{X_t = 1\}$  is an absorbing state.



## Appendix B

# Stationarity and Cointegration in Panels

Panel data combine a cross-section of units with time series. When panels are short, unit root and cointegration tests typically have low power and design-based identification considerations dominate. When panels are longer (for example,  $T \gtrsim 30$ ), or when outcomes exhibit macro-style persistence, stochastic trends and cointegration become empirically relevant. This appendix summarises the concepts, threats, diagnostics, and remedies most often used in panel settings.

### Definitions

Let  $Y_{it}$  denote an outcome for unit  $i$  at time  $t$ .

- **Stationarity ( $I(0)$ ).** A process is  $I(0)$  if it is covariance stationary (or can be rendered so by removing deterministic components such as a constant and seasonal dummies).
- **Unit root ( $I(1)$ ).** A process is  $I(1)$  if  $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$  is  $I(0)$ . In applied work, “near-unit-root” persistence can look similar to  $I(1)$  over finite samples.
- **Trend-stationarity.** A process is trend-stationary if  $Y_{it}$  becomes  $I(0)$  after removing a deterministic trend (and possibly seasonal components).
- **Cointegration.** Two (or more)  $I(1)$  variables are cointegrated if a linear combination is  $I(0)$ . Cointegration supports an error-correction representation, which separates short-run dynamics from long-run equilibrium adjustment.
- **Cross-sectional dependence.** Panels often share common shocks (macroeconomic conditions, platform policy changes, seasonal demand). Ignoring common factors can distort unit-root and cointegration diagnostics.

## Threats to Interpretation

Stationarity diagnostics matter because they affect both statistical interpretation and the credibility of counterfactual extrapolation.

- **Spurious regression.** Regressing persistent or  $I(1)$  series on each other can produce misleadingly precise estimates and high  $R^2$  even when there is no substantive relationship.
- **Structural breaks and evolving environments.** Breaks in mean, trend, or seasonality can mimic unit-root behaviour and can invalidate residual-based arguments that treat pre-period dynamics as representative of post-period dynamics.
- **Seasonality and deterministic components.** High-frequency marketing outcomes often contain strong deterministic seasonality; failing to model it can generate persistence and residual structure that is mistaken for stochastic trends.
- **Cross-sectional dependence.** Unmodelled common shocks can inflate apparent persistence and reduce the validity of tests that assume weak dependence across units.
- **Causal design remains primary.** Passing time-series diagnostics does not establish identification for causal effects. These tools address dynamic misspecification and spurious correlation, not selection into treatment.

## Diagnostics and Decision Path

Treat the following as diagnostics rather than certification. The goal is to decide whether to model outcomes in levels, in differences, or with an error-correction structure.

### 1. Assess panel dimensions ( $N$ vs $T$ ).

- **Short  $T$  ( $T \lesssim 20$ ):** Unit-root tests have low power. Prefer parsimonious specifications with unit fixed effects, time fixed effects, and explicit seasonality. Focus on stability diagnostics and sensitivity to window choice.
- **Longer  $T$  ( $T \gtrsim 30$ ):** Persistence diagnostics become informative. Proceed with plots and tests, paying attention to deterministic components and cross-sectional dependence.

### 2. Assess variable type (domain knowledge).

- **Bounded variables** (probabilities, market shares, rating scales): a literal unbounded random walk is incompatible with strict bounds, but bounded series can still be highly persistent. Consider transformations (for example, logit for proportions) and treat persistence as an empirical question.
- **Scale variables** (sales, revenue, spend): levels are often persistent; logs and growth rates can be closer to  $I(0)$ , depending on the environment and aggregation.
- **Ratios and rates:** can be mean-reverting or near-unit-root depending on institutional constraints and regime stability.

### 3. Plot and pre-process.

- Plot representative unit paths and the cross-sectional mean. Identify seasonality, breaks, and common shocks.
  - Decide whether to include deterministic components (constant, trend, seasonal dummies) and whether to remove common shocks via time fixed effects or demeaning.
4. **Run unit-root diagnostics.** If the unit-root null is not rejected, treat  $Y_{it}$  as highly persistent and avoid interpreting level regressions mechanically.
5. **If both  $Y$  and  $X$  appear  $I(1)$ , assess cointegration.** Cointegration tests operationalise whether a long-run equilibrium relationship is compatible with the data, accounting for heterogeneity and dependence depending on the chosen test.
6. **Choose a modelling strategy aligned with the substantive question.**
- **If series are  $I(0)$  (after appropriate deterministic components):** standard fixed effects or distributed-lag specifications are natural.
  - **If series are  $I(1)$  and cointegrated:** an error-correction representation is often conceptually aligned with long-run level effects.
  - **If series are  $I(1)$  and not cointegrated:** first differences can avoid spurious level relationships, but this changes the estimand toward short-run changes.

## Panel Unit Root Tests

Heterogeneous panels and cross-section dependence challenge classical tests. In practice, results can be sensitive to whether the unit-level regressions include intercepts, trends, and seasonal components, and to whether common shocks are removed (for example, via time fixed effects). Common options:

**Table B.1** Panel unit root tests: scope and assumptions

Test	Key feature	Notes
Levin–Lin–Chu (LLC, 2002)	Common autoregressive parameter across units	Allows heteroskedasticity across $i$ ; limited cross-section dependence handling.
Im–Pesaran–Shin (IPS, 2003)	Heterogeneous AR parameters; averages unit-level ADFs	More flexible than LLC; assumes weak cross-section dependence.
Pesaran CIPS (2007)	Cross-section dependence via common factors (CADF)	Robust under unobserved common factors; recommended with strong cross-dependence.

## Panel Cointegration

If variables are  $I(1)$  but share a long-run equilibrium, cointegration supports modelling in levels with an error-correction representation. In panel settings, cointegration tests differ in whether they allow heterogeneous cointegrating vectors and how they handle dependence and dynamics. Common tests:

**Table B.2** Panel cointegration tests

Test	Key feature	Notes
Pedroni (1999, 2004)	Within- and between-dimension tests; heterogeneous cointegrating vectors	Widely used; accommodates individual intercepts and trends.
Kao (1999)	Residual-based test; homogeneous cointegrating vector	Simpler but restrictive homogeneity.
Westerlund (2007)	Error-correction-based tests with robust small-sample properties	Powerful under general dynamics and some cross-dependence.

## Remedies and Modelling Choices

Remedies should be chosen to match both the diagnostics and the substantive object of interest.

- **Model deterministic structure explicitly.** Many marketing outcomes contain strong seasonality and calendar effects. Include seasonal dummies, holiday indicators, and time fixed effects before attributing persistence to stochastic trends.
- **Work in logs or rates when appropriate.** For scale outcomes, logs can stabilise variance and convert multiplicative growth into additive structure.
- **First differences.** Differencing can remove stochastic trends but shifts interpretation toward short-run changes. Use when long-run level relationships are not supported and the decision question concerns changes.
- **Error-correction models (ECM).** When cointegration is plausible, ECMs separate short-run dynamics from long-run equilibrium adjustment and preserve long-run level interpretation.
- **Structural breaks.** If plots or residuals suggest breaks, consider restricting the analysis window, adding break indicators, or stratifying by regimes.
- **Cross-sectional dependence.** When common shocks dominate, prefer diagnostics and models that accommodate common factors (for example, CIPS-type diagnostics) and use inference procedures that respect dependence.

## Implementation Notes

While this book does not prescribe specific software, the following map conceptual tests to standard implementations in major statistical packages.

**Table B.3** Implementation Map for Stationarity Diagnostics

Task	Method Class	Implementation Pointer
Unit Root (Basic)	IPS / LLC	<b>R:</b> <code>plm::purtest</code> (modes: "ips", "levinlin") <b>Stata:</b> <code>xtunitroot ips, xtunitroot llc</code>
Unit Root (Robust)	Pesaran CIPS	<b>R:</b> <code>punitroots::pbadf</code> <b>Stata:</b> <code>xtcips</code> (community-contributed)
Cointegration	Pedroni / Westerlund	<b>R:</b> <code>pco::pedroni, plm</code> extensions <b>Stata:</b> <code>xtcointtest pedroni, xtwest</code>
Error Correction	PMG / MG	<b>R:</b> <code>plm::pmg</code> <b>Stata:</b> <code>xtpmg</code>

## Marketing Scenarios

### Scenario A: Daily Category Sales (High Frequency)

- **Data:** Daily sales for 500 SKUs over 3 years ( $T \approx 1000$ ).
- **Diagnosis:** Strong weekly seasonality and holiday spikes; common shocks across products. Persistence may reflect deterministic seasonality and breaks rather than a stochastic trend.
- **Action:** Do not difference blindly. Model deterministic seasonality (day-of-week and holiday effects), consider time fixed effects for common shocks, and run persistence diagnostics on de-seasonalised residuals. If breaks cluster around launches or policy changes, treat them explicitly.

### Scenario B: Brand Tracking Survey (Bounded)

- **Data:** Monthly "Consideration" scores (0-100%) for 20 brands over 5 years ( $T = 60$ ).
- **Diagnosis:** The outcome is bounded, but can still exhibit near-unit-root persistence over finite samples, especially if measurement practices change or the survey instrument drifts.
- **Action:** Treat persistence as an empirical feature. Consider transformations (for example, logit after rescaling to (0, 1)), include seasonality if present, and use dynamic panel models when serial dependence is substantive. Cointegration language is typically less natural in this bounded setting.



## References

- Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.
- Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california’s tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.
- Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.
- Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.
- Naman Agarwal, Ravi Kenthapadi, Satyen Kale, and Suresh Theertha. Inference for linear regression with differential privacy. In *International Conference on Artificial Intelligence and Statistics*, 2021.
- Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.
- Seung C Ahn and Alex R Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3): 1203–1227, 2013.
- Alexander Almeida, Susan Athey, Guido W. Imbens, Eva Lestant, and Alexia Olaizola. Estimating variances for causal panel data estimators. *arXiv preprint arXiv:2510.11841*, 2025.
- Takeshi Amemiya and Thomas E. MacCurdy. Instrumental-variable estimation of an error-components model. *Econometrica*, 54(4):869–880, 1986.
- Donald WK Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856, 1993.
- Isaiah Andrews, James H. Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753, 2019.
- Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, 2009.
- Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

- Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011.
- Manuel Arellano. *Panel Data Econometrics*. Oxford University Press, 2003.
- Dmitry Arkhangelsky and Guido W. Imbens. Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 27(3):C1–C61, 2024.
- Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.
- Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017.
- Susan Athey and Guido W Imbens. The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier, 2017.
- Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.
- Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.
- Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.
- Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *Review of Economic Studies*, page rdaf087, 2025a.
- Susan Athey, Guido Imbens, Zhaonan Qu, and Davide Viviano. Triply robust panel estimators. *arXiv preprint arXiv:2508.21536*, 2025b. Presented as the Journal of Applied Econometrics lecture at the ASSA meetings, San Francisco, January 2025.
- Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioural research*, 46(3):399–424, 2011.
- Andrii Babii, Eric Ghysels, and Jonas Striaukas. Econometrics of machine learning methods in economic forecasting. *Handbook of Research Methods and Applications in Macroeconomic Forecasting*, page 246, 2024.
- Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.
- Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.
- Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.
- Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.
- Patrick Bajari, Brian Burdick, Guido W. Imbens, Lorenzo Masoero, James McQueen, Thomas S. Richardson, and Ido M. Rosen. Experimental design in marketplaces. *Statistical Science*, 1(1):1–19, 2023.
- Badi H Baltagi, Peter H Egger, and Michaela Kesina. Small sample properties and pretest estimation of a spatial hausman–taylor model. In *Essays in Honor of Jerry Hausman*, pages 215–236. Emerald Group Publishing Limited, 2012. doi: 10.1108/S0731-9053(2012)0000029013.

- Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.
- Frank M Bass. A new product growth model for consumer durables. *Management Science*, 15(5):215–227, 1969.
- Marc F. Bellemare and Casey J. Wichman. Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, 82(1):50–61, 2020.
- Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.
- Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.
- Eli Ben-Michael, Avi Feller, and Jesse Rothstein. Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):351–381, 2022.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.
- Jan Beran. *Mathematical Foundations of Time Series Analysis*. Springer, Cham, Switzerland, 2017.
- Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.
- Steven Berry, Amit Gandhi, and Philip Haile. Identification in differentiated products markets using market level data. *Econometrica*, 82(5):1749–1797, 2014.
- Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.
- Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- Thomas Blake, Chris Nosko, and Steven Tadelis. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1):155–174, 2015.
- Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 2022.
- Abhishek Borah and Gerard J Tellis. Halo (spillover) effects in social media: Do recalls of one brand hurt or help a rival brand? *Journal of Marketing Research*, 2016.
- Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, 2024.
- John Bound, Charles Brown, and Nancy Mathiowetz. Measurement error in survey data. In James J. Heckman and Edward Leamer, editors, *Handbook of Econometrics*, volume 5, pages 3705–3843. Elsevier, 2001. doi: 10.1016/S1573-4412(01)05012-7.
- Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001. doi: 10.1214/ss/1009213286.
- Simon Broadbent. Modelling with adstock. *Journal of the Market Research Society*, 26(4):295–312, 1984.

- Robert L Brown, James Durbin, and James M Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(2):149–163, 1975.
- Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- Brantly Callaway and Pedro HC Sant’Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 2021.
- Brantly Callaway, Derek Dyal, Pedro HC Sant’Anna, and Emmanuel S Tsyawo. Beyond parallel trends: An identification-strategy-robust approach to causal inference with panel data. *arXiv preprint arXiv:2511.21977*, 2025.
- Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.
- A. Colin Cameron and Pravin K. Trivedi. *Microeconomics: Methods and Applications*. Cambridge University Press, 2005.
- A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.
- Emmanuel J Candès. The power of convex relaxation: The surprising stories of matrix completion and compressed sensing. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1321–1321. SIAM, 2010.
- Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 2011.
- Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.
- Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition, 2006. doi: 10.1201/9781420010138.
- Matias D. Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021.
- Gary Chamberlain. Panel data. In *Handbook of Econometrics*. Elsevier, 1984.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. Exact and robust conformal inference methods for synthetic control and causal inference. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021.
- Judith A. Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

- Gregory C Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.
- Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.
- Carlos Cinelli, Avi Feller, Guido Imbens, Edward Kennedy, Sara Magliacane, and Jose Zubizarreta. Challenges in statistics: A dozen challenges in causality and causal inference. *arXiv preprint*, 2025. URL <https://arxiv.org/abs/2508.17099>.
- Darral G Clarke. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research*, 1976.
- Timothy G. Conley, Christian B. Hansen, and Peter E. Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.
- Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.
- Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021.
- Clément de Chaisemartin and Xavier d’Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.
- Clément de Chaisemartin and Xavier d’Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.
- Iván Díaz and Nima S Hejazi. Nonparametric causal effects based on incremental propensity score interventions. *Biometrics*, 76(4):1133–1142, 2020.
- Nikolay Doudchenko and Guido W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER*, 2016.
- Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1097–1104, 2011.
- Liran Einav and Jonathan Levin. Empirical industrial organization: A progress report. *Journal of Economic Perspectives*, 24(2):145–162, 2010.
- Gary Erickson and Robert Jacobson. Gaining stock market value through discretionary expenditures: The case of advertising and r&d. *Journal of Marketing Research*, 1992.
- Peter S. Fader, Bruce G. S. Hardie, and Ka Lok Lee. Counting your customers the easy way: An alternative to the pareto/nbd model. *Marketing Science*, 2005.
- Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221, 2021.
- Iván Fernández-Val and Martin Weidner. Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics*, 192(1):291–312, 2016.
- Christian Fong and Justin Grimmer. Discovery of treatments from text corpora. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1600–1609, 2016.
- Claes Fornell, Sunil Mithas, Forrest V Morgeson, and M S Krishnan. Customer satisfaction and stock prices: High returns, low risk. *Journal of Marketing*, 70:3–14, January 2006.
- Jonathan Fuhr and Dominik Papies. Double machine learning meets panel data – promises, pitfalls, and potential solutions. *arXiv preprint arXiv:2409.01266*, 2024.

- Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.
- Peter N Golder and Gerard J Tellis. Will it ever fly? modeling the takeoff of new consumer durables. *Marketing Science*, 16(3):256–270, 1997.
- Peter N. Golder, Debanjan Mitra, and Christine Moorman. What is quality? an integrative framework of processes and states. *Journal of Marketing*, 76(4):1–23, 2012. doi: 10.1509/jm.09.0416.
- Avi Goldfarb, Catherine Tucker, and Yuan Wang. Conducting research in marketing with quasi-experiments. *Journal of Marketing*, 86(3):1–20, 2022. doi: 10.1177/00222429221082977.
- Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.
- Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 2019.
- Giulio Grossi, Alessandra Mattei, and Georgia Papadogeorgou. Spatial vertical regression for spatial panel data: Evaluating the effect of the florentine tramway’s first line on commercial vitality. *arXiv preprint arXiv:2505.00450*, 2025.
- Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. *arXiv preprint arXiv:1906.07801*, 2020.
- Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15): e2014602118, 2021.
- Jinyong Hahn and Whitney Newey. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319, 2004.
- Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.
- Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.
- James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.
- John R Hauser, Gerard J Tellis, and Abbie Griffin. Research on innovation & new products: A review & agenda for marketing science. *Marketing Science*, 25(6):687–717, 2007.
- Jerry A. Hausman. Valuation of new goods under perfect and imperfect competition. In Timothy F. Bresnahan and Robert J. Gordon, editors, *The Economics of New Goods*. University of Chicago Press, 1996.
- Jerry A. Hausman and William E. Taylor. Panel data and unobservable individual effects. *Econometrica*, 49 (6):1377–1398, 1981.
- Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
- Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 73–84, 2004.
- Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960, 1986.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

- Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.
- Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the american statistical association*, 103(482):832–842, 2008.
- Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959, 2012.
- Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.
- Robert Jacobson and Natalie Mizik. The financial markets and customer satisfaction: Reexamining possible financial market mispricing of customer satisfaction. *Marketing Science*, 28(5):810–819, 2009.
- Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
- Masahiro Kato. Riesz regression as direct density ratio estimation. *arXiv preprint arXiv:2511.04568*, 2025a.
- Masahiro Kato. A unified theory for causal inference: Direct debiased machine learning via bregman-riesz regression. *arXiv preprint arXiv:2510.26783*, 2025b.
- Katherine Keith, David Jensen, and Brendan O’Connor. Text as data: A new framework for machine learning and the social sciences. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8663–8680, 2020.
- Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.
- Reinhold Kesler. The impact of apple’s app tracking transparency on app monetization. *Available at SSRN*, 2022.
- Frank Kleibergen and Richard Paap. Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1):97–126, 2006.
- Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
- Hans R Künsch. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241, 1989.
- Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- Retsef Levi, Elisabeth Paulson, Georgia Perakis, and Emily Zhang. Heterogeneous treatment effects in panel data. *arXiv preprint arXiv:2406.05633*, 2024.
- Randall A. Lewis and Justin M. Rao. The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics*, 130(4):1941–1973, 2015.
- Marvin B. Lieberman and David B. Montgomery. First-mover advantages. *Strategic Management Journal*, 9(S1):41–58, 1988.

- Yiqi Liu. Synthetic parallel trends. *arXiv preprint arXiv:2511.05870*, 2025.
- Malte Londschen. A statistician's guide to weak-instrument-robust inference in instrumental variables regression with illustrations in python. *arXiv preprint arXiv:2508.12474*, 2025.
- James G MacKinnon and Matthew D Webb. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254, 2017.
- Enno Mammen. Bootstrap, wild bootstrap and generalized bootstrap. *Probab. Th. Rel. Fields*, 1992.
- Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478.
- Philip Marx, Elie Tamer, and Xun Tang. Heterogeneous treatment effects via linear dynamic panel data models. *arXiv preprint arXiv:2410.19060*, 2024.
- Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.
- James Meldrum, Basem Suleiman, Fethi Rabhi, and Muhammad Johan Alibasa. New money: A systematic review of synthetic data generation for finance. *arXiv preprint arXiv:2510.26076*, 2025.
- Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.
- Natalie Mizik and Robert Jacobson. Valuing branded businesses. *Journal of Marketing*, 73(6):137–153, 2009.
- José Luis Montiel Olea and Carolin Pflueger. A robust test for weak instruments. *Journal of Business and Economic Statistics*, 37(4):638–649, 2019.
- José Luis Montiel Olea and Mikkel Plagborg-Møller. Simultaneous confidence bands: Theory, implementation, and an application to svrs. *Journal of Applied Econometrics*, 34(1):1–17, 2019.
- Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.
- Yair Mundlak. On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85, 1978.
- Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.
- Whitney K Newey. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of econometrics*, 36(3):231–250, 1987.
- Jerzy Neyman and Elizabeth L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.
- Stan Openshaw. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*, 38: 1–41, 1984.
- Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2 edition, 2009. ISBN 978-0-521-89560-6.
- Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.
- Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

- Vamsi K Potluru, Daniel Borrajo, Andrea Coletta, Niccolò Dalmasso, Yousef El-Laham, Elizabeth Fons, Mohsen Ghassemi, Sriram Gopalakrishnan, Vikesh Gosai, Eleonora Kreačić, et al. Synthetic data applications in finance. *arXiv preprint arXiv:2401.00081*, 2023.
- Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023.
- William S. Robinson. Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15(3):351–357, 1950. doi: 10.2307/2087176.
- Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.
- Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.
- David Roodman, Morten Ørregaard Nielsen, James G MacKinnon, and Matthew D Webb. Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60, 2019.
- Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.
- Jonathan Roth. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 2022.
- Jonathan Roth, Pedro H.C. Sant’Anna, Alyssa Bilinski, and John Poe. What’s trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244, 2023.
- Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.
- Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. doi: 10.2307/2287653.
- Pedro HC Sant’Anna and Qi Xu. Difference-in-differences with compositional changes. *Journal of Econometrics*, 253:106147, 2026.
- Pedro HC Sant’Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122, 2020.
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- Raj Sethuraman, Gerard J. Tellis, and Richard A. Briesch. How well does advertising work? generalizations from a meta-analysis of brand advertising elasticity. *Journal of Marketing Research*, 48(3):457–471, 2011. doi: 10.1509/jmkr.48.3.457.
- Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.
- Bradley T. Shapiro, Gunter J. Hitsch, and Anna E. Tuchman. Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica*, 2021.
- Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

- Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature human behaviour*, 4(11):1208–1214, 2020.
- Ashish Sood and Gerard J Tellis. Do innovations really payoff? total stock market returns to innovation. *Marketing Science*, 28(3):442–456, 2009.
- Asher Spector, Rina Foygel Barber, and Emmanuel Candes. Mosaic inference on panel data. *arXiv preprint arXiv:2506.03599*, 2025.
- Shuba Srinivasan and Dominique M Hanssens. Marketing & firm value: Metrics, methods, findings, & future directions. *Journal of Marketing Research*, 46(3):293–312, 2009.
- Shuba Srinivasan, Koen Pauwels, Jorge Silva-Risso, and Dominique M Hanssens. Product innovations, advertising, and stock returns. *Journal of Marketing*, 73(1):24–43, 2009.
- James H. Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression. In Donald W. K. Andrews and James H. Stock, editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pages 80–108. Cambridge University Press, 2005.
- Sebastiano Stramaglia, Tomas Scagliarini, Yuri Antonacci, and Luca Faes. Local granger causality. *Physical Review E*, 103(2):L020102, 2021.
- Liangjun Su and Xia Wang. Time-varying factor models with heterogeneous effects. *Journal of Econometrics*, 199(1):84–107, 2017.
- Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 2021.
- Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, and Xu Shi. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.
- Gerard J. Tellis and Philip Hans Franses. Optimal data interval for advertising response models. *Marketing Science*, 25(3):217–229, 2006. doi: 10.1287/mksc.1050.0168.
- Gerard J Tellis, Stefan Stremersch, and Eden Yin. The international takeoff of new products: Economics, culture & country innovativeness. *Marketing Science*, 22(2):188–208, 2003.
- Sesha Tirunillai and Gerard J Tellis. Mining marketing meaning from chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51(4), August 2014.
- Seshadri Tirunillai and Gerard J Tellis. Does chatter really matter? dynamics of user-generated content and stock performance. *Marketing science*, 31(2):198–215, 2012.
- Seshadri Tirunillai and Gerard J. Tellis. Does offline tv advertising affect online chatter? quasi-experimental analysis using synthetic control. *Marketing Science*, 36(6):862–878, 2017. doi: 10.1287/mksc.2017.1040.
- Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, 2 edition, 2009.
- Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.
- Aad W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.
- Jon Vaver and Jim Koehler. Measuring ad effectiveness using geo experiments. *Unpublished manuscript, Google Inc*, 2011.
- Victor Veitch, Dhanya Sridhar, and David M Blei. Adapting text embeddings for causal inference. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

- J Miguel Villas-Boas and Russell S Winer. Endogeneity in brand choice models. *Management science*, 45(10):1324–1338, 1999.
- Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- Meijia Wang, Ignacio Martinez, and P. Richard Hahn. Longbet: Heterogeneous treatment effect estimation in panel data. *arXiv preprint arXiv:2406.02530*, 2024.
- J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2010.
- Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.
- Zhiguo Xiao and Peikai Wu. Causal inference in panel data with a continuous treatment. *arXiv preprint arXiv:2506.23226*, 2025.
- Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.
- Zou Yang, Seung Hee Lee, Julia R. Köhler, and AmirEmad Ghassami. Causal data fusion for panel data without a pre-intervention period. *arXiv preprint arXiv:2410.16391*, 2024.
- Alwyn Young. Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, 134(2):557–598, 2019.
- Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.
- Anna Zaremba and Tomaso Aste. Measures of causality in complex datasets with application to financial data. *Entropy*, 16(4):2309–2349, 2014.
- Andrei Zeleniev and Weisheng Zhang. Tractable estimation of nonlinear panels with interactive fixed effects. *arXiv preprint arXiv:2511.15427*, 2025.
- Georgios Zervas, Davide Proserpio, and John W Byers. The rise of the sharing economy: Estimating impact of airbnb on the hotel industry. *Journal of Marketing Research*, 2018.
- Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014.
- Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965–993, 2019.
- José R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.



# Glossary

This glossary provides definitions for key terms used throughout this book. Terms are organised alphabetically within thematic categories.

## Causal Framework

**Average Treatment Effect (ATE)** The population-average difference between potential outcomes under treatment and control:  $\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$ . Represents the effect of treating a randomly selected unit.

**Average Treatment Effect on the Treated (ATT)** The average treatment effect among units that actually received treatment:  $\text{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$ . The primary estimand in most observational panel designs.

**Average Treatment Effect on the Untreated (ATU)** The average treatment effect among units that did not receive treatment:  $\text{ATU} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 0]$ . Relevant for policy extrapolation.

**Conditional Average Treatment Effect (CATE)** The treatment effect conditional on observable characteristics:  $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$ . Captures treatment effect heterogeneity.

**Counterfactual** The outcome that would have been observed for a unit under an alternative treatment assignment. Fundamentally unobservable; causal inference methods aim to construct credible estimates.

**Local Average Treatment Effect (LATE)** The average treatment effect for compliers in an instrumental variables design—units whose treatment status is affected by the instrument.

**Potential Outcomes** The set of outcomes a unit would exhibit under each possible treatment state. Denoted  $Y_i(d)$  for treatment level  $d$ , with  $Y_i(0)$  and  $Y_i(1)$  for binary treatments. The foundation of the Rubin Causal Model.

**Selection Bias** Systematic differences between treatment and control groups that confound the causal effect. Arises when treatment assignment is correlated with potential outcomes.

**SUTVA (Stable Unit Treatment Value Assumption)** The assumption that (i) each unit's outcome depends only on its own treatment, not others' treatments (no interference), and (ii) there is only one version of each treatment level (no hidden variations).

**Treatment Assignment Mechanism** The process determining which units receive treatment. May be randomised (experimental), as-if random (quasi-experimental), or confounded (observational).

## Panel Data Fundamentals

**Balanced Panel** A panel dataset where all units are observed for the same time periods. Contrast with unbalanced panels where observation periods vary across units.

**First Differences** A transformation that removes time-invariant unobserved heterogeneity by taking  $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$ . Alternative to fixed effects estimation.

**Fixed Effects (Unit)** Unobserved time-invariant characteristics of each unit, denoted  $\alpha_i$ . Absorbed by within-unit transformations or unit dummies.

**Fixed Effects (Time)** Common shocks affecting all units at each time period, denoted  $\lambda_t$ . Absorbed by time dummies or time-demeaning.

**Panel Data** Data with repeated observations on the same units over multiple time periods. The  $(i, t)$  structure enables identification strategies unavailable in pure cross-sections.

**Two-Way Fixed Effects (TWFE)** A regression model including both unit and time fixed effects:  $Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \varepsilon_{it}$ . The workhorse specification for panel causal inference.

**Within Transformation** The operation that demeans each observation by its unit mean:  $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$ . Removes unit fixed effects.

## Difference-in-Differences

**Canonical DiD** The classic  $2 \times 2$  design with two groups (treated/control) and two periods (before/after). Estimates ATT via double-differencing.

**Cohort** A group of units that adopt treatment at the same time. In staggered designs, each adoption timing defines a distinct cohort.

**Event Time** Time relative to treatment adoption, denoted  $k = t - G_i$  where  $G_i$  is unit  $i$ 's adoption date. Allows alignment of heterogeneous adoption timings.

**Forbidden Comparisons** In staggered DiD, problematic comparisons that use already-treated units as controls. Can bias TWFE estimates under heterogeneous effects.

**Never-Treated** Units that never receive treatment during the sample period. Often serve as the comparison group in staggered designs.

**Not-Yet-Treated** Units that eventually receive treatment but have not yet adopted at a given time. Can serve as controls for early adopters.

**Parallel Trends** The identifying assumption that treatment and control groups would have followed identical outcome trajectories in the absence of treatment.

**Pre-Trend Test** A diagnostic examining whether treatment and control groups exhibited similar trends before treatment. Failure suggests parallel trends may not hold.

**Staggered Adoption** A treatment design where different units adopt at different times. Common in policy rollouts and marketing interventions.

## Event Studies

**Dynamic Treatment Effects** Treatment effects that vary with time since treatment:  $\tau_k = \mathbb{E}[Y_{it}(1) - Y_{it}(0) | t - G_i = k]$ . Estimated via event-study specifications.

**Event-Study Plot** A graphical display of treatment effect estimates at each event time, typically with confidence intervals. Shows both pre-trends and dynamic post-treatment effects.

**Leads and Lags** Indicator variables for periods before (leads) and after (lags) treatment. Enable estimation of anticipation effects and dynamic responses.

**Reference Period** The omitted event-time category (typically  $k = -1$ ) that serves as the normalisation point in event studies. Effects are measured relative to this period.

## Synthetic Control Methods

**Convex Weights** Non-negative weights summing to one, constraining the synthetic control to be an interpolation (not extrapolation) of donor units.

**Donor Pool** The set of untreated units available for constructing a synthetic control. Selection requires substantive judgment about comparability.

**Pre-RMSPE** Root Mean Squared Prediction Error in the pre-treatment period. Measures synthetic control fit quality; used in placebo inference.

**Synthetic Control** A weighted combination of untreated units constructed to match the treated unit's pre-treatment trajectory. The estimated effect is the post-treatment gap.

**Synthetic Difference-in-Differences (SDID)** A hybrid method combining synthetic control weighting with DiD estimation. Provides doubly-robust identification and often outperforms either method alone.

## Factor Models and Matrix Methods

**Interactive Fixed Effects** A model where unit and time effects interact multiplicatively:  $Y_{it} = \sum_r \lambda_{ir} f_{rt} + \varepsilon_{it}$ . More flexible than additive TWFE.

**Latent Factors** Unobserved common factors  $f_t$  affecting all units with heterogeneous loadings  $\lambda_i$ . Capture patterns beyond additive fixed effects.

**Matrix Completion** Imputation of missing entries (counterfactuals) in the outcome matrix using low-rank structure. Nuclear norm minimisation is the convex relaxation.

**Nuclear Norm** The sum of singular values of a matrix,  $\|M\|_* = \sum_r \sigma_r$ . Regularisation with nuclear norm encourages low-rank solutions.

**Principal Component Analysis (PCA)** A method for estimating latent factors by extracting leading eigenvectors from the outcome covariance matrix.

## Inference and Uncertainty

**Cluster-Robust Variance Estimator (CRVE)** A variance estimator allowing arbitrary correlation within clusters (e.g., units) while assuming independence across clusters.

**Conformal Prediction** A distribution-free method for constructing prediction intervals with guaranteed finite-sample coverage under exchangeability.

**False Discovery Rate (FDR)** The expected proportion of false rejections among all rejections. Controlled by Benjamini-Hochberg and related procedures.

**Familywise Error Rate (FWER)** The probability of making at least one false rejection among a family of hypothesis tests. Controlled by Bonferroni and stepdown procedures.

**Randomisation Inference** Exact inference based on permuting treatment assignments under the sharp null hypothesis. Provides valid p-values without distributional assumptions.

**Wild Cluster Bootstrap** A resampling method for inference with few clusters. Perturbs residuals with random signs at the cluster level.

## Machine Learning Methods

**Cross-Fitting** A sample-splitting procedure in DML where nuisance functions are estimated on one fold and used to estimate treatment effects on another. Reduces overfitting bias.

**Double Machine Learning (DML)** A framework for causal inference with ML-estimated nuisance parameters. Combines Neyman orthogonality with cross-fitting for valid inference.

**Lasso** Least Absolute Shrinkage and Selection Operator. A penalised regression method ( $\ell_1$  penalty) that performs variable selection and regularisation simultaneously.

**Nuisance Parameter** A parameter required for estimation but not of primary interest. In causal inference, typically includes propensity scores and outcome models.

**Orthogonal Score** An influence function that is locally insensitive to perturbations in nuisance parameter estimates. Key to valid inference after ML selection.

**Ridge Regression** A penalised regression method ( $\ell_2$  penalty) that shrinks coefficients toward zero without setting them exactly to zero.

## Marketing Applications

**Adstock** The cumulative effect of advertising over time, accounting for carryover. Often modelled as geometric decay:  $A_t = \sum_{s=0}^{\infty} \lambda^s X_{t-s}$ .

**Attribution** The process of assigning credit for conversions to marketing touchpoints. Causal attribution requires valid identification strategies.

**Designated Market Area (DMA)** A geographic region used for television advertising measurement. Common unit of analysis in geo-experiments.

**Geo-Experiment** A quasi-experimental design where geographic regions are assigned to treatment and control conditions for marketing tests.

**Holdout** A randomly selected subset of units withheld from treatment to serve as a control group. The gold standard for incrementality measurement.

**Incrementality** The causal effect of a marketing intervention—the lift in outcomes attributable to the treatment beyond what would have occurred anyway.

**Marketing Mix Modelling (MMM)** Statistical analysis of sales and marketing inputs to estimate the effect of each marketing channel. Often uses aggregate time-series data.

**Return on Advertising Spend (ROAS)** The revenue generated per unit of advertising expenditure. Causal ROAS requires identification of advertising's incremental effect.

**Saturation** The phenomenon where additional advertising spending yields diminishing returns. Often modelled via Hill functions or logarithmic transformations.

**Switchback Experiment** A design where treatment alternates over time within units. Used in platform settings to handle interference and measure short-run effects.

## Threats and Diagnostics

**Anticipation Effects** Changes in outcomes before formal treatment implementation, arising when units anticipate treatment and adjust behaviour.

**Attrition** Loss of units from the sample over time. Differential attrition between treatment and control threatens validity.

**Composition Bias** Bias arising when treatment changes the composition of units being measured, rather than outcomes of fixed units.

**Interference** Violation of SUTVA where one unit's treatment affects another unit's outcome. Common in networks, platforms, and geographic spillovers.

**Overlap (Positivity)** The assumption that all units have positive probability of receiving each treatment level. Violation leads to extreme weights and extrapolation.

**Pre-Period Fit** The quality of counterfactual prediction in the pre-treatment period. Poor fit in synthetic control suggests potential bias.

**Sensitivity Analysis** Assessment of how conclusions change under violations of identifying assumptions. Quantifies robustness to unobserved confounding.

**Spillover** The indirect effect of treatment on untreated units, typically through geographic proximity, social networks, or market competition.

## Colour Key for Text Boxes

### Examples and Case Studies

Boxes with blue frames contain concrete examples, case studies, and worked problems. They illustrate how concepts apply to real-world marketing data.

### Practical Guidance

Boxes with green frames offer actionable advice, implementation checklists, and step-by-step workflows for practitioners.

### Warnings and Diagnostics

Boxes with orange or red frames highlight common pitfalls, critical distinctions, and diagnostic procedures to detect assumption violations.

### Reference and Notation

Boxes with gray frames provide notation guides, checklists, and formal definitions.



# Notation

## Basic Indices and Dimensions

$i$  Index for units (consumers, stores, markets, products).

$t$  Index for time periods (days, weeks, months, quarters, years).

$N$  Number of units in the panel.

$T$  Number of time periods.

$c$  Index for clusters (markets, DMAs, geo cells) when discussing clustered inference.

$G$  Number of independent clusters (distinct from  $G_i$  for adoption time). Used in asymptotic statements such as  $\sqrt{G}(\hat{\tau} - \tau_0)$ .

$\mathcal{C}_c$  Set of units in cluster  $c$ , where  $\mathcal{C}_c \subseteq \{1, \dots, N\}$ .

$s$  Index for strata or time blocks (for example, strata in clustered randomisation, day-of-week blocks in switchbacks).

$S$  Number of strata or blocks.

## Observables and Regression Components

$Y_{it}$  Observed outcome for unit  $i$  at time  $t$  (sales, clicks, revenue, churn indicator).

$D_{it}$  Treatment or exposure for unit  $i$  at time  $t$  (binary:  $D_{it} \in \{0, 1\}$ , or continuous:  $D_{it} \in \mathbb{R}$ ).

$X_{it}$  Column vector of observed covariates for unit  $i$  at time  $t$  (demographics, lagged outcomes, controls, seasonality).

$Z_{it}$  Instrumental variables for unit  $i$  at time  $t$  (cost shifters, policy shocks, randomised encouragements).

$\alpha_i$  Unit fixed effect capturing time-invariant heterogeneity (store quality, brand strength, persistent preferences).

$\lambda_t$  Time fixed effect capturing common shocks (seasonality, macro conditions, national campaigns). In factor models,  $\lambda_{ir}$  denotes unit-specific loadings for factor  $r$ .

$\varepsilon_{it}$  Idiosyncratic error term.

$\tau$  Treatment effect parameter (scalar in canonical TWFE). For functional forms, see  $\tau(g, t)$  and  $\tau(d)$  below.

$\gamma$  Coefficient vector for covariates  $X_{it}$  in regression models.

## Potential Outcomes

$Y_{it}(d)$  Potential outcome for unit  $i$  at time  $t$  under treatment level  $d$  (with  $Y_{it}(0)$  and  $Y_{it}(1)$  for binary treatments).

$\underline{d}_i^t$  Treatment history (dose path) for unit  $i$  up to time  $t$ :  $\underline{d}_i^t = (d_{i1}, d_{i2}, \dots, d_{it})$ . Used in dynamic potential outcomes  $Y_{it}(\underline{d}_i^t)$ .

$Y_i(d)$  Aggregate potential outcome for unit  $i$  under treatment level or path  $d$  (for example, a sum or average of  $Y_{it}(d)$  over a specified horizon).

$h_i(D_{-i,t})$  Exposure mapping for unit  $i$  summarising spillovers from other units' treatments. Spillover-aware potential outcomes are written as  $Y_{it}(d_{it}, h_i(D_{-i,t}))$ .

## Causal Estimands

**ATE** Average Treatment Effect:  $\text{ATE} = \mathbb{E}[Y_{it}(1) - Y_{it}(0)]$ , where the expectation is over an appropriate population of unit-period cells.

**ATT** Average Treatment Effect on the Treated:  $\text{ATT} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) | D_{it} = 1]$ .

**ATU** Average Treatment Effect on the Untreated:  $\text{ATU} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) | D_{it} = 0]$ .

**CATE** Conditional Average Treatment Effect:  $\text{CATE}(x) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) | X_{it} = x]$ , the average treatment effect conditional on covariates.

**LATE** Local Average Treatment Effect: the average treatment effect for compliers in instrumental-variables designs.

**LRM** Long-run multiplier for dynamic effects:  $\text{LRM} = \frac{\sum_{k=0}^K \theta_k}{\theta_0}$ , where  $K$  is chosen so that  $\theta_k$  has effectively dissipated.

## Staggered Adoption and Event Time

$G_i$  Treatment adoption time (cohort) for unit  $i$ : the first period in which the unit is treated in once-treated-always-treated designs.  $G_i = \infty$  denotes never treated.

$k$  Event time:  $k = t - G_i$ , where  $k = 0$  is the first treated period,  $k < 0$  are pre-treatment periods, and  $k > 0$  are post-treatment periods.

$D_{it}^k$  Event-time indicator:  $D_{it}^k = \mathbf{1}\{t - G_i = k\}$ , used in event-study specifications.

$\theta_k$  Event-time effect (dynamic response):  $\theta_k = \mathbb{E}[Y_{i,G_i+k}(G_i) - Y_{i,G_i+k}(\infty) \mid G_i < \infty]$ .

$\tau(g, t)$  Cohort-time effect in staggered adoption:  $\tau(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g]$  for  $t \geq g$ .

$\text{ATT}_t$  Time-specific Average Treatment Effect on the Treated:  $\text{ATT}_t = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid D_{it} = 1, t]$ .

## Dose-Response and Marginal Effects

$\mu(d)$  Average dose-response function:  $\mu(d) = \mathbb{E}[Y_{it}(d)]$ .

$\tau(d)$  Marginal effect in dose-response models:  $\tau(d) = \frac{\partial \mu(d)}{\partial d}$ , when defined.

## Weights and Propensity Scores

$w_{it}$  Observation weight for unit-time cell  $(i, t)$  (for example, in weighted estimators or bootstrap schemes). Lowercase  $w$  always denotes weights, not treatments.

$w_{gt}$  Aggregation weights for cohort-time cells  $(g, t)$  in staggered difference-in-differences.

$e(X_{it})$  Propensity score for binary treatment:  $e(X_{it}) = \mathbb{P}(D_{it} = 1 \mid X_{it})$ .

$r(d \mid X_{it}, \alpha_i, \lambda_t)$  Generalised propensity score for continuous or multi-valued treatment: the conditional density or mass function  $f_{D|X,\alpha,\lambda}(d \mid X_{it}, \alpha_i, \lambda_t)$ .

## Factor Models and Matrix Methods

$f_{tr}$  Latent factor  $r$  at time  $t$  in interactive fixed-effects and factor models.

$\lambda_{ir}$  Unit-specific loading on factor  $r$  for unit  $i$  (distinct from time fixed effect  $\lambda_t$ ).

$R$  Number of latent factors in factor models, with  $R \ll \min(N, T)$ .

$\mathbf{Y}$  Outcome matrix of dimension  $N \times T$  (bold uppercase for matrices).

**D** Treatment matrix of dimension  $N \times T$ .

## Inference and Influence Functions

**$\Psi_i$**  Unit-level influence function contribution for unit  $i$ .

**$\Psi_c$**  Cluster-level influence contribution:  $\Psi_c = \sum_{i \in C_c} \Psi_i$ .

**CRVE** Cluster-robust variance estimator, used to account for within-cluster dependence.

**HAC** Heteroskedasticity-and-autocorrelation-consistent standard errors.

## Generic Sets and Structures

**C** Generic index set of control unit-time cells, used to define placebo and variance estimators (chapter-specific variants are defined where used).

**$\mathbf{1}\{\cdot\}$**  Indicator function:  $\mathbf{1}\{A\} = 1$  if condition  $A$  is true, and 0 otherwise.

# Index

## A

ad-stock 473  
adoption time 174  
advertising 215, 793  
algorithmic confounding 707  
Anderson–Rubin inference 759  
anticipation 175, 215  
 $\text{ATT}(g,t)$  *see* group-time effects  
attribution 883  
augmented synthetic control 289–290  
average treatment effect on the treated (ATT) 122

## B

balance 761  
bias-variance trade-off 290  
binning 181  
block bootstrap 742  
bootstrap 741–742

## C

carryover 175  
Carryover effects 106  
CATE 559–608  
clustering 202  
two-way 202  
composition bias 174  
confidence interval 735  
conformal inference 248, 735, 745–746  
continuous treatment 659–701  
convex hull condition 230  
convex weights 230  
cross-fitting 559, 750

Curse of dimensionality 618

## D

decay half-life 216  
delta method 747  
diagnostics 761–787  
difference-in-differences 119–171  
distributed lag 473  
DMA (designated market area) 232  
donor pool 228  
dose-response 559, 659  
double machine learning 559–608  
inference 750–751  
doubly robust 291  
Dynamic panels 106  
dynamic treatment effects 471–513

## E

elastic net 611  
estimators  
Borusyak–Jaravel–Spiess 146  
Callaway–Sant’Anna 144, 181  
Sun–Abraham 145, 181  
event study 173–224  
event time 174

## F

factor loadings 352  
factor model 229, 351–408  
false discovery rate 752  
familywise error rate 752  
Feedback effects 106  
Fisher randomisation test 743

**G**

gap plot 228  
 geo-experiment 518  
 group-time effects 181

**H**

half-life 224  
 heterogeneity-robust estimators 143, 224  
 heteroskedasticity 737  
 high-dimensional controls 611–658  
 hybrid methods 289

**I**

Identification  
 Strategy-robust 958  
 identified set 243  
 impulse response 472  
 imputation 146  
 Incidental parameters problem 671  
 inference 735  
 instrumental variables 754–758, 796  
 Interactive Fixed Effects  
 Nonlinear 448  
 interactive fixed effects 353  
 interference *see* spillovers, 931

**K**

Kleibergen-Paap statistic 759

**L**

lasso 611  
 latent factors 352  
 leads and lags 181  
 leave-one-out 761  
 local exchangeability 739  
 long-run multiplier 472  
 loyalty programme 215, 793

**M**

marketing applications 793–882  
 Matrix Completion  
 Regularisation 354  
 matrix completion 351–408  
 measurement error 707

missing data 883  
 mosaic permutation 739–740  
 multiple testing 735, 752–753

**N**

negative weights 140  
 never-treated 134  
 Neyman orthogonality 559  
 Neyman-Scott bias 671  
 no anticipation 135, 240  
 no interference 240  
 Nonlinear Panels

  Interactive Fixed Effects 448  
 nonstationarity 931  
 not-yet-treated 134  
 nuclear norm 354

**O**

overlap 761

**P**

panel data 883–928  
 parallel trends 121, 176  
 conditional 125  
 partial identification 748  
 permutation test 246  
 placebo check  
 in-space 246  
 placebo diagnostic 761  
 platform 232, 793  
 policy learning 559  
 Potential Outcomes  
 SUTVA 9  
 pre-trend diagnostic 176  
 pre-trends 761  
 prediction intervals 745  
 pricing 793  
 privacy 883

**R**

ramp-up 224  
 randomisation inference 203, 735, 743–744  
 reference bin 181  
 Regularisation  
 Bias-variance tradeoff 230

Complexity penalisation 564  
Nuclear norm 354  
Ridge regression 230  
Sparsity assumption 618  
regularisation 291  
regularised synthetic control 290  
RMSPE 247  
Robustness  
  Multiply-robust 958  
  Triple-robustness 958  
Romano-Wolf procedure 753

**S**

seasonality 707  
simultaneous confidence bands 747–749  
specification curve 761  
spillovers 517–556  
  competitive 517  
  geographic 517  
  network 517  
staggered adoption 127  
standard errors

cluster-robust 735  
store 216  
SUTVA 121  
synthetic control 227–288  
synthetic difference-in-differences 290  
Synthetic Parallel Trends 243

**T**

tensor completion 409, 412  
threats to validity 707–729  
TWFE *see* two-way fixed effects  
two-way fixed effects 139, 181

**V**

variance estimation 737–738

**W**

weak instruments 759–760  
wild cluster bootstrap 735, 741  
window selection 181  
word-of-mouth 518