UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Methodological Foundations of the Synthetic Control Group Design: Formalizing Model Construction for Case Study Applications in the Social Sciences

Permalink

https://escholarship.org/uc/item/5gc1f373

Author

Bartos, Bradley Jerome

Publication Date

2020

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NoDerivatives License, available at https://creativecommons.org/licenses/by-nd/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Methodological Foundations of the Synthetic Control Group Design: Formalizing Model Construction for Case Study Applications in the Social Sciences

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Criminology, Law and Society

by

Bradley Jerome Bartos

Dissertation Committee:
Professor Richard D. McCleary, Chair
Associate Professor Bryan L. Sykes
Professor Charis E. Kubrin

2020

© 2020 Bradley Jerome Bartos

DEDICATION

To John and Susan Bartos,

Your patience, emotional support, and understanding kept my head above rising water for many years. Thank you for helping me back up after my many falls.

To Natalie,

I could not have accomplished this, or anything else, without your unwavering support and belief in me. Much like Dr. McCleary, you believed so strongly in my potential that it became a source of inspiration and motivation itself. Behind every great career, there is a supportive partner. It has been the greatest privilege I have ever experienced to live our lives together, and I am thrilled to be there for you, through success and struggle, just as you have for me.

"When you are old and grey and full of sleep, and nodding by the fire, take down this book…"

-W B Yeats

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	vi
ACKNOWLEDGEMENTS	vii
VITA	ix
ABSTRACT OF THE DISSERTATION	xv
CHAPTER 1: INTRODUCTION	1
Overview of Dissertation Structure	7
CHAPTER 2: Causal Inference from Observational Data	10
The "Gold Standard" of Experimentation	11
Campbell's Four Validities and Quasi-Experiments	12
Rubin's Causal Model	16
Econometric Methods for Causal Inference from Quasi-Experiments	18
CHAPTER 3: The Synthetic Control Group Design	21
Synthetic Control Group Algebra	23
Model Construction	26
CHAPTER 4: The Theory of Standardization	33
Population Rates	39
Opportunity Rates	41
Behavioral Rates	42
CHAPTER 5: To Fit? Or to Match? Covariates and the Influence of Pre-intervention Observations	59
Kaul et al.'s (2015) Position	60
Reanalysis of Abadie and Gardeazabal (2003) Basque	63
Reanalysis of Abadie et al. (2010) Proposition 99	69
Reanalysis of Abadie et al. (2015) German Reunification	76
Discussion	85
CHAPTER 6: Conclusions	121
REFERENCES	127
APPENDIX A: Chapter 4 Supplementary Figures and Tables	135
APPENDIX B: Chapter 5 Supplementary Figures and Tables	140

LIST OF FIGURES

		Page
Figure 4.1	Motor vehicle fatalities for 50 U.S. states reported in the National Highway Transportation Safety Administration's Fatality Analysis Reporting System	38
Figure 4.2	Motor vehicle fatalities "outliers" highlighted and labelled	38
Figure 4.3	Motor vehicle fatalities: Wyoming and California units of interest labelled	38
Figure 4.4	Motor vehicle fatalities scaled per 100,000 residents	40
Figure 4.5	Motor vehicle fatalities scaled per registered vehicle	40
Figure 4.6	Motor vehicle fatalities scaled per vehicle mile travelled	42
Figure 4.7	Motor vehicle fatalities scaled per 100,000 vehicle miles travelled	42
Figure 4.8	Synthetic Wyoming plot: Motor vehicle fatalities (raw frequencies)	42
Figure 4.9	Synthetic California plot: Motor vehicle fatalities scaled per 100,000 residents	43
Figure 4.10	Synthetic California plot: Motor vehicle fatalities scaled per registered vehicle	43
Figure 4.11	Synthetic California plot: Motor vehicle fatalities scaled per 100,000 vehicle miles travelled	44
Figure 4.12	Synthetic Wyoming plot: Motor vehicle fatalities scaled per 100,000 vehicle miles travelled	44
Figure 5.1	Synthetic Control Group Results Plot: Weights reported in Abadie & Gardeazabal (2003)	65
Figure 5.2	Abadie & Gardeazabal (2003) Replication model Compared to Terrorist Activity Series	66
Figure 5.3	Synthetic Control Group Results Plot: Gardeazabal (2003) covariates only, no pre-intervention observations of GDP per capita	66
Figure 5.4	Synthetic Control Group Results Plot: 2nd Order Taylor Expansion Only (no covariates)	66
Figure 5.5	Synthetic Control Group Results Plot: 2nd Order Taylor Expansion plus Abadie & Gardeazabal (2003) covariates	66
Figure 5.6	Synthetic Control Group Gap Plot: Abadie and Gardeazabal (2003) vs covariates only	68

Figure 5.7	Synthetic Control Group Gap Plot: Pre-intervention DV observations, 1st, and 2nd Order Differences (No covariates) vs Abadie and Gardeazabal (2003)	68
Figure 5.8	Synthetic Control Group Gap Plot: Pre-intervention DV observations, 1st, and 2nd Order Differences and original covariates vs Abadie & Gardeazabal (2003)	68
Figure 5.9	Synthetic California Plot: Weights Reported in Abadie Diamond and Hainmueller (2010)	70
Figure 5.10	Synthetic California Plot: Produced using Stata synth_ado replication code	70
Figure 5.11	Synthetic California Plot: Covariates only	71
Figure 5.12	Synthetic California Plot: Pre-1989 Cigarette Sales Observations and Differences Only	72
Figure 5.13	Synthetic California Plot: 2nd Order Taylor Series plus Covariates	73
Figure 5.14	Synthetic California Plot: Optimal Model using Pre-1989 sales observations differences and covariates	74
Figure 5.15	Synthetic California Gap Plot: Abadie et al. (2010) vs. Covariates only model	74
Figure 5.16	Abadie et al. (2010) vs. Pre-1989 Cigarette Sales Obs & Differences Only	74
Figure 5.17	Synthetic California Gap Plot: Abadie et al. (2010) vs. Synthetic California produced using Stata synth_ado code	75
Figure 5.18	Synthetic California Gap Plot: Abadie et al. (2010) vs. Optimal Synthetic California	75
Figure 5.19	Synthetic West Germany Plot: Weights Reported in Abadie Diamond and Hainmueller (2015)	78
Figure 5.20	Synthetic West Germany Plot: Produced using replication code in R from Harvard Dataverse Repository	80
Figure 5.21	Synthetic West Germany Plot: Produced using Stata translation of replication code	80
Figure 5.22	Synthetic West Germany Plot: Synthetic West Germany Plot: Covariates only	80
Figure 5.23	Synthetic West Germany Plot: Pre-1990 reunification GDP Observations and Differences Only	81
Figure 5.24	Synthetic West Germany Plot: Covariates and half-set of Pre-reunification GDP Observations	82
Figure 5.25	Synthetic West Germany Plot: Pre-1990 reunification GDP Observations, 1st-Differences, and Covariates	83

LIST OF TABLES

		Page
Table 4.1	Non-Zero Weights for Synthetic California Time Series	46
Table 5.1	Basque Predictors	65
Table 5.2	Synthetic Control Composition, Fit Statistics, and Intervention Estimates	66
Table 5.3	Predictor Values Compared to Original Model	69
Table 5.4	California Predictors	71
Table 5.5	Synthetic California Composition, Fit Statistics, and Intervention Estimates	72
Table 5.6	California Predictors (Alternative Models)	75
Table 5.7	Covariate Balance between West Germany, OECD sample mean, and Alternate compositions of Synthetic West Germany	79
Table 5.8	Synthetic West Germany Composition, Fit Statistics, and Intervention Estimates	79

ACKNOWLEDGEMENTS

I feel incredibly fortunate to have received such focused, thoughtful, and supportive mentorship from a small handful of faculty, staff, and fellow graduate students in the Criminology, Law and Society Department. Over the years, a rotating cast of characters have stepped in to make invaluable contributions to my professional and personal development, but a select group of faculty have truly changed the course of my life.

On my recruitment visit to UC-Irvine, I was lucky to meet Dr. Richard McCleary, and by the end of my first year at Irvine, Dr. McCleary had taken me under his wing, hired me as a research assistant, and invited me to join a writing project that would eventually materialize into a book. Beyond these conveniently tangible aspects of his mentorship, Dr. McCleary invested the most valuable resource he had in me, time. In my first few years at Irvine, we would spend 2+ hours at the start of each day meeting with each other, writing together on our book project, and I picked up as much of his wisdom as I could. After a year or two, Dr. McCleary threw me into the fire (in the best possible way) and had me lead pitch meetings about our population projections modelling to potential clients. This is just one example of how Dr. McCleary prioritized my growth and development as a scholar above his own short-term gain, but a full listing of those moments could fill another 100 pages. Yes, Dr. McCleary was a generous, attentive, and active mentor to me, but he was also the epitome of honesty (whether I wanted to hear it or not), compassion, and utter determination. In many ways, I owe Dr. McCleary an immense debt. A debt I intend to repay by continuing the research lines he brought me on to, just as Dr. McCleary did for Thomas Cook and Donald Campbell. I could spend the full-length of this dissertation outlining just how important my time with Dr. Richard McCleary has been to my holistic development, but put simply, Dr. McCleary changed my life, and inspired me to rise to the potential he saw in me. Dr. McCleary is like family to me now, a "west-coast father", having spent thousands of hours together working and supporting each other since 2013. I cannot begin to imagine where I would be without Dr. McCleary, and I only hope he still feels that I am worth his investment.

My faculty advisor, Dr. Bryan Lamont Sykes, has been another source of well-intentioned honesty, early career wisdom and guidance among many others. Dr. Sykes worked with Dr. McCleary and I on a number of projects, and his candor forthright advice led to more and more collaborations between Bryan and I. Perhaps most notably, Dr. Sykes helped me navigate a major personality conflict in my second year in a diplomatic manner, after which Dr. Sykes became my faculty advisor. Since then, he has been an unending well of information on topics such as publishing, early career scholar strategy, the tenure process, and he aided my job search greatly. He has always done whatever he can to support graduate students, including hiring me on as managing editor in 2019.

Dr. Charis Kubrin and I began collaborating on writing projects in the third or fourth year of my doctoral career, and since that moment, we have not stopped. Dr. Kubrin funneled all of the political backlash and criticism from our first co-authored paper through her, protecting me from

the onslaught while exposing me to the experience. She would call me into her office to show me the emails she received, and then write out her replies with me. This gave me a no-risk opportunity to gain experience interacting with journalists, members of the news media, and members of the public. Dr. Kubrin and I would draft op-eds, press releases, and response articles together, so I not only saw the finished product; I was integrated into her process. Dr. Kubrin clearly cared about my professional success as well as my family's well-being. Working together has been a genuine pleasure and we will surely continue to collaborate.

A handful of fellow graduate students in the Criminology, Law and Society Department also deserve special thanks. Matthew Lazarus Renner has been one of my closest collaborators and truest friends since we met in 2014, and I do not know if some of our consulting work would have been possible to complete without Renner on the team. Matt and I serve as each other's sounding board, often playing devil's advocate, free from the chilling effects of judgement or pride. Almost all of the ideas presented in my published works were first presented to and scrutinized by Matthew Renner, and my work has certainly benefitted from his input.

In addition to being some of my closest friends and comrades, Michelle Mioduszewski, Carol Newark, Christopher J. Bates, and Ben Forthun made valuable contributions the Irvine Simulation Modeling Lab and our population projections work. I must also thank Mary Underwood for her administrative support throughout my time at UCI. Mary is far from the only staff member worthy of thanks, but she is the only one that has been a friendly resource from day 1 through day 2,468.

VITA

BRADLEY J. BARTOS
Department of Criminology, Law and Society
University of California, Irvine

EDUCATION

Ph.D.	**University of California, Irvine** Criminology, Law & Society	Spring 2020 (Expected)
	Dissertation: Methodological Foundations of the Synthetic Control Group Design: Formalizing Model Construction for Case Study Applications in Social Science	
	Committee: Richard D. McCleary (chair), Bryan L. Sykes and Charis E. Kubrin	
M.A.	**University of California, Irvine** Social Ecology	2016
	Purdue University	2013
B.S.	Psychological Sciences	
B.A.	Law & Society	

SCHOLARLY INTERESTS

Time Series Analysis, Causal Inference, Quasi-Experiments, Synthetic Control Group Designs, Mass Incarceration, Discrete-Event Simulation, Forensic Populations, Drug and Firearm Policies

PUBLICATIONS

Books

McDowall, D., McCleary, R., & **Bartos, B.J.** (2019). *Interrupted Time Series Analysis*. Oxford University Press.

McCleary, R., McDowall, D., & **Bartos, B. J.** (2017). *Design and Analysis of Time Series Experiments*. Oxford University Press.[1]

[1] Kenneth Land recently published a review of DATSE:

Land, K. C. (2019). Book review: McCleary, R., McDowall, D., & Bartos, B. J. (2017). Design and Analysis of Time Series Experiments. Oxford University Press. *Journal of Experimental Criminology*, *15*(4), 719–723. https://doi.org/10.1007/s11292-019-09386-x

Articles Published in Refereed Journals

Kubrin, C. E., & **Bartos, B. J.** (2020). Sanctuary Status and Crime in California: What's the Connection?. *Justice Evaluation Journal*, 1-19.

Bartos, B. J., McCleary, R., Mazerolle, L., & Luengen, K. (2020). Controlling Gun Violence: Assessing the Impact of Australia's Gun Buyback Program Using a Synthetic Control Group Experiment. *Prevention Science*, *21*(1), 131–136**.** https://doi.org/10.1007/s11121-019-01064-8

Bartos, B.J. & McCleary, R. D., (2019) Tautology in Criminological Theory. Available at: https://ssrn.com/abstract=3418629

Bartos, B. J., Kubrin, C. E., Newark, C., & McCleary, R. (2019). Medical Marijuana Laws and Suicide. *Archives of Suicide Research*, 1–14. https://doi.org/10.1080/13811118.2019.1612803

Bartos, B. J., Newark, C., & McCleary, R. (2018). Marijuana medicalization and motor vehicle fatalities: A synthetic control group approach. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-018-9345-3

Bartos, B. J., & Kubrin, C. E. (2018). Can we downsize our prisons and jails without compromising public safety? Findings from California's Prop 47. *Criminology & Public Policy*, *17*(3).

Bartos, B.J., Renner, M., Newark, C., McCleary, R., & Scurich, N. (2017). Characteristics of forensic patients in California with dementia/Alzheimer's disease. *Journal of Forensic Nursing*.

Renner, M., Newark, C., **Bartos, B. J**., McCleary, R., & Scurich, N. (2017). Length of stay for 25,791 California patients found incompetent to stand trial. *Journal of Forensic* & *Legal Medicine*, *51*, 22–26.

Bartos, B.J., & McCleary, R. (2015). Generic Discrete-Event Simulation Model of a Prison. Available at SSRN: http://ssrn.com/abstract=2622783

Technical Reports

Renner, M., **Bartos, B. J.,** Newark, C., & McCleary, R. (2018). *6 Month validation and accuracy assessment for UCI DSH projection model version 2.0.* Report submitted to California Department of State Hospitals.

Bartos, B. J., Newark, C., Renner, M., &McCleary, R. (2015). *The DSH prototype model: year-one report.* Report submitted to California Department of State Hospitals.

Bartos, B. J., Newark, C., Renner, M., &McCleary, R. (2015). *The LASD prototype model: year-one report.* Report submitted to Los Angeles County Sheriff's Department.

McCleary, R., Erlanger, C., & **Bartos, B.J.** (2014). *SOMS model version 1.0: 48-month population Projections.* Report submitted to the California Department of Corrections and Rehabilitation Office of Research.

Works In-Progress

Bartos, B. J., McCleary, R., McDowall, D. (Drafting). Synthetic Control Designs: Methodological Foundations for Social Science Applications. Oxford University Press.

Bartos, B. J., McCleary, R., Renner, M. (Drafting). Projecting Institutional Populations through Discrete Event Simulation: An Application of SAS Simulation Studio to State-level Forensic Mental Health Systems. SAS Press.

Bartos, B. J., McCleary, R., Kubrin, C. The Greek Sovereign Debt Crisis, austerity measures, and Suicide rates among vulnerable groups: A synthetic control approach.

Bartos, B. J. The diminishing returns of incarceration: Evidence from California's Substance Abuse and Crime Prevention Act (SACPA).

Other Publications

Kubrin, C., & **Bartos, B. J.** (2018). The Myth that Crime Rises as Prisons Shrink. *Governing Magazine*. Retrieved from https://www.governing.com/gov-institute/voices/col-california-proposition-47-downsizing-prisons-crime-rates.html

PROFESSIONAL PRESENTATIONS

Bartos, B. J. (2019). *Scaling Transformations and Standardization: Implications for Time-series Matching Methods.* Presented at the American Society of Criminology Annual Conference, San Francisco, CA, November 14.

Bartos, B. J., Mazerolle, L., & McCleary, R. (2018). *Life and Death Decisions: Estimating the Causal Impact of Australia's Gun Buyback Program on Homicide and Suicide.* Presented at the American Society of Criminology – 74th Annual Meeting, Atlanta, GA, November 14.

Bartos, B. J., Renner, M., & McCleary, R. (2018). *Discrete-event simulation models: Forensic Applications.* Presentation to the National Association of State Mental Health Program Directors annual Board of Directors meeting, Washington D.C., July 24.

Bartos, B. J., Mioduszewski, M., Renner, M., & McCleary, R. (2017). *An application of discrete event simulation for planning and resource allocation in a state hospital system servicing both criminal and civil commitments* (pp. 4509–4511). Presented at the Winter Simulation Conference Hosted by IEEE.

Bartos, B. J. (2017). *Synthetic control group methodology and the effects of marijuana medicalization in California*. Presented at the American Society of Criminology – 73rd Annual Meeting, Philadelphia, PA, November 16.

Bartos, B. J. (2017). *The Effect of Marijuana Medicalization in California on Motor Vehicle Fatalities.* Presented at the Western Society of Criminology Annual Conference, Las Vegas, NV, February 10-11.

Bartos, B. J. (2016). *Medical Marijuana and Motor Vehicle Fatalities: A Synthetic Control Approach to California's Compassionate Use Act.* Presented at the Orange County Long Beach Chapter of the American Statistical Association's Annual Poster Symposium, Irvine, CA, December 10.

Bartos, B. J and McCleary, R. (2016). *Medical Marijuana and Motor Vehicle Fatalities: A Synthetic Control Approach to California's Compassionate Use Act.* Presented at the American Society of Criminology – 72nd Annual Meeting, New Orleans, LA, November 18.

Bartos, B. J., & McCleary, R. (2016). *Marijuana Medicalization in California: Public Health & Safety Outcomes.* Presented at the Association for Criminal Justice Research- California Annual Fall Conference, Irvine, CA, October 20-22.

Bartos, B. J. (2015). The Impact of California's Public Safety Realignment Act (PSRA) on Crime Rates: A Synthetic Control Time Series Experiment. In panel titled: New Developments in Statistical Modeling and Causal Analysis. Presented at the American Society of Criminology - 71st Annual Meeting, Washington, DC.

Bartos, B. J, Erlanger, C., & McCleary, R. (2015). *Unintended Consequences of California's Public Safety Realignment.* Presented at the Western Society of Criminology Annual Conference, Phoenix, AZ, February 19-21.

Bartos, B. J. (2014). *Understanding the Impact of California's Realignment of Corrections on Race and Ethnic Minority Group Offenders.* Presented at the American Society of Criminology Annual Conference, San Francisco, CA, November 17-20.

RESEARCH APPOINTMENTS

2014-2019 Model Architect, Graduate Student Researcher
California Department of State Hospitals Population Projections Contract.
With Richard D. McCleary and Bryan L. Sykes.

2014-2017 Model Architect, Graduate Student Researcher
Los Angeles County Sheriff's Department Population Projections Contract.
Principal Investigator: Richard D. McCleary.

2013-2014 Graduate Student Researcher
California Department of Corrections Population Projections Contract.
With Richard D. McCleary, Susan F. Turner, and Bryan L. Sykes.

GRANTS, FELLOWSHIPS, AWARDS, AND STIPENDS

DSH Research Contract- 2015- 2019
LASD Research Contract - 2015- 2017
CDCR Research Contract - 2014- 2015

Social Ecology Dean's Dissertation Fellowship- 2019
Martha Newkirk Award for Excellence in Graduate Student Research- 2018-19
Departmental Commendation (annual performance) - 2014, '16, '17, '18, '19
Kitty Calavita Award for outstanding pre-dissertation research project- 2016
Graduate Student Mentoring Award- 2014
Martha Newkirk Fellowship- 2013

TEACHING EXPERIENCE

Spring 2019 "Statistical Data Analysis II (Regression/Econometrics)." With John R. Hipp.
Graduate course with two labs. Statistical software used: Stata.

Spring 2017 "Statistical Data Analysis II (Regression/Econometrics)." With John R. Hipp.
Graduate course with two labs. Statistical software used: Stata.

Spring 2014 "Criminological Theory." With Bryan Burton.
Undergraduate Course with two discussion sections.

Winter 2014 "Introduction to Criminology, Law & Society." With William Thompson.
Undergraduate Course with ~400 enrolled students.

PROFESSIONAL SERVICE

Editorial Service
2019 - present Managing Editor- *Sociological Perspectives*

Referee Activities
Criminology
Journal of Quantitative Criminology
Justice Quarterly
Accident Analysis and Prevention

Experimental Criminology *Punishment & Society*
Criminology & Public Policy

Departmental Service and Research Lab Affiliations

2015 - 2016 Graduate Student Representative
2015 - present Irvine Simulation & Modeling Laboratory

ABSTRACT OF THE DISSERTATION

Methodological Foundations of the Synthetic Control Group Design: Formalizing Model Construction for Case Study Applications in the Social Sciences

by
Bradley Jerome Bartos
Doctor of Philosophy in Criminology, Law and Society
University of California, Irvine, 2020
Professor Richard McCleary, Chair

To estimate the effect of an exogenous intervention on a treated unit, a control unit is necessary. Control time-series are chosen to render threats to internal validity implausible. While a literature review can often reveal the most plausible threats to the internal validity for a particular research design or substantive outcome, literature reviews rarely suggest an appropriate control time-series. Sometimes, appropriate control time series simply do not exist. When an appropriate control time-series is not available, it may be possible to construct a close approximation of an *ideal* control series from a set of control units that are less than appropriate individually. I describe the statistical theory underlying the synthetic control group design, outline the modeling procedure to construct a synthetic control group, and identify aspects of the model construction process that are in need of formalization. The importance of choosing an appropriate scaling method for the outcome of interest, and the inclusion of pre-intervention observations versus covariates as predictors are discussed at length in Chapters 4 and 5. Beyond those prescriptive conclusions, the early chapters trace the development of causal reasoning, contextualize synthetic control methods within the econometric wave, and serve as an algebraic reference. The dissertation concludes with a discussion of some of some remaining unknowns regarding synthetic control applications, including predictor "trimming", fuzzy onset treatments, and restricting the analysis timeframe. Guidelines for social science applications of the synthetic control method are outlined, and readers are cautioned to avoid applying the method in a few specific circumstances.

Chapter 1 - Introduction

"Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analysis seriously." – Edward Leamer (1983)

Natural Experiments, Quasi-experiments, and Case Studies

Over the past decade, social scientists have grappled, often publicly, over the relative standing of knowledge produced from randomized controlled trial experiments (RCTs) versus observational studies. While RCTs are widely regarded as the "gold standard" of causal inference by virtue of the rigor applied to minimize threats to internal validity, effective policymaking can be undermined by our failure to consider issues of external validity as well.

The essential features of RCTs include random assignment of participants to treatment/control conditions, blinding researchers to participants' treatment/control group status, and a controlling the study setting. Random assignment of participants to treatment and control groups minimizes the potential bias stemming from participants self-selecting into the treatment or control group due to the presence of a confounding factor that conditions their response to the treatment (i.e. confounding or allocation bias). Similarly, if a researcher, knowingly or unwittingly, assigns participants to treatment or control groups due to some internal preference, the resulting treatment and control samples may not reflect the population from which they were sampled (i.e. selection bias). The bias-reduction strategies employed in modern RCTs are far too sophisticated and numerous than the narrow subject of this dissertation. However, it is important to note the value that random treatment assignment contributes to RCTs in order to assess the consequences of non-random treatment assignment for quasi-experimental research designs. Without random assignment, treatment assignment cannot be assumed to be independent from treatment response.

Random assignment of participants to treatment and control groups is impractical or even impossible in many cases, however. Although the ethical and logistical obstacles to RCTs in criminological research are numerous, Weisburd (2010) argues that these barriers are overstated and too readily accepted as fact. Weisburd may be rightfully seen as one of the leaders of criminology's "experimental turn", as coined by Sampson (2010). Immodest claims that experiments enhance scientific quality, evidence-based policy, and causal inference, led Angus Deaton (2009) to coin the term "*Randomistas*" to describe economists and social scholars that adhered to the belief threat RCTs are the best approach to sustained scientific progress. Deaton's "*Randomistas*" label conveys the degree of condescension with which these arguments were presented.

Very few policy interventions or experimental treatments that interest researchers in criminology, development economics, political science and public health, are practical, ethical, or politically acceptable to randomly assign in practice. Weisburd's (2010) suggestion that ethical barriers to experimentation are less common in criminology than practical constraints, such as funding limitations and uncooperative practitioners, may be an accurate and constructive insight. However, these constraints are no less problematic than the ethical obstacles overstated. Regardless of whether experimental research in criminal justice has been inhibited by ethical concerns versus political, financial, or academic structure constraints (such as the risk experimental studies present to junior faculty if, heaven forbid, the endeavor does not yield a large number of publications); it remains true that experimental evaluations of criminal justice policies are frequently impractical or impossible.

Weisburd's insight, that evaluation researchers cite ethical barriers to experimentation less frequently than other practical constraints, fails to recognize that political, financial, and other constraints that restrict the opportunities for experimental criminological research are no easier to

resolve than ethical constraints. Whether the random assignment of a welfare program designed to reduce poverty is ethically unacceptable to scientists or politically unacceptable to stakeholders; researchers are left in the similar position of abandoning the endeavor or adopting a, dare I say *lesser*, research methodology that does not assume random assignment. In addition to the ethical and political obstacles to random assignment, anticipation effects, spillover (i.e. contamination of controls) and treatment heterogeneity further complicate causal evaluations of social policy. Estimating public health outcomes for interventions such as regulations limiting lead levels in gasoline and vaccination campaigns, are notoriously difficult tasks because treating a sample of the population would impact health outcomes in the control group as well. All hope for uncovering causal relationships is not lost, however.

The term "Natural Experiment" refers to an intervention that is not designed for research purposes but the assignment of subjects unfolded "as if" it were randomly assigned, thereby allowing treatment assignment and response to be considered independent and by extension, causal inferences to be made regarding the intervention. Truly random variation in exposure (i.e. variation in treatment exposure that is entirely independent of treatment response) is rare outside RCTs, so this assumption requires caution and special justification. Natural experiments are used extensively in epidemiology and public health to investigate environmental causes of disease by exploiting unplanned and apparently random variation in exposure (Craig et al., 2012). Interventions that satisfy the *"as if" randomly assigned* assumption are rare, and it would be naive to expect a suitable natural experiment to be available in every case where RCTs are impractical, unpalatable, or impossible. When unplanned interventions (i.e. not designed for research purposes) cannot assume "as if" random assignment (e.g. California voters pass legislation legalizing recreational marijuana in

the state) econometric methods such as instrumental variables, regression discontinuity, and difference-in-differences designs are often utilized to induce quasi-randomization.

Natural and quasi-experimental designs considerably expand the range of interventions that can be usefully evaluated. Although non-random treatment exposure in natural and quasi-experiments introduces a threat to internal validity that RCTs rigorously guard against (i.e. selection bias), RCTs lack the external validity and essential timeliness that quasi-experiments can provide. Cartwright (2007) describes how satisfying the assumptions underlying RCTs, in the ideal case, produces an appropriate causal conclusion at the great cost of narrowness of scope (Cartwright, 2007 p.11).

Worrall (2007, p.995) provides a sobering example of the immense practical consequences that can arise when RCTs are conducted without adequate attention to external validity. A nonsteroidal anti-inflammatory treatment for arthritis and musculoskeletal pain passed RCTs with flying colors. However, the RCTs were explicitly restricted to adult participants aged 18-65 while musculoskeletal pain is most common among the elderly. The narrow scope of the RCTs failed to account for the drug's target population being older, on average, than the RCT participants. As a result, there were a significant number of deaths from renal failure among elderly patients once the drug was introduced to the market. After significant loss of life, the pharmaceutical "Opren" was eventually withdrawn. Worrall's (2007) illustration of the non-trivial costs associated with the narrow scope of RCTs is not presented to suggest that quasi-experimental inferences would have avoided the fatalities produced by Opren, but rather to demonstrate how effective policy can be undermined just as easily by failing to account for threats to both external and internal validity. To be clear, a more cautious investigator might have selected a sample of participants that better represented the intended target population for "Opren." However, this critique appears more obvious than it might

have been at the time given the clarity afforded by hindsight. The narrowness of RCTs and controlled experimental conditions may leave substantive variation unobserved. Cronbach (1982) argued that interactions between local contexts and treatment heterogeneity greatly limit the generalizability of experimental results. If a result cannot be generalized beyond the experimental context, efforts to protect against threats to internal validity are rendered irrelevant.

Although RCTs can be impractical under various conditions, observational studies, quasi-experiments, and natural experiments can similarly produce ineffective results when the population affected by an intervention is too small, multiple interventions occur over too short of time intervals, or there is insufficient variation in treatment exposure between treatment and control groups. An equally thorough unpacking of the conditions under which observational studies and quasi-experiments perform poorly relative to RCTs could easily fill the space dedicated in this dissertation to the opposite case. While I would argue that findings produced from RCTs are not superior in some inherent manner which would justify its characterization as "the gold standard" of inference, interpreting this critique as an endorsement of observation research methods over RCTs would be misguided. As Cartwright (2007) and Deaton (2009) put it, "Randomization is not a gold standard because *there is no gold standard*."

When the goal is the production of effective policy, it appears as though RCTs and observational studies exhibit complementary strengths and weaknesses. When the aim of research is the production of generalizable inferences that can be applied to previously encountered combinations of conditions, however, the debate becomes much murkier.

Pawson and Tilley (1997) argue that thirty years of project evaluation in sociology, education and criminology has been largely unsuccessful because evaluations tend focused on whether projects

produced the desired outcome rather than why they work (i.e. too often are mechanism agnostic). A similar argument has been made by James Heckman (1992) within the field of economics. Both critiques seem to accept that failures to identify exception-less regularity between events (i.e. causes and their effects) reflect a shortcoming of epistemology utilized in social science research (rather than acknowledging the complexity of such relationships, scholars universal measurement problems, or perhaps the troublingly stochastic reality of human behavior). Scholarly efforts that fall short of uncovering universal laws of human or societal behavior surely cannot be considered failures if they suggest previously unconceived mechanisms by which an intervention would cause a change in an outcome. Just as ethnographic and qualitative research methods complement quantitative hypothesis testing by identifying potentially complex and nuanced relationships between factors for which quantitative measures may not yet exist; natural experiments, quasi-experiments, and observational studies serve a host of non-negligible intermediate goals (Greed, 1994, p.125-127).

One such intermediate goal particularly relevant to research on criminal justice policies is to produce a body of empirical observations from which new theory can be constructed or existing theories refined. Uncovering potential causal mechanisms by which similar interventions produce varied outcomes is a common avenue by which theory is generated and refined across disciplines. As Donald Campbell (1991[1971]) argued, we do not advance the scientific enterprise by confirming hypothesized relationships in research labs or the real world. Instead, we only learn about causality when our predictions fail to reproduce themselves in the real world. When a hypothesized pattern does not occur, we are able to ask, "why not?" and interrogate the situations where our theoretical understanding of the phenomenon falls short.

In recent years, a new quasi-experimental method for causal inference has emerged out of Econometrics which aims to construct an approximation of a counterfactual group from a weighted

combination of less-than-ideal control units. Known as the Synthetic Control Method, Synthetic Case Control Design, and Synthetic Control Group Modelling; the method was developed by Alberto Abadie (Harvard),Javier Gardeazabal (Universidad del Pais Vasco) , Jens Hainmueller (Stanford), and Alexis Diamond (International Finance Corporation), as a means to better satisfy the "parallel trends" assumption from difference-in-differences models. These synthetic control methods are flexible, require relatively few strong assumptions, and can quickly produce causal estimates from observational case studies. The remaining dissertation traces their emergence through the historical regimes of causal inference. Once the design has been contextualized, the remainder of the dissertation dives deeply into the technical model construction process. Through example analyses as well as replications of Abadie et al.'s foundational articles that introduced the design, I illustrate the importance of proper scaling, predictor selection, and covariate balancing. The results from Chapter 5 largely contradict the arguments made by another scholar regarding synthetic control group predictor selection. Throughout this dissertation, I highlight aspects of synthetic control methods that remain open questions at the time of this publication, and where appropriate I weigh-in on those ongoing debates.

Overview of Dissertation Structure

The remainder of this dissertation is structured as follows; I begin Chapter 2 with a short history of causality, describing its development from Aristotle, Hume, and Campbell to the current regime, commonly referred to as the "Neyman-Rubin *Potential Outcomes* Model". I then describe classes of econometric models derived from the *Potential Outcomes* causal framework aiming to produce causal inferences from observational data. I explain how these methods approximate random treatment assignment, or the independence between response and treatment assignment, by

treating the intervention "as if" it were randomly assigned. I then unpack how each of these new models is able to assume "as if" randomness and draw causal inferences.

In Chapter 3, I describe the statistical theory underlying the synthetic control group design, outline the modeling procedure to construct a synthetic control group, and identify aspects of the model construction process that are in need of formalization. Appropriate methods for scaling the outcome of interest and the inclusion of pre-intervention observations versus covariates as predictors are the two aspects in need of formalization that will be the subject of investigation in chapters four and five.

Chapter 4 examines the impact of marijuana medicalization in California (1996) alongside an untreated control state Wyoming, on motor vehicle fatalities within each state in order to illustrate the impact that unit scaling can have on synthetic control group construction. I describe multiple standardization methods for the outcome of interest (i.e. motor vehicle fatalities) intended to move high and low frequency outliers within the boundaries of the observed convex hull. Put simply, I evaluate a number of methods for reducing variance in the panel; with particular attention devoted to two emblematic outliers (only one of the standardization methods appears to work for both outliers). I then demonstrate how each of the described scaling transformations produces different results with slightly different interpretations. I complete the section by recommending methods for treating count outcomes in synthetic control group models which can apply to criminological applications with both Public Health and crime outcomes.

In Chapter 5, I replicate the analysis from Abadie's (2003, 2010, 2015) foundational articles, in which the synthetic control group methodology is defined and subsequently refined. In each case, when pre-intervention observations of the outcome of interest are excluded from the set of original

predictor variables, the quality of the pre-intervention fit diminishes to such a degree that it prohibits meaningful inference. However, when the covariates are excluded and all observed pre-intervention observations of the outcome of interest are included as predictors, the pre-intervention fit between the treated and constructed synthetic control remains similar to, and in one case, improves upon the results published in Abadie (2003, 2010, 2015). When all pre-intervention observations of the outcome of interest are included as predictors alongside the covariates, the results remain similar in sign and magnitude to the published results, however, the covariate weights approach, but do not quite reach, zero. Through these re-estimations, I demonstrate the limited impact that covariate matching imposes on synthetic control group construction in practice. These results challenge the conventional practice of matching on covariates to eliminate confounds. Since the optimal synthetic control unit is not improved (i.e. determined) by the inclusion of covariate predictors, covariate balance should be presented as a post-hoc validity check, not a metric of "fit" as it has been discussed conventionally.

Chapter 6 will conclude the dissertation with a discussion of the utility and fragility of the synthetic control group design for intervention analysis across social science disciplines. I discuss the particular usefulness of synthetic control group applications in disciplines where RCT experiments and random treatment assignment are impractical or unethical (i.e. Criminology, Public Health, Political Science, and Economics). The dissertation concludes with an accounting of some unresolved methodological issues with the synthetic control group design, including covariate balance, trimming the analysis time frame, and subsequent interventions in the results period.

Chapter 2 – Causal Inference from Observational Data

Counterfactual Causality

The notion of causality was first introduced by Plato and Aristotle, but substantial revisions from David Hume, John Stuart Mill, R.A. Fisher, Neyman, and most recently, Donald T. Campbell, Donald B. Rubin, and Judea Pearl, comprise modern *Counterfactual Causality*. In recognition of Rubin's contributions, Holland (1986) labelled modern causality "Rubin's Model". However, the model is more commonly called the *Neyman-Rubin Model* to acknowledge Neyman's (1923, 1934) contribution of the *potential outcomes* framework. While *Counterfactual Causality* intuitively reflects the logical foundation of modern causality, Rubin (1990) instead refers to it as the *Potential Outcomes* model. I will use Rubin's label where necessary, and the more illustrative label *Counterfactual Causality* generally. Pearl (2000) posits a graphical alternative to the probability-based language of Neyman-Rubin. While Pearl's notion of causal graphs necessarily falls outside the scope of this dissertation, the approaches have been proven to be equivalent when random assignment can be assumed (Pearl, 2000, p.184).

A *counterfactual* can be thought of as a "what if" statement about what would have occurred in the absence of an experienced treatment or in the presence of an unexperienced treatment. For example, if I had not had children I would have achieved greater career success. Put another way, if the control subjects had been treated, would they respond similarly to the treated subjects? We imagine ourselves as cartographers of our "roads not taken", but they remain more than a time machine away. We can never know how treated subjects would have behaved if they had not received the treatment, nor how the control subjects would have responded to the treatment.

Neyman (1923) introduced the concept of observed and unobserved *potential outcomes* in his master's thesis describing field trials of crop yield. For a binary treatment (*w*), unit (*j*) and response (*Y*) two *potential outcomes* are postulated:

$$(Y_j|_{w=1})=Y_{j1} \text{ and } (Y_j|_{w=0})=Y_{j0}$$

The "Gold Standard" of Experimentation

Under the *Potential Outcomes* model, the causal effect of an exogenous binary intervention *t[0,1]* on outcome *Y* for unit *i* is the difference between two *potential outcomes* Y_{i0} and Y_{i1} . Where Y_{i0} represents the outcome for unit *i* given the absence of the treatment, and Y_{i1} represents the outcome for unit *i* given the presence of the treatment. Only one of these two outcomes is ever observable, however. The fact that only one of Y_{i0} and Y_{i1} can ever be observed is what Holland (1985) calls "the fundamental problem of causal inference". This is distinction may seem trivial or semantic, but its importance should become clear when we return to this topic near the end of this chapter.

Randomized controlled trial experiments (RCTs) attempt to overcome the "fundamental problem" by controlling for exogenous environmental factors and observing participant demographics and characteristics. While only one of Y_{i0} and Y_{i1} can be observed for an individual participant in an RCT, random assignment of participants to treatment and control conditions allows researchers to assume that an individual's treatment status is *independent* from their response to the treatment. When treatment assignment and response are known to be independent, the unobserved potential outcome can be estimated from the pool of participants similar in recorded demographics and characteristics that were assigned to the opposite condition.

Absent random assignment, it is difficult to eliminate the possibility that the difference in outcome between treated and control groups was caused by *selection bias* or *assignment bias* rather than

the effect of the examined treatment. Even in tightly controlled experiments, independence between group assignment and treatment response requires that neither the researcher nor participant determine treatment or control status. In order to protect against these threats to internal validity, double-blind designs are used in which neither the participants nor researchers are aware of participants' treatment versus control group status.

Rubin Causality and the *Potential Outcomes* framework illustrate the essential contribution of random assignment to causal inference. The unobserved outcome Y_{j0} can only be estimated when participants in the control group are comparable to the treatment group in every way except treatment assignment.

Campbell's Four Validities and *Quasi-Experiments*

Campbell's approach to validity is implicitly causal. An intervention is assumed to cause a change in the level of a time series if the enumerable threats to validity can be rendered implausible or ruled out entirely. Campbell's (1957) early interest in empirical inference came from two related but unresolvable problems. While non-laboratory research seemed unreliable, they were easily generalized across situations. At the same time, results of laboratory research seemed reliable, but were difficult to generalize. While reliability and generalizability appear at odds, causal validity requires both. Campbell named these problems of reliability and generalizability, "internal" and "external" validity. Campbell's two-validity system, and its elaboration by Campbell and Stanley (1963), were essentially reactions to the manner in which analog experiments had come to dominate psychological research at the time (Campbell, 1986). While many of the threats to internal validity identified by Campbell and Stanley (1963) can be addressed through randomization (researcher blinding, random sampling, and random assignment), Campbell (1957, p. 297) defined external validity as residual category containing threats to "representativeness, or generalizability". In other

words, "to what populations, settings, and variables can this effect be generalized?" (Campbell, 1957). Cronbach (1982) argued that interactions between local contexts and interventions greatly limit the generalizability of experimental results. Cronbach (1982) defined a notational system *UTOS,* to refer to the instances of units (*U*), treatments (*T*), observations (*O*), and settings (*S*) that constrain the inferences drawn from an experiment probing a causal hypothesis. If a result cannot be generalized beyond the experimental context, then internal validity is irrelevant.

In an attempt to generalize the two-validity system to program evaluations in the social sciences, Cook and Campbell (1976; 1979) proposed two new categories of validity (statistical conclusion validity and construct validity, which blurred the crisp boundary between internal and external validity). While once considered an element of internal validity, Cook and Campbell (1976; 1979) grouped questions of confidence and power into a new category called, "Statistical Conclusion Validity". Questions of confounding were implicitly included as threats to external validity under Campbell and Stanley's (1966) two-validity system; however, Cook and Campbell (1979) grouped confounding into a new category "Construct Validity". Many years later, Will Shadish joined Tom Cook and Don Campbell (2002) to authoritatively define Cook and Campbell's (1979) four validities and list their associated threats.

Shadish et al. (2002) coin the term "quasi-experiment" to describe research designs that lack random assignment, but involve some structures and procedures used to rule out threats to internal validity. Shadish et al. (2002) also link the notion of ruling out threats to internal validity to Mackie's (1974) analysis of causation in philosophy, "Typically, we infer from an effect to a cause (inus condition) by eliminating other possible causes" (Mackie, 1974, p.67). There is substantial variation among quasi-experiments, however, in terms of the number of procedures utilized to rule out threats to internal validity, and the sophistication of those procedures.

Campbell and Stanley (1963) list eight common threats to internal validity, and this list was expanded to thirteen by Cook and Campbell (1979). Of these, *history* and *maturation* are probably the most difficult threats to internal validity to address in otherwise well-designed time-series experiments. As threats to internal validity, *history* refers to the possibility that an impact may be the result of some coincidental historical event rather than the studied intervention; and *maturation* refers to the possibility that an impact attributed to the intervention may be the result of some natural change process or uncontrolled trend. Threats such as *instrumentation, regression,* and *selection* also arise frequently in time-series designs. These threats will be discussed at length in the dissertation, and Chapter 3 will refer to this section repeatedly when outlining how synthetic control group models render some of these threats to internal validity implausible.

Quasi-experimental research can, at times, approach the rigorous internal validity protection that true experiments gain through random assignment. Threats to internal validity are not uniformly eliminated through random assignment, however. Rubin (1974, p.690) recognized that causal inference can be problematic even in true experiments. Our ability to draw causal inferences from simple two-group pre-post intervention designs is threatened by potential carryover effects (e.g. effect of first treatment lingers) and the maturation of subjects (e.g. adolescent participants performed better on the second fine motor task due to their development over time rather than the effect of the treatment).

While Shadish, Cook, and Campbell (2002) devote much of their work to quasi-experimental procedures that supplement the design's lack of random assignment, sometimes unanticipated events (e.g. natural disasters, disruptions from novel technologies, etc) can approximate the effects of random assignment (i.e. ensuring the independence of treatment assignment and treatment response). The internal validity of these "natural experiments" are then determined by the

unanticipated event's ability to approximate the elements of random assignment. As Holland (1986) notes, however, ideal "natural experiments" are exceedingly rare in practice. In fact, the inability of normal scientists to simply wait for a specific natural experiment to occur was a primary reason Francis Bacon advocated for experimental manipulation. While scientists cannot afford to wait for a specific natural experiment to occur as Bacon argues, opportunities for knowledge production nonetheless arise when natural experiments eventually occur (Kirk, 2009).

Campbell was an early proponent of strong quasi-experimental designs, arguing that quasi-experiments are acceptable when random assignment is not possible. However, Campbell believed that there are few (if any) situations in which random assignment is categorically unfeasible. Given Campbell's experience with cynical policy-makers, supporting the use of quasi-experimental designs "where random assignment is not possible" (Campbell, 1991[1971], p. 232) seemed like an invitation for policy makers to choose weak and self-serving quasi-experimental designs. In retrospect, it is my impression that Campbell was afraid of the types of research findings an ill-intentioned pseudo-scholar or motivated operative may be able to produce using quasi-experimental designs. All quasi-experiments are not created equal, and Campbell feared that if the design was unleashed upon the world without careful boundaries set on its appropriate use, it could be used to deliver the death knell to whichever welfare, education, or juvenile services program authorities grew tired of funding.

Both Campbell (1957) and Rubin (1974) recognized the potential trade-off between internal and external validity, and further agreed that quasi-experimental results can, at times, be interpreted causally.

The Rubin Causal Model (Rubin Causality)

Rubin (1974) defines the *causal treatment effect (CTE)* for the *j*th subject as a realization of two *potential outcomes*.

$$CTE = Y_{j1} - Y_{j0}$$

However, only one of these outcomes can be observed in practice. This is what Holland calls the "fundamental problem of causal inference". Even in a randomly sampled and assigned experiment, each subject's counterfactual outcome is unobservable. Of course, in most cases, we are not interested in the causal effect of a treatment on a single subject *j*. In clinical trials, for example the causal effect of the treatment on subject *j* would be less important than the treatment effect across all subjects *J*.

Even if both Y_{j1} and Y_{j0} were observable (which they certainly are not), when j is one of many J units, we now have multiple causal treatment effects which may vary in sign and magnitude between units J. Under the Neyman-Rubin Causal Model, the average causal effect (ACE) of treatment can be calculated by taking the difference between expected values for the treated and control units.

$$ACE = E(Y_{j1} - Y_{j0})$$

In its expanded form,

$$ACE = E(Y_{j1}) - E(Y_{j0}) = E(Y_j | w=1) - E(Y_j | w=0)$$

For the treated units,

$$ACE_{w=1} = E(Y_j | w=1) = E(Y_{j1} | w=1) - E(Y_{j0} | w=1)$$

For the control units,

$$ACE_{w=0} = E(Y_j | w=0) = E(Y_{j1} | w=0) - E(Y_{j0} | w=0)$$

The Average Causal Effect (ACE) for the two groups combined gives the ACE across all units *j*.

$$ACE = ACE_{w=1} - ACE_{w=0}$$

By substitution,

$$ACE = E(Y_{j1} | w=1) - E(Y_{j0} | w=1) + E(Y_{j1} | w=0) - E(Y_{j0} | w=0)$$

To estimate the average causal effect of a treatment, however, two assumptions must be made. First, the observed potential outcome for one unit should be unaffected by the assignment of treatments to the other units. This is called the Stable Unit Treatment Value Assumption (SUTVA). SUTVA would be violated, for example, when examining the effect of not vaccinating children in a school district. See Cox (1958, chapter 2) or Rubin (1978) for a discussion of possible threats to this assumption. In this case, the treatment of one district could easily affect the occurrence of diseases like measles and whooping cough in other districts through reducing herd-immunity. Since vaccinations reduce infection rates both among the vaccinated and the unvaccinated (due to reduced disease prevalence among the vaccinated), reductions in the control group infection rate would lead researchers to erroneously partial out these reductions from the vaccination's treatment effect.

Second, Rubin's causal model incorporates many of the threats to internal validity described by Campbell by assuming that treatment assignment (w) and treatment response (ΔY_i) are independent. This assumption can reasonably made when subjects are randomly assigned to treatment and control conditions in randomized control trial experiments. However, independence may also be assumed when treatments are assigned randomly to populations, treatment and control

groups are defined by arbitrary criteria unrelated to the outcome, or there is an instrument unrelated to the treatment that can be used to determine the degree of "treated-ness".

Econometric Methods for Causal Inference from Quasi-Experiments

Regression Discontinuity & Instrumental Variables

In recent years, three econometric quasi-experimental designs that enable researchers to infer causal relationships from observational data have gained widespread use in the social sciences. Instrumental variables regression (IV) and regression discontinuity (RD) designs exploit features of the intervention under examination to satisfy the assumption of independence between group assignment and treatment response. In IV designs, an instrument is defined as a third variable z which only affects the outcome y through its impact on the studied predictor x. In order to estimate the causal effect of smoking on general health outcomes, tobacco taxes can be used as an instrument that only affects health through changes in tobacco consumption (Angrist and Krueger, 2001). In this classic example of an IV design, if the researcher finds that tobacco taxes are correlated with health outcomes, one would conclude that cigarette smoking causally impacts general health outcomes. Levitt (1997) employed an instrumental variables regression framework, using electoral cycles and firefighter hiring as instruments for police strength in order to address the endogeneity of police staffing and crime due to simultaneity bias in structural equations.

RD designs exploit arbitrary boundaries across which treated and untreated units can be assumed to be otherwise alike. When assignment to a treatment depends deterministically on a quantified threshold on some continuous assignment variable, the selection process is fully observed and can be modeled to produce an unbiased estimate of the causal impact of the treatment on an outcome (Cook, 2008). Classic examples of such boundaries focus on political, bureaucratic, and

administratively formalized thresholds, including: American youths, for whom alcohol consumption is legally prohibited the day before their 21st birthday but permitted the following day, students scoring just above or just below a scholarship threshold, or household incomes just above and just below the poverty threshold getting differential access to government services like childcare. Rubin (1977) is often cited as the origin of RD designs, most likely due to the formal proof it contained and its publication in a prominent statistics journal. However, the first publication on RD designs was an application by Thistlethwaite and Campbell (1960) in the field of education, followed shortly thereafter by Campbell and Stanley (1963). For a more complete history of RD designs and Donald Campbell's role in the method's development, see Tom Cook's "Waiting for Life to Arrive" (2008).

Difference-in-differences

Unlike IV and RD designs, which exploit some arbitrary boundary or treatment threshold allowing researchers to assume independence between treatment assignment and treatment response, Difference-in-Differences (DnD) models mimic experimental research designs by comparing the treated units pre/post change in the outcome to the pre/post change observed in an untreated comparison group to account for time trends unrelated to the intervention's effect. DnD models are particularly useful in the context of natural experiments, where treatment assignment can most confidently be assumed to be independent from unit characteristics that may influence both treatment status and response.

DnD designs assume that, absent the treatment, average outcomes for the treated and control units would follow *parallel paths* over time. DnD models subtract the pre-post difference in outcome Y for the control group from the treatment group's pre-post difference to estimate the effect of the treatment. In order to attribute the post-intervention change in distance between time-

series as the impact of the intervention, however, one must assume that the impact of unobserved confounders is *time invariant* for both units. Thus, the distance between time-series in the pre-intervention period can be treated as a constant which can be partialled out. Card and Krueger's (1994) study, which examined the impact of a minimum wage increase in New Jersey on employment in the fast food industry, stands as one of the most famous applications of the DnD design.

The utility of DnD models in criminological research and observational studies generally, however, has been the necessary assumption of "parallel trends". That is, the treated and control groups/units must follow parallel paths throughout the time series of pre-intervention observations. While this may be achievable in studies involving a limited number of pre-intervention observations or very smooth time-series, the parallel trends assumption becomes unattainable as the pre-intervention time-series gets longer, and the pool of potential control time-series gets smaller.

When examining a state-level intervention's effect on motor vehicle thefts in 2010, for example, it becomes highly unlikely that any individual state or the average of untreated states would follow "parallel trends" across forty years of pre-intervention observations. Although almost infinitely unlikely to occur in nature, an ideal control time series would be co-integrated with the treated time-series prior to the intervention. When an ideal control time-series is unavailable, however, it can sometimes be approximated as a weighted combination of less-than-ideal untreated units.

The weighted combination of untreated units, called a synthetic control group, addresses what is perhaps the most common barrier to DnD estimation of treatment effects, the general unavailability of comparison units that satisfy the parallel trends assumption. In Chapter 3, I define

the synthetic control group design as a special class of ($2 \times T$) DnD panel design, describe the conventional procedure for synthetic control group construction, and outline aspects of synthetic control group construction and inference that are in need of formalization.

Chapter 3 – The Synthetic Control Group Design

"Arguably the most important innovation in the evaluation literature in the last fifteen years is the synthetic control approach developed by Abadie et al. (2010, 2015) and Abadie and Gardeazabal (2003)." -- Athey and Imbens (2017)

Introduction

To estimate the effect of an exogenous intervention on a treated unit, a control unit is necessary. Control time-series are chosen to render threats to internal validity implausible. While a literature review can often reveal the most plausible threats to the internal validity for a particular research design or substantive outcome, literature reviews rarely suggest an appropriate control time-series. An *ideal* control time-series resembles the treated unit's time-series in every way except the experience of the intervention. Comparing a treated unit's time-series to its *ideal* control series would fit nicely into Rubin's (1974) *potential outcomes* model as the unobservable counterfactual, while neatly rendering all of the threats to internal validity discussed in Chapter 2 implausible. I use the term *ideal* deliberately here, as it invokes the unattainability of ideal states under Plato's Theory of Forms/Idealism. Just as each chair is distinct from the abstract form of chair-ness. Each control time-series is distinct from the abstract *ideal* control series. An *ideal* control time series serves as the standard of perfection that allows us to identify the imperfections in all of the control time series at

our disposal. While no *ideal* control series exists, the appropriateness of any series as a control series is determined by its ability to approximate the ideal.

One common solution to control time-series selection has been to dis-aggregate the treated time-series by some temporal or spatial dimension across which the intervention's impact on the outcome is expected to differ. Ross et al. (1971) estimate the impact of an intervention aimed at deterring intoxicated driving in Britain by comparing Britain's time-series of vehicle fatalities during the night-time hours when bars were open to Britain's vehicle fatality time series after all bars had closed for the night. Since the intervention targeted intoxicated drivers and the reduction observed post-intervention during the operating hours of British bars was not observed post-intervention time-series after bars had closed, the authors were able to rule out a change in *instrumentation* as a threat to internal validity.

When disaggregation is unfeasible, outside units of similar scale can be used for controls (e.g. other states in the U.S., other OECD nations, other school districts). While these comparisons are generally convenient and available to researchers, they are less useful for ruling out threats to internal validity than disaggregation, and rarely satisfy the "parallel trends" assumption. The average of all untreated units would seem to be another intuitive and convenient control series, but averaged or aggregated groups of untreated units are similarly limited in their ability to render threats to internal validity implausible.

Sometimes, appropriate control time series simply do not exist. When an appropriate control time-series is not available, it may be possible to construct an approximation of an *ideal* control series from a set of control units that are less than appropriate individually.

Synthetic Control Group Algebra

In this section, I define the synthetic control group estimation procedure guided by the motivating model described in Abadie et al. (2010) and adopting their notation. Readers who are familiar with this material may skip to the Model Construction section below. For those interested in a more detailed treatment of this subject, see Abadie et al. (2010; 2015). Suppose that we observe $J+1$ units, and that only the first unit was exposed to the intervention. The remaining J units serve as a "donor pool" of potential controls. Let Y_{it}^{N} be the outcome that would be observed for unit i at time t in the absence of the intervention and let Y_{it}^{I} be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention. Let $\alpha = Y_{it}^{I} - Y_{it}^{N}$ be the effect of the intervention for unit i at time t, and let D_{it} be a dichotomous indicator that takes value one if unit i is exposed to the intervention at time t, and value zero otherwise. The observed outcome for unit i at time t is, therefore,

$$Y_{it} = Y_{it}^{N} + \alpha_{it} D_{it}$$

Because only one unit is exposed to the intervention and only after period T_0, we have that

$$D_{it} = \begin{cases} 1 & \text{if } i = 1 \text{ and } t > T_0 \\ 0 & \text{otherwise} \end{cases}$$

We aim to estimate the effect of the intervention on unit i for all times $t > T_0$.

$$\alpha_{1t} = Y_{1t}^{I} - Y_{1t}^{N} = Y_{1t} - Y_{1t}^{N}$$

Because Y_{1t}^{I} is observed, to estimate α_{1t} we just need to estimate the unobserved component Y_{1t}^{N}.

Suppose that Y_{it}^{N} is given by a factor model

$$Y_{it}^{N} = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_t + \varepsilon_{it} \quad (1)$$

where δ_t is an unknown common factor with constant factor loadings across units, $\mathbf{Z}_i$ is a (r×1) vector of observed covariates that are unaffected by the intervention, $\boldsymbol{\theta}_t$ is a (1×r) vector of unknown parameters, $\boldsymbol{\lambda}_t$ is a (1×*F*) vector of unobserved common factors, $\boldsymbol{\mu}_t$ is an (*F*×1) vector of unknown factor loadings, and the error terms ε_{it} are unobserved transitory shocks at the region level with zero mean.

Let $\mathbf{W} = (w_2, \ldots, w_{J+1})$ be a (*J*×1) vector of donor pool unit weights where $\mathbf{W}$=1and $w_1, \ldots, w_J \geq 0$. The weights are chosen, via a quadratic programming optimization routine, to minimize $(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})'\boldsymbol{V}(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})$ where $\boldsymbol{X}_1$ is a (*K*×1) vector of pre-intervention values of *K* predictive factors of outcome *Y* for the treated unit, $\boldsymbol{X}_0$ is a (*K*×*J*) matrix of the same K predictive factors for the untreated units, $\boldsymbol{W}$ is a matrix of unit weights, and $\boldsymbol{V}$ is a (*K*×*K*) positive definite matrix whose diagonal elements are weights reflecting the relative importance of the factors in $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$. The quadratic programming matching algorithm minimizes the *Root Mean Square Prediction Error* (RMSPE) of the outcome variable across the pre-intervention periods. In other words, the quadratic programming routine identifies a matrix of unit weights that minimizes the differences between Yt and Yt under linear constraints ($\mathbf{W}_1+\mathbf{W}_2+\ldots\mathbf{W}_J=1$ and $w_1, \ldots, w_J \geq 0$). Although predictor selection is generally guided by our theoretical understanding of an outcome's structural covariates and the practical availability of the covariates, the choice of which predictors to include in the matching algorithm is arbitrary. Predictor selection will be discussed at length in Chapter 5, which examines the practice of including pre-intervention values of the outcome as predictors.

Each value of the vector $\boldsymbol{W}$ represents a potential synthetic control (i.e. a weighted combination of control units). The value of the outcome variable Y for each synthetic control indexed by $\boldsymbol{W}$ is

$$\sum_{j=2}^{J+1} w_j Y_{jt} = \delta_t + \boldsymbol{\theta}_t \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt}.$$

Therefore,

$$Y_{it}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N = \boldsymbol{\theta}_t \left(\boldsymbol{Z}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j \right) + \boldsymbol{\lambda}_t \left(\boldsymbol{\mu}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j \right) + \sum_{j=2}^{J+1} w_j \left(\varepsilon_{1t} - \varepsilon_{jt} \right)$$

As Abadie, Diamond, and Hainmueller (2010) note, we assume that the error terms εit are independent across units and time. However, the unobserved residual $u_{it}=\boldsymbol{\lambda}_t\boldsymbol{\mu}_t+\varepsilon_{it}$ may be correlated across units and time regardless of the assumed independence of ε_{it} because of the presence of the term $\boldsymbol{\lambda}_t\boldsymbol{\mu}_t$. If the term $\boldsymbol{\lambda}_t$ is assumed to be constant for all times t, then equation (1) produces the traditional fixed-effects difference-in-differences model. Since the model presented in Abadie, Diamond, and Hainmueller (2010) allows the effect of unobserved confounders to vary over time, taking time differences does not eliminate the potential bias of unobserved confounders $\boldsymbol{\mu}_t$.

Suppose that there are $\left(w_2^*, \ldots, w_{J+1}^*\right)$ such that

$$\sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \quad \sum_{j=2}^{J+1} w_j^* Y_{j2} = Y_{12}, \ldots,$$

$$\sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}, \text{ and } \sum_{j=2}^{J+1} w_j^* Z_j = Z_1 \tag{3}$$

then the effect of the intervention can be estimated as the difference between the observed outcome variable Y_{1t}^{I} and the estimated outcome for the synthetic control unit:

$$\hat{\alpha}_{1t} = Y_{1t} + \sum_{j=2}^{J+1} \left(w_j^* Y_{jt} \right) \tag{4}$$

Equation (3) can hold exactly only if (Y_{11},...,Y_{1T0}, $\boldsymbol{Z_1}'$) belongs to the convex hull of {(Y_{21},...,Y_{2T0}, $\boldsymbol{Z_2}'$), ..., (Y_{J+11},...,Y_{J+1T0}, $\boldsymbol{Z_{J+1}}'$)}. Often, no set of weights exists that would allow for Equation (3) to hold exactly. When that is the case, weights are selected so that Equation (3) holds approximately. The degree to which treated unit's outcome time series belongs to the convex hull of donor pool outcome vectors under various standardizations of the outcome is discussed at length in chapter 4. I will use a series of intuitive and available denominators for standardizing motor vehicle fatalities at the state-level to show how the treated unit's location within the convex hull can change drastically depending on which denominator is utilized. To the best of my knowledge, no study has addressed this issue directly.

Valid causal inference assumes, nevertheless, that the actual and "Counterfactual" time series have parallel trends prior to the intervention and that the trends would have continued absent the intervention.

Model Construction

The procedure for constructing a synthetic control group involves a number of features that render various threats to internal validity implausible. By excluding units that experienced some portion of the intervention being studied, we ensure that our estimate of the causal impact of the treatment is not being masked by a control unit that also experienced the treatment to some degree.

Cook and Campbell (1979) label this threat to internal validity as *treatment diffusion*, whereas Shadish, Cook and Campbell instead categorize *treatment diffusion* as a threat to *construct validity*. Furthermore, excluding units from the donor pool that are not sufficiently comparable to the treated unit ensures that our synthetic control will not be constructed from some peculiar combination of outliers. A state-level analysis might opt to exclude Washington D.C., for example, since it is governed differently than other federalist states (i.e. no senate representation, diminished sovereign authority, etc). Similarly, an analysis involving a treatment in a large developed economy might exclude smaller and/or less developed nations from its donor pool. Units should also be excluded from the donor pool if they experienced any large idiosyncratic structural shocks to their outcome time series within the analysis timeframe. If a shock is experienced universally (i.e. by all units to some degree), the synthetic control method conveniently partials-out that shock from the treatment effect (since it will be represented in both the treated unit and donor pool units). If the shock is unique to an individual or set of donor pool units, but not the treated unit, however, it may be mis-attributed to the intervention under investigation. In essence, these front-end donor pool exclusions ensure that the constructed synthetic control group serves as an optimally realistic approximation of the treated unit's unobservable counterfactual (i.e. the treated unit had the intervention not occurred).

While researchers have a great deal of latitude when deciding what to provide the quadratic programming optimization routine as predictors, the convention established by Abadie and Gardeazabal (2003) and reinforced by Abadie et al. (2010, 2015) has been to include covariates thought to structurally impact the outcome as guided by theory and the existing literature. These covariates can be included as an average across the pre-intervention period, an average across a pre-intervention segment, a single pre-intervention observation, or even multiple pre-intervention observations as individual predictors. In addition to these structural covariates, pre-intervention

observations of the outcome of interest are almost always included as predictors. In general, pre-intervention observations of the outcome are given large V weights by the matching algorithm, and the inclusion of at least one pre-intervention observation of the outcome appears to be necessary in order to produce an interpretable (i.e. sufficiently precise) pre-intervention fit. While more impactful interventions can still be identified when the pre-intervention fit is less precise, the capacity of synthetic control group models to identify smaller (but nonetheless significant) interventions, is directly determined by the quality of pre-intervention fit. Put simply, pre-intervention fit determines the type 2 error rate (i.e. statistical power) of synthetic control group models.

While Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015) all use varying combinations of pre-intervention observations of the outcome and structural covariates as predictors, it is unclear how the authors arrived at each combination of predictors. In theory, including a structural covariate of the outcome as a predictor in the matching algorithm should produce a synthetic control group which has a similar value for the covariate as the treated unit. When the treated and synthetic unit have similar values for a covariate thought to impact the outcome, the covariate is ruled out as a cause of any post-intervention gap that emerges between the treated and synthetic unit. Thus, matching on structural covariates provides a protection against the threat that *maturation* and *selection* pose to internal validity.

As mentioned above, however, it appears as though it is necessary to include one or more pre-intervention observations of the outcome as predictors in order for the matching algorithm to converge on a synthetic control unit that is interpretably similar to the treated unit's outcome time series prior to the intervention. Each pre-intervention observation of the outcome can be expressed by the factor model defined by equation 1. If the structural covariates included as predictors explained nearly all of the observed variance of the outcome, and the relationship between the

covariates and the outcome were constant across units and over time, pre-intervention observations of the outcome would add little predictive power to the matching algorithm. Although this sort of complete causal knowledge may be approachable in disciplines such as chemistry and physics, it would be a career-making revelation to identify such a phenomenon in the social sciences. Gross domestic product, homicide, and other complex economic and social phenomenon are, to out contemporary knowledge, driven by a less predictable and dynamic set of factors (measurement error aside).

Since the known covariates and their parameters are not the only factors that determine the outcome, including only covariates as predictors ignores the impact of unknown common factors, their factor loadings, and measurement error. Pre-intervention observations of the outcome, however, reflect the combined impact of all of the factors on the right-hand side of equation 1. Since pre-intervention levels of the outcome reflect the combination of the known structural covariates and unknown/unobserved factors, each pre-intervention observation of the outcome of interest that is added as a predictor will share some of the variance it explains with the included structural covariates. To scholars that primarily focus on regression-based inference, the idea that the variance in the outcome explained by a given covariate is nested in another included predictor seems like an obvious collinearity problem. However, the predictor loadings produced by the quadratic optimization algorithm are not the object of interest in this analysis. If the parameter estimated for a given covariate diminishes as pre-intervention observations of the outcome are added as predictors it would not bias estimate of the intervention effect.

A small group of methodologists have begun to debate the appropriateness of including some or all pre-intervention observations as predictors in synthetic control group models. McCleary, McDowall, and Bartos (2017) find that including all-available pre-intervention observations of the

outcome as predictors invariably produces a better fitting synthetic control time-series than covariate-only models and covariate models with only a few pre-intervention observations of the outcome of interest as predictors. On the other end of this debate, however, Kaul et al. (2015) argue that using all available pre-intervention observations of the outcome as predictors leads to very small covariate weights V produced by the matching algorithm. They argue that the utility of the covariates to rule out alternative explanations for the post-intervention effect is lost when pre-intervention observations of the outcome receive larger V weights than the included covariates. This argument ignores what is gained, however.

In chapter 5 of this dissertation, I demonstrate that in each of the foundational synthetic control group analysis published by Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015), the V weights associated with structural covariates inevitably approach zero as pre-intervention observations of the outcome (and first-differences) are included as predictors. Through re-estimation of Abadie's seminal works, I further demonstrate that the V weights associated with the covariates diminish as each pre-intervention observation of the outcome is added as a predictor. However, the pre-intervention RMSPE also reduces with each added pre-intervention observation of the outcome. As the pre-intervention RMSPE diminishes, more precise estimates of the intervention's effect on the outcome can be made. In many situations, this added precision can make the difference between an estimate that cannot be distinguished from null, and an estimate with a clear sign and magnitude. The remainder of Chapter 5 explores the practical trade-off between ruling out threats to internal validity through covariate similarity and producing a better approximation of the treated unit's pre-intervention outcome time-series. Secondary considerations, such as the flexible application of synthetic control group models matched on pre-intervention values of the

outcome to various unit-levels where covariate time series are unavailable, are also discussed in chapter 5.

Post-Estimation Testing and Inference

A post-intervention gap cannot be categorized into objective thresholds of small, medium, or large effect sizes. The magnitude of the estimated intervention effect relative to the idiosyncratic year-to-year variation of the outcome time series, and the likelihood of misattributing a spurious artifact as the impact of the studied intervention can only be assessed in comparison to estimates of the intervention effect produced from the untreated donor pool units. By iteratively reassigning the treatment condition to untreated donor pool units, this "in-sample placebo test" provides a distribution of observed placebo effects among the donor pool units (Abadie et al, 2010; 2015). If the estimated treatment effect is smaller than or comparable to the distribution of placebo effects, it would suggest that the intervention's effect may be a spurious artifact. If the gap estimated via SCG comparison is larger than any of the gaps estimated in units that never experienced the treatment, we can be confident that the estimated effect of the intervention on the outcome was larger than the idiosyncratic shocks to the outcome in untreated units, thereby rendering spuriousness implausible. By ranking the treated unit and donor pool by the ratio of post-intervention gap over pre-intervention RMSPE, this "in-sample placebo test" is akin to Fisher's Exact Test.

The second conventional post-estimation test, known as the "Leave One Out" sensitivity test, is used to assess the dependency of the estimated effect on certain donor pool units receiving a non-zero weight (Abadie et al., 2015). Depending on the comprehensiveness of the donor pool restrictions employed on the front end, this test can serve various purposes. When rigorous front-end exclusions are made on donor pool membership, the Leave One Out test communicates, in essence, that the result would not have changed even if they included a unit in the donor pool that

was inappropriate. Because heavy front-end exclusions have already been made, it would seem implausible that multiple inappropriate donor pool units met the strict inclusion criteria. Thus, the Leave One Out test can complement the front-end donor pool exclusions and further protect against *selection* threats to internal validity.

Given a relatively unique or complex intervention, however, determining which potential donor pool units experienced one or more elements of the intervention can involve a great deal of abstraction. In the context of state-level criminal justice policies, for example, estimating the effect of an omnibus sentencing reform in California via synthetic control would require researchers to exclude U.S. states from the donor pool that experienced any aspect of the California omnibus reform. If the elements of the omnibus criminal justice reform can be reduced to a sufficient level of abstraction (i.e. eliminated sentencing enhancement for prior offenses) then some donor pool exclusions should be possible. However, as interventions get more complex, and address conditions that are somewhat unique to the treated unit, front-end donor pool restrictions are less and less likely to be comprehensive. Since we are less confident that our front-end donor pool restrictions were comprehensive, and the likelihood that multiple inappropriate units were included in the donor pool, the Leave One Out Test should be expanded to supplement, rather than compliment donor pool exclusions. Instead of limiting our sensitivity test to one donor pool unit at a time, donor pool units can be iteratively removed until the estimated post-intervention effect changes sign (i.e. + or -) or the pre-intervention fit becomes too weak to rule out a null effect. By reporting the donor pool units that must all be excluded for the interpretation of the intervention effect to change, researchers provide the audience with all of the information needed to personally critique the synthetic unit's composition, as well as the context to evaluate whether addressing the critique would change our interpretation (Bartos and Kubrin, 2018).

Chapter 4 – The Theory of Standardization

Introduction

In many cases, the treated unit and its suitably similar donor pool of untreated comparison units vary considerably in geographic area, total resident population, population density, transportation infrastructure, non-resident travel, and many other factors that can limit or expand each unit's potential frequency of a given count outcome. As discussed in Chapter 3, the donor pool weights assigned by the quadratic programming matching algorithm must be non-negative and sum to one. These constraints prohibit SCG models from extrapolating beyond the range of outcomes observed among untreated donor pool units (Abadie and Gardeazabal, 2003). The use of negative weights, however, may be more flexible than summing to unity. Scholars continue to debate the appropriateness of allowing negative weights to be assigned to donor pool units (i.e. allowing inverse trends to contribute to synthetic control units).

Interventions that occur within a single U.S. state conveniently imply a donor pool comprised of other U.S. states, as their relatively crisp geographic boundaries, distinct internal bureaucracies/operations, and shared federal system they are all nested within, plausibly satisfy the Stable Unit Treatment Value Assumption (SUTVA) in many cases (McCleary et al., 2017). However, U.S. states vary considerably in geographic area, resident population, population density, and many other factors which can influence the frequency of the observed outcome. When an intervention occurs in a state with the highest or lowest observed frequency of the outcome of interest, no combination of untreated states (with weights $\mathbf{W_1+W_2+\ldots W_J}=1$) can approximate the treated unit's pre-intervention time series.

For example, the state of California's Gross Domestic Product is greater than any other state, and if one were to attempt constructing a synthetic control unit that approximates California, had some economic policy intervention not occurred, no linear combination of other (lower GDP) states would be able to approximate California's pre-intervention GDP time series with donor pool weights that sum to one. In this case, the count outcome must be standardized, or scaled by a denominator term intended to make California less of an outlier.

Which factor would be most appropriate denominator in our standardizing transformation depends entirely on the outcome examined and the scope of the research question. For mortality and other health outcomes, population rates such as "per 100,000 resident population" are commonly used for state, county, and city units of analysis. Per capita rates – with a few adjustments – work quite well for health outcomes where every living human has the potential to be included as an occurrence of the count outcome. In the case of a state-level public health intervention aiming to reduce obesity, the total number of humans living in each state provides an intuitive denominator for standardizing the frequency of obese residents, because each resident had the potential to be counted in the numerator. When public health interventions target a specific subpopulation such as residents 65+ years old, the total resident population of a state becomes much less useful as a standardizing denominator. If, for example, we were interested in the impact a shingles vaccine rollout in North Carolina had on hospitalizations for residents 65+ years old, a more appropriate denominator would be each states 65+ year old resident population. These types of population rates should be familiar to most readers, as per capita standardizations are the accepted convention in many fields, and convenient measures of the total population denominator are available at a variety of unit levels (nation, state, county, city, village, census block, etc) and across decades of relatively gapless reporting.

It should be no surprise then, that the conventional use of population rates extends – questionably so – into situations where neither population nor a subpopulation would reflect the number of times people are "exposed" to the numerator. In other words, the ideal scaling term, or denominator by which the frequency count numerator is transformed into a rate, should reflect the number of instances in which one had the potential to contribute to the numerator. If the outcome of interest requires that an individual has access to some functional object, such as an automobile or firearm, in order to produce an instance of the outcome (e.g. motor vehicle fatality or suicide via firearm), the frequency of that condition within each unit would be a more appropriate standardization denominator than the total population. In other words, if some material, social, or environmental *opportunity* is required for the outcome of interest to occur, then the frequency of the *opportunity* within each unit would be a better standardizing factor. Similarly, if the outcome of interest is a potential outcome each time a certain risky *behavior* is undertaken, then the frequency of that *behavior* would be the most appropriate factor to use as the denominator in a standardizing transformation.

These standardization methods are discussed at length in the remainder of this chapter; however, I mention them now to demonstrate that there is no categorically "correct" standardization method. There is no statistical test that can be utilized to determine the "correct" standardizing transformation (public health generally chooses the transformation with lower variance). Instead, the appropriate standardization method for a given research question is guided by our theoretical understanding of how the outcome of interest is produced and how we would expect an intervention to impact that process. Although outside the scope of the present discussion, public health scholars working with mortality data from geographic units of varying sizes, generally use direct and indirect standardization methods. Direct standardization is widely used by the Centers for

Disease Control and World Health Organization to adjust mortality rates by calculating group-specific (e.g. age groups) rates weighting them to mimic the age structure of a standard population. Indirect Standardization can be used to calculate the expected outcome rate for an index population, given age specific outcome rates from a reference population. Although public health scholars have developed sophisticated and well-defined standardization methods intended to make mortality rates and disease incidence rates comparable across nations that can vary widely in age-structure, outside of public health however, for many outcomes the size of the population within each age group fails to reflect the population's total exposure to the outcome.

It is also worth noting that when evaluating the impact of an intervention on an outcome, the mechanism by which the intervention would affect the outcome cannot be used as the standardizing factor. If an intervention is expected to reduce the frequency of an outcome through a reduction in the denominator used to scale the outcome, the standardization would mask the intervention's effect. For example, if a public health researcher wanted to examine the impact of a tobacco tax increase on lung cancer diagnoses within a state, using statewide cigarette sales (in packs) as the denominator for a standardizing transformation would mask the impact of the intervention on the outcome. Rather than assessing if the tax increase reduced the frequency of lung cancer diagnoses in the state, an outcome standardized in this way would reflect whether the number of lung cancer diagnoses *per pack of cigarettes purchased* in the state decreased following the tax increase. In this hypothetical case, the tax increase on tobacco is intended to reduce the statewide frequency of lung cancer diagnoses by making tobacco consumption more expensive and therefore, less frequent. Standardizing the frequency of lung cancer diagnoses using consumption as the denominator is inappropriate for the hypothetical tobacco tax intervention because its impact on lung cancer should, according to our theoretical understanding, be achieved by reducing tobacco consumption.

If our hypothetical intervention were not a tobacco tax increase, but instead, a state-level policy establishing regulations prohibiting the sale of tobacco containing certain chemical additives, this standardization would be perfectly suited to address the research question. The fundamental difference between these two interventions that makes this standardization appropriate for enacting regulations on chemical additives but inappropriate for a tax increase, is that regulating chemical additives would be expected to change the probabilistic relationship between packs of cigarettes consumed and the occurrence of lung cancer. Whereas the tobacco tax would be expected to affect lung cancer through reduced cigarette consumption, regulating the chemical content of cigarettes sold in the state would be expected to reduce lung cancer occurrence when cigarette consumption remains constant.

The remainder of this chapter examines the types of standardization methods available to researchers and illustrates the non-trivial consequences that researchers' choice of scaling denominator (i.e. exposure term) can produce in synthetic control models. Under each method, the shape of the convex hull changes, and the location of the treated unit within the convex hull may, but does not always, change as well. Many standardization methods may appear equally sufficient and differences between them can seem inconsequential when the treated unit is similar in scale and outcome frequency to the majority of untreated donor pool units. However, as I will demonstrate below using state-level motor vehicle fatality data, different standardization methods can (but do not always) produce synthetic controls with varying interpretations. Three considerations remain essential when selecting a denominator for an outcome rate. First, the standardized outcome should move the treated unit into the convex hull of observed variation among the donor pool of untreated units. Second, the outcome rate should avoid masking the mechanism by which an intervention

would impact the outcome. Finally, the best standardization method generally involves scaling by an exposure term (i.e. the number of times people have had the chance to contribute to the numerator).

[Insert Figure 4.1 about here]

The examples presented in this section utilize a 50 state panel dataset containing annual motor vehicle fatalities between 1975 and 2014 gathered from the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS). The panel also includes total population by state (census), vehicle miles traveled, registered vehicles, and licensed drivers gathered from the Federal Highway Administration's Highway Statistics series (Appendix 4, Table 1 displays summary statistics for the panel dataset). Figure 4.1 plots the number of motor vehicle fatalities that occurred in each state between 1975 and 2014. California and Wyoming are plotted as solid black lines, and the remaining 48 states are plotted as dashed grey lines. As Figure 4.2 shows, there are a number of outliers besides California and Wyoming.

[Insert Figure 4.2 about here]

California and Wyoming will be highlighted throughout the remaining examples in Chapter 4, as they represent two ends of the population density spectrum and the standardization methods will affect their positions in the 50 state panel differently. While this chapter devotes much attention to these two extremes, for synthetic control group construction, it is beneficial to reduce variance in the outcome and bring any outliers into the panel's overlap range. When the ideal standardization method is displayed later in this chapter, however, all of the outliers are made less extreme and fall within the panel's overlap range.

[Insert Figure 4.3 about here]

Standardization Methods

Population Rates

To say the use of population-based standardizations is the conventional approach across social science disciplines would be a severe understatement. The ubiquity of population rates extends beyond the conventions of social scientists, as most of the world would be familiar with the phrase "per capita" even if they were unable to define its precise meaning. This should come as no surprise, however, as population rates tend to be conveniently available and often most appropriate in order to compare resident demographics across states, nations, etc. They are also ideally suited to epidemiological analyses; even if the denominator for these rates must often be limited to the subpopulation at risk for the outcome (e.g. Alzheimer's incidence per 100,000 U.S. adults aged 50+). Population rates are often age adjusted to account for the portion of the population that would not plausibly contribute to the numerator count outcome (e.g. age-adjusted divorce rate per capita). In general, population rates are most appropriate when each individual member of a population (or sub-population) has the opportunity to contribute to the count outcome. The various adjustments commonly applied to population rates generally account for segments of the population that are unable to contribute an observation to the numerator for some reason (e.g. fertility rates only include the segment of the population within child-bearing range).

A state agency compiling an annual report on highway safety, for example, could easily be forgiven for reporting motor vehicle fatalities per 100,000 state residents. A per capita vehicle fatality rate still accounts for the size of the state and the number of drivers to a degree. Taking this a step further and weighting each state so that they all have the same aged 18 and over exposure denominator would be analogous to direct standardization used in public health. Although a per capita rate based upon a state's total resident population would include residents younger than the

legal minimum driving age in the state, including them in the denominator is defensible to some degree, since one doesn't need to be the driver to be killed in a motor vehicle accident. While the frequency of a mortality outcome in a geographic unit such as a state would seem to be bounded by the total population of that state, a measure of this type can become quite misleading if individuals not included in the total population denominator are able to produce an occurrence of the numerator. A motor vehicle fatality may occur in California without any California residents being involved for example. Truck drivers, tourists, and other non-resident drivers can contribute greatly to a state's frequency of motor vehicle fatalities without being represented in the denominator of a per capita standardization.

[Insert Figure 4.4 about here]

Figure 4.4 displays annual motor vehicle fatalities per 100,000 residents for fifty U.S. states. Considering that Wyoming jumps from the bottom of the distribution in Figure 4.3 to a high-end outlier when scaled per capita, motor vehicle fatalities per 100,000 residents appears problematic. While many governors might not notice a problem with this metric from simply looking at a panel of vehicle fatalities per 100k residents by state, the governor of Wyoming would likely protest. Using resident population as a denominator, Wyoming holds the highest per capita motor vehicle fatality rate of any U.S. state. Before Wyoming scrambles to initiate drunk driving awareness campaigns and replace the guardrails on every highway in the state in an effort to relinquish its "most dangerous roads in the United States" status, other standardization methods should be considered.

Opportunity Rates

[Insert Figure 4.5 about here]

In some circumstances, the occurrence of a count outcome requires some material object, opportunity, or pre-condition. In order for an individual to commit suicide with a handgun, for example, one must have access to a handgun. For a motor vehicle fatality to occur there must be a motor vehicle and a driver. In most cases, vehicles tend to be registered and drivers tend to be licensed. Therefore, the number of licensed drivers and registered motor vehicles in a state may seem like convenient and attractive opportunity-based standardization/exposure terms. Since the denominator used to standardize the outcome is intended to reflect the frequency of opportunities in which a motor vehicle fatality was able to take place, the number of registered vehicles seems like promising candidate for an exposure term. Figure 4.5 displays the motor vehicle fatality rate per 100,000 registered vehicles for fifty states between 1980 and 2014, with Wyoming in red and California in green. Similar to the per capita transformation in Figure 4.3, Wyoming remains at the high end of the panel throughout, although no longer an outlier, while California again appears close to the center of the distribution.

Behavioral Rates

When an outcome of interest only occurs following a particular antecedent behavior, a measure of that behavior's frequency may be a more appropriate standardizing denominator than a crude population-based exposure term. A motor vehicle fatality can result every time someone operates a motor vehicle, and a motor vehicle must be driven in order to produce a motor vehicle fatality. As long as the fatality occurs within a given state's borders the fatality contributes to that state's motor vehicle fatality count, regardless of which state the vehicle(s) were registered within, the residence of the driver(s), and the residence of the victim. Since the drivers, victims, and vehicles involved in a motor vehicle fatality can all come from outside the state where the fatality occurs, none of these factors are ideal exposure terms for standardizing the outcome. The only pre-requisite

condition to a motor vehicle fatality occurring seems to be that a motor vehicle is driven, therefore, an ideal standardizing denominator would reflect the total amount of driving in the state, such as the vehicle miles travelled.

[Insert Figures 4.6 and 4.7 about here]

Wyoming appears to be a particularly dangerous place to drive when standardizing by resident population and registered vehicles as the state has relatively few residents, drivers, and vehicles. Each of these exposure denominators fail to capture the motor vehicle fatalities that can result from out of state drivers and interstate truck driving. When motor vehicle fatalities are instead standardized using vehicle miles travelled as the exposure term(as they are plotted in Figures 4.6 and 4.7), the resulting rate incorporates those previously unaccounted for sources of non-resident motor vehicle fatalities. These potential exposures may only explain a small, even negligible, portion of fatalities in larger states with greater resident populations. In Wyoming, however, the resident population is quite small, and the motor vehicle fatalities attributable to out of state drivers and interstate trucking comprise the majority of the state's motor vehicle fatalities. When motor vehicle fatalities are standardized using vehicle miles traveled as the denominator, Wyoming ceases to be an outlier, and can be approximated by a synthetic control unit without extrapolation.

[Insert Figure 4.8 about here]

When an intervention intended to impact motor vehicle fatalities occurs in a state with as small of a population and pool of licensed drivers as Wyoming, the inappropriateness of per capita and per licensed driver rates should be inescapably obvious to the researcher. When the same intervention is examined in a state as large as California, however, none of the standardizations examined in this chapter would necessarily seem inappropriate. Figures 4.8-4.10 illustrate the impact

these standardization methods can have on synthetic control construction using California (a particularly large and active state) and Wyoming (a particularly low population state considering its large space). In 1996, California enacted Proposition 215, which enabled qualifying patients to possess and cultivate small amounts of cannabis for medicinal purposes with physician approval. Since Prop 215 occurred in California and not Wyoming, the remaining examples in this chapter are expected to show an intervention effect in California following 1996, while no effect is anticipated for Wyoming. Without any scaling/standardization method the matching algorithm is unable to converge on a synthetic control group for California, as it is the high end outlier in terms of raw frequencies (see Appendix Figure A4.1). As expected, results for Wyoming (see Figure 4.8) suggest no 1996 intervention effect when modeled as untransformed frequencies. Figure 4.9 plots California and its synthetic control when standardized per 100,000 resident population. The post-1996 gap between California (solid black line) and its Synthetic control (dashed black line) suggests that the implementation of Prop 215 caused California's per capita vehicle fatality rate to be lower than it would have been without Prop 215 from 1996 to 2014. Just as the matching algorithm could not converge on a synthetic control for untransformed California, Appendix Figure A4.2 shows the same failure to converge for per capita Wyoming (now a high-end outlier). Figure 4.10 displays the synthetic control results for California when the outcome is scaled using registered vehicles as the exposure denominator. The Synthetic California displayed in Figure 4.10 is similarly well-fit throughout the pre-1996 segment to per capita version in Figure 4.9, and imply comparable reduction in vehicle fatalities due to Prop 215. Synthetic Wyoming could not be constructed when resident population or registered vehicles were chosen as exposure terms (i.e. scaling denominators).

[Insert Figures 4.9 and 4.10 about here]

Of the exposure terms explored in this chapter, standardizing motor vehicle fatalities by vehicle miles travelled is the only scaling denominator that allowed the synthetic control matching algorithm to converge on a synthetic control for both California and Wyoming. Looking back at the panel transformation plots in Figure 4.3 through 4.7, it should come as no surprise that vehicle miles travelled functioned as the best exposure term for California and Wyoming, since neither stood out as outliers in Figures 4.6 and 4.7. Figure 4.11 plots California and its synthetic control when standardized per vehicle mile travelled. Although the post-1996 gap still suggests that Prop 215 caused a reduction in motor vehicle fatalities, the reduction appears much smaller than Figures 4.9 and 4.10. Figure 4.12 plots Wyoming and its synthetic control when standardized per vehicle mile travelled. Since Proposition 215 did not impact Wyoming, the null post-1996 gap indicated in Figure 4.12 is expected, and aligns with our knowledge of the intervention studied. For California's Prop 215, these alternative scaling methods reinforce the conclusion that Prop 215 caused a reduction in statewide motor vehicle fatalities, although Figure 4.11 indicates the reduction may be less permanent and of lesser magnitude than Figures 4.9 and 4.10 might have suggested.

[Insert Figures 4.11 and 4.12 about here]

Overall, the various exposure terms explored in this chapter varied in their ability to bring California and Wyoming near the center of the fifty state convex hull. While the line plots in Figures 4.8 and 4.9 are much more clustered and reflect the mortality risk of driving in each state much more accurately, vehicle miles travelled is still a lacking measure of exposure. If a vehicle travels a mile, but carries three passengers in addition to the drive, a fatal crash at the end of the mile would produce four fatalities per mile travelled. Considering that four people travelled a mile, a more precise measure of risk/exposure might be "passenger miles travelled". Considering the number of

passengers at one time that can be on public buses, for example, further reveals the shortcomings of the vehicle miles travelled denominator.

The exposure term and standardization method researchers employ can greatly impact the results of synthetic control models, and some standardization methods can completely prohibit synthetic control analysis, at least for certain units. Standardization, in general, exponentially expands the range of applications in which constructing synthetic control units is possible. Without standardization, unit size would be the primary determinate of synthetic control composition, rather than longitudinal trend similarity in the outcome. As the models in this chapter demonstrate, proper standardization can allow for units varying extremely in geographic size and population to contribute to each other's synthetic counterfactuals. However, the small variation in the synthetic control estimates for Prop 215's impact on motor vehicle fatalities California suggests that researchers should examine a variety of standardizations prior to drawing confident conclusions. In this case, one could be quite confident in the sign(-) of Prop 215's effect on motor vehicle fatalities (i.e. Prop 215 caused a reduction), but would be less confidence about the precise magnitude of said reduction. Table 4.1 lists the donor pool weights comprising the Synthetic Californias plotted in Figures 4.9-4.11.

Chapter 4 Tables

Table 4.1 Non-Zero Weights for Synthetic California Time Series

Motor Vehicle Fatalities Per 100k Population		**Motor Vehicle Fatalities per registered vehicle**		**Motor Vehicle Fatalities per vehicles miles travelled.**	
State	$w_i>0$	*State*	$w_i>0$	*State*	$w_i>0$
AZ	0.047	AZ	0.081	AZ	0.031
CT	0.068	CT	0	CT	0
ID	0.068	ID	0	ID	0.008
IL	0	IL	0	IL	0.007
IN	0	IN	0	IN	0.025
MA	0.215	MA	0	MA	0.005
NH	0.339	NH	0.145	NH	0.122
NY	0	NY	0.451	NY	0.579
NC	0	NC	0	NC	0.134
SC	0.119	SC	0.066	SC	0
UT	0	UT	0	UT	0.054
VA	0.026	VA	0.225	VA	0
WV	0.099	WV	0.032	WV	0
RMSPE	0.429699		0.006886		0.0002332
Mean Post-intervention gap	2.419996		0.031825		0.0009946
Post/pre Ratio	5.631842		4.621434		4.2640981

Chapter 4 Figures

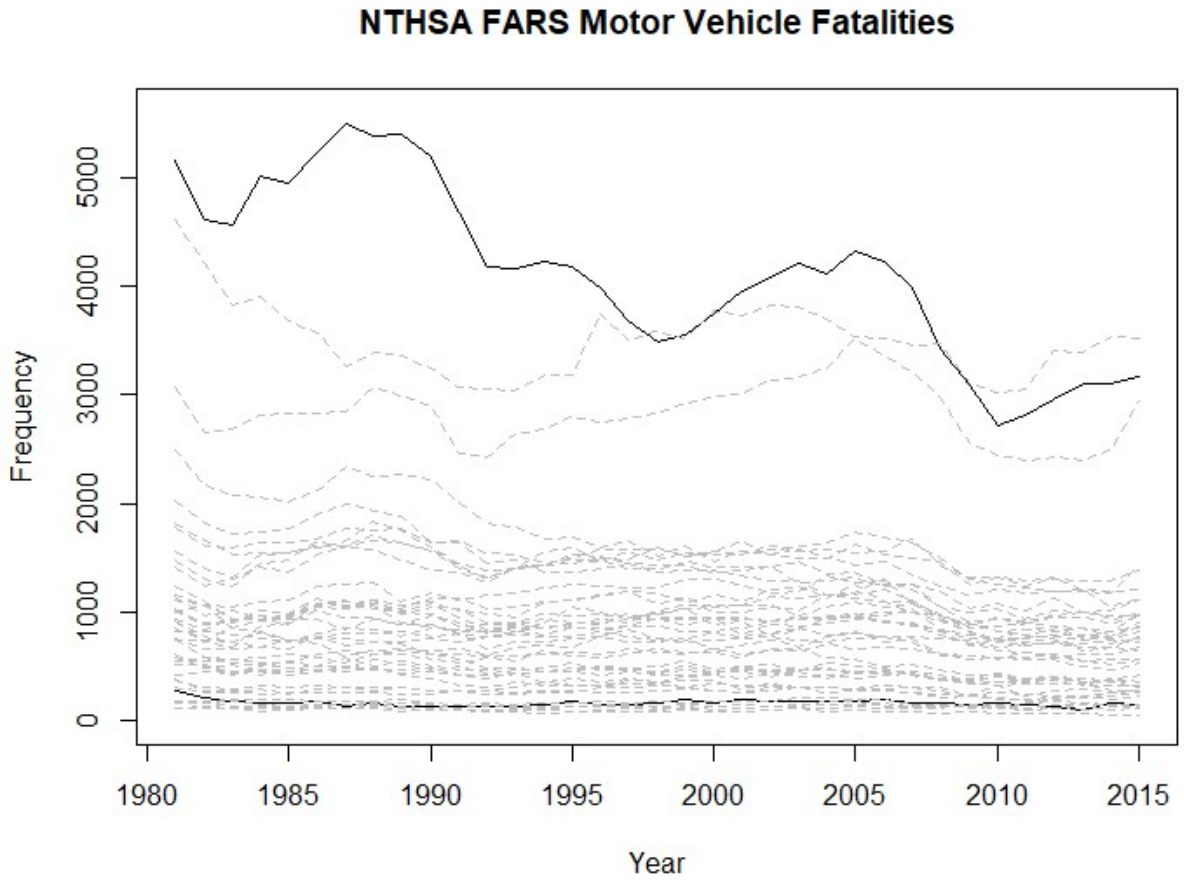

Figure 4.1 Motor vehicle fatalities for 50 U.S. states reported in the National Highway Transportation Safety Administration's Fatality Analysis Reporting System

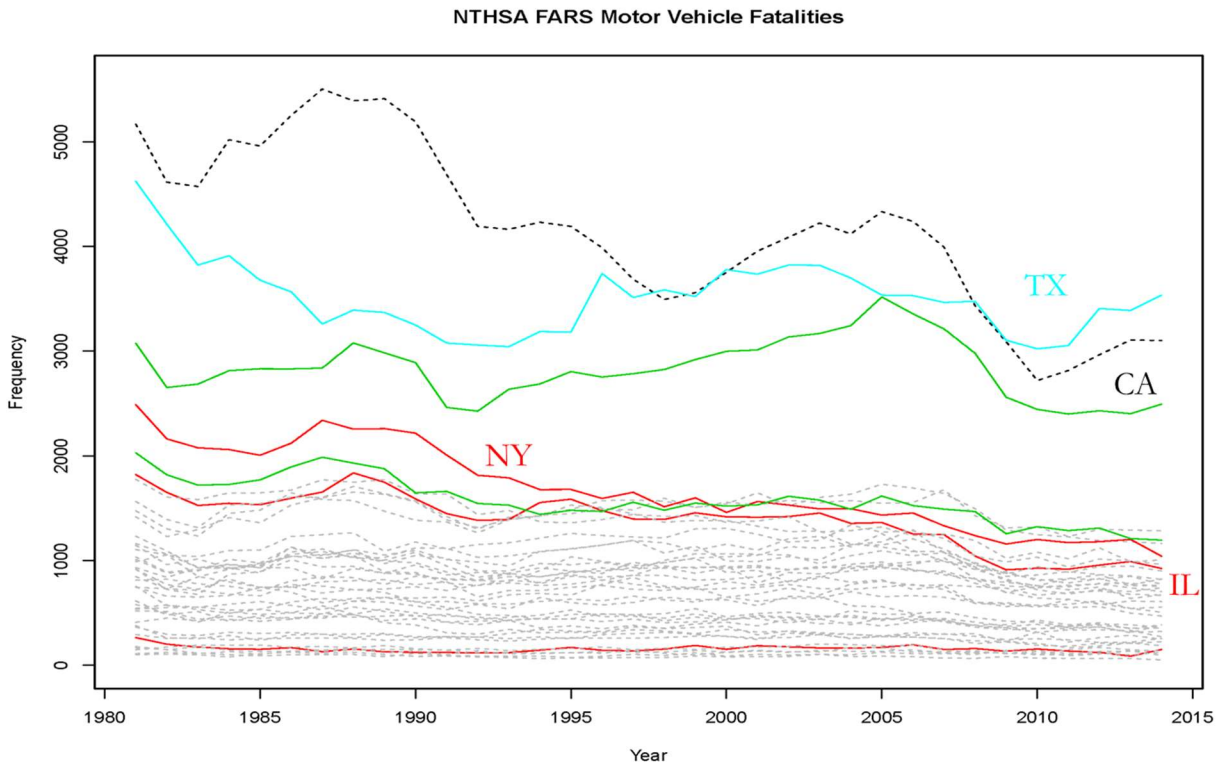

Figure 4.2 Motor vehicle fatalities "outliers" highlighted and labelled

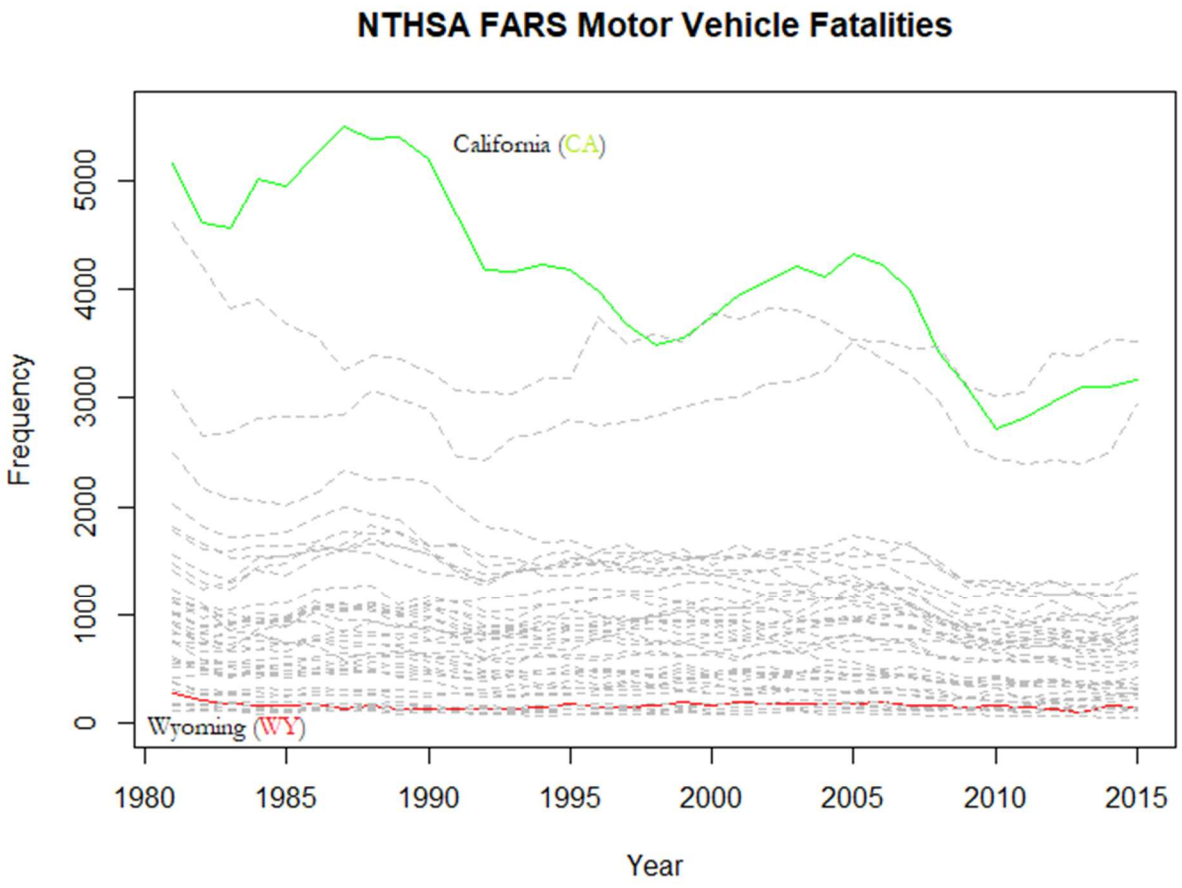

Figure 4.3 Motor vehicle fatalities: Wyoming and California units of interest labelled

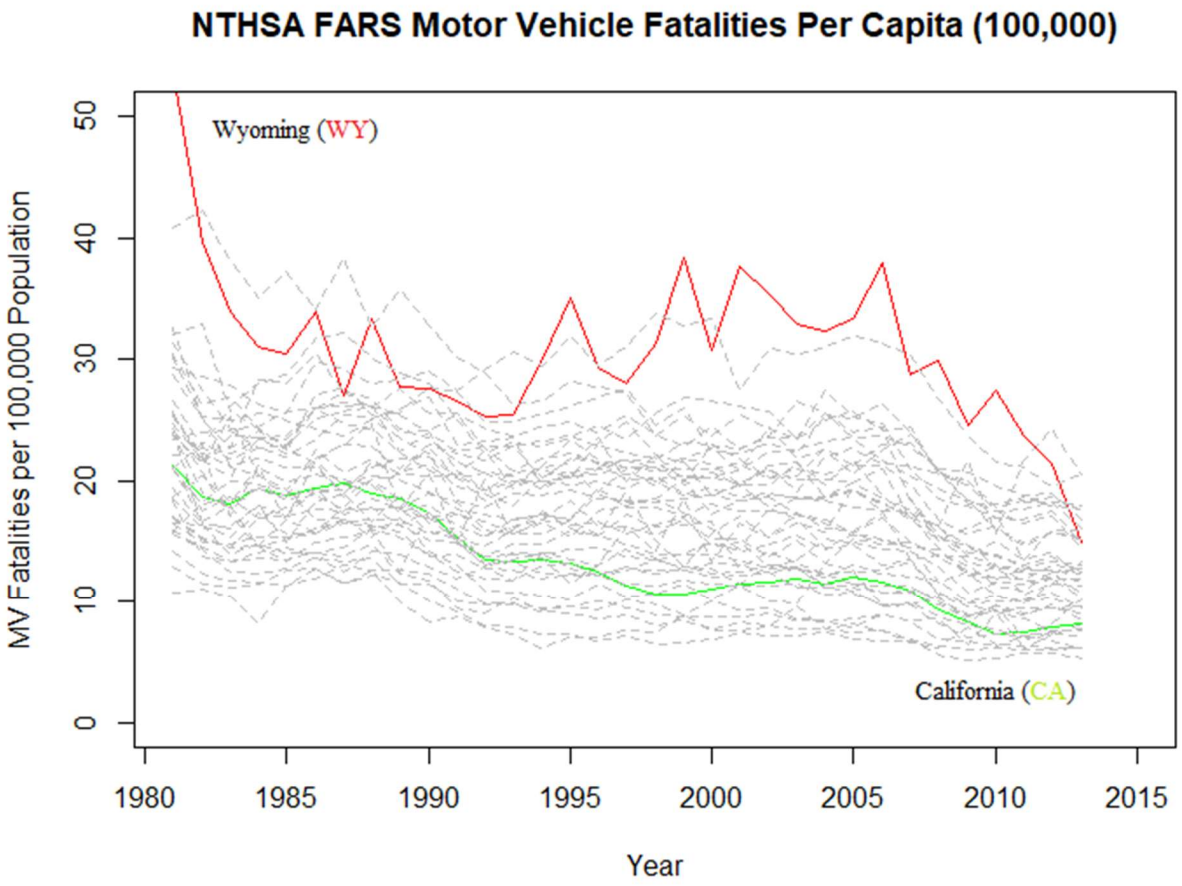

Figure 4.4 Motor vehicle fatalities scaled per 100,000 residents

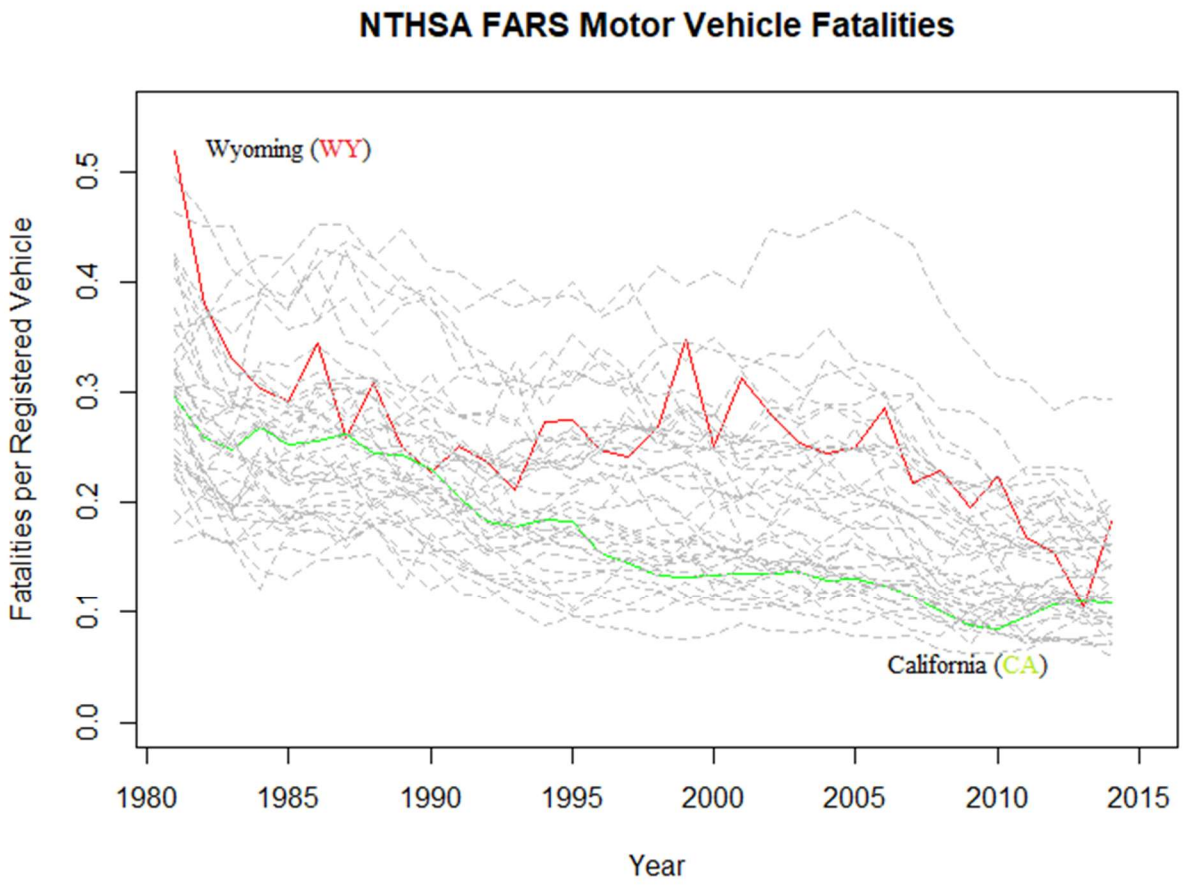

Figure 4.5 Motor vehicle fatalities scaled per registered vehicle

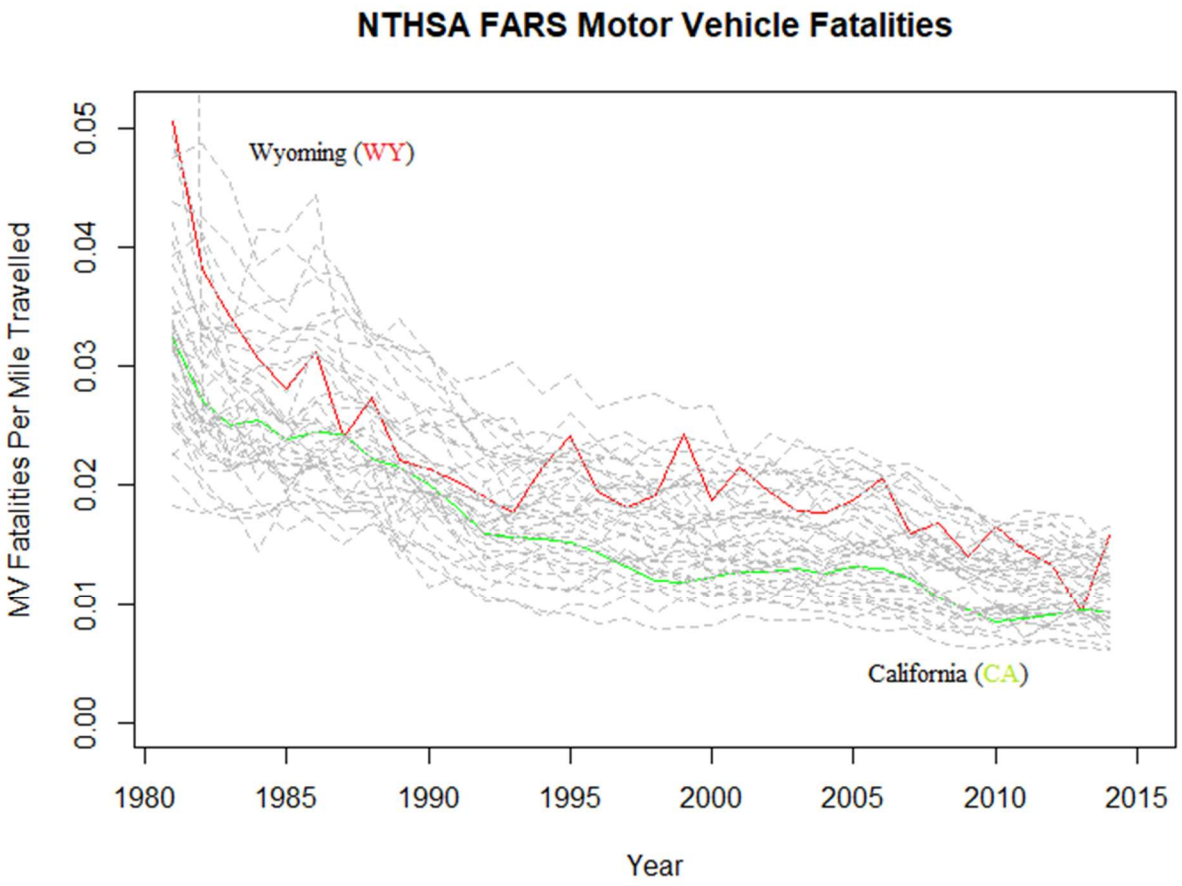

Figure 4.6 Motor vehicle fatalities scaled per vehicle mile travelled

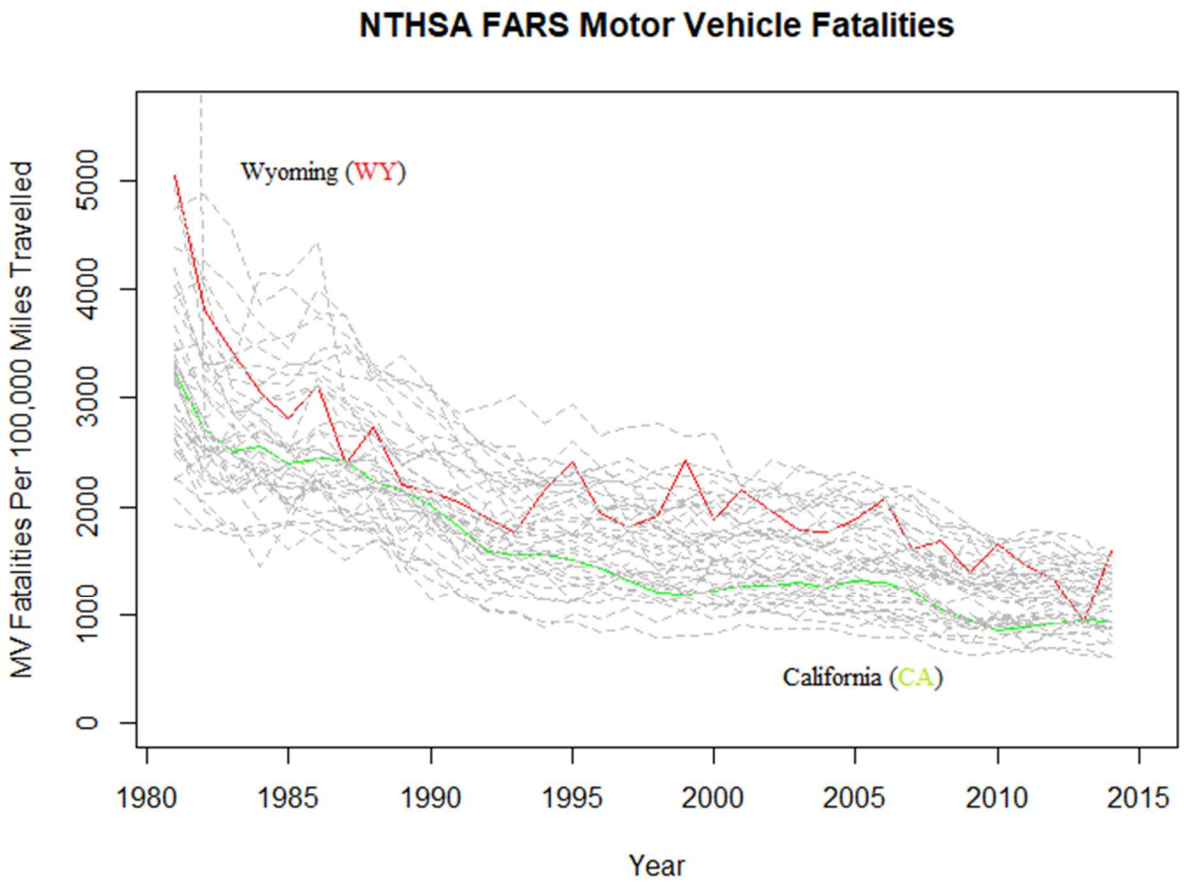

Figure 4.7 Motor vehicle fatalities scaled per 100,000 vehicle miles travelled

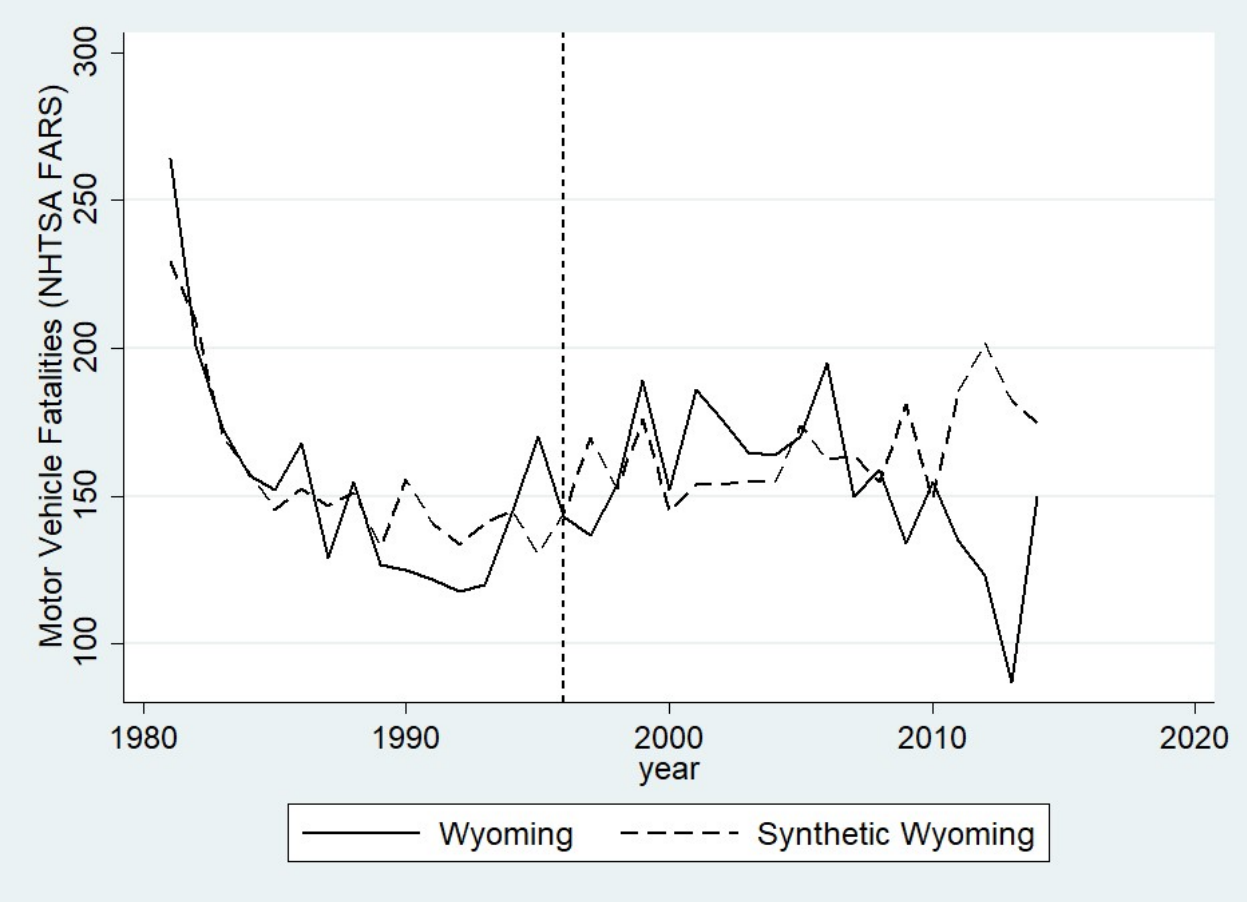

Figure 4.8 Synthetic Wyoming plot: Motor vehicle fatalities (raw frequencies)

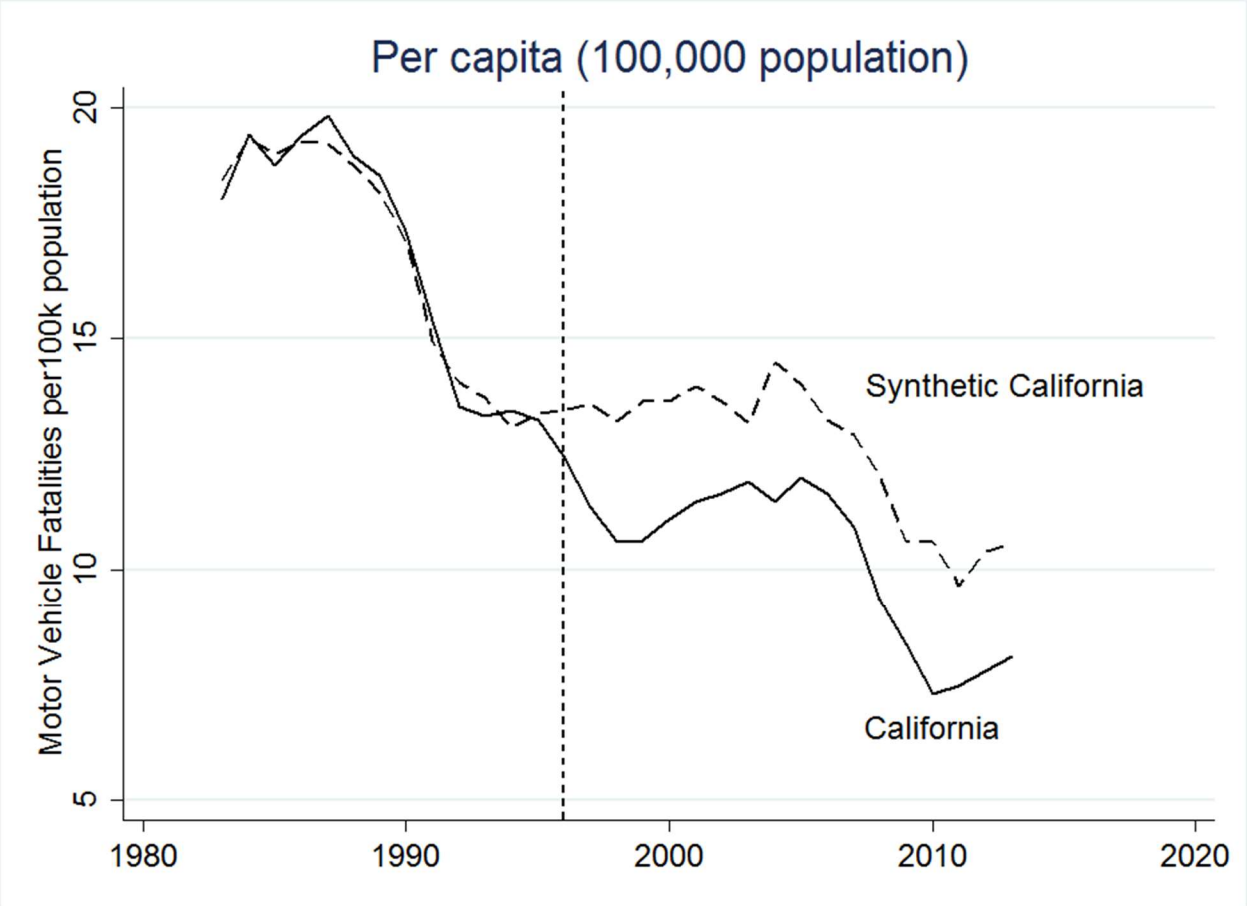

Figure 4.9 Synthetic California plot: Motor vehicle fatalities scaled per 100,000 residents

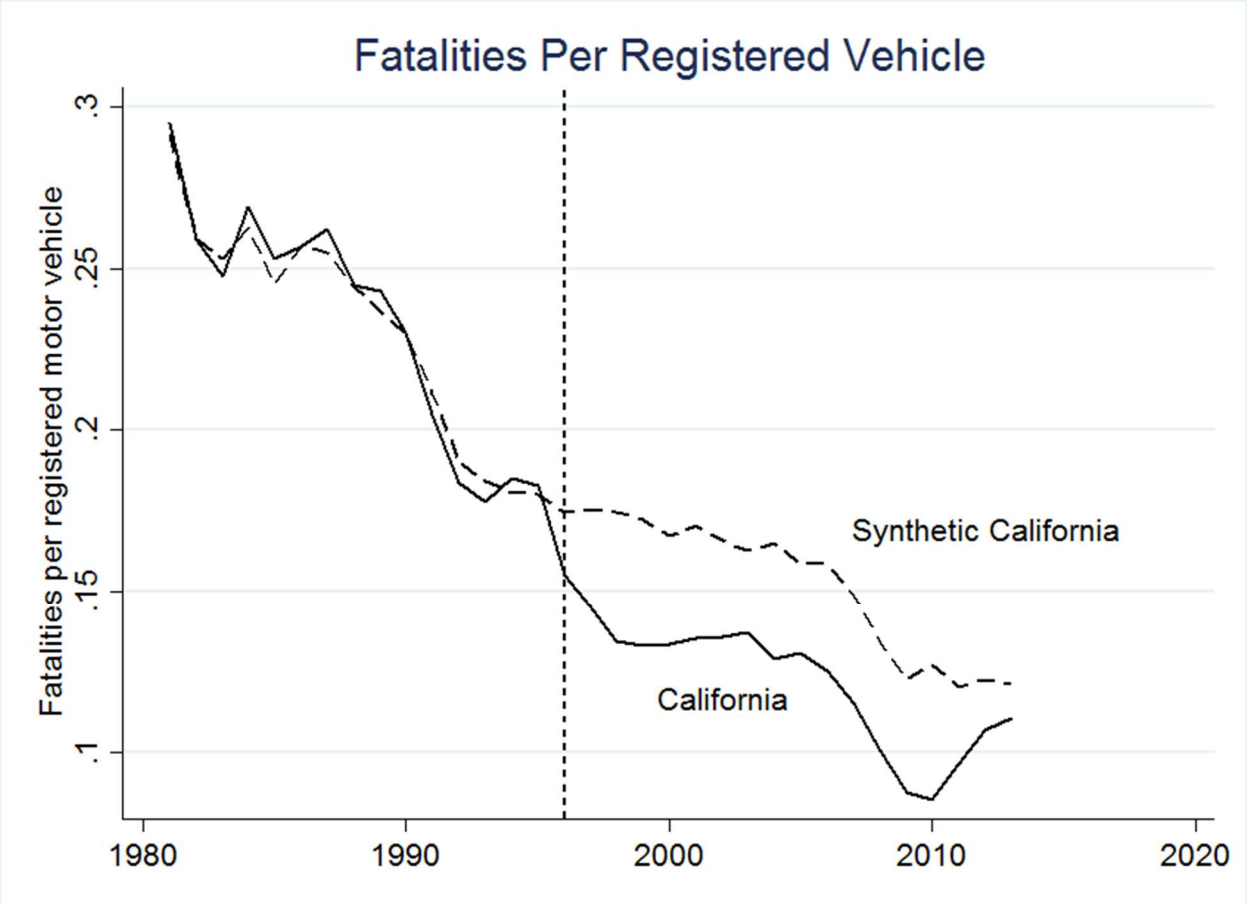

Figure 4.10 Synthetic California plot: Motor vehicle fatalities scaled per registered vehicle

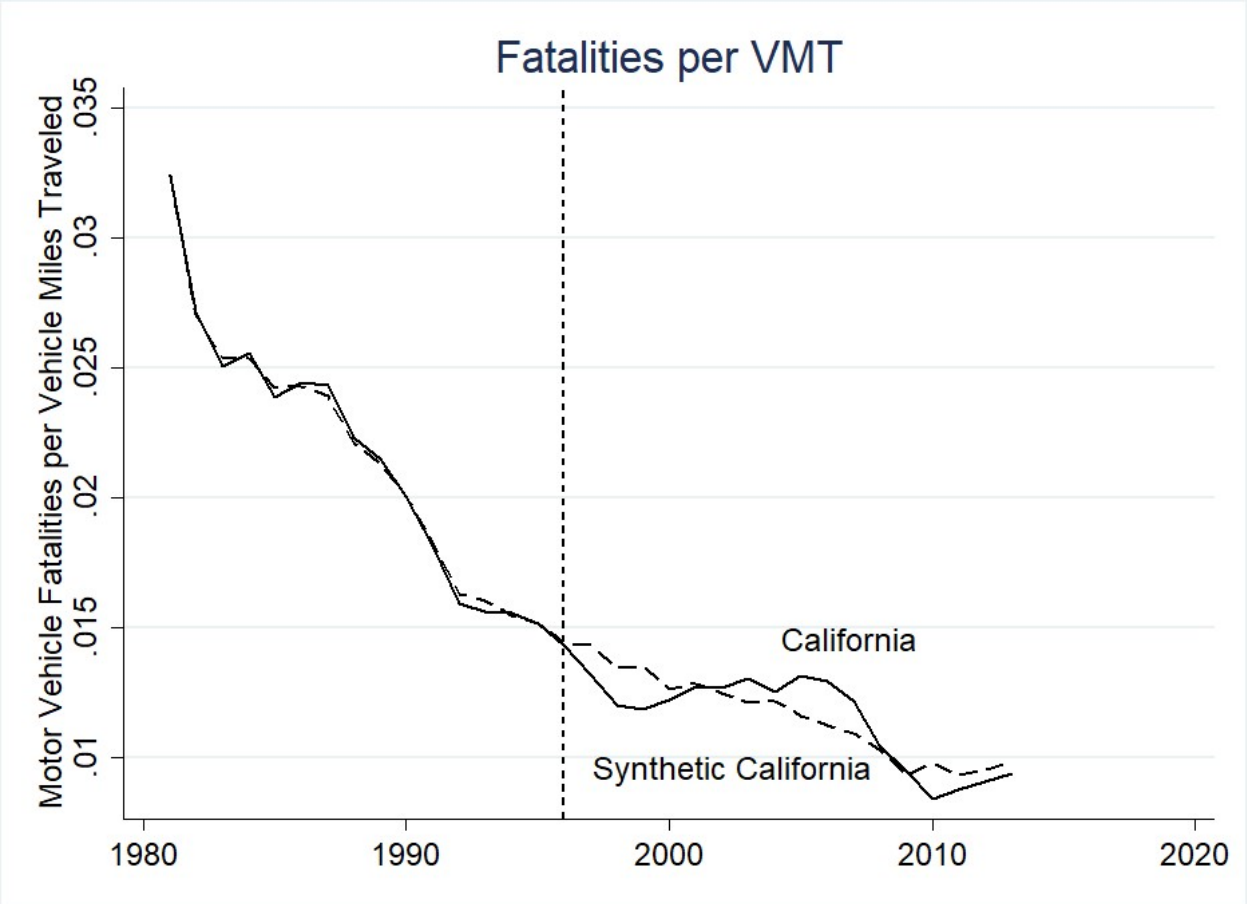

Figure 4.11 Synthetic California plot: Motor vehicle fatalities scaled per 100,000 vehicle miles travelled

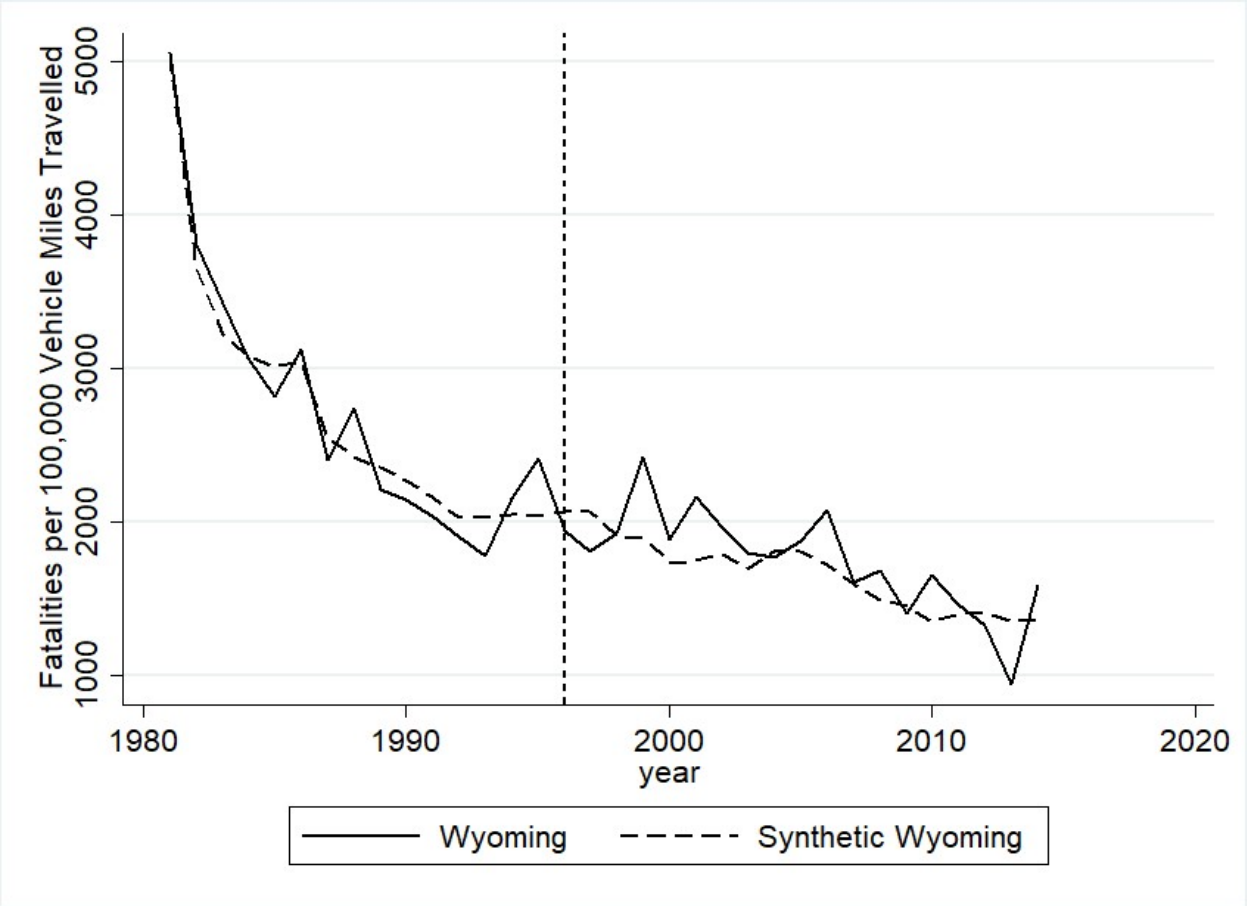

Figure 4.12 Synthetic Wyoming plot: Motor vehicle fatalities scaled per 100,000 vehicle miles travelled

Chapter 5 – To Fit? Or to Match?: Covariates and the Influence of Pre-intervention Observations

With few exceptions (Billmeier and Nannicini, 2013; Bohn et al., 2014; Lofstrom and Raphael, 2013; McCleary et al., 2017), the conventional approach to selecting predictors in research utilizing the synthetic control group design has been to include economic and demographic variables that have predictive power for explaining variation in the dependent variable -- referenced as "covariates" in the remainder of this text -- as well as a few pre-intervention observations of Y. Including covariates as predictors serves two purposes. First, if a factor is thought to causally impact the outcome $Y_{i,}$ then including the factor as a predictor in the SCG optimization algorithm should improve pre-intervention fit of the constructed synthetic control. Second if the treated and synthetic units have similar values for the included covariates, researchers can be more confident that any difference between Y_i and $\bar{Y}_i$ is due to the experience of the treatment rather than some change in economic or demographic trends.

This convention should be familiar to most, as omitted variable bias and collinearity are two of the primary sources of bias covered at length in most introductory statistics courses at the graduate level. Taking linear regression as our guiding analogy, omitting a predictive independent variable from the set of predictors used in the SCG optimization algorithm might seem like a similarly problematic source of bias. However, unlike linear regression, the estimate we are interested in producing is not an unbiased regression coefficient or *ceteris paribus* effect of a one-unit change in an independent variable. Instead, the goal of a SCG model is to construct a synthetic control unit that optimally reproduces the treated unit's pre-intervention outcome time series by minimizing the mean squared prediction error prior to the intervention. When the constructed synthetic control

follows the treated unit's pre-intervention outcome trends precisely, it serves as an approximation of the treated unit's outcome time series *had the intervention not occurred*. The post-intervention difference between the treated unit and its synthetic control provides an unbiased estimate of the change in the outcome caused by the treatment experience. Including collinear independent variables in a linear regression leads to biased β coefficients and threatens our interpretation of IV significance. However, collinearity does not bias the predicted outcome ($\bar{Y}_i$), only the significance of the betas. When collinear predictors are included in the SCG matching algorithm, one or both of the predictors will have smaller V weights than they would have if they had been included alone. However, any explanatory power that is not shared among the collinear factors will be ignored if only one of them is included as an SCG predictor. Therefore, when the inclusion of one factor diminishes the V weight otherwise associated with another factor, the treatment effect estimated by a SCG model is not biased in the same way that linear regression parameters would be.

This chapter of the dissertation argues, with considerable empirical support, that omitting (or ignoring) covariates identified in the literature from the SCG matching algorithm does not bias the estimated treatment effect when the pre-intervention time-series is longer than a few observation. As the number of pre-intervention observations increases and the time series exhibits greater variation in the pre-intervention time period, the predictive contributions from covariates quickly approach zero. In other words, including an omitted covariate generally does not change the composition of a synthetic control unit unless the pre-intervention time series is limited to a small number of observations.

Kaul et al.'s (2015) Argument

As applications of the SCG method abound in the fields of economics, public health, criminology, and political science, there have been instances where the conventional approach to

covariates has been somewhat ignored (examples include: Bartos & Kubrin, 2018; Bilgel and Galle, 2015; Billmeier and Nannicini, 2013; Bohn et al., 2014; Hinrichs, 2012; Kreif et al., 2016; Kubrin & Bartos, 2020; Liu, 2015; Nannicini and Billmeier, 2011; O'Neill et al., 2016; and Stearns, 2015). In all of these examples, the researchers used all of the available pre-intervention observations of the outcome as separate predictors when constructing their SCG. Kaul et al., (2015) argue that including all available pre-intervention observations of the outcome of interest causes covariate weights in the matching algorithm to approach zero. They further argue that, because covariate weights approach zero as more pre-intervention observations of the outcome are included as predictors, models that include all available pre-intervention observations of the outcome essentially optimize the pre-intervention fit of the outcome variable with very little, or zero, regard to whether the treated and synthetic unit resemble one another in terms of pre-intervention values of predictors (including covariates).

As empirical support for their argument, Kaul et al., (2015) replicate the Billmeier and Nannicini (2013) study under two alternatives, one version including the average value of the outcome variable across the pre-intervention time series as a predictor, and another including only the last observed pre-intervention value of the outcome as a predictor. They then compare the constructed SCGs under each constraint to the original estimate produced using all pre-intervention observations of the outcome as predictors. Across a series of models included in the Billmeier and Nannicini (2013) analysis, Kaul et al. (2015) show how limiting the inclusion of pre-intervention observations of the outcome to the pre-intervention average of the outcome or the last observation pre-intervention produces synthetic control groups that vary in their interpretations. They interpret the variation across these constraints as evidence that including all available pre-intervention observations of the outcome as predictors can lead to biased estimates of the treatment effect.

In this chapter I argue that the patterns observed by Kaul et al., (2015) can be interpreted as support for including pre-intervention observations of the outcome as SCG predictors, rather than evidence against their inclusion.

The primary flaw in Kaul's reasoning stems from their focus on constraining the pre-intervention outcome time-series to a single predictor term (i.e. pre-intervention average or last pre-intervention observation). For researchers, the appeal of including multiple pre-intervention observations of the outcome as predictors stems from the pre-intervention observations of the outcome reflecting the impact of (1) known and included covariates, (2) known but unavailable covariates, (3) unknown factors, and most importantly (4) the (unknown) varying relationship between these factors and the outcome over time (see Equation 1 in Chapter 3). When the average pre-intervention value of the outcome or the last pre-intervention outcome are included on their own as predictors, the predictor contributes what essentially amounts to an intercept or scaling term.

As one would expect, the covariates included as SCG predictors are assigned much larger weights when pre-intervention observations of the outcome are limited to a single predictor term. If each pre-intervention observation of the outcome reflects the changing relationship between covariates and the outcome over time, then limiting these observations to a single predictor term functionally eliminates a great deal of collinear terms from the model. This comes at a great cost, however, as the elimination of collinearity also excludes all of the unknown variation and heterogeneity in the relationship between the covariates and the outcome that the pre-intervention time series of the outcome reflects. In other words, the alternative models presented in Kaul involve a great deal of information loss and are therefore much less precise than the models presented in the original studies.

In the remainder of this chapter, I replicate the synthetic control group analysis presented in three of Abadie's (2003, 2010, 2015) foundational articles which defined and developed the synthetic control group design. In addition to replicating the original models presented in Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015), I construct three synthetic control units for each study, using (1) covariates as they were specified in the original model as predictors, excluding pre-intervention observations of the outcome entirely (2) all pre-intervention observations of the outcome of interest as well as first differences as predictors, excluding covariates, and (3) all pre-intervention observations of the outcome, first differences, and all of the original covariates. The pattern of results from the SCG models matched on alternative sets of predictors demonstrates how the inclusion of pre-intervention observations of the outcome leads to increasingly precise estimates of the treatment effect, and that restricting or eliminating these observations from the set of SCG predictors invariably results in a worse fitting pre-intervention match (i.e. increased RMSPE) as well as less precise estimates of the intervention's impact on the outcome. While models including all available pre-intervention observations of the outcome of interest and models with few or no pre-intervention observations included as predictors can produce post-intervention gaps with different interpretations, these disagreements only appear when the pre-intervention fit drastically improves.

Abadie 2003- Basque replication

Abadie and Gardeazabal (2003) foundational study analysed the impact that increased terrorist activity in the late 1970s and 1980s had on GDP growth in Spain's Basque Country. Drawing direct inspiration from Paul R. Rosenbaum's (1995) statistical matching methods for observational studies, Abadie and Gardeazabal (2003) were the first to publish a study utilizing synthetic control group analysis. Spain's Basque Country was a particularly interesting economic case study, as the region had been one of the richest in Spain prior to the start of separatist terrorist

activity in the early 1970s. Thirty years later, Basque Country had fallen from third richest to sixth. A pure time-series analysis model would fail in this case, because the impact of Basque terrorism on GDP growth would be contaminated by the economic downturn Spain experienced nation-wide during the height of Basque terrorism activity (i.e. late 1970's and early 1980's). Comparing Basque's GDP growth time-series to Spain as a whole would also be insufficient, as Basque Country differed from other Spanish regions in terms of factors thought to be related to economic growth potential at the outset of terrorism in the region. Therefore, the authors sought to construct a better control time-series than was available in reality.

Abadie and Gardeazabal (2003) utilized a dataset containing Annual GDP values (in 1986 US Dollars) for all 17 regions in Spain spanning 1955 to 1997. In addition to the GDP outcome time-series for Spain's 17 regions, the authors gathered 1969 population density, annual Gross Total Investment between 1964 and 1969, sectoral shares of total production for six industries from 1961-1969, and four metrics of "human capital" (i.e. literacy and educational attainment among working-age population) from 1964 to 1969.

Replication Model

Abadie and Gardeazabal (2003) supplied the synthetic control group matching algorithm with a set of predictors listed in Table 5.1. Along with the dependent variable (per capita GDP) averaged from 1960-1969, the authors also include the average ratio of Gross total investment over GDP from 1964 to 1969, population density in 1969, sectoral shares over total production averaged from 1961 to 1969 for six industry sectors (Agriculture, Energy, Industry, Construction, Marketable services, and Nonmarketable services) and human capital (using educational attainment as a proxy) broken into four categories (Illiterate, primary school, high school, greater than high school). Table 1 displays the value of these predictors for Basque Country, the synthetic control for Basque Country

reported in Abadie and Gardeazabal (2003), and the mean value across regions of Spain (i.e. the sample mean). From a simple visual comparison of the predictor values in Table 1, it appears as though the synthetic control unit reported by Abadie and Gardeazabal (2003) outperforms the crude sample mean for each predictor.

[Insert Figure 5.1 and Table 5.1 about here]

Using the predictors specified by Abadie and Gardeazabal (2003), the replication model produced a Synthetic Basque unit comprised of Catalonia (w=0.8508) and Madrid (w=0.1492). The results of the replication model are plotted in Figure 5.1. Unlike the Abadie models I address in the two subsequent sections, Abadie and Gardeazabal (2003) neglect to include a vertical reference line in 1970 corresponding to the treatment time period. This may be to be an intentional decision, as Figure 5.2 shows, terrorist activity in Basque Country began to increase in 1970 but it did not reach its annual peak until 1981, slowly returning to early 1970s levels by 1997. Given the pattern of terrorist activity plotted in Figure 5.2, assuming that terrorist activity in Basque acted as an abrupt and complete intervention (i.e. an intervention which neatly splits the time series into completely untreated pre-intervention observations and fully treated post-intervention observations) appears problematic. As Figure 5.2 shows, the height of terrorist activity in Basque lasted from 1978 to around 1987. This suggests that Basque did not experience the full dosage of the terrorist treatment until 1978. Therefore, it is understandable why the authors would be hesitant about specifying a clear pre/post time point on their results figure. To their credit, the per capita GDP trends for Basque and the Synthetic Basque reported in Abadie and Gardeazabal(2003) do not diverge substantially until 1978. This leads to the same interpretation -- that terrorist activity in Basque led to a reduction in per capita GDP which gradually became a larger reduction from 1978 to 1983 and stabilized from there -- as a synthetic control group model with a pre-intervention period lasting

from 1955 to 1978 would produce. In effect, Abadie and Gardeazabal (2003) constructed their Synthetic Basque unit using a censored pre-intervention time series. I revisit the issue of treatment definition and splitting the time series into pre- and post-intervention segments when I replicate the Abadie, Diamond, and Hainmueller (2010) model in the next subsection of Chapter 5.

[Insert Figure 5.2 about here]

Alternative Model Specifications

The synthetic control group model reported in Abadie and Gardeazabal (2003) used only one pre-intervention observation of the outcome of interest (per capita GDP averaged from 1960 to 1969) as a predictor in addition to the covariates displayed in Table 1. In order to evaluate the arguments presented by Kaul et al. (2015), I reconstruct Synthetic Basque Country using three sets of predictors. First, the GDP predictor term is excluded, and Synthetic Basque Country is reconstructed using only the covariate predictors supplied by Abadie and Gardeazabal (2003). Second, I exclude Abadie and Gardeazabal (2003) covariates and reconstruct Synthetic Basque Country using all available pre-intervention observations of per capita GDP, as well as first and second order differences as predictors. Third, I reconstruct Synthetic Basque Country using all available pre-intervention observations of per capita GDP, first difference scores, as well as Abadie and Gardeazabal original set of covariates as predictors.

[Insert Figures 5.3-5.5 and Table 5.2 about here]

Figure 5.3 plots the results of the Synthetic Basque Country constructed using only covariate predictors reported in Abadie and Gardeazabal (2003). The solid line represents Basque Country's actual per capita GDP time series and the dashed line represents the Synthetic Basque unit constructed excluding the pre-intervention observation of the DV time series from the predictor set and using only covariate predictors. Table 5.2 displays the donor pool weights for the original

Synthetic Basque Country reported in Abadie and Gardeazabal (2003) alongside the weights for the three alternative models described above. Considering the small change in the composition of Synthetic Basque when the pre-intervention GDP per capita predictor is excluded, it should come as no surprise that Figures 5.1 and 5.3 are difficult to distinguish side by side. The pre-intervention fit suffered somewhat (original model RMSPE=0.0881, covariates only model RMSPE=0.0907, see Table 5.3), however, the sign and magnitude of the estimated intervention impact on per capita GDP did not change.

Figure 5.4 plots the results of the Synthetic Basque Country constructed using only pre-1970 observations of per capita GDP and first and second order differences as predictors. The solid line represents Basque Country's actual per capita GDP time series and the dashed line represents the Synthetic Basque unit constructed excluding the covariates reported by Abadie and Gardeazabal (2003) from the predictor set and using only pre-1970 observations of the outcome and differences as predictors. Although the composition of Synthetic Basque changed more substantially than when the covariates were excluded in Figure 5.3, the pre-1970 observations of the dependent variable and difference scores still performed well. Again, Figures 5.1 and 5.4 are difficult to distinguish side by side. Although the pre-intervention fit suffered somewhat (original model RMSPE=0.0881, pre-1970 DV observations and differences only model RMSPE=0.0909, see Table 5.3), the sign and magnitude of the estimated intervention impact on per capita GDP did not change substantively.

Figure 5.5 plots the results of the Synthetic Basque Country constructed using all pre-1970 observations of per capita GDP and first and second order differences as well as the full set of covariates reported by Abadie and Gardeazabal (2003) as predictors. Although the composition of Synthetic Basque changed very little from the original model, the pre-1970 observations of the dependent variable, difference scores, and covariates outperformed the original Abadie and

Gardeazabal (2003) model in terms of pre-intervention fit (original model RMSPE=0.0881; covariates, pre-1970 DV observations, and differences model pre-intervention RMSPE=0.0861, see Table 5.3).

In order to make the visual comparison of the three alternative model results against the original model, Figures 5.6-5.8 plot the difference (or "gap") between Basque Country's actual per capita GDP time series and the Synthetic control unit constructed by Abadie and Gardeazabal (2003) alongside the gap between Basque and Synthetic Basque constructed using only the covariate predictors, using only pre-intervention observations and difference scores as predictors, and using all pre-intervention observations and difference scores as well as covariates as predictors, respectively.

[Insert Figures 5.6, 5.7 and 5.8 about here]

While the pre-1970 fit was marginally worse than the original model reported in Abadie and Gardeazabal (2003) according to the RMSPE values displayed in Table 5.3, neither the covariates only model or the pre-intervention levels and differences only model produced substantive changes in the interpretation of the synthetic control model results. Similarly, the model constructed using all pre-1970 observations, differences, and covariates marginally outperformed the Abadie and Gardeazabal (2003) model (see Table 5.3), the interpretation remained nearly identical to the original model (original model Mean Squared Error Post-1969=0.7152; covariates, pre-1970 DV observations, and differences model Mean Squared Error Post-1969=0.7167, see Table 5.3). Table 5.3 shows the covariate balance between Basque, the Abadie and Gardeazabal (2003) Synthetic Basque, and the three alternative Synthetic controls I outlined above. Contrary to the argument made by Kaul et al. (2015), across all the models in Table 5.3 the covariate balance only varies marginally. The models excluding pre-intervention observations of the dependent variable and covariates performed slightly worse than the original model and the model including both types of

predictors. While small variances did appear across the original and alternative Synthetic Basque models in this subsection, all of these models would lead researchers to the same conclusion. While this is not necessarily the case for the subsequent Abadie studies examined in this Chapter, these results suggest that including pre-intervention observations of the dependent variable as predictors a) produces a better fitting synthetic control unit than covariates alone and b) does not produce synthetic control units with poor covariate balance and c) leads to an improvement in pre-intervention fit without sacrificing covariate balance when included alongside covariate predictors.

[Insert Table 5.3 about here]

Abadie 2010- Proposition 99 Replication

Although Abadie, Diamond, and Hainmueller's (2010) study examining the impact California's 1988 tobacco control program on statewide tobacco consumption was not the first to utilize the synthetic control group design, it is arguably the most influential and widely recognized application of the method. Abadie et al (2010) extend the idea of placebo studies to produce quantitative inferences in comparative case studies. Abadie and Gardeazabal (2003) repeated the Basque analysis for untreated Catalonia as a single unit "falsification test". Abadie et al (2010) propose an iterative placebo test similar to the framework for permutation inference. By iteratively reassigning the treatment condition to donor pool units, authors assess whether the estimated effect is large relative to effects estimated in untreated units. When researchers present the results of SCG models, Abadie et al's influence (2010) is often reflected in the organization of the results and the interpretation of the in-sample placebo test results.

The 1988 California voter initiative known as Proposition 99, was a large-scale tobacco control program that increased taxes on tobacco sales, directed the resulting tax revenue to fund anti-smoking education and media campaigns, and spurred local clean air ordinances statewide

(Siegel, 2002). Abadie et al (2010) estimated Proposition 99's impact on tobacco consumption in California, using an outcome time series of annual cigarette sales (in packs) per capita as a proxy for tobacco consumption generally. The authors included three pre-intervention observations of the outcome (i.e. cigarette sales in 1975, 1980 and 1988) as well as a set of four structural covariates as predictors in the SCG matching algorithm. The included covariates -- Logged GDP per capita, percent of state population aged 15 -24, retail price of a pack of cigarettes (in cents), and per capita beer consumption (in gallons) were all averaged from 1980 to 1988.

Replication Model

[Insert Figures 5.9 & 5.10 about here]

The synthetic control group results presented in Abadie et al. (2010) are plotted in **Figure 5.9**, with California's actual cigarette sales (in packs) time series represented by the solid line, and its synthetic control represented by the dashed line. The covariate balance and donor pool composition (i.e. state weights, $W_1+W_2+\ldots W_J=1$) for Abadie et al.'s (2010) Synthetic California are listed in **Table 5.4** and **Table 5.5** respectively. According to the article text and the Stata syntax provided as part of the Stata "Synth" routine (Hainmueller, Abadie, and Diamond, 2014), the authors supplied the matching algorithm the set of predictors listed in the leftmost column of Table 5.4. Unlike the Abadie and Gardeazabal (2003) model, Abadie et al. (2010) include three single year pre-intervention observations of the outcome time series, rather than a single term for the average value of the outcome over a range of pre-intervention years. In addition, the authors included a set of covariates thought to be causal factors that influence smoking prevalence. The covariates include: the average price of a pack of cigarettes (averaged between 1980 and 1988), logged per capita state GDP (averaged between 1980 and 1988), the percent of the population aged 15-24 (averaged between 1980 and 1988), and per capita beer consumption (averaged between 1984 and 1988).

[Insert Table 5.4 about here]

It is worth noting that the example data and syntax from the Stata "synth" ado package construct a similar, but comparably worse fitting Synthetic California than what the donor pool weights reported in Abadie et al. (2010) produce. While the difference is not substantial enough to change the authors' overall finding, it is curious that there is any difference between the reported model and the replication code model whatsoever. Figure 5.10 plots the synthetic control results produced by the example data and code from the Stata "synth" ado package. While Figures 5.9 and 5.10 may appear the same at first glance, their composition (displayed in columns two and three of Table 5.5) varies considerably. The most notable difference being the large increase in the weight assigned to Colorado (.164 reported, .261 via replication syntax) and an equivalent reduction in the remaining non-zero donor states. Since the solid line falls below the dashed line following the Proposition 99 intervention in Figures 5.9 and 5.10, both models, regardless of their compositional difference, suggest that Proposition 99 caused per capita cigarette sales (in packs) to decline in the state.

In addition to the manner in which pre-intervention observations of the outcome series were included as predictors, the Abadie et al. (2010) model departs from the Abadie and Gardeazabal (2003) model by specifying a precise time period in which the intervention occurred and labelling their results with a vertical reference line. This decision was likely due to the continuous, fuzzy onset of Basque terrorist activity compared to the discrete, "effective as of" precision of California's Proposition 99.

Alternative Model Specifications

[Insert Figure 5.11 about here]

The Abadie et al. (2010) model includes three separate, single-year pre-intervention observations of the outcome of interest (i.e. per capita cigarette sales in packs) as predictors in the matching algorithm. To determine the dependency of the Abadie et al. (2010) model on the contribution of the three pre-intervention observations of the outcome, Figure 5.11 plots the synthetic control unit produced when the three cigarette sales observations pre-Prop 99 are dropped, and Synthetic California is constructed using the covariate predictors alone. The black and dashed lines in Figure 5.11 are clearly not co-integrated up to the point of the intervention, and Synthetic California performs poorly from 1975 through the intervention year. Clearly, the 1975, 1980, and 1988 observations of per capita cigarette sales are essential predictors, and without their inclusion, the resulting synthetic control unit is so poorly fit, it becomes uninterpretable.

[Insert Figure 5.12 and Table 5.5 about here]

To determine the relative contribution that the pre-intervention observations of the outcome make to Synthetic California's pre-intervention fit, the Synthetic California plotted in Figure 5.12 is constructed from pre-intervention observations of the outcome of interest and first-differences alone. With the retail price per pack of cigarettes, logged per capita state GDP, the proportion of residents aged 15-24 years, and beer consumption excluded from the synthetic control matching algorithm, the Synthetic California plotted in Figure 5.12 performs quite well throughout the pre-intervention period. Unlike the version constructed using covariates alone, the covariates-excluded model in Figure 5.12 largely maintains a parallel relationship with California from 1970 through 1989, and the post-intervention gap produces a similar estimate of the intervention's impact (in terms of sign and magnitude) to the model reported in Abadie et al. (2010). In this case, it appears that the covariates alone are unable to produce a well-fit synthetic control unit, whereas the pre-1989 observations of per capita cigarette sales (in packs) produced a Synthetic California similar to the

version reported in Abadie et al. (2010) both in terms of composition (see Table 5.5), and estimated intervention effect (see row labelled "Mean Squared Error Post 1988 at the bottom of Table 5.5).

[Insert Figure 5.13 about here]

Figure 5.13 plots the Synthetic California produced when the matching algorithm is provided with the pre-1989 observations of the outcome and first-differences as predictors (predictors used to construct Figure 5.12) alongside the covariate predictors used to construct Figure 5.11. Like Figures 5.9 and 5.12, the solid and dashed lines track together quite well up to the intervention and suggest a substantial decrease in per capita cigarette sales was caused by Proposition 99. One noteworthy covariate, the average retail price of a pack of cigarettes, appears to provide a distinct source of variation in cigarette sales which the pre-intervention observations and other covariates do not. When the retail price of cigarettes is excluded from the predictor set, the pre-intervention fit between California and Synthetic California suffers. Across the versions of Synthetic California presented in Figures 5.9-5.13, there is a tendency for the first three pre-intervention years (i.e. 1970-1973) and last two pre-intervention years (i.e. 1987-1988) tend to be the poorest fit (i.e. largest gap) years. With these two patterns in mind, Figure 5.14 plots Synthetic California constructed from a set of predictors including: all pre-intervention values of the outcome, first-differences, the covariate terms from Abadie et al. (2010), and seven single year predictors for the retail price of cigarettes in years 1970, 1971, 1972, 1973, 1986, 1987, and 1988. Beyond the patterns noted above, the retail price of cigarettes would be expected to be a strong predictor of cigarette sales, as the retail price would not only reflect the immediate shock of a tobacco sales tax increase (a component of Proposition 99), but also serves as a limiting factor for the number of packs people can afford to purchase. Therefore, providing the matching algorithm with single-year predictor terms for retail cigarette price in the pre-intervention years with the most persistent gap might improve the pre-

intervention fit between California and Synthetic California. Although Figure 5.14 is difficult to distinguish from Figure 5.9 visually, the composition and fit statistics presented in Table 5.5 indicate that Figure 5.14 provides a more precise pre-intervention fit than the versions plotted in Figures 5.9-5.13. While the general interpretation of Abadie et al.'s (2010) original model and the model with the best pre-intervention fit (Figure 5.14) are largely the same, the version of Synthetic California from Figure 5.14 suggests a slightly smaller reduction in per capita cigarette sales (in packs).

[Insert Figures 5.14 and 5.15 about here]

Through the alternative Synthetic Californias constructed under various sets of predictors presented in this section, perhaps the most striking finding was that covariate-alone are insufficient to produce a synthetic control unit that is reasonably well-fit throughout the pre-intervention period. Figure 5.15 plots the differences (or gap) between California and the Synthetic California reported in Abadie et al. (2010) as a solid black line alongside the gap between California and the synthetic control constructed using only covariates as predictors as in Figure 5.10 (as the dashed line). Clearly, per capita cigarette sales in 1975, 1980, and 1988 are essential predictors, as the solid line hovers near a zero-gap from 1972 to 1987 while the dashed line produces a large gap by 1972 and never returns to zero.

[Insert Figure 5.16 about here]

Figure 5.16, on the other hand, plots the gaps between California and its synthetic control for the Synthetic California constructed using only pre-intervention observations of the outcome and first differences as predictors (as in Figure 5.11) as a dashed line with the Abadie et al. (2010) gap again as a solid black line. In comparison to the two gap plots presented in Figure 5.15, the two series in Figure 5.16 appear to be distinguished only by noise. While the pre-intervention fit was

better in the Abadie et al. (2010) model (as the Pre-1989 RMSPE row in Table 5.5 shows), the interpretation of the two models is largely the same even though all of the original covariate predictors (i.e. causal factors that influence smoking prevalence) are excluded. From Figures 5.15 and 5.16, including pre-intervention observations of the outcome of interest appears to be necessary in order to produce synthetic control unit that is well-fit throughout the pre-intervention period. This finding detracts from Kaul et al.'s (2015) argument, as covariates alone cannot produce an interpretable synthetic control unit. Furthermore, when pre-intervention observations of the outcome of interest and first differences alone are supplied as predictors, the resulting Synthetic California is quite comparable to Abadie et al.'s (2010) synthetic control, both in terms of composition and general interpretation. In comparison to the huge divergence shown in Figure 5.15, the two trends in Figure 5.16 appear almost trivially different from each other. Figure 5.17 plots the gap between California and Synthetic California from Abadie et al. (2010) alongside the gap plot from the model resulting from the syntax included with the Hainmueller et al. (2014) Stata "synth" routine. Figure 5.17 contextualizes the gaps plotted in Figures 5.15 and 5.16. While Figures 5.15 and 5.16 demonstrate that pre-intervention observations of the outcome and first differences are the primary determinants of pre-intervention fit, Figure 5.17 illustrates the range of gaps that can result from minor technical differences in computing power (likely RAM), rounding to different decimal points, and other hidden modelling mechanisms can produce.

[Insert Figures 5.17 and 5.18 as well as Table 5.6 about here]

Finally, Figure 5.18 compares the gap between California and Synthetic California for Abadie et al.'s original model alongside the gap plot for the best fitting model presented in Figure 5.14 above. The solid and dashed lines appear almost indistinguishable in Figure 5.18, even in comparison to the plots in Figure 5.17. However, as the Pre-1989 RMSPE row of Table 5.5

indicates, the model presented in Figure 5.14 outperforms Abadie et al.'s (2010) model in terms of pre-Prop 99 fit. Through these gap plots, the relative contributions to pre-intervention fit provided by different classes of predictors are illustrated, with pre-intervention observations of the outcome of interest playing the largest role for pre-intervention fit. More importantly, however, the only model to produce an intervention effect estimate that varied in substance (rather than mere precision) from the Abadie et al. (2010) model, was the model constructed using covariate predictors alone. Furthermore, once all pre-intervention observations of cigarette sales were included in the predictor set, the matching algorithm began to converge on the optimal composition for Synthetic California. As covariates and additional terms were added back into the predictor set alongside pre-intervention observations of the outcome, the pre-intervention fit became more and more precise, but these additions seem to yield quickly diminishing returns. Importantly for Kaul et al. (2015) the stability of Synthetic California in terms of composition, pre-intervention fit, and the sign and magnitude of the post-Prop 99 gap, across the models presented in Figures 5.9, and 5.11-5.14 do not suggest that including pre-intervention observations of the outcome introduces bias to the estimate. In fact, the practicality of Kaul et al.'s (2015) headline argument, "never use all pre-intervention outcomes together with covariates as predictors", appears wholly impractical in this case.

Abadie 2015- German Reunification replication & Taylor Series Expansion

In their most recent study utilizing a SCG design, Abadie, Diamond, and Hainmueller (2015) estimate the impact of Germany's 1990 reunification on West Germany's GDP trend. SCG procedures introduced in Abadie and Gardeazabal (2003) and Abadie et al (2010) are presented as part of a more complete and organized framework for the application of SCG designs in comparative politics and comparative case studies more generally. In addition to a clear synthesis of their previous SCG design innovations, the authors introduce two additional post-estimation tests in

their analysis of Germany's 1990 reunification. First, the authors evaluate the sensitivity of their estimated intervention effect by iteratively removing donor pool units that received non-zero weights and reconstructing synthetic West Germany from the remaining units. If the post-intervention GDP gap remains comparable under alternative compositions of synthetic West Germany, then the estimated intervention effect is not dependent on the contribution of specific donor pool units. This "Leave one out test", as discussed in Chapter 3, provides an additional protection against *selection* as a threat to internal validity.

The second post-estimation test introduced in Abadie et al (2015) builds upon the in-sample placebo test introduced in Abadie et al (2010) by applying the same permutation inference framework to the time dimension rather than the unit/spatial dimension. The "In-time placebo test" presented in Abadie et al (2015) iteratively reassign the 1990 German Reunification intervention to years prior to 1990 and reconstruct Synthetic West Germany to produce estimated intervention effect on GDP years before Germany reunified. In order for an in-time placebo test to be applicable, however, researchers must assume that the treated unit did not experience any structural shocks to its outcome time-series in pre-intervention years. This assumption cannot be reasonably made for most interventions and outcomes, so its applications tend to be limited to unique treatments and natural experiments.

The authors construct synthetic West Germany's GDP time series from a donor pool of 16 OECD member countries. Abadie et al (2015) include a set of standard economic growth predictors in the synthetic control group matching algorithm. In addition, a pre-intervention average of the outcome (i.e. GDP per capita) the authors include inflation rate, industry share of value added, investment rate, schooling, and trade openness averaged for the 1981-1990 period as predictors.

Replication Model

Unlike the previous two Abadie studies replicated in this chapter, Abadie et al. (2015) split the pre-intervention period into two segments: a training period (1971-1980) and a validation period (1981-1990). The authors fit a synthetic control group for West Germany using predictor observations from the training period, and select **V** (i.e. the matrix of predictor variable weights) so that it minimizes the Root Mean Squared Prediction Error in the validation period. Then, the **V** weights selected in the previous model are imposed upon the synthetic control group matching procedure, now using predictor observations from the validation period (1981-1990) to estimate West Germany. Abadie et al. (2015) argue that the choice of **V** (i.e. the matrix of predictor variable weights) could be subjective and based in theory, but they opt for an approach which "intuitively minimizes out-of-sample errors" (Abadie et al., 2015, p.502). The authors argue that a **V** matrix, supplied with predictor observations from 1971-1980 and selected to minimize prediction error from 1981-1990, has more out-of-sample validity than a model which minimizes the whole pre-intervention time segment. While there may be situations where an expected causal linkage between predictors and the outcome should be pre-specified or assumed to be stable over time. In practice, however, supplying a **V** matrix never produces a synthetic control unit with a smaller Root Mean Squared Prediction Error term than an analogous model without the V constraint.

[Insert Figure 5.19 about here]

The synthetic control group results reported in Abadie et al. (2015) are plotted in Figure 5.19, with West Germany's Purchasing Power Parity (PPP)-adjusted GDP time series represented by the solid line, and its synthetic control represented by the dashed line. The covariate balance and donor pool composition (i.e. state weights, $W_1+W_2+\ldots W_J=1$) for Abadie et al.'s (2015) Synthetic West Germany are listed in Table 5.7 and Table 5.8 respectively. According to the article text and

syntax provided as part of the R Synth routine (Hainmueller et al., 2014), the authors supplied the matching algorithm the set of predictors listed in the leftmost column of Table 5.7. Similar to the Abadie and Gardeazabal (2003) model, Abadie et al. (2015) include a single term for the average value of the outcome over a subset of the available pre-intervention observations. In addition, the authors included a set of covariates thought to be causal factors that influence long-term economic growth. The covariates include: the ratio of real domestic investment (private plus public) to real GDP (averaged between 1980 and 1985), the percentage of the population aged 25 years or older that completed secondary school (averaged between 1980 and 1985), industry share of value added, annual percent change in consumer prices, and trade openness (export plus imports as a percent of GDP (averaged between 1981 and 1990).

[Insert Tables 5.7 & 5.8 about here]

It is again worth noting that the example data and syntax from Hainmueller et al. (2014) do not perfectly reproduce the Synthetic West Germany reported in Abadie et al. (2015) when the procedure is performed in Stata. However, unlike the Abadie et al. (2010) replication package, the R-code does perfectly reproduce the reported resulting in Abadie et al (2015). It is quite suprising then, that the version of Synthetic West Germany produced using Stata to perform the training and validation procedure described in the article text, has a much smaller Root Mean Squared Prediction Error (RMSPE) than the reported model. Although the Stata translation of the R-script presented in Hainmueller et al. (2014) has a RMSPE 50% smaller than the original model (14955.4 in R, 9966.9 in Stata, see Table 5.8), the original model plot in Figure 5.19, the R-replication in Figure 5.20, and the state version of Abadie et al.'s (2015) modelling procedure in Figure 5.21 are difficult to distinguish visually. Although the improvement in pre-intervention fit is difficult to assess visually, the Stata-produced version of Synthetic West Germany plotted in Figure 5.21 suggests a smaller post-

reunification gap than the version described in Abadie et al. (2015). While Figures 5.19 and 5.20 are, in fact, precisely the same (successful replication), Figure 5.21 outperforms them both. The composition of Stata's Synthetic West Germany, displayed in column three of Table 5.8, varies considerably from the original model in column 2. The most notable difference being a non-zero weight assigned to the United Kingdom. Overall, the original model, the R-based reconstruction, and the Stata version, all support the conclusion that the 1990 reunification of West Germany caused GDP to be lower than it would have West Germany remained separate from East Germany. The R-script and Stata routine disagree, however, on the precise magnitude of the lost GDP.

[Insert Figure 5.20 and Figure 5.21 about here]

Alternative Model Specifications

[Insert Figure 5.22 about here]

The authors' decision to select a **V** matrix on only a subset of the pre-intervention segment undoubtedly explains, at least to a degree, why every alternative model I construct in the remainder of this section greatly outperforms the model reported in Abadie et al. (2015), in terms of pre-intervention fit at least. The other likely explanation behind the ease at which well-fit synthetic controls can be found for this outcome series, is the relative smoothness of the time series compared to the California cigarette sales time series in the previous section of Chapter 5. When the outcome series is smooth, there are, in effect, fewer extreme points on the time series that can only be reproduced via a certain set of donor pool units. Figure 5.22 plots Synthetic West Germany constructed using only the covariates supplied as predictors by Abadie et al. (2015). When the pre-intervention observation of GDP is excluded from the matching algorithm, the resulting West Germany has a smaller pre-intervention RMSPE (111.99) than the model reported in Abadie (122.29) et al. (2015). Table 5.8 lists the Fit Statistics and estimated intervention effects below

Synthetic West Germany's various compositions. The synthetic controls constructed from a set of predictors containing only covariates (no pre-intervention observations of the DV) did not produce a better pre-intervention fit (i.e. smaller RMSPE) than the models reported in Abadie and Gardeazabal (2003) and Abadie et al. (2010). Therefore, the covariate predictors only version of Synthetic West Germany producing a smaller RMSPE than the original model from Abadie et al. (2015) seems to be a byproduct of the authors use of a training and evaluation period, rather than the predictive utility of the covariates themselves. The Synthetic West Germany in Figure 5.22 appears to be quite well-fit throughout the pre-intervention segment, but the West Germany (PPP)-adjusted GDP time series is so smooth and free of inflection points that the Synthetic West Germany plotted in Figure 5.22 is actually the second worst pre-intervention fit presented in this Abadie et al. (2015) subsection.

[Insert Figure 5.23 about here]

To compare the ability of causal covariates and pre-intervention observations of the outcome series to produce a version of Synthetic West Germany that is well-fit pre-intervention, the synthetic control displayed in Figure 5.23 was constructed using only pre-intervention observations of the outcome series, first-differences, and second differences (similar to a Maclaurin series). The RMSPE values listed in Table 5.8 indicate that while the covariates-only model from Figure 5.22 improved upon the Synthetic West Germany reported in Abadie et al. (2015), the version constructed from exclusively pre-1990 observations and differences clearly outperforms both models in terms of pre-intervention fit. Although in this application the covariates alone produced a surprisingly well-fit synthetic control, pre-intervention observations of the outcome of interest produced synthetic controls with better pre-intervention fits than covariates-only models in each of the three replications presented in this chapter. At the very least, it is clear that pre-intervention

observations of the outcome series must be included (in some form) in the predictor set. Table 5.8 also lists the donor pool weights (i.e. composition of Synthetic West Germany) for all of the models in this subsection. Although the models in Figures 5.19-5.22 are all comprised of different mixtures of the same five or six nations (Austria, Netherland, Switzerland, Japan, United States, and the United Kingdom), the Synthetic West Germany constructed using only pre-intervention levels of the outcome series and differences (see Figure 5.23) assigned small weights to Denmark, Italy, Norway, and Greece and zero-weights to the Netherlands and Japan. Of course, a drastic reduction in the average pre-intervention gap (see Table 5.8) cannot occur without the donor pool weights changing, at least somewhat. From the results of the models in Figures 5.22 and 5.23, it appears as though the pre-intervention observation predictors are better able to produce synthetic controls that are well-fit pre-intervention than covariate predictor terms.

[Insert Figure 5.24 about here]

To evaluate the extent to which the covariate predictors and pre-intervention DV observations provide independent and exclusive utility vs shared/collinear utility to the matching algorithm, the Synthetic West Germany plots in Figure 5.24 and Figure 5.25 are constructed using the covariates from Abadie et al. (2015) as predictors alongside half of the available pre-intervention DV observations and all of the available pre-intervention observations and differences, respectively. Keep in mind the RMSPE and average pre-intervention gap values presented in Table 5.8. For reference, the model reported in Abadie et al. (2015) had an average pre-intervention gap of nearly 15,000, the model using only covariate predictors in Figure 5.22 brought the average pre-1990 gap down to ~12,540, and the model using only pre-intervention observations of the outcome and differences as predictors brought the average pre-1990 gap down to a staggering 3,815 (see Table 5.8). Figure 5.24 plots Synthetic West Germany constructed from the full set of Abadie et al. (2015)

covariates as well as half of the available pre-1990 DV observations. The pre-intervention observation predictors were able to outperform the covariates alone by a wide margin, but when only half of those predictors are included, it produces a Synthetic West Germany with an average pre-1990 gap (pre-1990 gap=~4652) that is about one-third of the covariates only model from Figure 5.22 (avg. pre-'90 gap= ~12540). A few conclusions can be drawn from Figure 5.24's fit statistics in Table 5.8. First, the pre-intervention DV observation predictors that were not included in the half-set of predictors for Figure 5.24 had predictive utility beyond the included pre-1990 DV observations as well as the covariates. This is made clear by the average pre-1990 gap approaching, but not reaching the pre-intervention fit from Figure 5.23 when half of the pre-intervention DV observations are supplied as predictors. Since the pre-1990 DV observations and differences alone produced a synthetic control unit with an average pre-1990 gap smaller than a half set of pre-1990 DV observations plus the original covariates, the excluded pre-intervention DV observations would be expected to reduce the average pre-1990 gap to 3825.8 (see Table 5.8), or perhaps even lower.

[Insert Figure 5.25 about here]

Figure 5.25 plots the Synthetic West Germany produced from including all original covariates and all pre-1990 GDP observations (n=30) and differences. If the covariate predictors do not have any exclusive predictive utility beyond the collinear pre-intervention observations of the outcome of interest, then the fit statistics for Figure 5.25 listed in Table 5.8 would be expected to be identical to the model in Figure 5.23 when covariates were excluded entirely and pre-1990 DV observations were supplied as predictors. As Table 5.8 shows, however, the average pre-intervention gap for the Synthetic West Germany in Figure 5.25 is 3710.38, which improves upon the fit from the model with covariates excluded from the predictor set. Therefore, in this application the

covariate predictors do provide utility to the matching algorithm beyond the utility they share with the pre-intervention DV observations.

Returning briefly to the authors' decision to impose a "training" and "validation" period on the pre-intervention segment. Once this procedure is dropped and the matching algorithm minimizes the gaps across the full pre-intervention segment; whereas the original model reported in Abadie et al. (2015) shows essentially no gap between West Germany and Synthetic West Germany from 1990 through 1993, all of the alternative compositions of West Germany in this chapter suggest, quite surprisingly, that PPP-adjusted GDP was actually greater than it would have been had reunification not occurred. Of course, the original model and all of the alternatives examined reflect Abadie et al.'s (2015) conclusions from 1993 onward. It is curious, however, that the authors' chose to introduce the training/validation split on an application which would have been more complex to interpret had they stuck to convention. Perhaps the authors were reluctant to weigh in on interpreting post-intervention gaps that vary in terms of both sign and magnitude in the results period. How should the small GDP increase from 1990 through 1993 be interpreted? German reunification perhaps caused a small but short-lived bump in GDP prior to the costs of reintegration were truly realized. Perhaps a second, unidentified intervention took place in 1993 and confounded the reunification effect. Synthetic case control designs are not particularly well suited for distinguishing between dynamic intervention effects over time and subsequent unidentified interventions biasing the estimate. Without an established convention for interpreting the 1990-1993 GDP bump, their decision to report a version of their model without the bump is quite understandable.

Discussion

Scholars critical of including all pre-intervention observations of the outcome as SCG predictors would likely argue that some sort of additional construct validity check should be performed, since the standard covariate balance approach is rendered useless when covariate V-weights are close to zero or zero, to assess rule out changes in theoretically identified structural covariates as an alternative explanation.

If including pre-intervention observations of the outcome of interest as predictors in the synthetic control matching algorithm produces the synthetic control with the most precise pre-intervention fit, but the covariate balance isn't quite as close as when few or no pre-intervention observations are included as predictors, a tradeoff is made. The better pre-intervention fit allows the model to detect more precise intervention effects, rank highly in post-estimation placebo testing, and move on to ruling out alternative factors as a post hoc internal validity check. In the replication of Abadie et al.'s (2010) model, it became clear that observations of the retail price per pack of cigarettes covariate were useful predictors. Even with all of the pre-intervention observations of the outcome included as predictors, retail price was still able to produce a model with a better pre-intervention fit than levels of the outcome and differences alone. Therefore, there may be cases in which covariates that are directly related to the outcome (i.e. retail price per pack-> Packs sold) add information to the matching algorithm that the pre-intervention DV observations do not contain. On the other hand, many social science outcomes have a complex causal nexus that can simply bloat the computing cost for the synthetic control matching algorithm. Therefore, both pre-intervention observations of the outcome and covariates thought to cause the outcome should be supplied as predictors for researchers' first pass at the model. They should then drop all of the covariates, determine if the synthetic control unit is worse fit pre-intervention than the previous model. If

dropping the covariates did not worsen the pre-intervention fit, then they should instead be used only as a post-hoc internal validity check. By including them as predictors, they appear to be a priori, but by separating them out and labelling the comparison as post-hoc, their (non)role in contributing to the synthetic control construction is made transparent. If a well-fit model with no covariates presents a post-hoc covariate balance check that appears even somewhat similar, then causal inferences can be drawn. If the opposite were to occur, and a covariates-only model produced a mediocre pre-intervention fit but the treated unit and synthetic control were perfectly balanced on six covariates, almost nothing of value can be concluded from the model. Of course, if the model with no covariates presents a post-hoc covariate balance check that shows huge differences in important causal covariates, it can either be evidence suggesting that the causal covariates included are not as strong of a causal factor as theory would have predicted it to be, or that the relationships between the causal covariates and the outcome are heterogeneous across units and/or time.

It is not necessary to assume complete (or even partial) knowledge of the causal mechanism by which an intervention produced a change in an outcome in order to estimate an intervention's causal impact. However, it does require the assumption that the data generating process for the outcome prior to the intervention *changed* in response to the treatment. An intervention changes the data generating process that produced the pre-intervention trends, however, while the post-intervention trends for the treated unit are driven by the *changed* post –intervention data generating process, the synthetic control maintains the pre-intervention data generating process though the results period.

The cost of such an assumption, however, is that the difference between the treated and synthetic control groups can most confidently be attributed solely to the intervention in the first post-intervention observation. Each subsequent post-intervention difference between treated and

control groups can be less confidently attributed to the intervention alone than the prior. The rate at which an intervention's impact decays can determine the post-intervention results cut-off point. Expectations guided by theory about the rate at which the impact of an intervention on an outcome would decay, if at all; determine the appropriate length of the post-intervention analysis period. We become less confident in our attribution of the post-intervention gap as the causal impact of the treatment due to the accumulating threat of confounding interventions and idiosyncratic shocks on the outcome of interest. As results are extended further and further out from the intervention, the threat of confounds and unidentified interventions grow precipitously.

A convenient example of this type of problem arises from Abadie's (2010) examination of California's Proposition 99 described above. As is surely the case in most, if not all, policymaking, Proposition 99 did not occur in a vacuum devoid of supplementary policies aiming to reduce tobacco consumption. Proposition 99 approved by voters in 1988, but a statewide ban on smoking in the workplace shortly followed in 1995 and expanded to restaurants in 1998. It would be disingenuous to attribute the entirety of the post-intervention gap to the 1988 intervention when the subsequent smoking bans in workplaces and restaurants surely had their own impacts on cigarette consumption. However, it is unknown whether the voter approval of Proposition 99 encouraged further tobacco legislation in the form of workplace and restaurant bans, or these would have occurred similarly in the absence of Proposition 99.

Finally, while the analyses in this chapter generally support including all available pre-intervention observations of the outcome as predictors in the synthetic control matching algorithm, the approach lacks parsimony, and future research should investigate the tradeoff between efficiency and precision that comes from excluding pre-intervention observations from the predictor set, as well as any bias the exclusions introduce.

Chapter 5 Tables

Table 5.1 Basque Predictors

	Basque	Sample Mean	Synthetic Basque (Abadie & Gardeazabal, 2003)
Real per capita GDP[a]	5.285	3.581	5.271
Investment ratio (%)[b]	24.647	21.424	21.583
Population density[c]	246.89	99.414	196.288
Sectoral Shares (%)[d]			
Agriculture, forestry, and fishing	6.844	21.353	6.179
Energy and water	4.106	5.31	2.76
Industry	45.082	22.425	37.636
Construction and engineering	6.15	7.276	6.952
Marketable services	33.754	36.528	41.104
Nonmarketable services	4.072	7.111	5.371
Human capital (5)[e]			
Illiterates	3.321	10.983	7.645
Primary or without studies	85.893	80.911	82.285
High school	7.522	5.427	6.965
More than high school	3.264	2.679	3.105

Sources: Data reported by Abadie and Gardeazabal(2003) gathered from Matilde Mas et al.(1998) and Fundacion BBV (1999).

[a] 1986 USD, average for 1960-1969.

[b] Gross Total Investment/GDP, average for 1964-1969

[c] Persons per square kilometer, 1969.

[d] Percentages over total production, 1961-1969.

[e] Percentages over working-age population, 1964-1969.

Table 5.2 Synthetic Control Composition, Fit Statistics, and Intervention Estimates

	Synthetic Basque (Abadie & Gardeazabal, 2003)	Synthetic Basque (covariates only)	Synthetic Basque (Pre-intervention lvls and differences, no covariates)	Synthetic Basque (All Pre-1970 obs, differences, and covariates)
Spain (Espa)	0	0	0	0
Andalucia	0	0	0	0
Aragon	0	0	0	0
Principado De Asturias	0	0	0	0
Baleares (Islas)	0	0	0.146	0
Carias	0	0	0	0
Cantabria	0	0	0	0
Castilla Y Leon	0	0	0	0
Castilla-La Mancha	0	0	0	0
Catalu	0.8508	0.862	0.711	0.83
Comunidad Valencia	0	0	0	0
Extremadura	0	0	0	0
Galicia	0	0	0	0
Madrid (Comunidad De)	0.1492	0.138	0.143	0.17
Murcia (Region de)	0	0	0	0
varra (Comunidad Foral De)	0	0	0	0
Rioja (La)	0	0	0	0
Pre-1970 gap (avg.)	0.007760802	0.008235552	0.00827	0.007417
RMSPE	0.088095411	0.090749944	0.09095	0.086124
Post-1969 gap (avg.)	0.511506382	0.51035959	0.84	0.513725
Mean Squared Error Post-1969	0.715196744	0.714394562	0.91652	0.716746
Post/Pre RMSPE Ratio	8.118433557	7.872121206	10.0771	8.322223

Table 5.3 Predictor Values Compared to Original Model

	Basque	Synthetic Basque (Abadie & Gardeazabal, 2003)	difference	Synthetic Basque (covariates only)	difference	Synthetic Basque (Pre-intervention lvls and differences only)	difference	Synthetic Basque (Pre-1970 obs, differences, and covariates)	difference
Real per capita GDP[a]	5.285	5.271	0.014	5.261505392	0.023	5.27572241	0.009	5.272291284	0.013
Investment ratio (%)[b]	24.647	21.583	3.064	21.65293833	2.994	21.5712225	3.076	21.5712225	3.076
Population density[c]	246.89	196.288	50.602	193.04754	53.842	196.80883	50.081	196.80883	50.081
Sectoral Shares (%)[d]									
Agriculture, forestry, and fishing	6.844	6.179	0.665	6.236064	0.608	7.115028	-0.271	6.170128	0.674
Energy and water	4.106	2.76	1.346	2.768772	1.337	2.647358	1.459	2.758294	1.348
Industry	45.082	37.636	7.446	37.817364	7.265	34.553746	10.528	37.606478	7.476
Construction and engineering	6.15	6.952	-0.802	6.933976	-0.784	7.174084	-1.024	6.955452	-0.805
Marketable services	33.754	41.104	-7.350	40.951148	-7.197	42.871242	-9.117	41.128546	-7.375
Nonmarketable services	4.072	5.371	-1.299	5.295228	-1.223	5.641406	-1.569	5.383706	-1.312
Human capital (5)[e]									
Illiterates	3.3210%	7.6450%	-0.043	7.769%	-0.044	7.811%	-0.045	7.745%	-0.044
Primary or without studies	85.8930%	82.2850%	0.036	82.581%	0.033	82.413%	0.035	82.532%	0.034
High school	7.5220%	6.9650%	0.006	6.665%	0.009	6.750%	0.008	6.715%	0.008
More than high school	3.2640%	3.1050%	0.002	2.985%	0.003	3.026%	0.002	3.008%	0.003

Sources: Data reported by Abadie and Gardeazabal(2003) gathered from Matilde Mas et al.(1998) and Fundacion BBV (1999).

[a] 1986 USD, average for 1960-1969.

[b] Gross Total Investment/GDP, average for 1964-69.

[c] Persons per square kilometer, 1969.

[d] Percentages over total production, 1961-1969.

[e] Percentages over working-age population, 1964-1969.

Table 5.4 California Predictors

	California	Mean of 38 Donor Pool States	Synthetic California (Abadie et al., 2010)	Synthetic California (optimal alt. model)
Ln(GDP per capita)[a]	10.08	9.86	9.86	9.84
Percent aged 15-24[b]	17.4%	17.29%	17.4%	17.57%
Retail Price (per pack)[c]	89.42	87.27	89.41	89.33
Beer consumption per capita[d]	24.28	23.75	24.2	22.77
Cigarette sales per capita (1988)[e]	90.1	114.2	91.62	90.94
Cigarette sales per capita (1980)[e]	120.2	136.58	120.43	119.90
Cigarette sales per capita (1975)[e]	127.1	132.81	126.99	126.40

Sources:

[a]Bureau of the Census, United States Statistical Abstract. Converted to 1997 dollars using the Consumer Price Index

[b]U.S. Census Bureau

[c]Orzechowski and Walker (2005). Price figures include state sales taxes where applicable.

[d]Beer Institute's Brewer's Almanac. Measured as the per capita consumption of malt beverages (in gallons).
[e]Orzechowski and Walker (2005). These data are based on the total tax paid on sales of packs of cigarettes in a particular state divided by its total population

Table 5.5 Synthetic California Composition, Fit Statistics, and Intervention Estimates

	Synthetic California (Abadie et al., 2010)	Synth Abadie Stata	Covariates only	Synth_Pre-'89 DV obs and differences	Synth_DV lvls difs and covariates	synth bestmodel obs dif and covs plus retail price
Alabama	0	0	0	0	0	0
Arkansas	0	0	0	0	0	0
Colorado	0.164	0.261	0	0	0.18	0.144
Connecticut	0.069	0.007	0.517	0.055	0.046	0.069
Delaware	0	0	0	0	0	0
Georgia	0	0	0	0	0	0
Idaho	0	0	0	0	0.046	0
Illinois	0	0	0	0	0	0
Indiana	0	0	0	0	0	0
Iowa	0	0	0	0	0	0
Kansas	0	0	0	0	0	0
Kentucky	0	0	0	0	0	0
Louisiana	0	0	0	0	0	0
Maine	0	0	0	0	0	0
Minnesota	0	0	0	0	0	0
Mississippi	0	0	0	0	0	0
Missouri	0	0	0	0	0	0
Montana	0.199	0.174	0	0.232	0.161	0.106
Nebraska	0	0	0	0	0	0
Nevada	0.234	0.218	0	0.134	0.161	0.192
New Hampshire	0	0	0	0.103	0.011	0.054
New Mexico	0	0	0	0	0.215	0
North Carolina	0	0	0	0	0	0
North Dakota	0	0	0	0	0	0
Ohio	0	0	0	0	0	0
Oklahoma	0	0	0	0	0	0
Pennsylvania	0	0	0	0	0	0
Rhode Island	0	0	0	0	0	0
South Carolina	0	0	0	0	0	0
South Dakota	0	0	0	0	0	0
Tennessee	0	0	0	0	0	0
Texas	0	0	0	0	0	0
Utah	0.334	0.34	0	0.477	0.178	0.434
Vermont	0	0	0	0	0	0
Virginia	0	0	0.483	0	0	0
West Virginia	0	0	0	0	0	0
Wisconsin	0	0	0	0	0	0
Wyoming	0	0	0	0	0	0
Pre-1989 gap (avg.)	3.089199757	3.800976329	201.8451784	3.411268574	5.196709606	2.963182685
RMSPE	1.757611947	1.949609276	14.20722275	1.846961985	2.279629269	1.721389754
Post-1988 gap (avg.)	396.6014747	386.8446433	1147.975241	446.8650469	397.427928	383.3641708
Mean Squared Error Post-1988	19.91485563	19.66836657	33.88178332	21.13918274	19.9355945	19.57968771
RMSPE Ratio	11.33063283	10.08836325	2.384828049	11.44538053	8.745103762	11.37434893

Table 5.6 California Predictors

	California	Synthetic California (Abadie et al., 2010)	difference	Synthetic California (optimal model)	difference	Synthetic California (Covariates only)	difference	Synthetic California (Pre-1989 cigsales and differences only)	difference
Ln(GDP per capita)[a]	10.08	9.86	0.22	9.840083743	0.240	10.076654	0.003	9.8385669	0.2414331
Percent aged 15-24[b]	17.4	17.4	0	17.57730992	-0.177	17.029768	0.370	17.785614	-0.385614
Retail Price (per pack)[c]	89.42	89.41	0.01	89.33058889	0.089	89.621044	-0.201	88.172278	1.247722
Beer consumption per capita[d]	24.28	24.2	0.08	22.7765	1.504	21.8109	2.469	21.99166	2.28834
Cigarette sales per capita (1988)[e]	90.1	91.62	-1.52	90.9426	-0.843	116.7301	-26.630	89.465698	0.634302
Cigarette sales per capita (1980)[e]	120.2	120.43	-0.23	119.9008	0.299	132.9247	-12.725	119.7663	0.4337
Cigarette sales per capita (1975)[e]	127.1	126.99	0.11	126.407	0.693	130.7275	-3.628	125.773	1.327

Sources:

[a]Bureau of the Census, United States Statistical Abstract. Converted to 1997 dollars using the Consumer Price Index

[b]U.S. Census Bureau

[c]Orzechowski and Walker (2005). Price figures include state sales taxes where applicable.

[d]Beer Institute's Brewer's Almanac. Measured as the per capita consumption of malt beverages (gallons).

[e]Orzechowski and Walker (2005). These data are based on the total tax paid on sales of packs of cigarettes in a particular state divided by its total population

Table 5.7 Covariate Balance between West Germany, OECD sample mean, and Alternate compositions of Synthetic West Germany

	West Germany	Abadie et al. (2015) Synthetic West Germany	OECD Sample (n=16) Mean	Optimal Synthetic West Germany
GDP per capita[a]	15808.9	15802.2	8021.1	15292.4
Trade openness[a]	56.8	56.9	31.9	49.5
Inflation rate[a]	2.6	3.5	7.4	6.1
Industry share[a]	34.5	34.4	34.2	33.0
Schooling[b]	55.5	55.2	44.1	52.8
Investment rate[b]	27	27	25.9	25.5

Notes: The last column reports a population-weighted average for the 16 OECD countries in the donor pool.

[a]GDP per capita, inflation rate, trade openness, and industry share are averaged for the 1981–90 period.

[b] Investment rate and schooling are averaged for the 1980–85 period.

Table 5.8 Synthetic West Germany Composition, Fit Statistics, and Intervention Estimates

	Synth Abadie reported	Abadie Synth routine	Covs only	levels and difs only	Covs & half-set lvls	Synth lvls difs and covs
USA	0.22	0.14	0.13	0.342	0.237	0.328
UK	0	0.038	0	0	0.042	0
Austria	0.42	0.431	0.458	0.336	0.379	0.295
Belgium	0	0	0	0	0	0.023
Denmark	0	0	0	0.006	0	0.019
France	0	0	0	0	0	0.011
Italy	0	0	0	0.08	0	0.084
Netherlands	0.09	0.093	0.092	0	0.182	0
Norway	0	0	0	0.025	0.071	0.039
Switzerland	0.11	0.177	0.177	0.11	0.089	0.111
Japan	0.16	0.121	0.142	0	0	0
Greece	0	0	0	0.101	0	0.09
Portugal	0	0	0	0	0	0
Spain	0	0	0	0	0	0
Australia	0	0	0	0	0	0
New Zealand	0	0	0	0	0	0
Pre-1989 gap (avg.)	14955.3911	9966.89067	12540.6	3815.8527	4652.73	3710.382
Root Mean Squared Prediction Error (RMSPE)	122.292236	99.8343160	111.985	61.772588	68.2109	60.91291
Post-1989 gap (avg.)	3824317.76	2908928.14	278417	3529149.9	6124048	3336581.
Mean Squared Error Post-1988	1955.5863	1705.55801	1668.58	1878.6031	2474.68	1826.631
RMSPE Ratio	15.9910911	17.0838854	14.9000	30.411599	36.2798	29.98758

Chapter 5 Figures

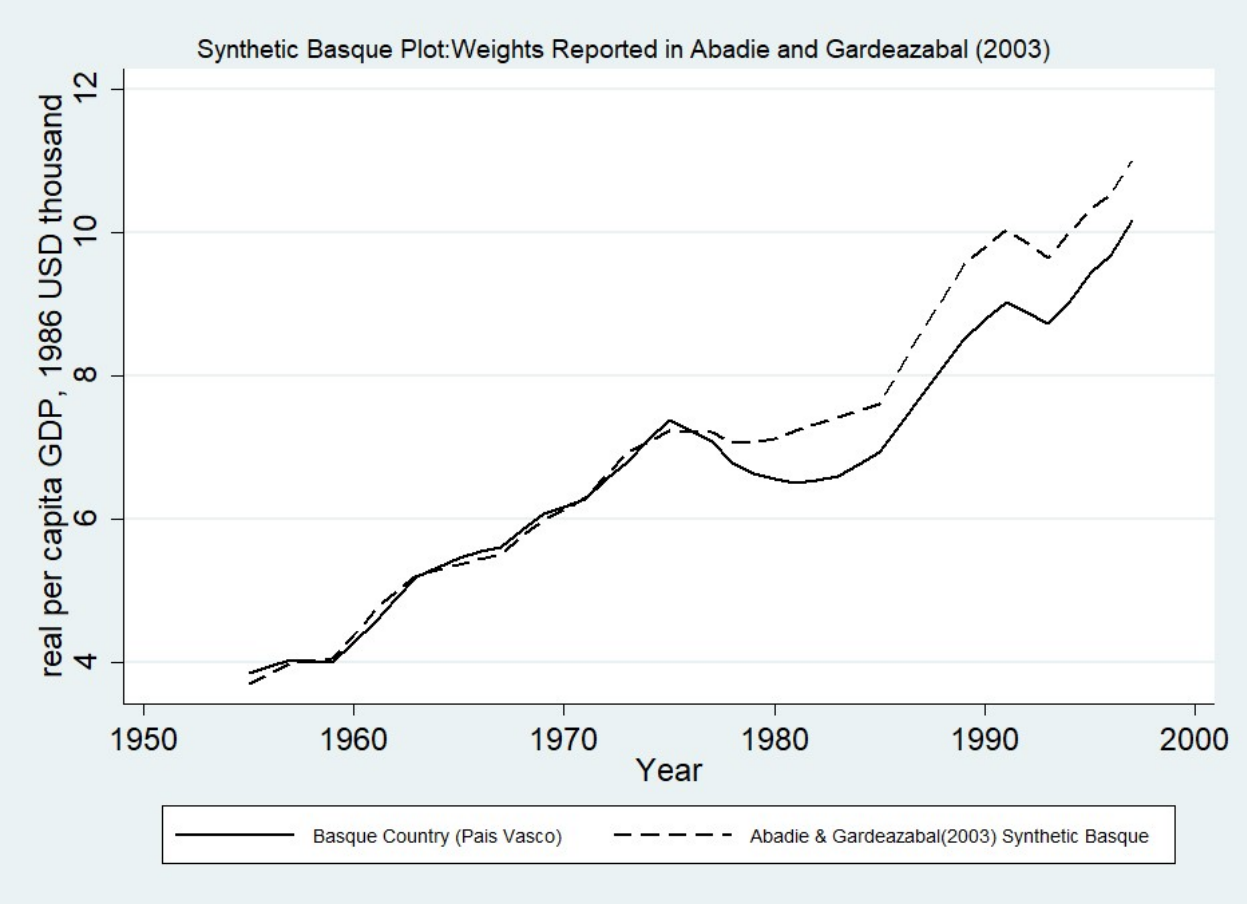

Figure 5.1 Synthetic Control Group Results Plot: Weights reported in Abadie & Gardeazabal (2003)

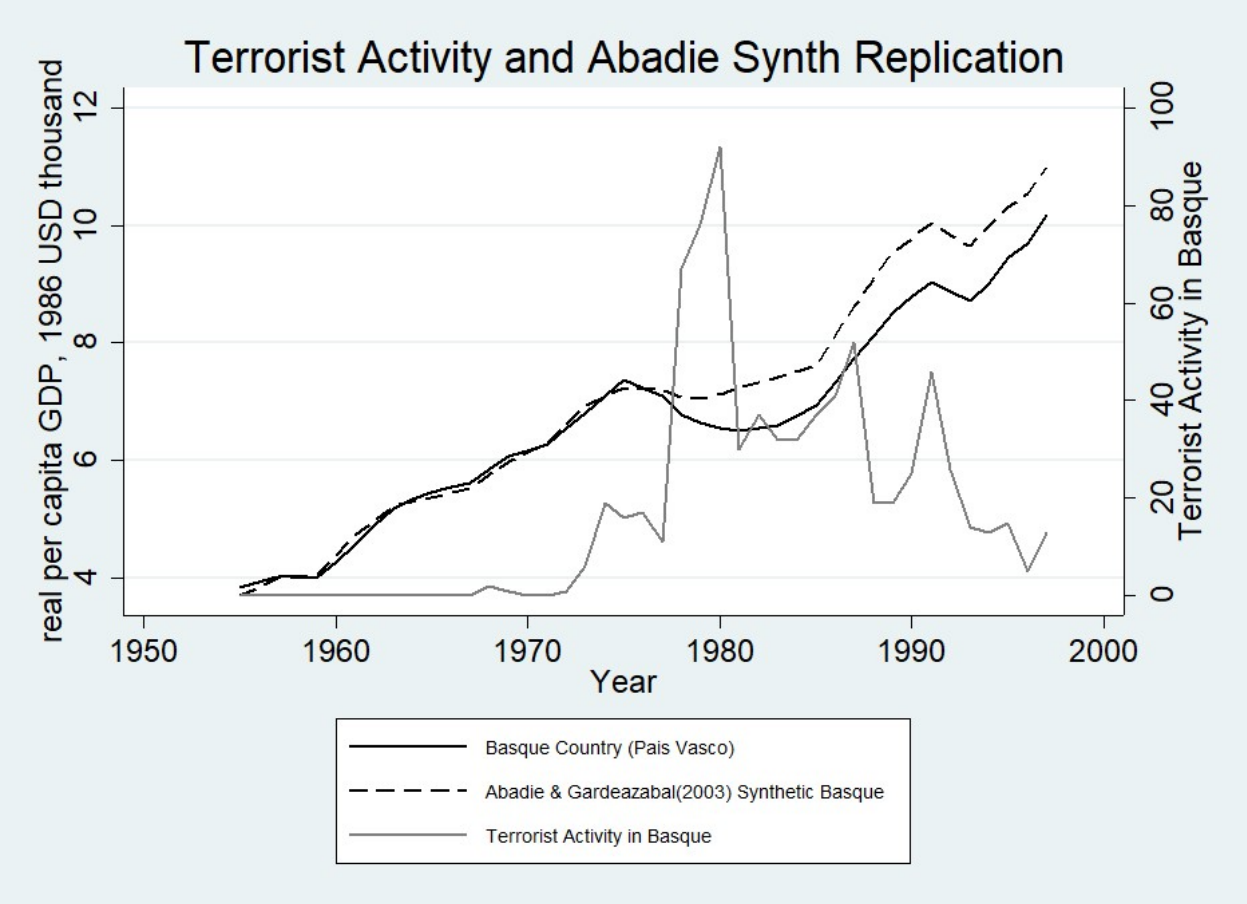

Figure 5.2 Abadie & Gardeazabal (2003) Replication model Compared to Terrorist Activity Series

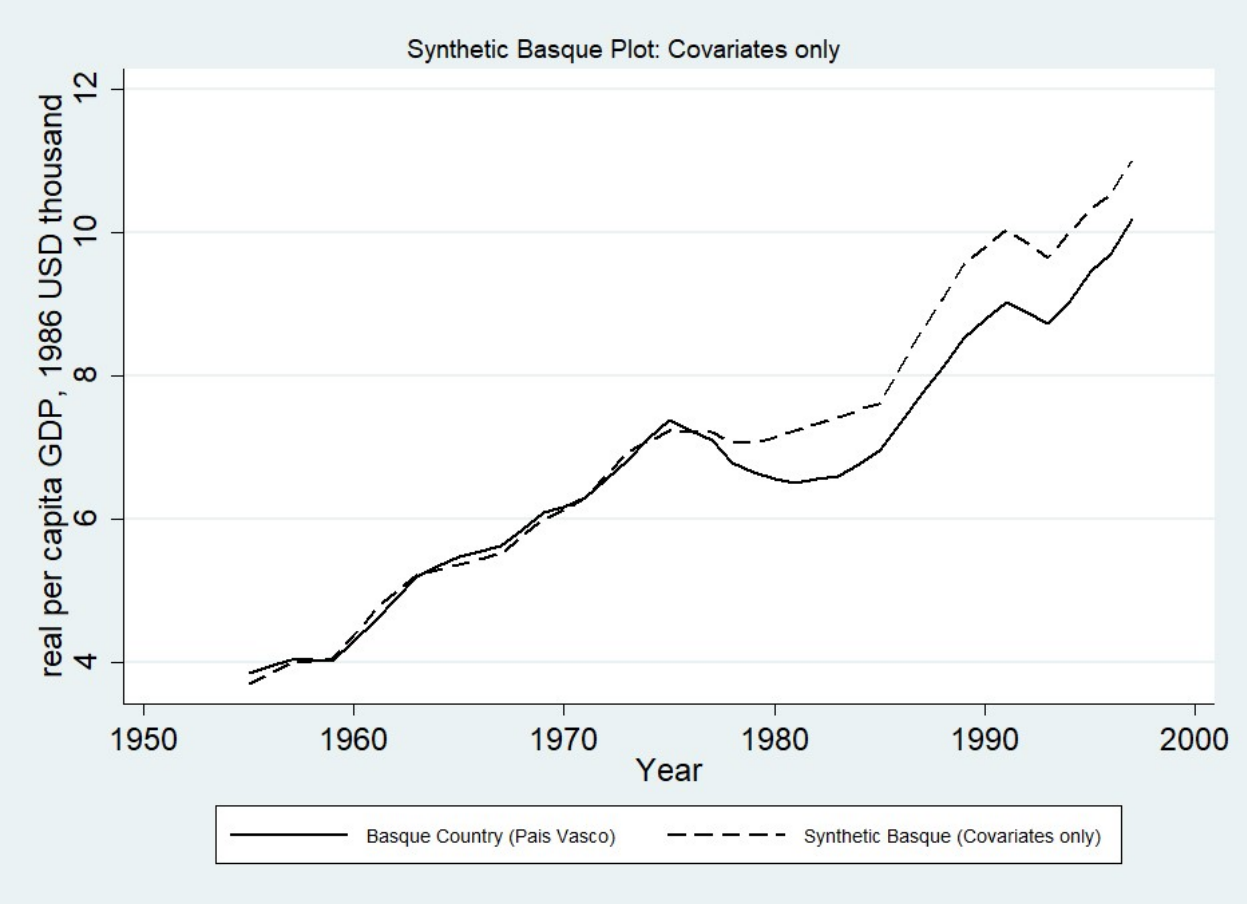

Figure 5.3 Synthetic Control Group Results Plot: Gardeazabal (2003) covariates only, no pre-intervention observations of GDP per capita

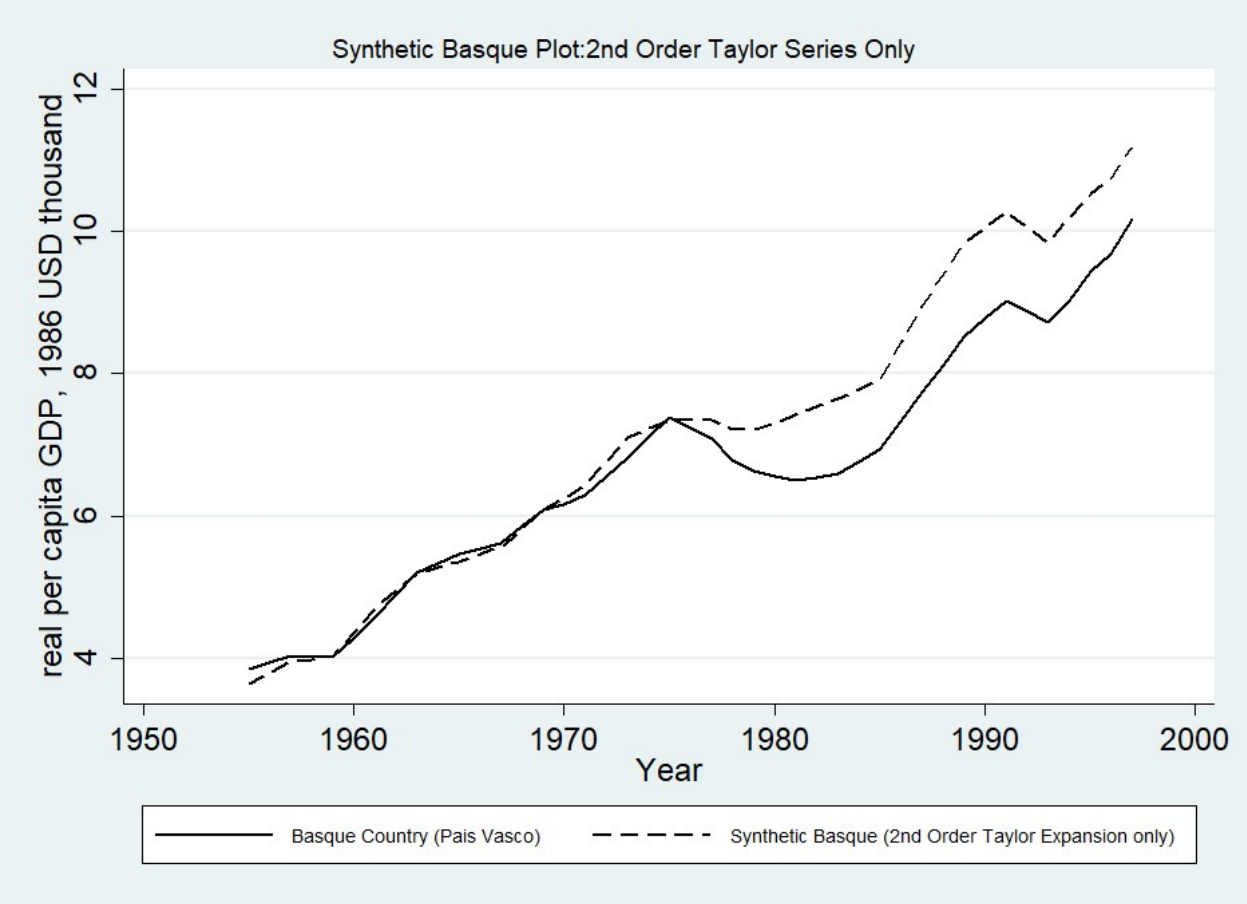

Figure 5.4 Synthetic Control Group Results Plot: 2nd Order Taylor Expansion Only (no covariates)

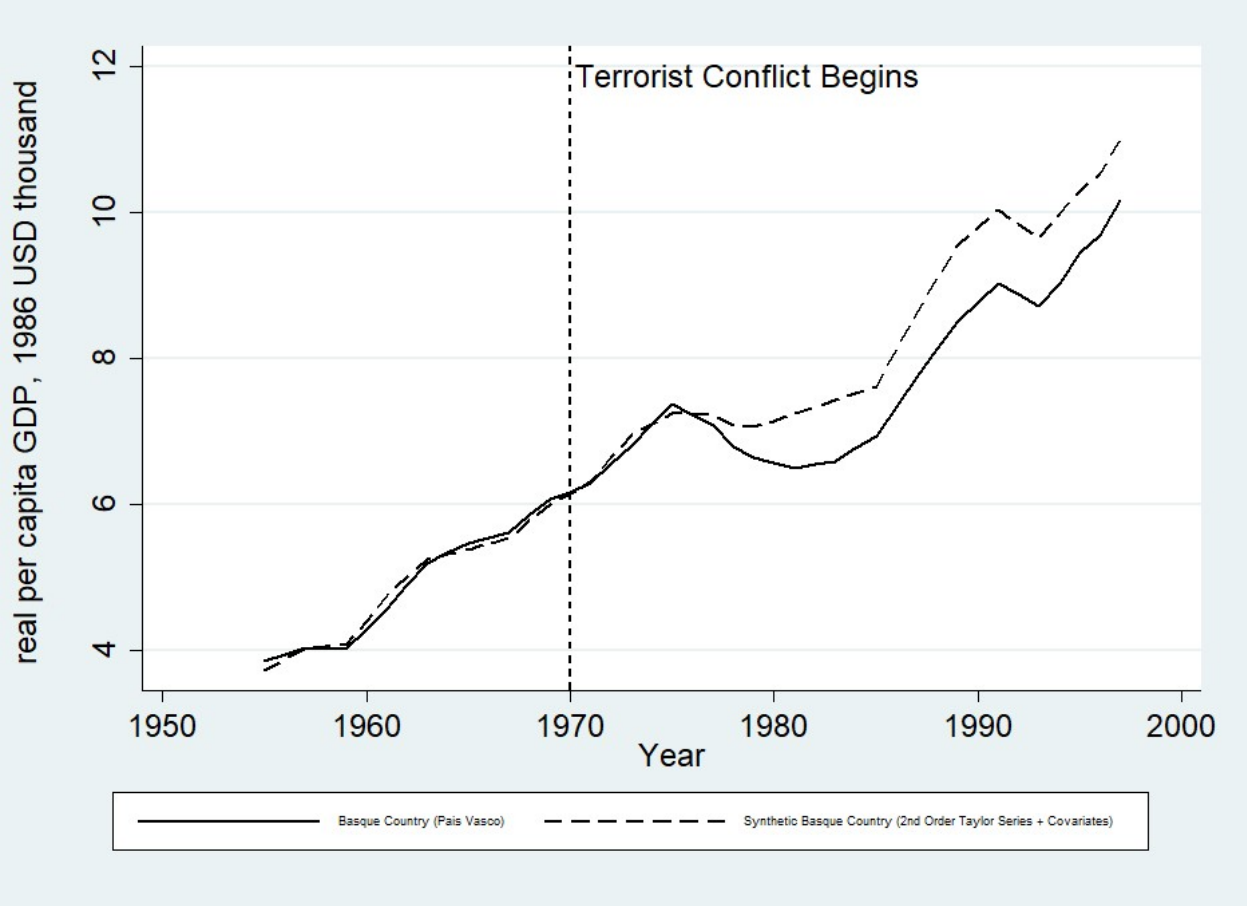

Figure 5.5 Synthetic Control Group Results Plot: 2nd Order Taylor Expansion plus Abadie & Gardeazabal (2003) covariates

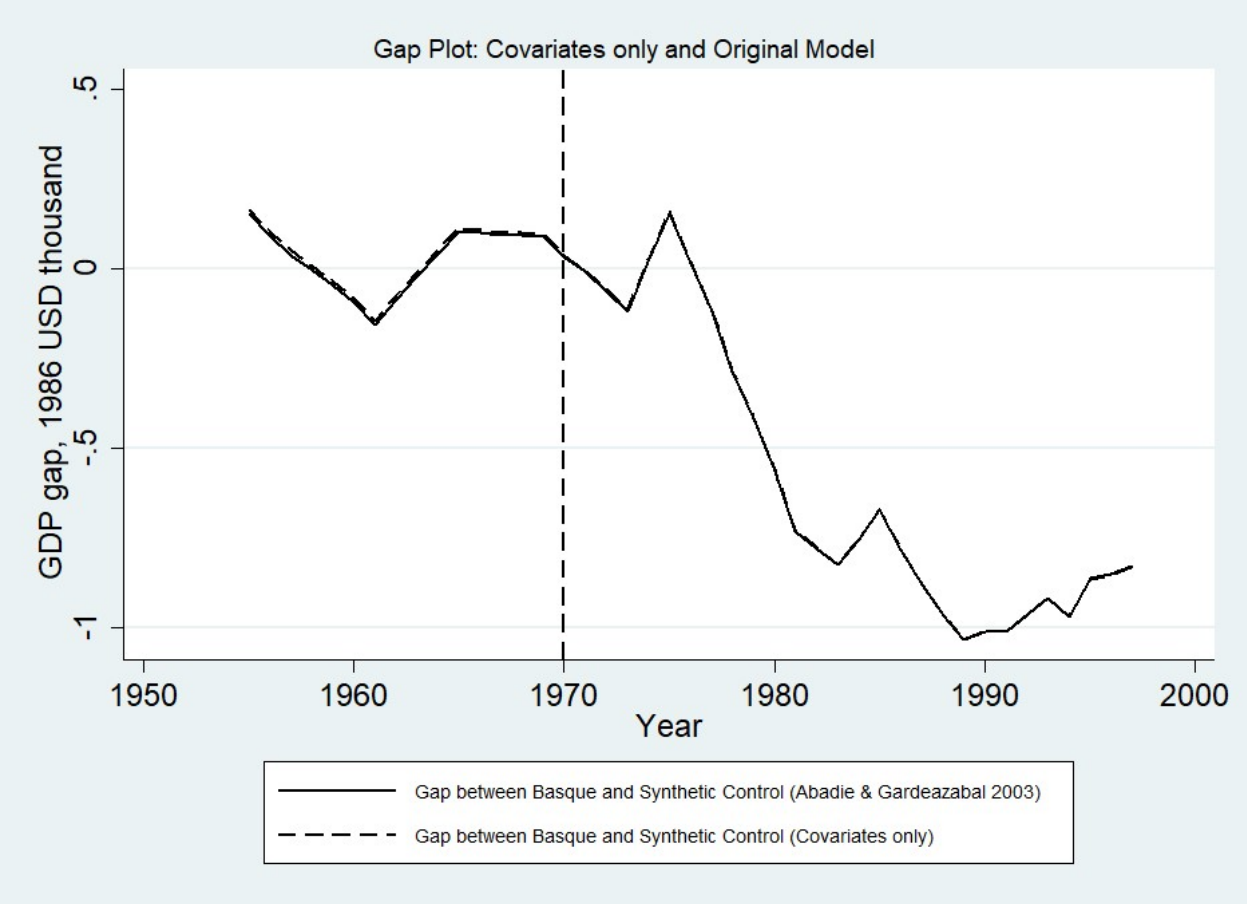

Figure 5.6 Synthetic Control Group Gap Plot: Abadie and Gardeazabal (2003) vs Covariates only

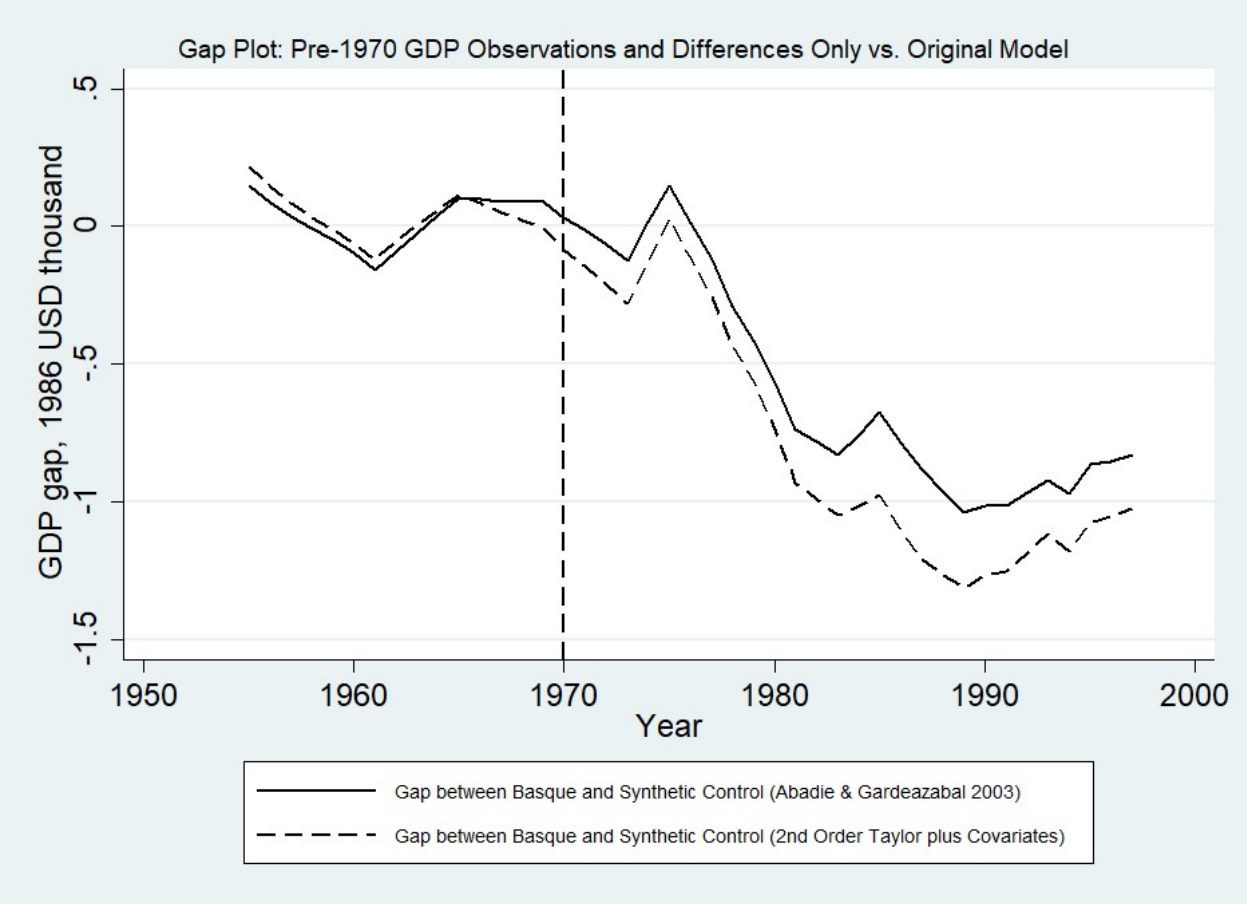

Figure 5.7 Synthetic Control Group Gap Plot: Pre-intervention DV observations, 1st, and 2nd Order Differences (No covariates) vs Abadie and Gardeazabal (2003)

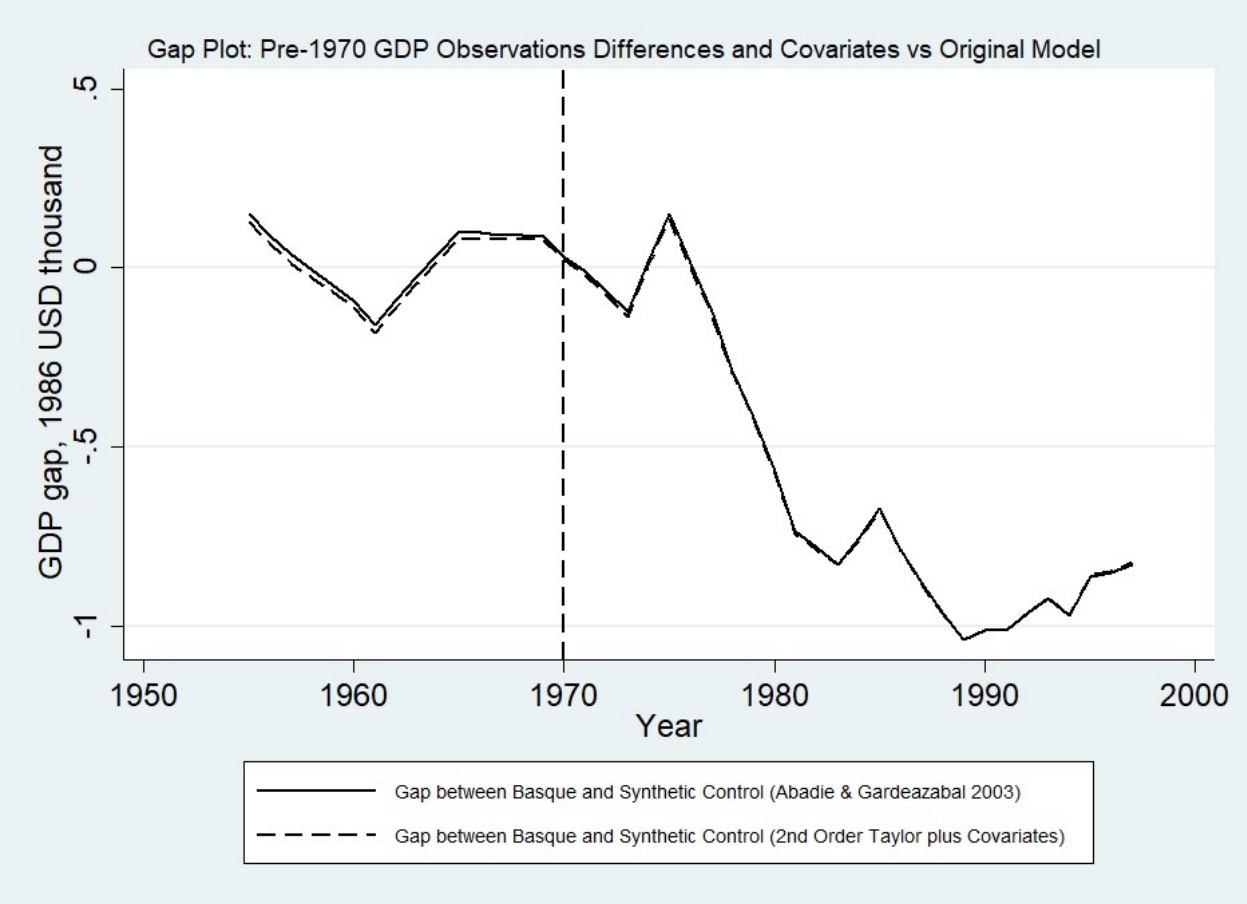

Figure 5.8 Synthetic Control Group Gap Plot: Pre-intervention DV observations, 1st, and 2nd Order Differences and original covariates vs Abadie and Gardeazabal (2003)

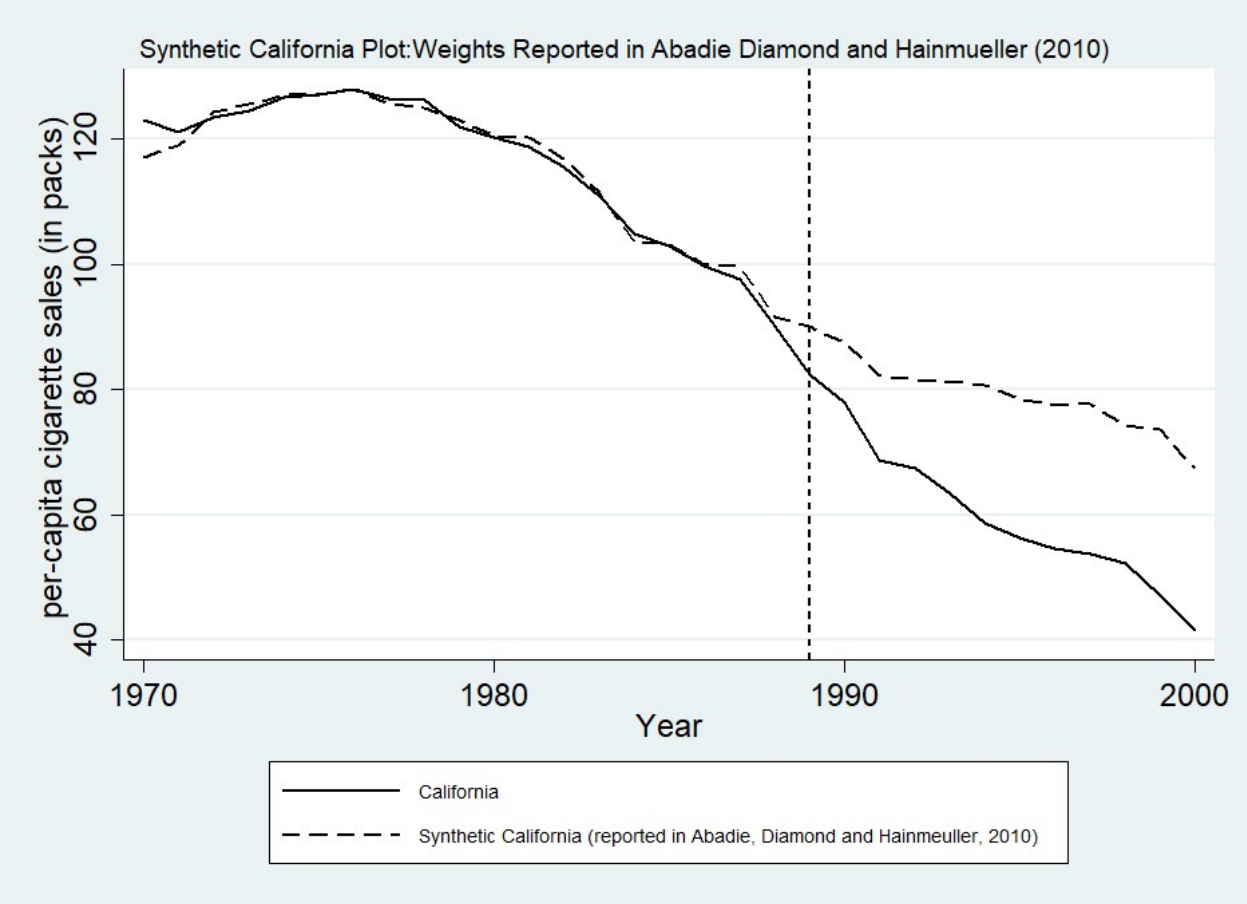

Figure 5.9 Synthetic California Plot: Weights Reported in Abadie Diamond and Hainmueller (2010)

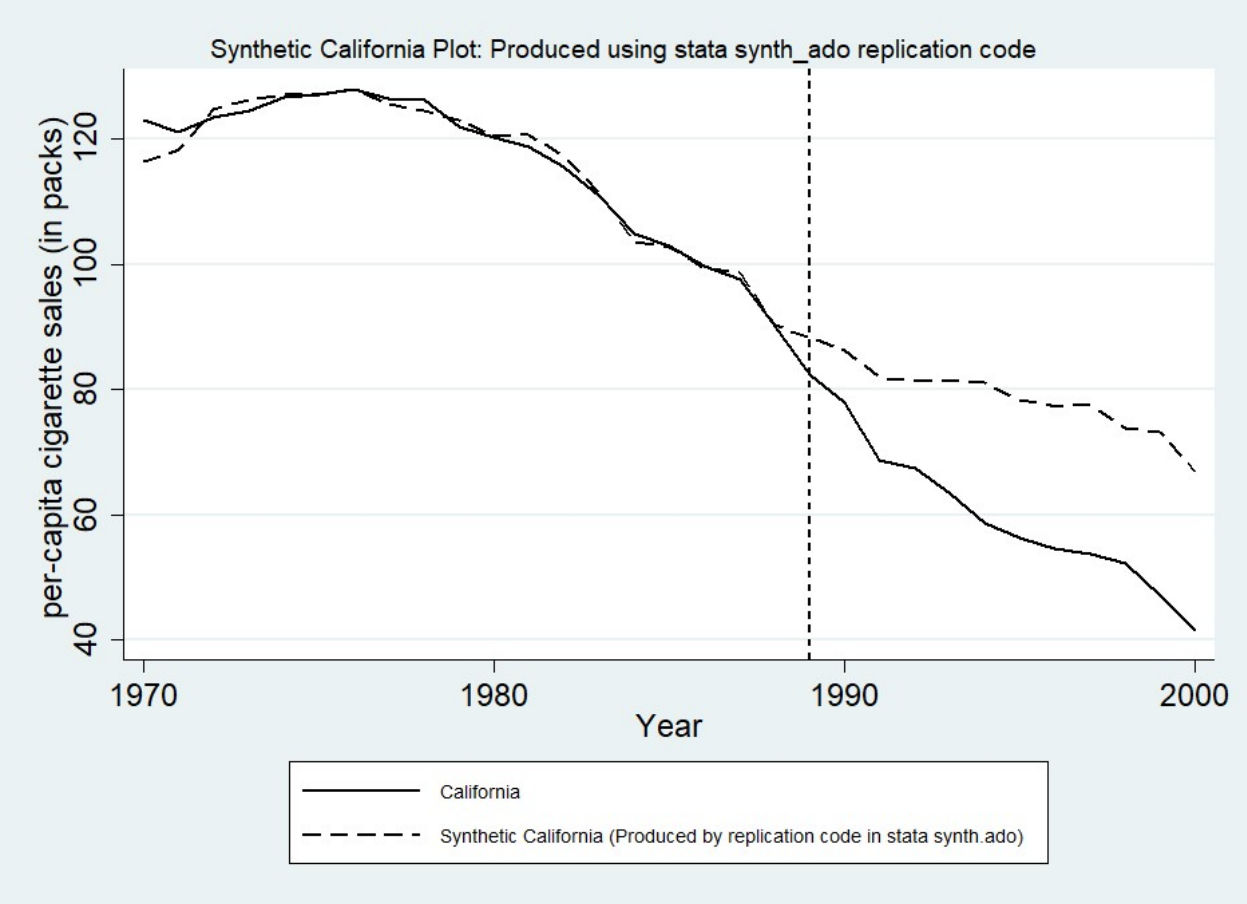

Figure 5.10 Synthetic California Plot: Produced using Stata synth_ado replication code

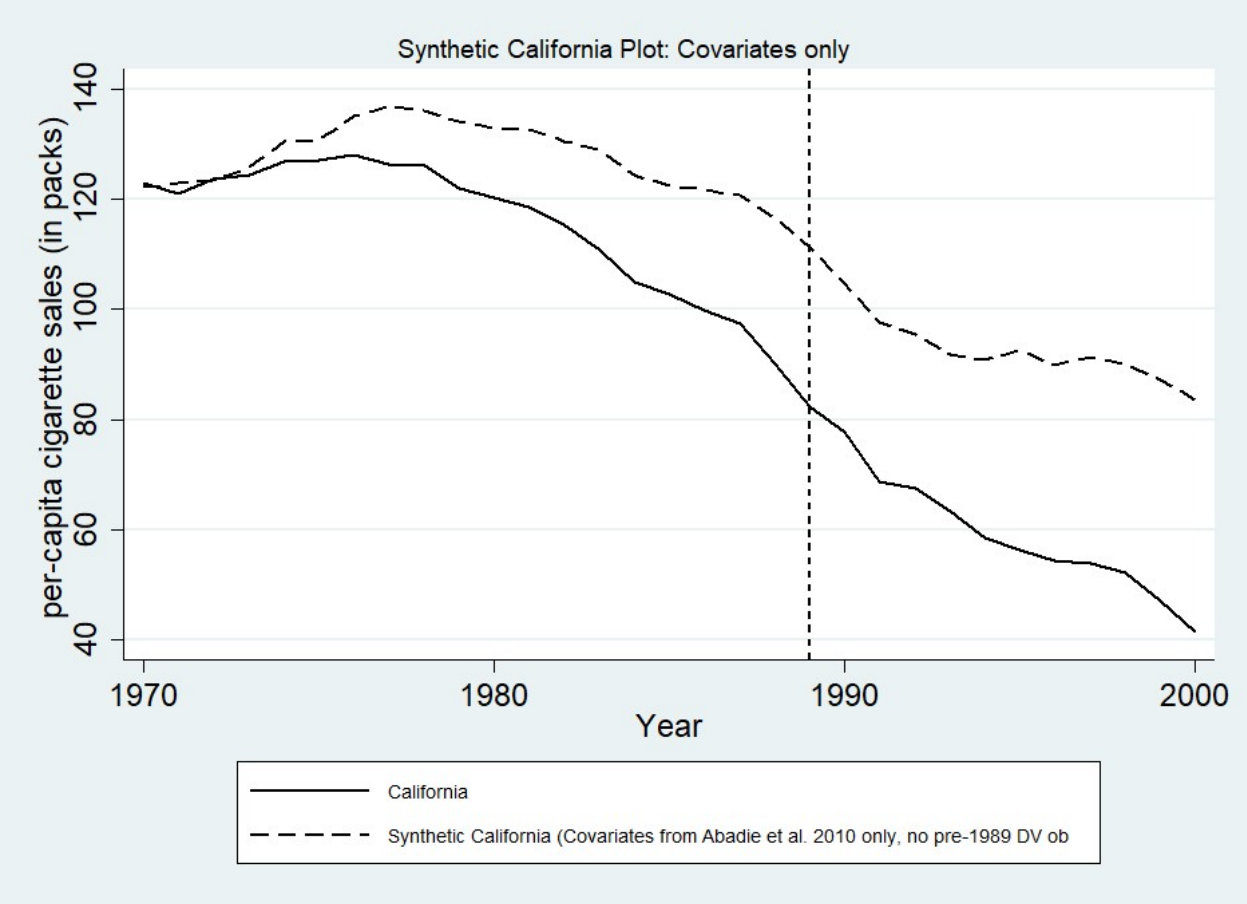

Figure 5.11 Synthetic California Plot: Covariates only

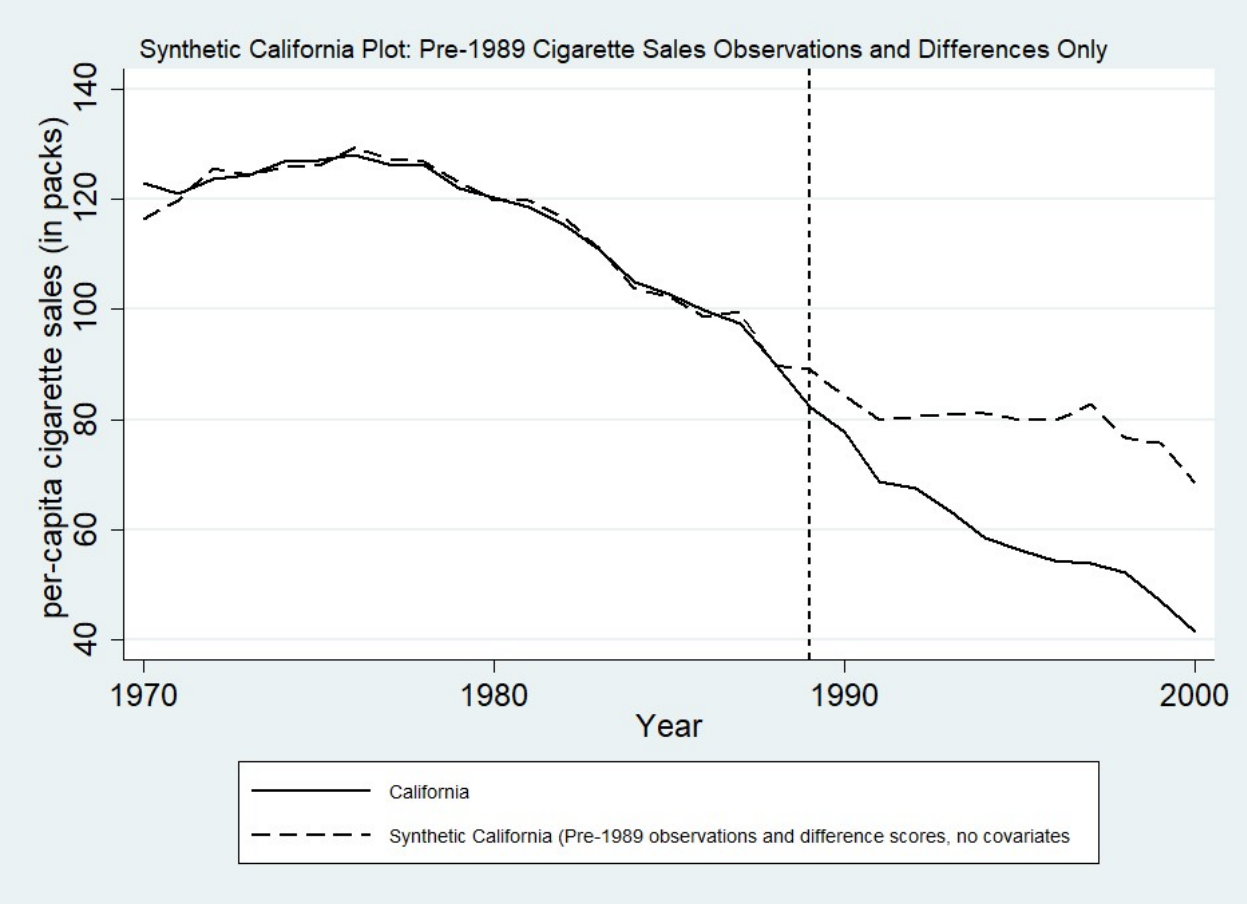

Figure 5.12 Synthetic California Plot: Pre-1989 Cigarette Sales Observations and Differences Only

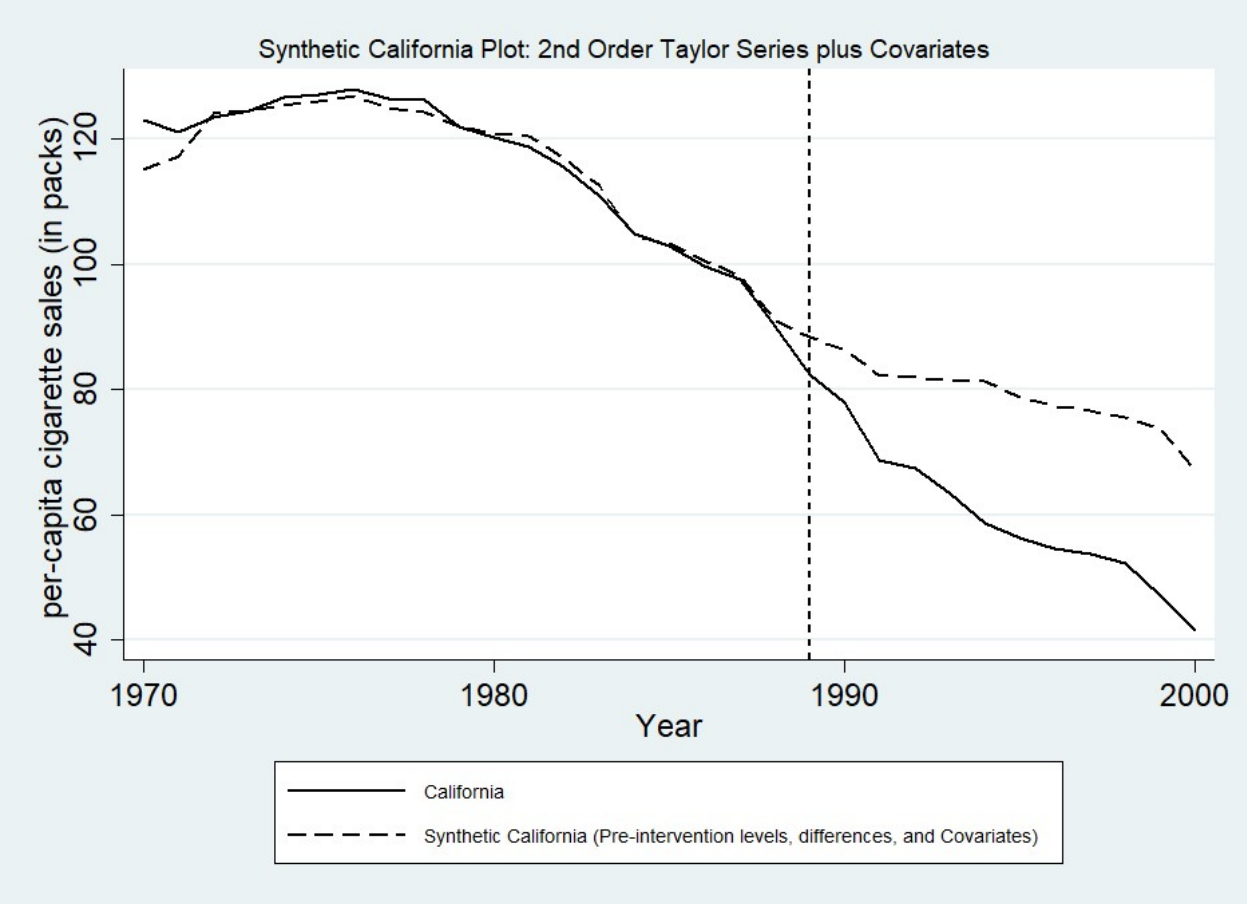

Figure 5.13 Synthetic California Plot: 2nd Order Taylor Series plus Covariates

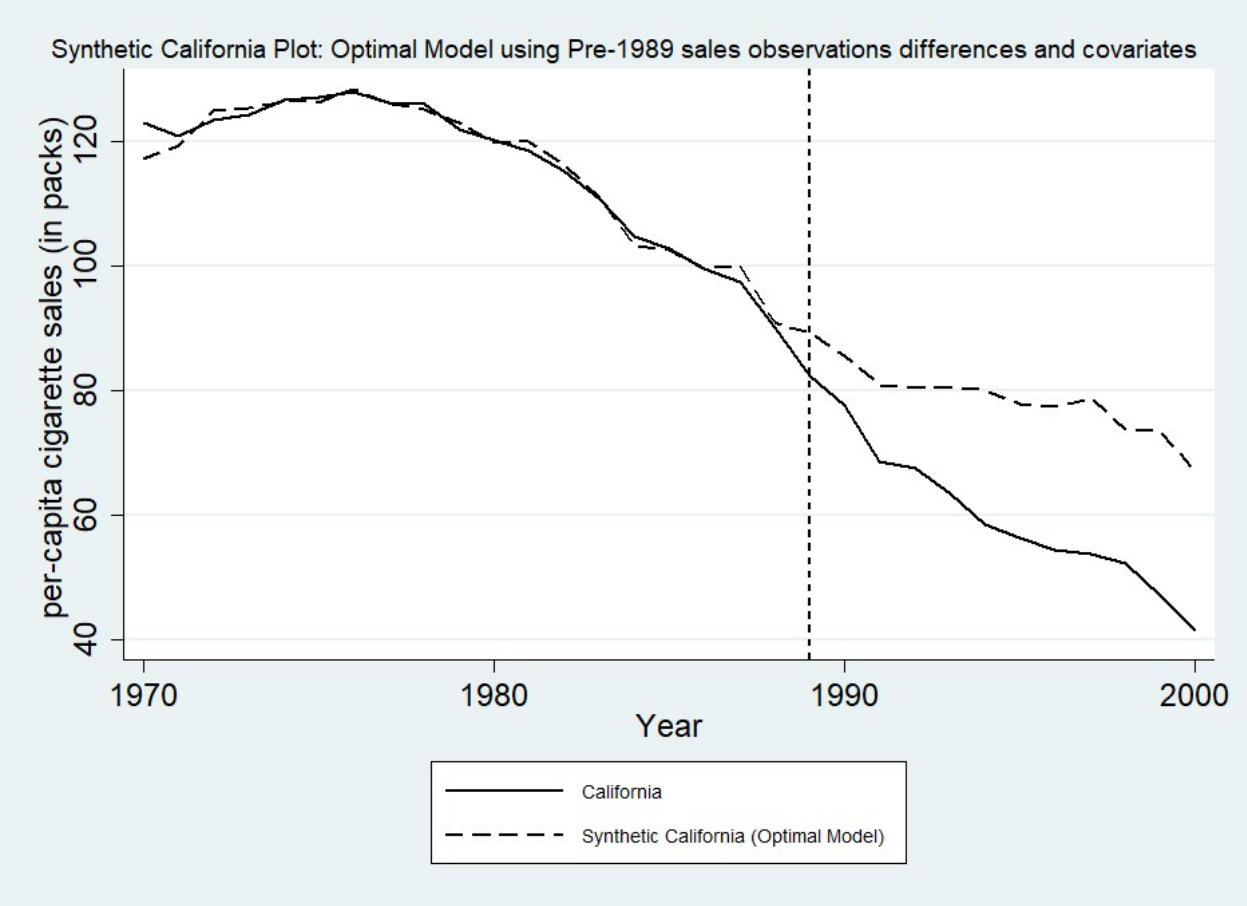

Figure 5.14 Synthetic California Plot: Optimal Model using Pre-1989 sales observations differences and covariates

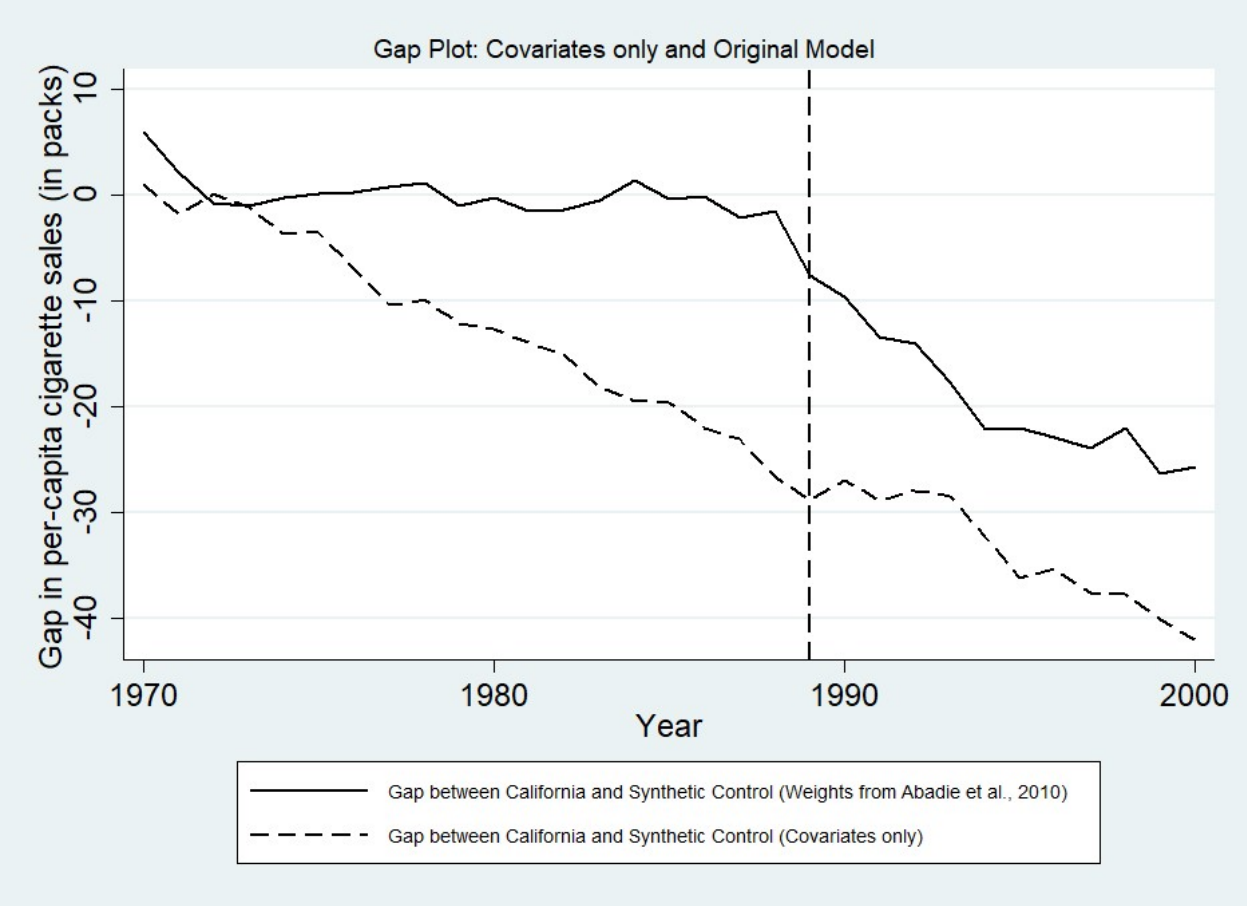

Figure 5.15 Synthetic California Gap Plot: Abadie et al. (2010) vs. Covariates only model

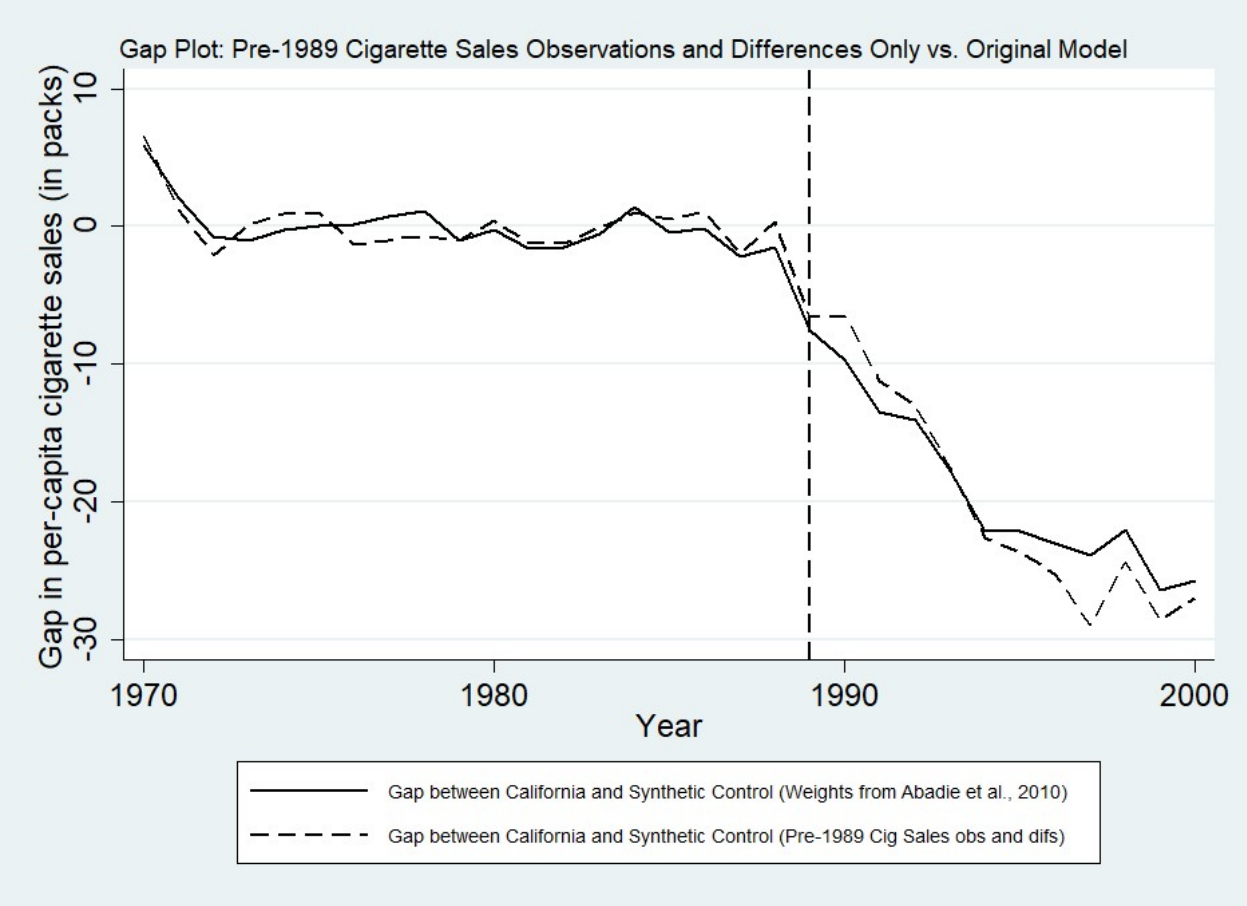

Figure 5.16 Synthetic California Gap Plot: Abadie et al. (2010) vs. Pre-1989 Cigarette Sales Observations and Differences Only

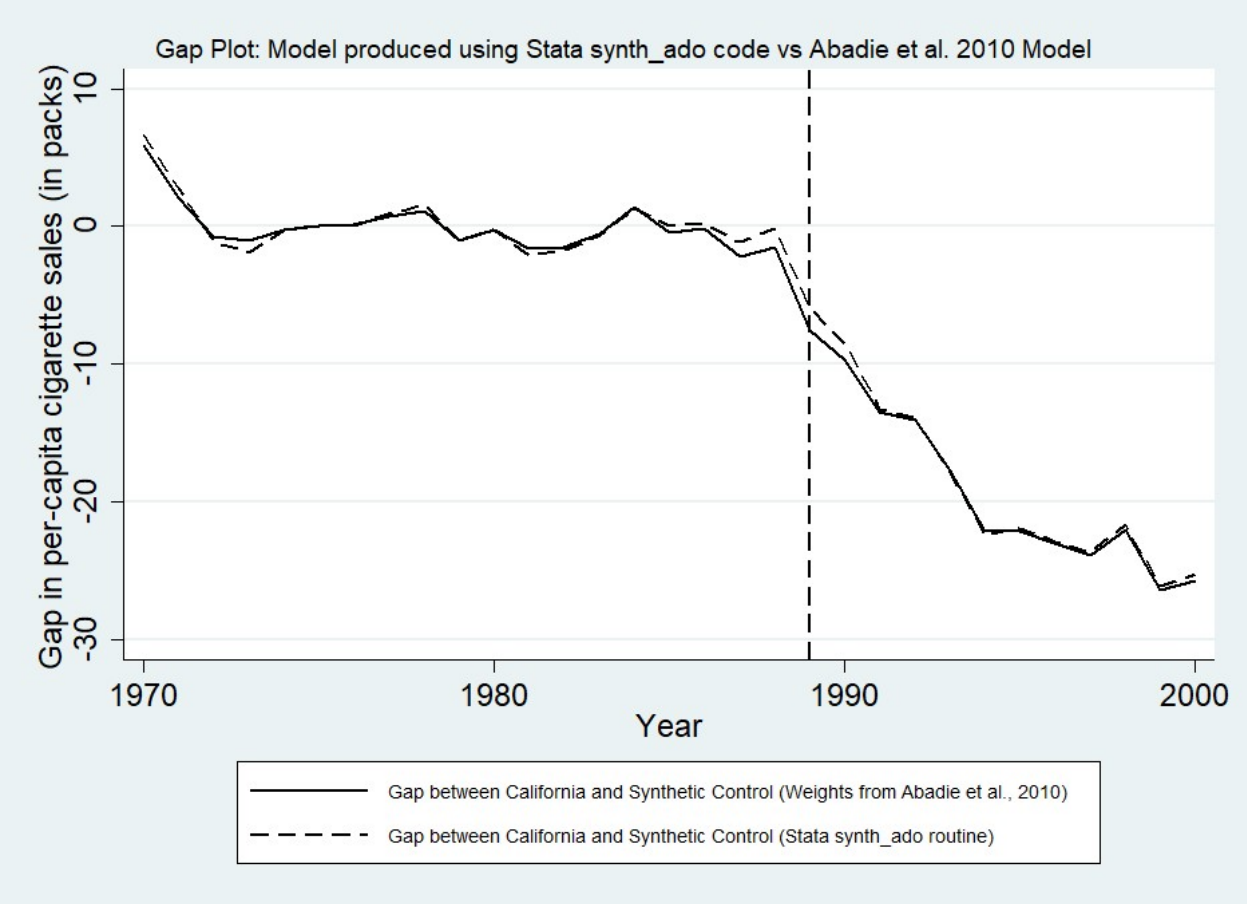

Figure 5.17 Synthetic California Gap Plot: Abadie et al. (2010) vs. Synthetic California produced using Stata synth_ado code

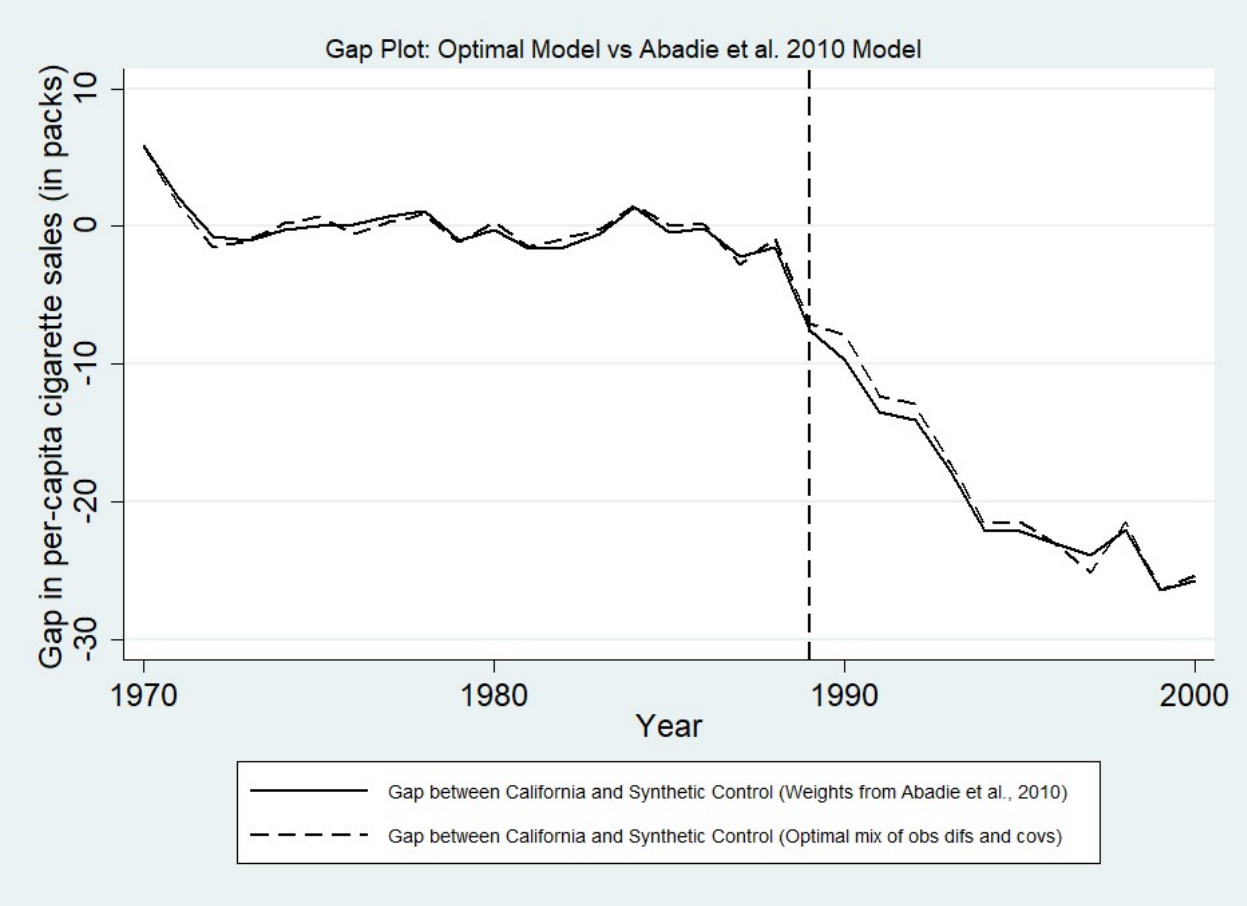

Figure 5.18 Synthetic California Gap Plot: Abadie et al. (2010) vs. Optimal Synthetic California

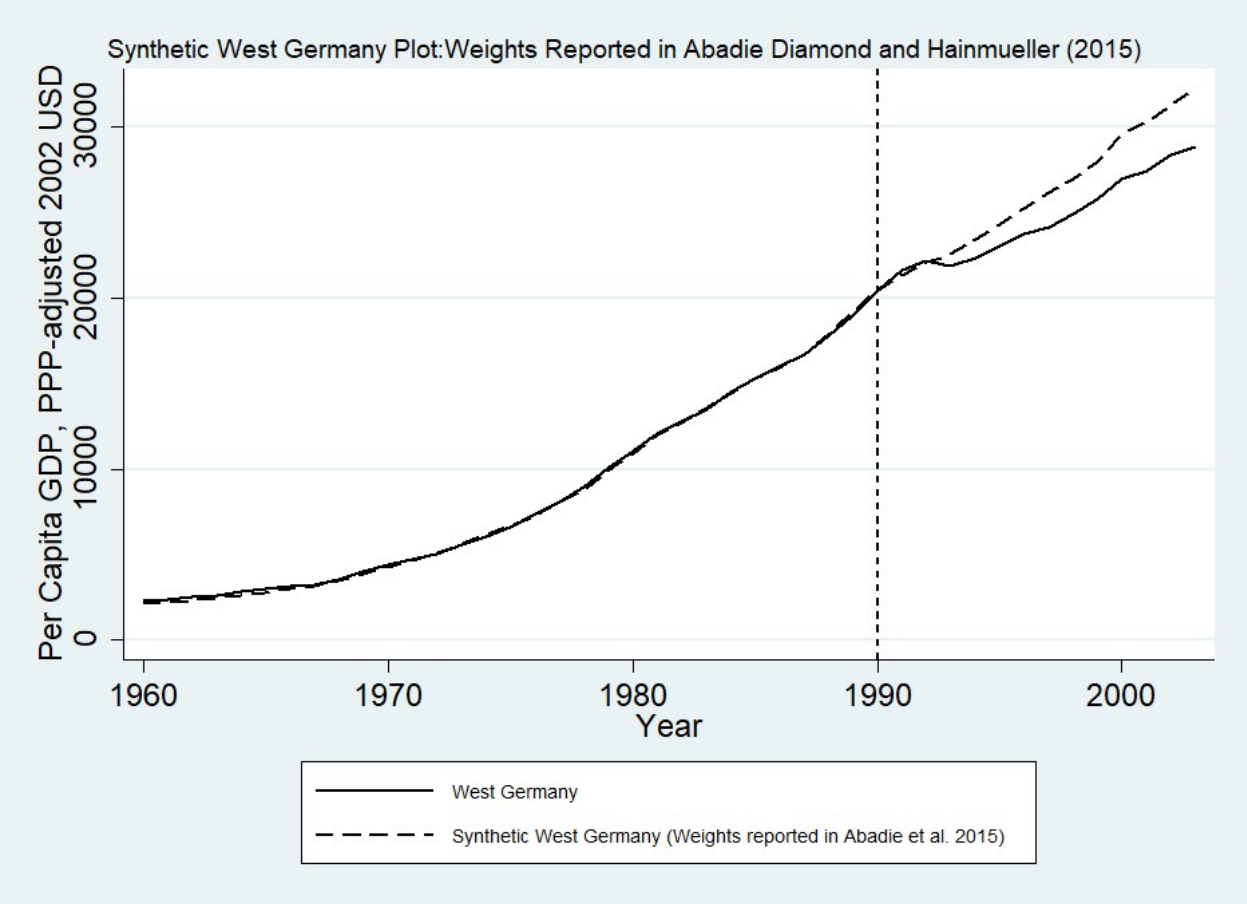

Figure 5.19 Synthetic West Germany Plot: Weights Reported in Abadie Diamond and Hainmueller (2015)

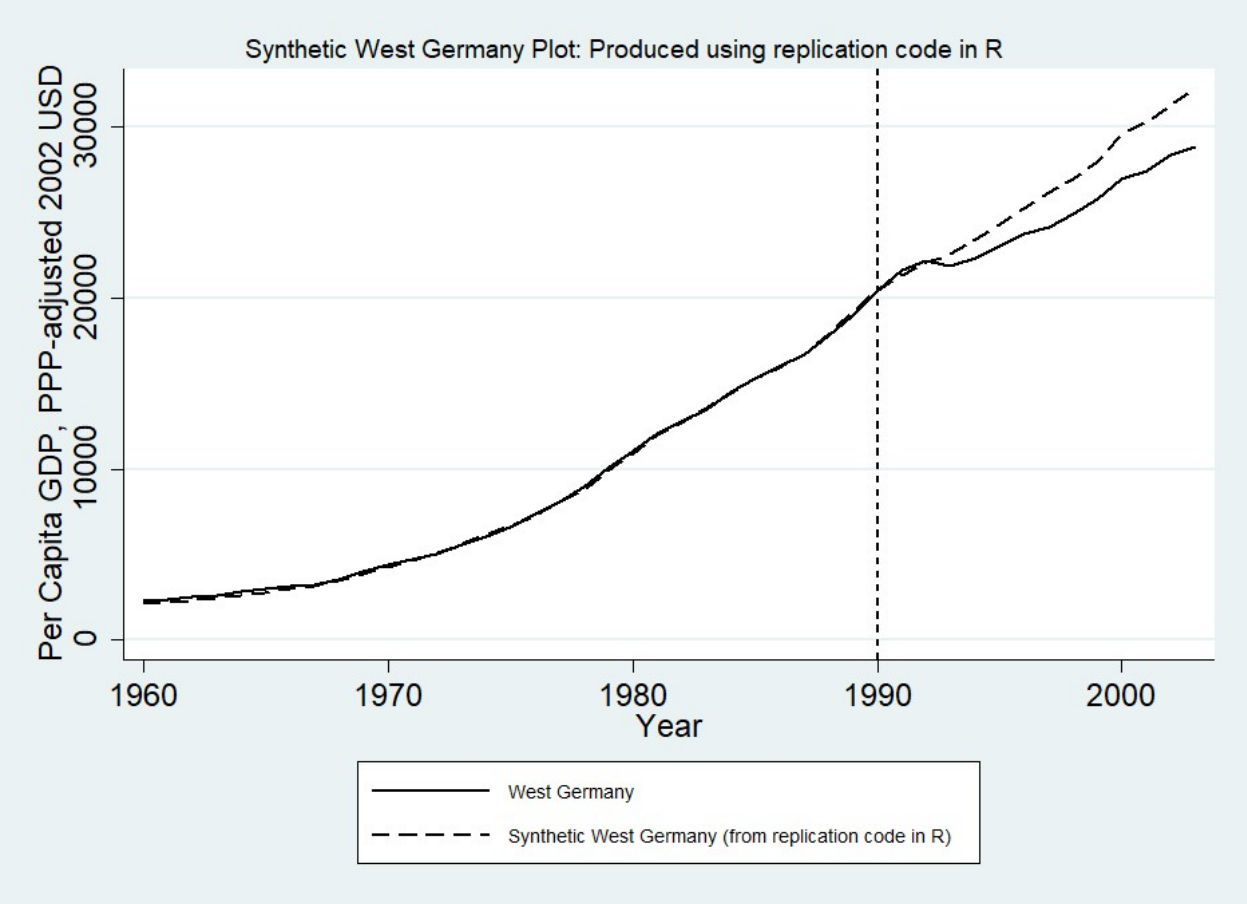

Figure 5.20 Synthetic West Germany Plot: Produced using replication code in R from Harvard Dataverse Repository

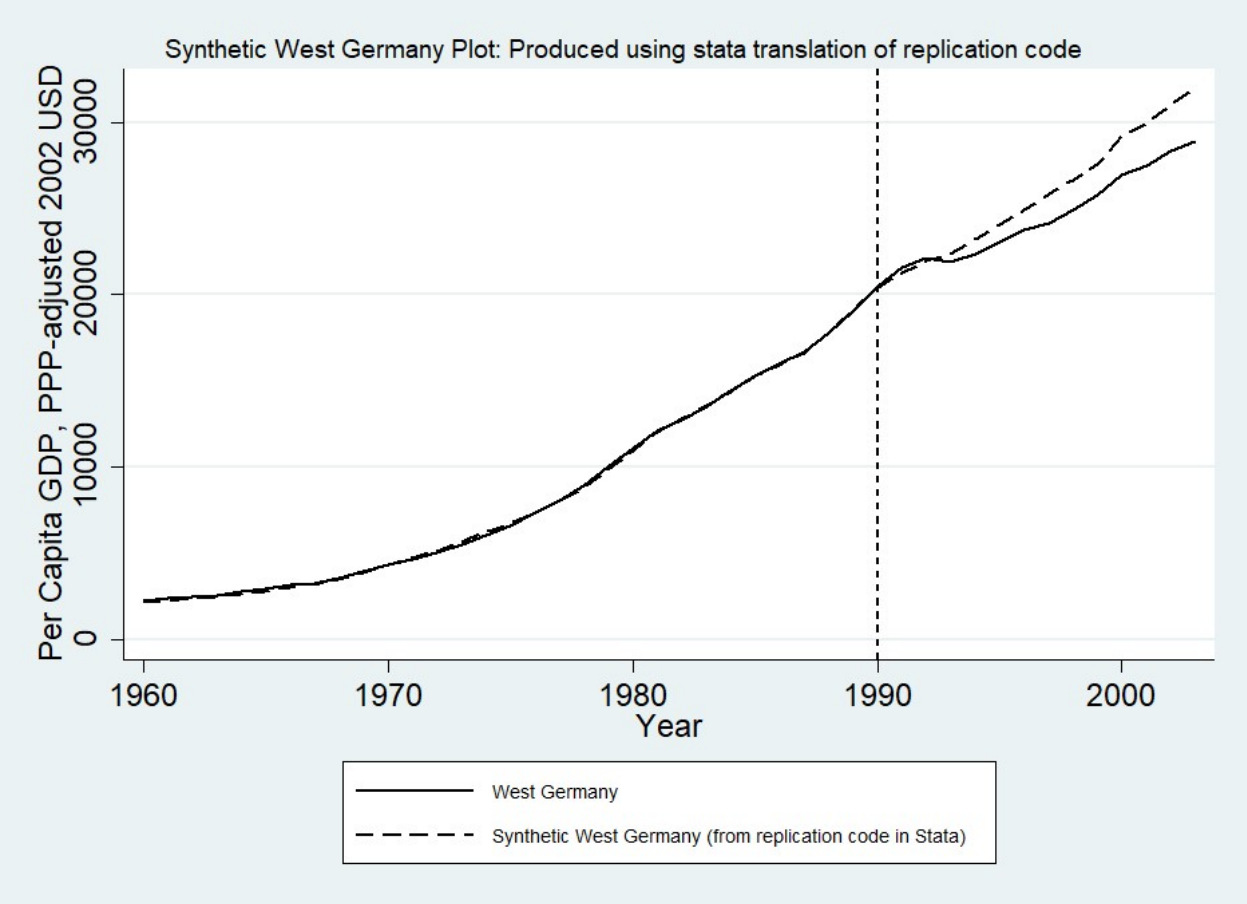

Figure 5.21 Synthetic West Germany Plot: Produced using Stata translation of replication code

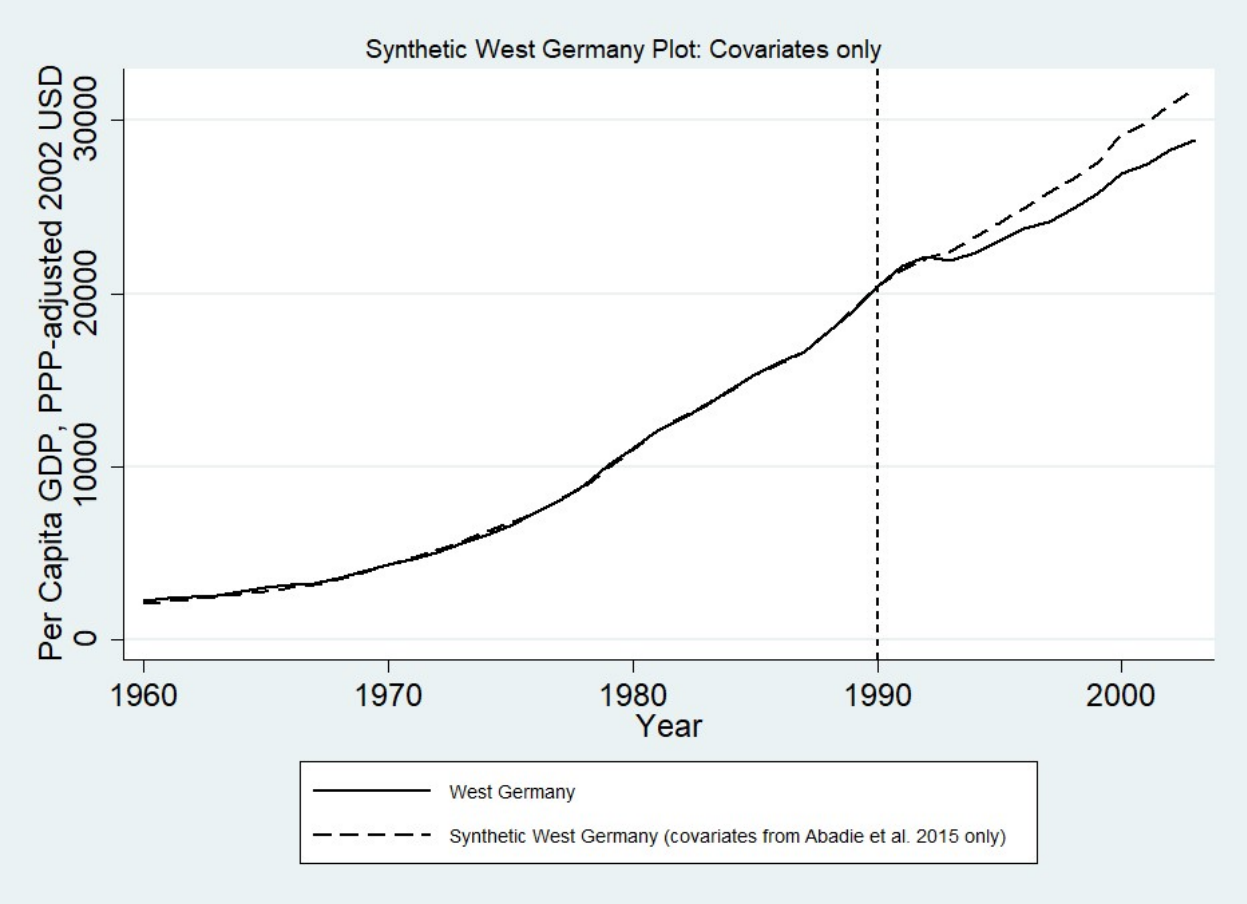

Figure 5.22 Synthetic West Germany Plot: Synthetic West Germany Plot: Covariates only

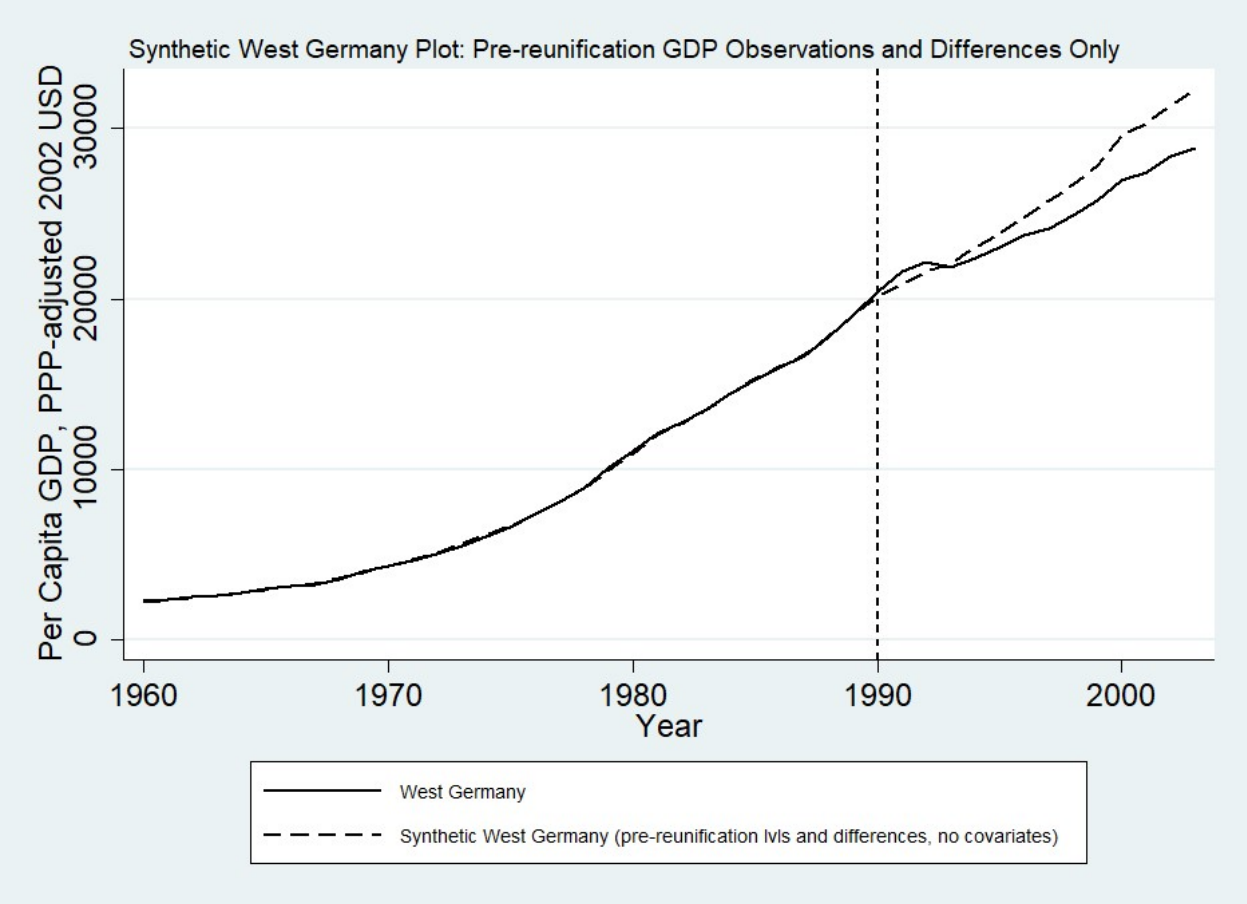

Figure 5.23 Synthetic West Germany Plot: Pre-1990 reunification GDP Observations and Differences Only

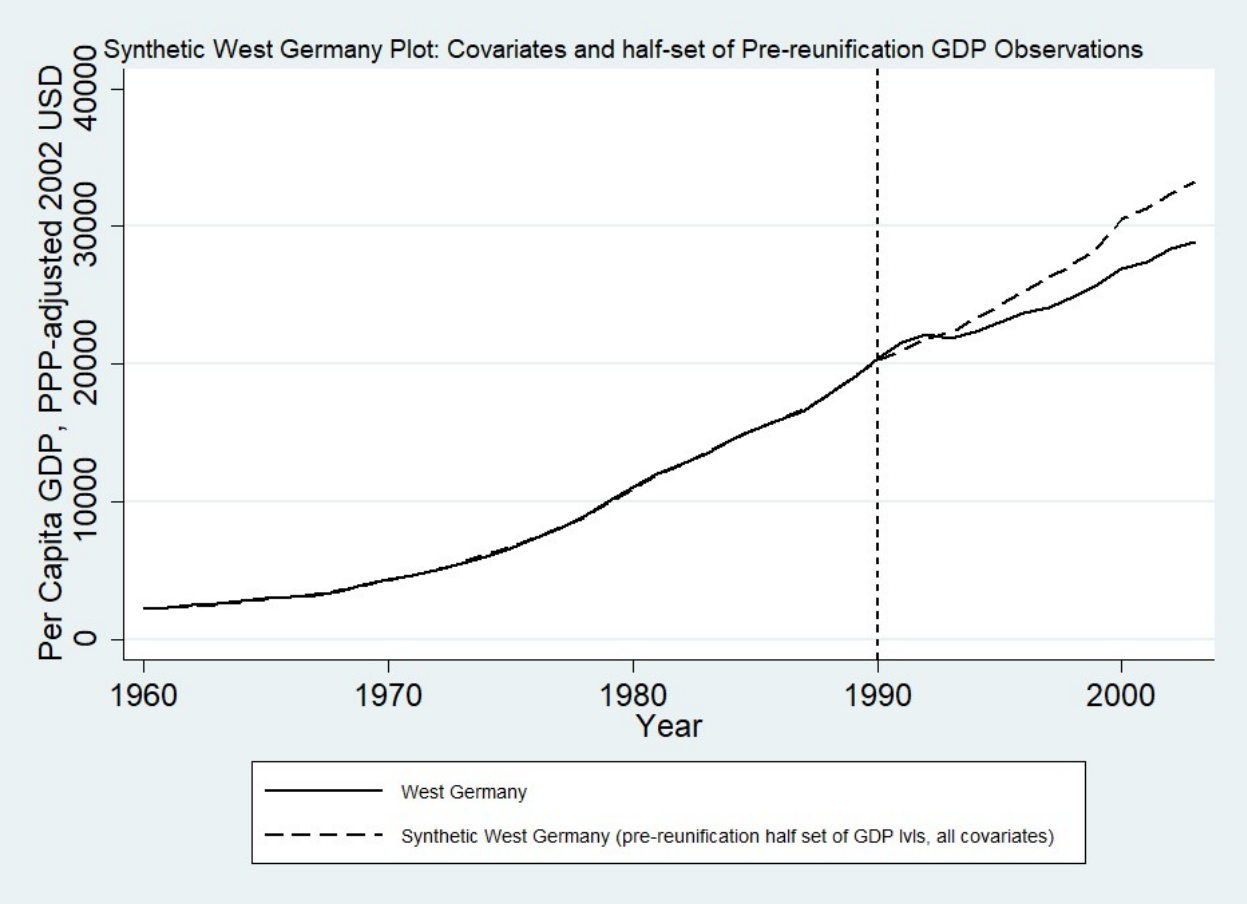

Figure 5.24 Synthetic West Germany Plot: Covariates and half-set of Pre-reunification GDP Observations

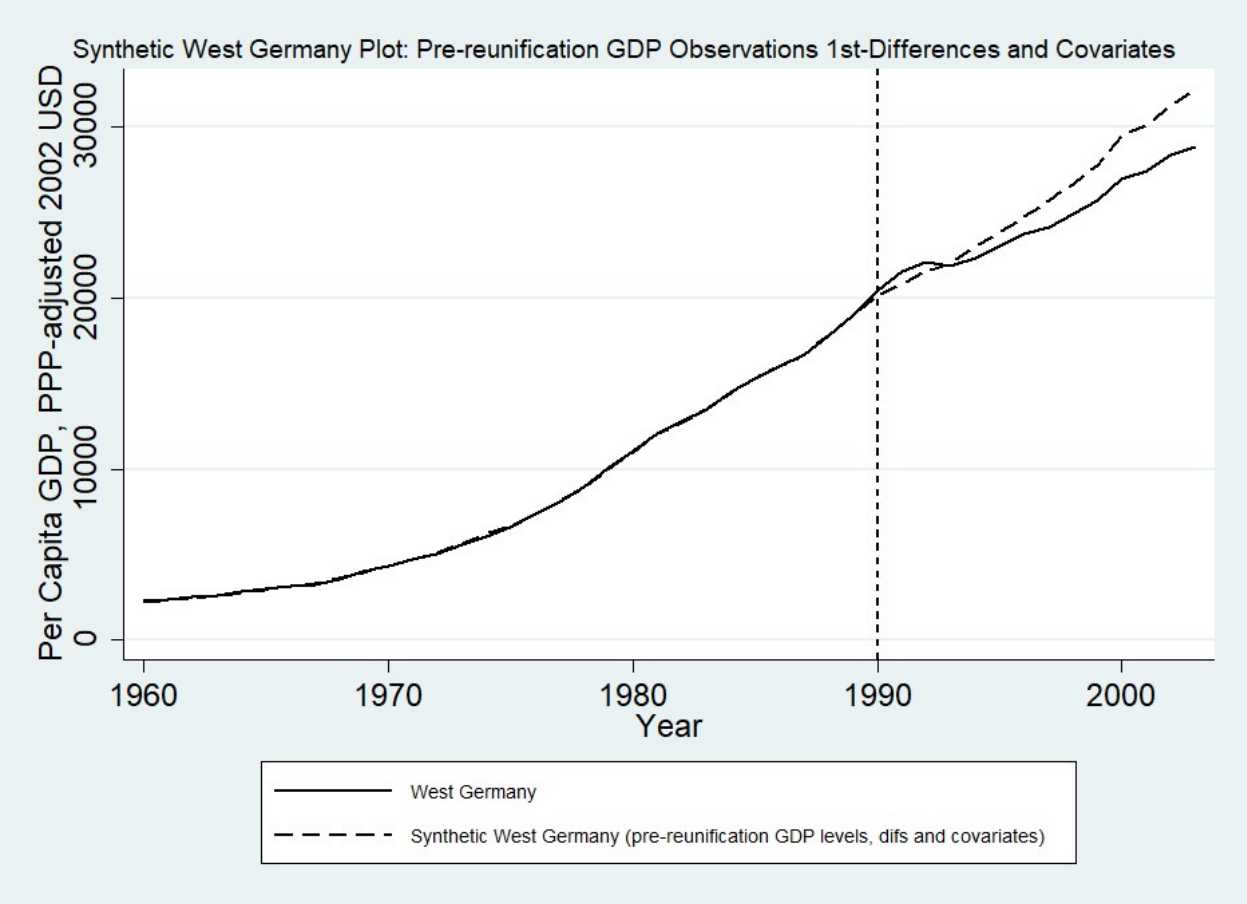

Figure 5.25 Synthetic West Germany Plot: Pre-1990 reunification GDP Observations, 1st-Differences, and Covariates

Chapter 6 – Conclusion

Previous chapters of this dissertation described the historical development of counterfactual causality, quasi-experimentation, econometric models exploiting "as if" randomness and the emergence of the contemporary synthetic case control design . After a careful outline of the underlying statistical rationale, Chapter 4 illustrates the impact that various standardization methods for the dependent variable can have on synthetic control construction, and Chapter 5 presents a series of alternate constructions from high profile synthetic control design publications as evidence against Kaul et al.'s (2015) argument regarding the use of pre-intervention levels of the outcome series as synthetic control group predictors.

Findings from Chapter 4 suggest that the ideal exposure term (i.e. scaling denominator) reflects the number of instances with the potential to produce a numerator observation over a time frame. For motor vehicle fatalities, vehicle miles travelled was the best exposure term available; however, passenger miles travelled would be even better since vehicle miles travelled does not account for non-drivers (which are eligible to contribute to the numerator in the event of a crash). It is worth noting that the estimated impact of California's Proposition 215 is smaller when standardized using vehicle miles travelled, but it still implies a clear reduction.

Chapter 5 presents a series of alternative constructions for Abadie and Gardeazabal's (2003) Basque Country, Abadie, Diamond, and Hainmeuller (2010) and Abadie et al. (2015), with the aim of demonstrating that Kaul et al.'s (2015) argument fails to recognize the predictive utility of pre-intervention observations of the outcome series, and overstates the ability of covariates alone to produce interpretable (i.e. well enough fit prior to the intervention for the post-intervention gap to

be distinguishable from the pre-intervention error) synthetic controls. To these ends, Chapters 4 and 5 largely accomplish these goals.

The various exposure terms and standardization methods evaluated in Chapter 4 produced minor differences in the quality of models pre-intervention fit, consistently indicated that Prop 215 caused a reduction in motor vehicle fatalities (varying in magnitude), and demonstrated that some theoretically reasonable scaling terms (e.g. per capita) made the synthetic control matching algorithm unable to converge on a synthetic control altogether. From these results, an ideal exposure term would reduce the range of panel values in each time point, place the treated unit of interest near the center of that distribution, and overall limit the presence of outliers in the panel. An ideal standardization removes the possibility that a synthetic control is being driven by only a small subset of the donor pool with small or large enough values to approximate the relatively large or small treated unit. In other words, if the treated unit is near the edge of the dependent variable overlap range with the donor pool, the resulting synthetic control with be constructed, out of sheer necessity, using donor pool units that also sit near the high end of the observed DV range. Therefore, this issue can be stated as, avoid fitting synthetic controls when the treated unit is close to the low-end or high-end range of the panel. Otherwise, the result will be biased by over contributions from its nearest neighbors.

Regarding Kaul et al.'s (2015) argument against supplying pre-intervention observations of the outcome in synthetic control predictor sets, none of the foundational Abadie synthetic control models in Chapter 5 produced synthetic controls with better pre-intervention fits that when pre-intervention observations of the outcome were included. The Abadie et al. (2010) model (see Figure 5.11) constructed on covariates alone performed particularly poorly. While Abadie and Gardeazabal (2003) and Abadie et al. (2015) did not perform quite as poorly as Abadie et al. (2010) when

constructed on covariate predictors alone, each of the three replication models produced a synthetic control with better pre-intervention fit when supplied only pre-intervention observations of the outcome compared to their covariates only counterparts. For each foundational study, I identify synthetic controls that outperform the original models in terms of pre-intervention fit by supplying both covariates and a full slate of pre-intervention DV observations and first-differences between DV observations. In effect, the matching algorithm is then supplied with a quadratic function for each pre-intervention point (i.e. level and slope).

From the findings in Chapter 5, it is clear that pre-intervention observations of the DV series and causal covariates generally share the bulk of their predictive utility and contain only a small amount of utility that isn't collinear with the other predictors. Kaul et al.'s (2015) fear that covariate imbalance suffers when pre-intervention observations are included as predictors, in my assessment, prioritizes covariate balance over the precision of the intervention effect estimate. In the Abadie et al. (2010) analysis, the retail price per pack of cigarettes held a substantial amount of predictive utility on its own, and was the only covariate among the three replicated models to improve Synthetic California's pre-intervention fit beyond what was achieved using all pre-intervention observations and original covariates when individual years of retail price per pack were added to the predictor set. Excluding that covariate reduces the quality of Synthetic California's pre-intervention fit, irrespective of how many other predictors were supplied along with it. In this case, the retail price per pack, California's Prop 99 increasing the state tax per pack of cigarette, and the number of packs sold (per capita) are directly related to one another. If we assume that retailers passed along the Prop 99 tax to consumers and didn't absorb any of the cost per pack themselves, then the retail price per pack of cigarettes in California would serve as an indicator of the timing and universal adoption of Prop 99 across California.

While Abadie et al. (2015) discuss some of the criteria used to exclude OECD nations from their donor pool, to my knowledge, no uniform measure of similarity or difference has been suggested in the SCG design literature. At the moment, the threshold of unit comparability is often known only to the researcher. Another unresolved methodological issue was raised at the conclusion of Chapter 5. When multiple treatments occur in within a short span of time and the treatments are all expected to impact the outcome time-series, parsing their impacts can be problematic. Abadie et al. (2010) cite Siegel (2002) as evidence linking the wave of local clean-air ordinances that followed the studied Proposition 99 intervention as being "spurred" by Proposition 99's voter approval in 1988.

In recent years, researchers have attempted to extend the synthetic control design for multiple treated units, either by pooling the treated units into a single aggregate unit (Kreif et al., 2015), through a factor model that predicts counterfactual outcomes (Xu, 2015), or through pooling treatment estimates from multiple units (Dube and Zipperer, 2013). While these approaches each have their merits and drawbacks, applying synthetic control methods analyses involving multiple treated units seems to be a somewhat misguided aim. Synthetic control methods were developed as an approach for quantitative causal analysis in single case observational studies. For the design to generalize to situations involving multiple treated units, strong (and often untenable) assumptions must be made about the treatment/intervention as well as the causal nexus of the outcome. Namely, researchers must assume that each treated unit experienced nomothetic (i.e. interchangeable) versions of the same intervention, and that the causal nexus of the outcome variable is consistent across treated units. Any variation in the policy itself across treated units would undermine the approaches to generalizing synthetic controls described above (Kreif et al., 2015; Xu, 2015; Dube and Zipperer, 2013). If the causal nexus of the outcome variable is inconsistent across treated units,

then the individual intervention effects will contain varying amounts of bias, an the intervention should be expected to impact the outcome in treated units to a varying degree. In both instances, the added value of pooled treatment effect estimates would be lost entirely.

As the synthetic case control design is still in its figurative adolescence a number of unresolved questions remain. While this dissertation expands the methodological basis for synthetic control applications in social science, issues related to its randomization inference procedure, the tendency for synthetic control results to vary (albeit slightly) across statistical computing platforms, and the length of the pre-intervention and post-intervention segments remain largely open. Without breaching these levees, however, longer pre-intervention segments are less likely to have equivalent solutions with different compositions and the probability of finding a high-quality pre-intervention fit greatly diminishes. This raises the possibility of trimming the pre-intervention segment shorter, but the technical argument in favor of trimming does not translate very well for causal inference. If each outcome observation is driven by the factor model in Equation [1], then a well fit synthetic control that is fit using all available observations of the outcome prior to the intervention can be more strongly assumed to be driven by the same factor model with the exception of the intervention. As such, synthetic controls constructed from an abbreviated pre-intervention segment should be interpreted with caution, and healthy suspicion.

In sum, this dissertation addresses a number of yet-to-be formalized aspects of synthetic case control design application, and as applications of this design proliferate and articles utilizing synthetic controls are published at an accelerating rate, the largely hidden researcher decisions addressed in this dissertation will likely become points of public disagreement as producers and consumers of synthetic case control designs become more familiar with the technical aspects of synthetic control construction. Synthetic Control Methods allow researchers to draw quantitative

causal inferences from single-case research designs. Although that newfound agency may very well precipitate a quantitative case study boom, researchers riding the synthetic control group wave would be well-served to take note of the nuances involved in their application which this dissertation addresses.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490).

Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, *42*(13).

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 113–132.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69-85.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of Economic Perspectives*, *24*(2), 3–30.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Bartos, B. J., McCleary, R., Mazerolle, L., & Luengen, K. (2020). Controlling Gun Violence: Assessing the Impact of Australia's Gun Buyback Program Using a Synthetic Control Group Experiment. *Prevention Science*, *21*(1), 131–136**.** https://doi.org/10.1007/s11121-019-01064-8

Bartos, B. J., Kubrin, C. E., Newark, C., & McCleary, R. (2019). Medical Marijuana Laws and Suicide. *Archives of Suicide Research*, 1–14. https://doi.org/10.1080/13811118.2019.1612803

Bartos, B. J., Newark, C., & McCleary, R. (2018). Marijuana medicalization and motor vehicle fatalities: A synthetic control group approach. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-018-9345-3

Bartos, B. J., & Kubrin, C. E. (2018). Can we downsize our prisons and jails without compromising public safety? Findings from California's Prop 47. *Criminology & Public Policy*, *17*(3).

Bilgel, F., & Galle, B. (2015). Financial incentives for kidney donation: A comparative case study using synthetic controls. *Journal of Health Economics*, *43*, 103–117.

Billmeier, A., & Nannicini, T. (2012). Assessing economic liberalization episodes: A synthetic control approach. *The Review of Economics and Statistics*, *95*(3), 983–1001.

Bohn, S., Lofstrom, M., & Raphael, S. (2013). Did the 2007 legal Arizona workers act reduce the state's unauthorized immigrant population? *The Review of Economics and Statistics*, *96*(2), 258–269.

Brady, H. E., Herron, M. C., Mebane, W. R., Sekhon, J. S., Shotts, K. W., & Wand, J. (2001). "Law and data": The butterfly ballot episode. *PS: Political Science and Politics*, *34*(1), 59–69.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312.

Campbell, D.T. (1986) Relabeling internal and external validity for applied social scientists. Pp.67-77 in W.M.K. Trochim (ed.), *Advancements in Quasi-Experimental Design and Analysis, New Directions for Program Evaluation, No. 31*. San Francisco: Jossey-Bass.

Campbell, D.T. (1991[1971]) Methods for the experimenting society. *American Journal of Evaluation*, 12, 223-260.

Campbell, D.T. and J.C. Stanley (1963) *Experimental and quasi-experimental designs for research*. Skokie: Rand-McNally.

Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market. *ILR Review*, *43*(2), 245–257.

Card, D., & Krueger, A. B. (1993). *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania*. National Bureau of Economic Research.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*(1), 11–20. https://doi.org/10.1017/S1745855207005029

Cook, T.D. and D.T. Campbell (1976) The design and conduct of quasi-experiments and true experiments in field settings. Pp 223-336 in M.D. Dunnette (ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Cook, T.D. and D.T. Campbell (1979) *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Cox, D. R. (1958). *Planning of experiments*. New York: Wiley.

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., … Thompson, S. (2012). Using natural experiments to evaluate population health interventions: New MRC guidance. *Journal of Epidemiology and Community Health*, *66*(12), 1182–1186.

Cronbach, L.J. (1982) *Designing evaluations of educational and social programs*. San Francisco:Jossey-Bass.

Deaton, A. S. (2009). *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development* (Working Paper No. 14690). National Bureau of Economic Research.

Dube, A. & Zipperer, B. (2013), Pooled Synthetic Control Estimates for Recurring Treatments: An Application to Minimum Wage Case Studies. Available at:

http://www.irle.berkeley.edu/events/spring14/zipperer/dubezipperer_pooledsyntheticcontrol.pdf

Greed, C. (1994). The place of ethnography in planning: Or is it 'real research? *Planning Practice & Research*, *9*(2), 119-127.

Hainmueller, J., Abadie, A., & Diamond, A. (2014). Synth Package for Stata.

Heckman, J. J. (1991). *Randomization and social policy evaluation* (Working Paper No. 107). National Bureau of Economic Research.

Hinrichs, P. (2010). The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. *The Review of Economics and Statistics*, *94*(3), 712–722.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Holland, P. W., Glymour, C., & Granger, C. (1985). Statistics and causal inference. *ETS Research Report Series*, *1985*(2), i-72.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *The American Political Science Review*, *105*(4), 765–789.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kaul, A., Klößner, S., Pfeifer, G., & Schieler, M. (2015). *Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates* (No. 83790). University Library of Munich, Germany.

Kirk, D. S. (2009). A natural experiment on residential change and recidivism: Lessons from Hurricane Katrina. *American Sociological Review*, *74*(3), 484-505.

Kreif N., Grieve R., Hangartner D., Turner A.J., Nikolova S., & Sutton M. (2015). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, *25*(12), 1514–1528.

Kubrin, C. E., & Bartos, B. J. (2020). Sanctuary Status and Crime in California: What's the Connection?. *Justice Evaluation Journal*, 1-19.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effects of police on crime. *American Economic Review*, *87*(3), 270-290.

Liu, S. (2015). Spillovers from universities: Evidence from the land-grant program. *Journal of Urban Economics*, *87*, 25–41.

Mackie, J. L. (1980). *The cement of the universe: A study of causation.* Oxford, England: Oxford University Press

McCleary, R., McDowall, D., & Bartos, B. J. (2017). *Design and analysis of time series experiments*. Oxford University Press.

McDowall, D., McCleary, R., & Bartos, B.J. (2019). *Interrupted Time Series Analysis*. Oxford University Press.

Nannicini T., & Billmeier A. (2011). Economies in transition: How important is trade openness for growth?. *Oxford Bulletin of Economics and Statistics*, *73*(3), 287–314.

Neyman J. S. (1923). On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. (Tlanslated and edited by D. M. Dabrowska and T. P. Speed, Statistical Science (1990), 5, 465-480). *Annals of Agricultural Sciences*, *10*, 1–51.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, *16*(1–2), 1–21.

Ozanne-Smith, J., Ashby, K., Newstead, S., Stathakis, V. Z., & Clapperton, A. (2004). Firearm related deaths: The impact of regulatory reform. *Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention*, *10*(5), 280–286.

Pearl, J. (2000). *Causality : models, reasoning, and inference Judea Pearl.* Cambridge, U.K. ;: Cambridge University Press.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Rosenbaum, P. R. (2002). Overt bias in observational studies. In *Observational studies* (pp. 71-104). Springer, New York, NY.

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

Ross, H. L., Campbell, D. T., & Glass, G. V. (1970). Determining the social effects of a legal reform: The British" breathalyzer" crackdown of 1967. *American Behavioral Scientist*, *13*(4), 493–509.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–58.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*(3), 808–840.

Sampson, R. J. (2010). Gold standard myths: Observations on the experimental turn in quantitative criminology. *Journal of Quantitative Criminology*, *26*(4), 489–500.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, *12*(1), 487–508.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Siegel, M. (2002). The effectiveness of state-level tobacco control interventions: A review of program implementation and behavioral outcomes. *Annual Review of Public Health*, *23*(1), 45–71.

Stearns, J. (2015). The effects of paid maternity leave: Evidence from temporary disability insurance. *Journal of Health Economics*, *43*, 85–102.

Weisburd, D, Mazerolle, L., & Petrosino, A. (2007). The academy of experimental criminology: Advancing randomized trials in crime and justice. *The Criminologist*, *32*(3), 3–7.

Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: Challenging folklore in evaluation research in crime and justice. *Journal of Experimental Criminology*, *6*(2), 209–227.

Winship, C., & Sobel, M. (2004). *Handbook of data analysis*. Sage Publications.

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science*, *58*(3), 451–488.

Xu, Yiqing. 2017. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25 (1): 57–76.

APPENDIX Chapter 4

Table A4.1 Synthetic Control Table of Results for California Motor Vehicle Fatalities Scaled Per Capita

Year	California	Synthetic_California	Difference	Diff^2			
1983	18.032318	18.433992	-0.401674	0.161342002			
1984	19.423946	19.324911	0.099035	0.009807931			
1985	18.758665	18.975975	-0.21731	0.047223636			
1986	19.382162	19.269326	0.112836	0.012731963			
1987	19.814846	19.227636	0.58721	0.344815584			
1988	18.943058	18.736155	0.206903	0.042808851			
1989	18.522726	18.143881	0.378845	0.143523534			
1990	17.330053	17.086474	0.243579	0.059330729			
1991	15.385254	14.940861	0.444393	0.197485138			
1992	13.533644	14.037378	-0.503734	0.253747943			
1993	13.314178	13.720818	-0.40664	0.16535609			
1994	13.441562	13.084645	0.356917	0.127389745			
1995	13.225401	13.370678	-0.145277	0.021105407	Mean pre-diff	Pre RMSPE	
1996	12.458292	13.457446	-0.999154	0.998308716	0.184641234	0.429699003	
1997	11.352579	13.586025	-2.233446	4.988281035			
1998	10.591834	13.21849	-2.626656	6.899321742			
1999	10.624133	13.63785	-3.013717	9.082490156			
2000	11.080063	13.652737	-2.572674	6.61865151			
2001	11.466291	13.97031	-2.504019	6.270111152			
2002	11.641406	13.635086	-1.99368	3.974759942			
2003	11.903805	13.171557	-1.267752	1.607195134			
2004	11.478306	14.47839	-3.000084	9.000504007			
2005	11.992091	14.028018	-2.035927	4.144998749			
2006	11.629964	13.224454	-1.59449	2.54239836			
2007	10.929272	12.916712	-1.98744	3.949917754			
2008	9.3425236	12.001405	-2.6588814	7.069650299			
2009	8.360013	10.582818	-2.222805	4.940862068			
2010	7.3012381	10.605263	-3.3040249	10.91658054			
2011	7.4710989	9.6468907	-2.1757918	4.734069957			
2012	7.7967625	10.398336	-2.6015735	6.768184676	Mean Post-diff	Post RMSPE	RMSPE Ratio
2013	8.1053896	10.56518	-2.4597904	6.050568812	5.856385053	2.419996912	5.631842039

Table A4.2 Synthetic Control Table of Results for California Motor Vehicle Fatalities Scaled Per Registered Vehicle

Year	California	Synthetic_California	Difference	Diff^2		
1981	0.29492018	0.2915405	0.00337968	1.14222E-05		
1982	0.25901484	0.25959794	-0.0005831	3.40006E-07		
1983	0.2475657	0.25304863	0.00548293	3.00625E-05		
1984	0.26888072	0.26246185	0.00641887	4.12019E-05		
1985	0.25290489	0.24521103	0.00769386	5.91955E-05		
1986	0.25661936	0.25685798	-0.00023862	5.69395E-08		
1987	0.26234889	0.25512613	0.00722276	5.21683E-05		
1988	0.24507055	0.24448016	0.00059039	3.4856E-07		
1989	0.24308284	0.23683058	0.00625226	3.90908E-05		
1990	0.23007625	0.22990973	0.00016652	2.77289E-08		
1991	0.2046794	0.21125376	-0.00657436	4.32222E-05		
1992	0.18365549	0.19022179	-0.0065663	4.31163E-05		
1993	0.17786874	0.18422612	-0.00635738	4.04163E-05		
1994	0.18485513	0.1806241	0.00423103	1.79016E-05		
1995	0.18254817	0.17991716	0.00263101	6.92221E-06	Mean pre-diff	Pre RMSPE
1996	0.15497427	0.17429453	-0.01932026	0.000373272	4.74228E-05	0.006886424
1997	0.14520064	0.17494928	-0.02974864	0.000884982		
1998	0.1343628	0.17467425	-0.04031145	0.001625013		
1999	0.13288756	0.17198383	-0.03909627	0.001528518		
2000	0.13333844	0.16699018	-0.03365174	0.00113244		
2001	0.13516423	0.17044178	-0.03527755	0.001244506		
2002	0.13557062	0.16560514	-0.03003452	0.000902072		
2003	0.13709725	0.16244847	-0.02535122	0.000642684		
2004	0.12864497	0.16478777	-0.0361428	0.001306302		
2005	0.1306584	0.15838885	-0.02773045	0.000768978		
2006	0.12504368	0.15829857	-0.03325489	0.001105888		
2007	0.11512071	0.14864685	-0.03352614	0.001124002		
2008	0.10028955	0.13398497	-0.03369542	0.001135381		
2009	0.08776058	0.12231547	-0.03455489	0.00119404		
2010	0.08560409	0.12706038	-0.04145629	0.001718624		
2011	0.09651538	0.12019343	-0.02367805	0.00056065		
2012	0.1070675	0.12242254	-0.01535504	0.000235777	Mean Post-diff	Post RMSPE
2013	0.11066795	0.12108133	-0.01041338	0.000108438	0.001012841	0.031825162
					RMSPE Ratio	
					4.621434932	

Table A4.3 Synthetic Control Table of Results for California Motor Vehicle Fatalities Scaled Per Vehicle Mile Travelled

Year	California	Synthetic_California	Difference	Diff^2		
1981	0.03233271	0.03240915	-7.644E-05	5.84307E-09		
1982	0.02714722	0.0270099	0.00013732	1.88568E-08		
1983	0.02503668	0.02535235	-0.00031567	9.96475E-08		
1984	0.02554226	0.02534863	0.00019363	3.74926E-08		
1985	0.0238921	0.02424425	-0.00035215	1.2401E-07		
1986	0.02444245	0.02427928	0.00016317	2.66244E-08		
1987	0.02432159	0.02392378	0.00039781	1.58253E-07		
1988	0.02232019	0.02211316	0.00020703	4.28614E-08		
1989	0.02152043	0.02133088	0.00018955	3.59292E-08		
1990	0.02005206	0.02006095	-8.89E-06	7.90321E-11		
1991	0.01817223	0.01830324	-0.00013101	1.71636E-08		
1992	0.0159666	0.01631133	-0.00034473	1.18839E-07		
1993	0.01563016	0.01604767	-0.00041751	1.74315E-07		
1994	0.01556209	0.01550165	6.044E-05	3.65299E-09		
1995	0.01516802	0.01519089	-2.287E-05	5.23037E-10	Mean pre-diff	Pre RMSPE
1996	0.0143467	0.01442704	-8.034E-05	6.45452E-09	5.4409E-08	0.000233257
1997	0.01321409	0.01436128	-0.00114719	1.31604E-06		
1998	0.01202216	0.01346288	-0.00144072	2.07567E-06		
1999	0.01186072	0.01355331	-0.00169259	2.86486E-06		
2000	0.01223875	0.0126602	-0.00042145	1.7762E-07		
2001	0.01273766	0.01287534	-0.00013768	1.89558E-08		
2002	0.01270765	0.01249401	0.00021364	4.5642E-08		
2003	0.01305347	0.0121231	0.00093037	8.65588E-07		
2004	0.01252596	0.01220342	0.00032254	1.04032E-07		
2005	0.01315953	0.01163783	0.0015217	2.31557E-06		
2006	0.01294743	0.0112613	0.00168613	2.84303E-06		
2007	0.0121683	0.01094473	0.00122357	1.49712E-06		
2008	0.01049235	0.01031322	0.00017913	3.20876E-08		
2009	0.00952275	0.00936064	0.00016211	2.62797E-08		
2010	0.00842499	0.00980211	-0.00137712	1.89646E-06		
2011	0.00877849	0.00936142	-0.00058293	3.39807E-07		
2012	0.00909058	0.00956963	-0.00047905	2.29489E-07	Mean Post-diff	Post RMSPE
2013	0.00942847	0.00984044	-0.00041197	1.69719E-07	9.89293E-07	0.000994632
					RMSPE Ratio	
					4.26409815	

```
Data Setup successful

              Treated Unit: California
            Control Units: Alabama, Arizona, Arkansas, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas,
                           Kentucky, Louisiana, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Nebraska, New Hampshire, New
                           Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Pennsylvania, South Carolina, South Dakota,
                           Tennessee, Texas, Utah, Virginia, West Virginia, Wisconsin, Wyoming

         Dependent Variable: farsfats
  MSPE minimized for periods: 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996
Results obtained for periods: 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003
                           2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

               Predictors: farsfatsdif(1982) farsfatsdif(1983) farsfatsdif(1984) farsfatsdif(1986) farsfatsdif(1987) farsfatsdif(1988)
                           farsfatsdif(1989) farsfatsdif(1990) farsfatsdif(1991) farsfatsdif(1992) farsfatsdif(1993) farsfatsdif(1994)
                           farsfatsdif(1995) farsfatsdif(1996) farsfats(1981) farsfats(1982) farsfats(1983) farsfats(1984) farsfats(1985)
                           farsfats(1986) farsfats(1987) farsfats(1988) farsfats(1989) farsfats(1990) farsfats(1991) farsfats(1992)
                           farsfats(1993) farsfats(1994) farsfats(1995) farsfats(1996)

Unless period is specified
predictors are averaged over: 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995

Second Step: Run Optimization

Nested optimization requested
Starting nested optimization module
could not calculate numerical derivatives
flat or discontinuous region encountered
r(430);
```

Figure A4.1 Synthetic Control Routine Stata Log for Unscaled California Motor Vehicle Fatalities

```
. synth per100kcapita per100kdif(1982)  per100kdif(1983) per100kdif(1984) per100kdif(1985) per100kdif(1986) per100kdif(1987) per100kdif(1988) per
> 100kdif(1989) per100kdif(1990) per100kdif(1991) per100kdif(1992) per100kdif(1993) per100kdif(1994) per100kdif(1995) per100kdif(1996)  per100kca
> pita(1983) per100kcapita(1984) per100kcapita(1985) per100kcapita(1986) per100kcapita(1987) per100kcapita(1988) per100kcapita(1989) per100kcapit
> a(1990) per100kcapita(1991) per100kcapita(1992) per100kcapita(1993) per100kcapita(1994) per100kcapita(1995) per100kcapita(1996) , trunit(51) tr
> period(1996) mspeperiod(1982(1)1996)  resultsperiod(1983(1)2013) unitnames(state) figure nested allopt keep(per100kcapita_51_test)

Synthetic Control Method for Comparative Case Studies

First Step: Data Setup

Data Setup successful

               Treated Unit: Wyoming
             Control Units: Alabama, Arizona, Arkansas, California, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa,
                            Kansas, Kentucky, Louisiana, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Nebraska, New Hampshire,
                            New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Pennsylvania, South Carolina, South Dakota,
                            Tennessee, Texas, Utah, Virginia, West Virginia, Wisconsin

        Dependent Variable: per100kcapita
  MSPE minimized for periods: 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996
Results obtained for periods: 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005
                            2006 2007 2008 2009 2010 2011 2012 2013

                Predictors: per100kdif(1982) per100kdif(1983) per100kdif(1984) per100kdif(1985) per100kdif(1986) per100kdif(1987)
                            per100kdif(1988) per100kdif(1989) per100kdif(1990) per100kdif(1991) per100kdif(1992) per100kdif(1993)
                            per100kdif(1994) per100kdif(1995) per100kdif(1996) per100kcapita(1983) per100kcapita(1984) per100kcapita(1985)
                            per100kcapita(1986) per100kcapita(1987) per100kcapita(1988) per100kcapita(1989) per100kcapita(1990)
                            per100kcapita(1991) per100kcapita(1992) per100kcapita(1993) per100kcapita(1994) per100kcapita(1995)
                            per100kcapita(1996)

Unless period is specified
predictors are averaged over: 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995

Second Step: Run Optimization

Nested optimization requested
Starting nested optimization module
could not calculate numerical derivatives
flat or discontinuous region encountered
```

Figure A4.2 Synthetic Control Routine Stata Log for Wyoming Motor Vehicle Fatalities per capita

APPENDIX Chapter 5

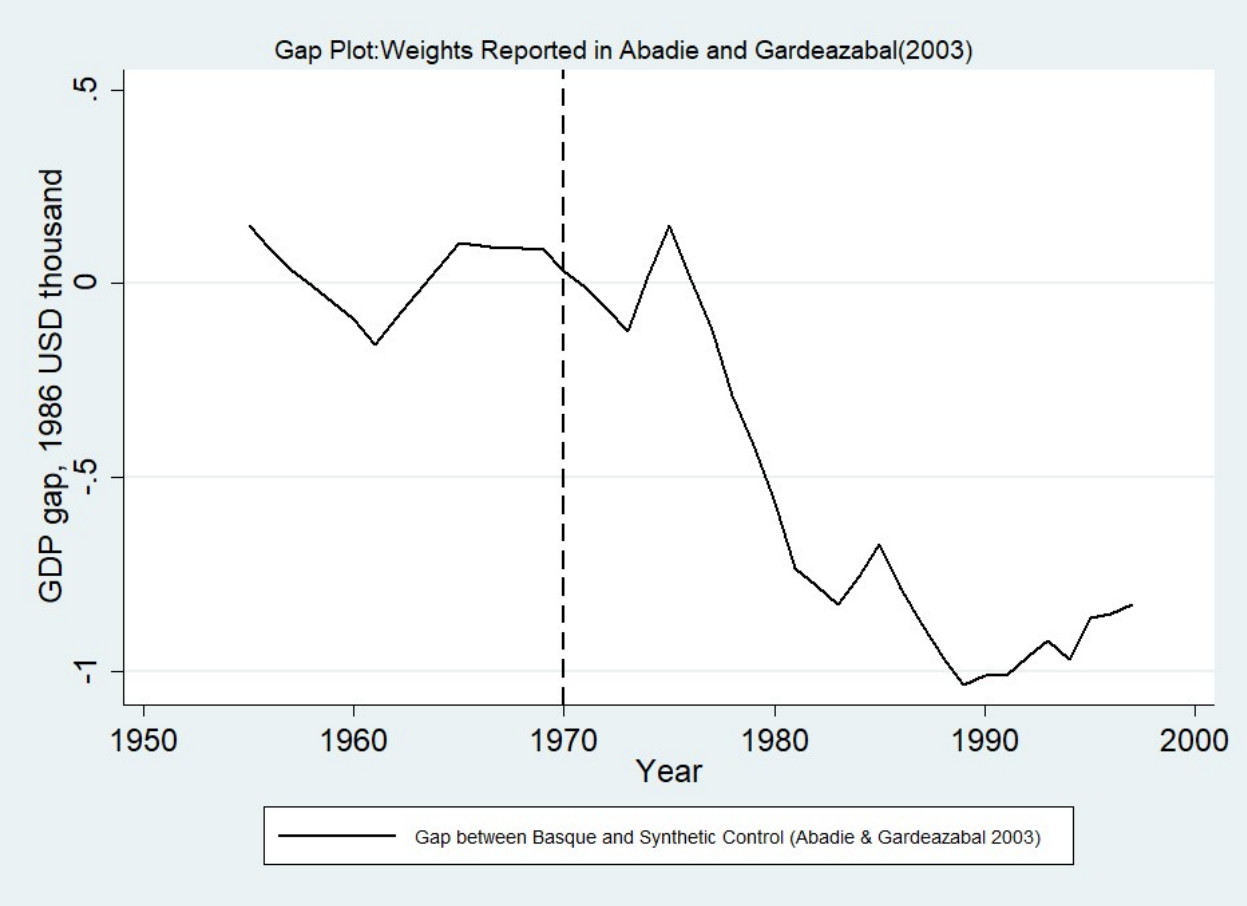

Figure A5.1 Synthetic Control Group Gap Plot: Weights reported in Abadie and Gardeazabal (2003)

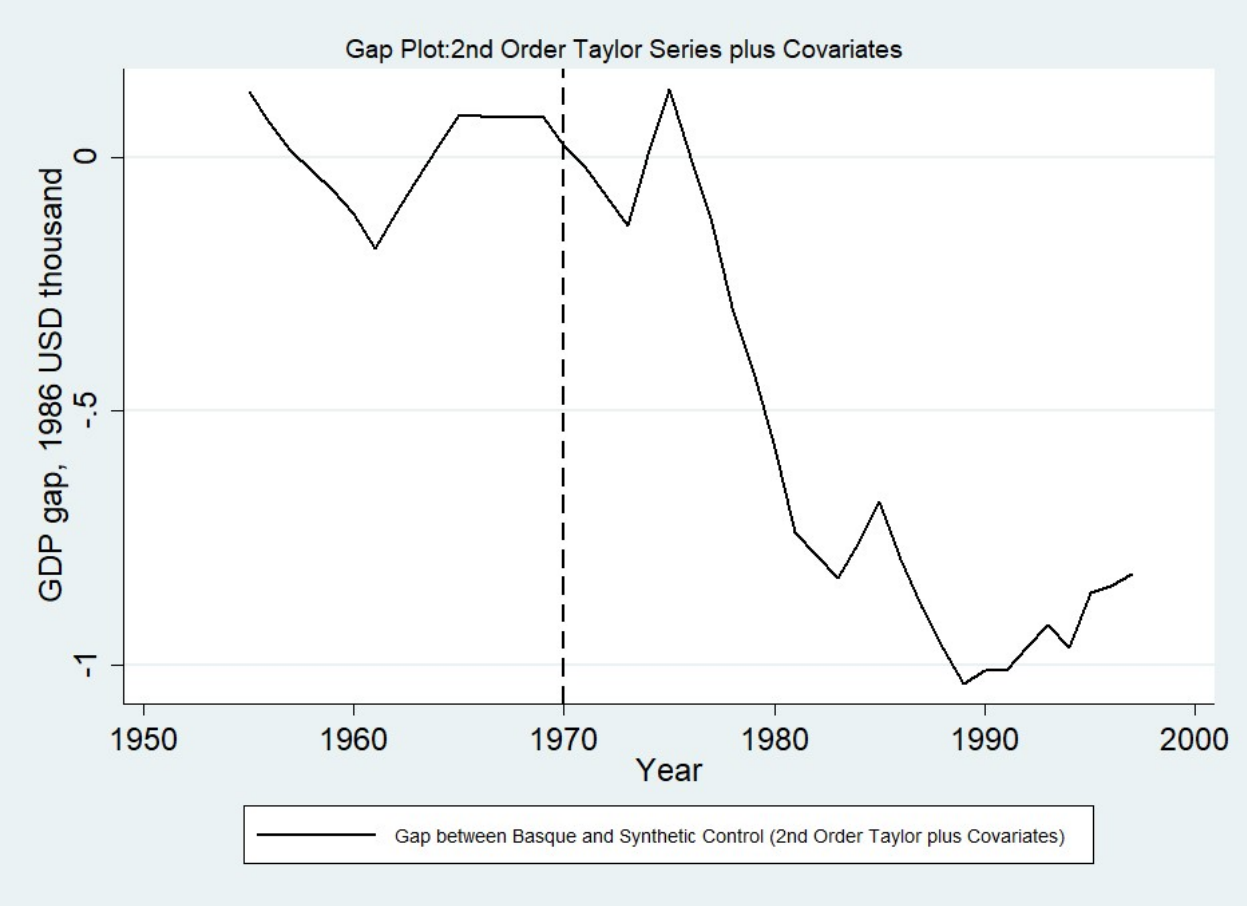

Figure A5.2 Synthetic Control Group Gap Plot: 2nd Order Taylor Expansion Plus Covariates

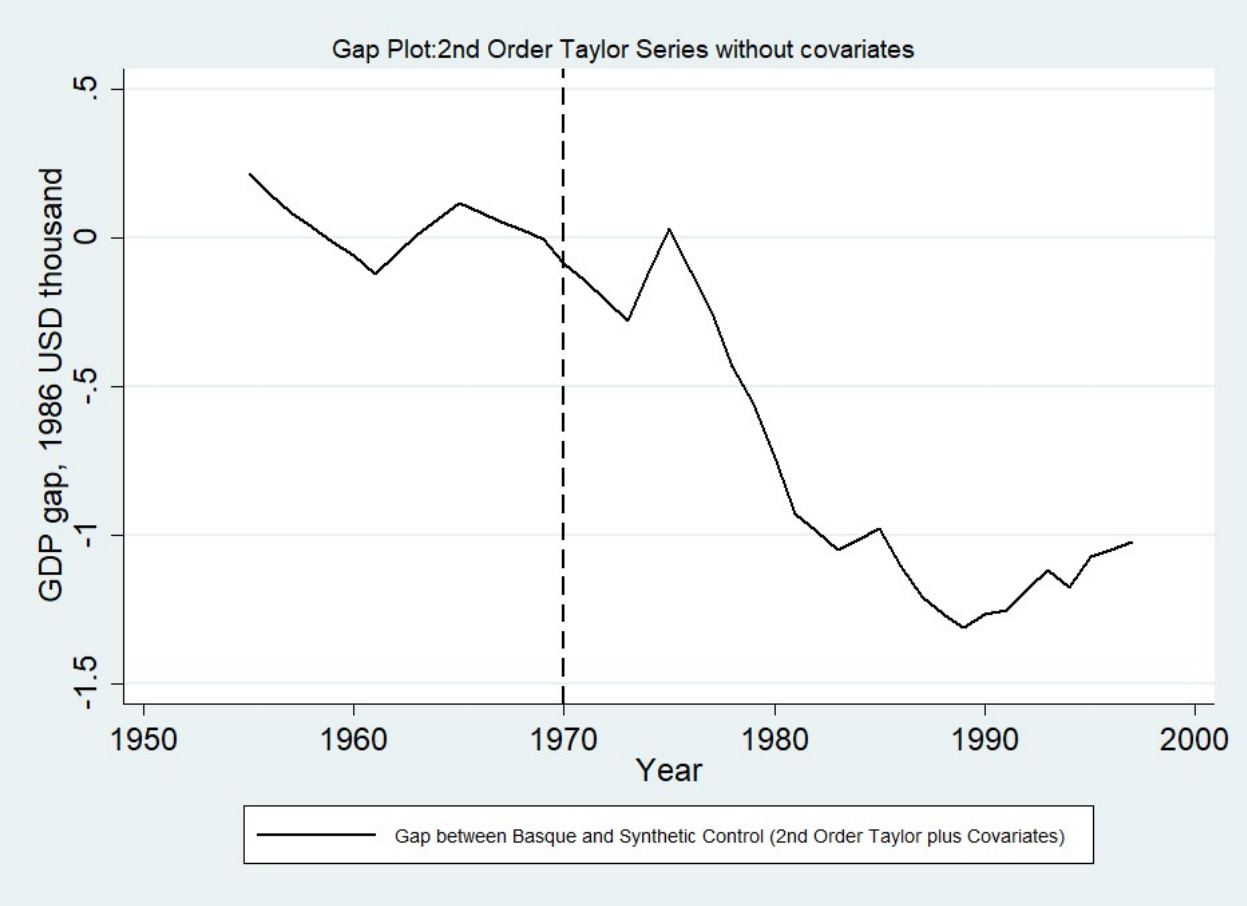

Figure A5.3 Synthetic Control Group Gap Plot: 2nd Order Taylor Expansion Without Covariate Predictors

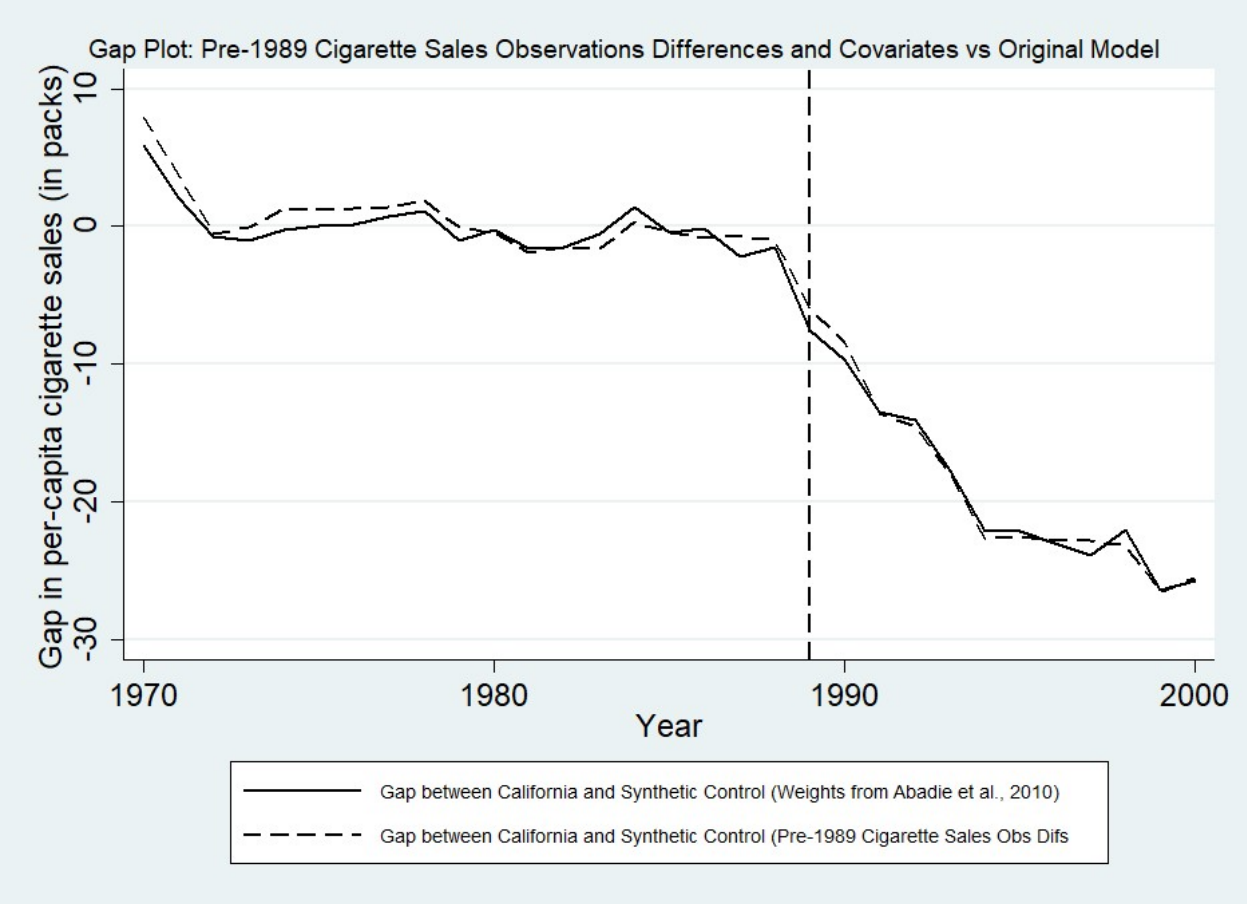

Figure A5.4 Synthetic California Gap Plot: Abadie et al. (2010) vs. Pre-1989 GDP Observations Differences and Covariates