The Legislative State and the Republican Revolution: A Causal Machine Learning Approach

By

JACK T. RAMETTA

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Political Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Christopher Hare, Chair

Ryan Hübert

Erik J. Engstrom

Scott MacKenzie

Lauren Peritz

Committee in Charge

2025

To Sarah

Acknowledgments

First and foremost I have to thank my partner Sarah for her support throughout graduate school. Without her support I surely would have failed. Thanks as well to my parents and sister for their support along the way.

I of course need to give special thanks to my advisors Chris Hare and Ryan Hübert for shepherding me through my graduate work. I learned an enormous amount from you both, and my perspective on politics has been shaped by your courses and feedback. I owe you both a great debt for any current or future success. My thanks to my other committee members as well: Erik Engstrom, Scott MacKenzie, and Lauren Peritz, for their guidance and feedback.

I also must thank Sam Fuller for his support and collaboration over the years. Last, and undoubtedly most important, I must acknowledge the support of my dogs Jaco and Stella, who, in their own way, supported (or perhaps hindered) my graduate work.

Dissertation Chair: Christopher D. Hare Jack T. Rametta (ID: 917850880)

The Legislative State and the Republican Revolution: A Causal Machine Learning Approach

Abstract

In this dissertation, I interrogate the effects of the Republican Revolution of 1994 on Congressional oversight capacity using a combination of traditional causal inference methods and causal machine learning approaches. The venue for this investigation is a novel dataset that comprises the universe of published and publicly available material from the Government Accountability Office (GAO), Congress' understudied "watchdog" oversight and auditing agency. Through my investigation I reveal the deleterious effects of the Republican Revolution on oversight capacity at GAO. In the first chapter, I explore the effects of the Republican Revolution in terms of several different dependent variables of interest. In the second chapter, I explore heterogeneity in these main effects using a novel causal machine learning estimation procedure. Finally, in the third chapter, I explore the broader applicability of causal machine learning methods in the social sciences in light of recent debates. My results contribute to a growing literature that is skeptical of Congress' ability to effectively oversee the executive branch. This dissertation also contributes to a growing literature that applies blackbox predictive algorithms for inference.

Contents

Introduction

Congress' capacity to craft public policy and oversee the executive branch is increasingly dubious. To put it bluntly, in the eyes of many the modern Congress is simply "overwhelmed" (LaPira, Drutman, and Kosar 2020); outmanned and outgunned by an increasingly powerful executive branch and a presidency armed with ambiguous unilateral powers (Moe and Howell 1999; Lowande and Rogowski 2021). Incoming members of Congress are inexperienced, more than half have no background in electoral politics (Porter and Treul 2024). How did we get here? Scholars have built a laundry list of potential causes for Congressional dysfunction, including for instance, increasing elite polarization (Hare and Poole 2014) and the decreasing benefits of holding Congressional office (Hall 2019). An understudied aspect of the decline in Congressional capacity is the diminished role and capabilities of what we might call the "Legislative State:" a small but complex network of institutions, rules, and people that assist in the day-to-day work of information acquisition, policymaking and oversight in Congress. Crucial in this regime are Congress' support agencies: the Congressional Budget Office, the Congressional Research Service, and the Government Accountability Office.[1] These critical support agencies, created in the early 20th century and reaching peak output in the 1970s and 1980s, have waned in both size and capability in recent decades (Kosar 2018). This decline was precipitated, in large part, by the 1994 Republican Revolution (Rubin 2002). After retaking the House for the first time in 40 years, the 94' Revolutionaries lashed out at the support agencies, cutting budgets and reducing staff levels dramatically, especially at GAO and CRS.

My dissertation will answer the following question: how did the Republican Revolution affect capacity in the legislative state? In particular, how did the 94' change in majority party, and subsequent resource reductions, alter these agencies ability and willingness to assist in Congressional oversight and policymaking? To answer this question, I gather a novel dataset: the universe of

[1] A complete list would also include the broader Library of Congress (of which CRS is a part), the Architect of the Capitol, and Government Printing Office.

public reports from the GAO. With these datasets, I am able to paint high resolution images of the productive capacity of this agency before and after the 94' revolution. My dissertation focuses on GAO as an understudied tool of centralized, "police patrol" oversight. The GAO is the largest single Congressional oversight body, and the only that employs a nonpartisan staffing regime. I detail overlooked pathways for GAO's recommendations to enter the executive policymaking process, both through the Congress and directly to executive agencies. Moreover, I examine the agency's response to severe scrutiny and budget cuts during the Republican Revolution of 1994. I provide general evidence that the Republican Revolution hindered GAO's ability to provide effective oversight of the executive branch on behalf of Congress, and further, potentially hampered the agency's long term capabilities as an oversight body. I discuss the implications of these findings for theories of Congressional oversight and policymaking. Methodologically, I employ both classical causal inference approaches and causal machine learning methods to reach my these conclusions. In the final chapter, my co-author and I argue for the applicability of causal machine learning methods more widely in the social sciences and provide a framework for scholars to conceptualize these approaches in their own work. The subject matter of this dissertation is both timely and of increasing normative importance. In recent years, the small group of nonpartisan agencies that continue to support Congress in its policymaking and oversight duties have once again come under fire. These attacks illustrate the ongoing importance of understanding the roots of attacks on legislative capacity in the modern Congress.

Outline

The first chapter discusses the GAO's response to the Republican Revolution, employing a regression discontinuity in time design to estimate the causal effects of majority party turnover on GAO's productive oversight outputs. The second chapter builds on the empirical analysis on Chapter 1 by interrogating the possibility of heterogeneity in the treatment effects identified in the first chapter. This analysis requires building out a custom analytical scaffolding that tailors causal machine learning methods to the peculiarities of the regression discontinuity in time setting. Finally the third chapter explores the more generally utility of causal machine learning methods for social scientists, presenting wide-ranging simulation evidence that compares these new approach to traditional methods for the analysis of cross-sectional data.

Bibliography

Hall, Andrew B. 2019. *Who wants to run?: How the devaluing of political office drives polarization.* University of Chicago Press.

Hare, Christopher, and Keith T Poole. 2014. "The polarization of contemporary American politics." *Polity* 46 (3): 411–429.

Kosar, Kevin R. 2018. "The atrophying of the congressional research service's role in supporting committee oversight." *Wayne L. Rev.* 64:149.

LaPira, Timothy M, Lee Drutman, and Kevin R Kosar. 2020. *Congress Overwhelmed: The Decline in Congressional Capacity and Prospects for Reform.* University of Chicago Press.

Lowande, Kenneth, and Jon C Rogowski. 2021. "Presidential unilateral power." *Annual Review of Political Science* 24:21–43.

Moe, Terry M, and William G Howell. 1999. "Unilateral action and presidential power: A theory." *Presidential Studies Quarterly* 29 (4): 850–873.

Porter, Rachel, and Sarah Treul. 2024. "Evaluating (In) Experience in Congressional Elections." *American Journal of Political Science (Forthcoming).*

Rubin, Irene S. 2002. *Balancing the federal budget: Trimming the herds or eating the seed corn?* CQ Press.

CHAPTER 1

Did the Republican Revolution Hamstring Congressional Oversight?

A Case Study of the Government Accountability Office

Jack T. Rametta

University of California, Davis

Legislative support agencies provide crucial policymaking and oversight assistance to Congress. However, their authority, influence, and scope has diminished in recent decades. The Republican Revolution of 1994 precipitated this decline, slashing support agency budgets and staff. But how exactly did the new Republican majority affect the oversight capacity through the support agencies? Did the '94 Revolution diminish Congressional oversight capacity? To answer this question, I compile and analyze a novel dataset: the universe of published reports and testimonies from the Government Accountability Office, Congress' "watchdog" oversight agency. This dataset contains more than 55,000 unique items spanning back to the creation of the agency in the early 20th century. Employing a regression discontinuity design, I investigate the effects of the Republican Revolution on GAO's public outputs. I find the Republican Revolution corresponds with a significant reduction in the volume of policy recommendations made by GAO to executive agencies and to Congress. These findings stand in contrast with the picture painted by GAO's reports to Congress during this period, which suggest increasing productivity despite party control and subsequent budget cuts. I reconcile this contrast by outlining qualitative evidence that GAO manipulated reported productivity statistics to avoid additional scrutiny from its new principal.

"...if you look at Price Waterhouse in New York, they got 5,000 people [auditors] there for the city of New York in their area and we [GAO] got what do we got now? 3500 or something like that. For the whole federal government. The whole nation, you know. 3100. Yeah, ridiculous."

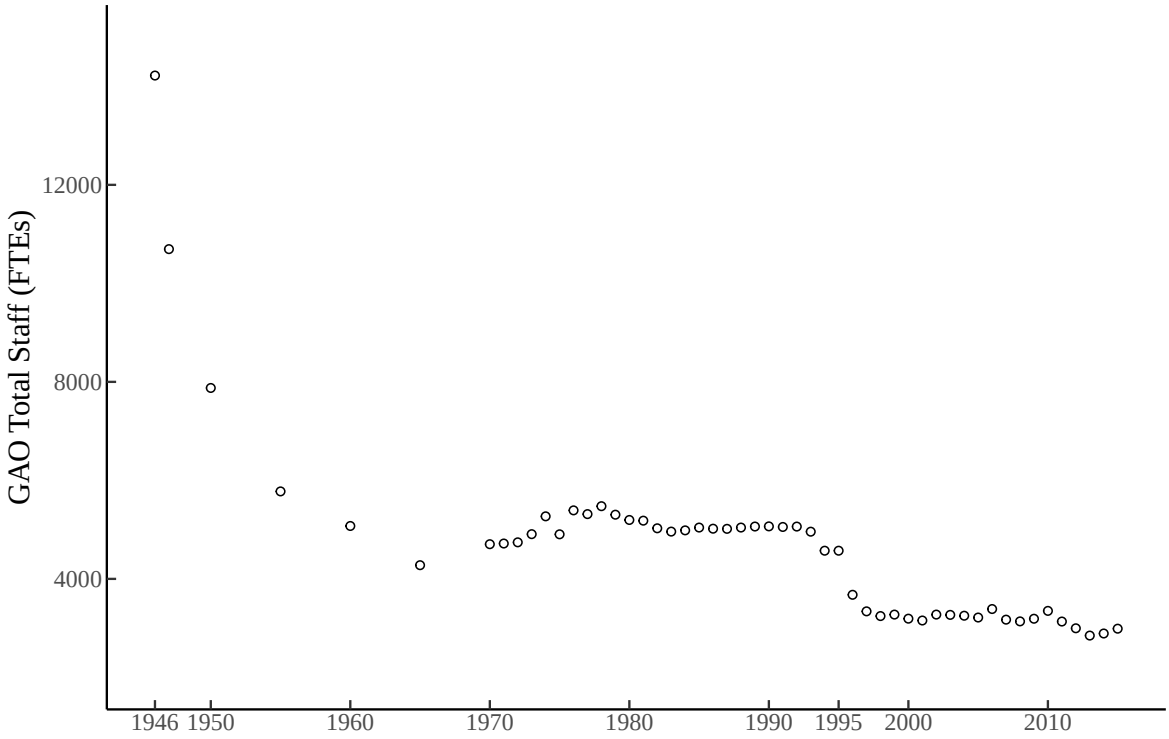
– Chuck Bowsher, GAO Comptroller General 1981-1996 (Bowsher 2021)

Introduction

On January 4, 1995, Newt Gingrich was sworn in as the first Republican Speaker of the House of Representatives in 40 years. Gingrich led the electoral blockbuster "Republican Revolution" (subsequently RR) in 1994, successfully flipping 54 House seats and 8 Senate seats. The '94 revolutionaries ran on the "Contract with America," an expansive agenda with a smorgasbord of ambitious policy goals. Among them was a desire to fundamentally transform the institution of the United States Congress (Gingrich 1994). This desire partially stemmed from a skepticism of what we might call the "legislative state:" a small but complex network of institutions, rules, and people that assist in the day-to-day work of policymaking and oversight in Congress. Many Republicans viewed this ecosystem as fundamentally opposed to their interests; a system designed to benefit rank and file Democratic members and powerful Democratic committee chairmen.

Acting on this view, '94 Republicans acted decisively against legislative institutions that existed at the time. They defunded the Office of Technology Assessment, a small support agency that provided Congress with policy information related to science and technology (Bimber et al. 1996). They disbanded legislative service organizations, structures that allocated funds to groups of legislators pursuing policy aims or political initiatives, and allowed those groups to hire additional policy staff (Clarke 2020). The revolutionaries enacted large spending cuts to legislative support agencies, in particular, the Government Accountability Office (GAO); Congress' "watchdog" agency built to assist Congressional oversight of the vast executive administrative state. These budget cuts precipitated large staffing reductions at GAO. As shown in Figure 1.1, GAO staff dwindled in fiscal year (FY) 1995 from 4,572 down to 3,677 by the end of FY96, and fell further to 3,245 by FY98. GAO staffing levels continued to slowly dwindle in the period since, resting around 3,000 in the 2000s and 2010s.

FIGURE 1.1. Government Accountability Office Staffing by Fiscal Year



Note: Data retrieved from (Reynolds 2023), originally gathered by (Ornstein, Mann, and Malbin 2002).

This leads to the central question of this chapter: How did the change in majority party affect Congressional oversight capacity via the Government Accountability Office? Put another way, how did the GAO's ability and willingness, to assist in Congress' oversight role change in response to the Republican Revolution?

To answer this question, I gather a novel dataset that comprises the publicly available universe of GAO reports over the period from 1922-2022. Following extensive data extraction, I employ a regression discontinuity design on these data to investigate the possible effects of both factors, a change in party management and budget cuts, on GAO's productive output. I find disparate results. On one hand, GAO's aggregate production of all public reports appears flat, or even increasing, over this tumultuous period. On the other hand, I find a marked decline in GAO's production of recommendations for executive policymaking that coincides directly with the Republican Revolution. To reconcile these findings, I outline strong qualitative evidence that GAO officials manipulated aggregate statistics in response to Congressional scrutiny. Specifically, agency officials

attempted to undermine the justification for *further* budget cuts—or worse, the total dissolution of their agency a la the Office of Technology Assessment—by strategically adjusting the mixture, timing, and nature of agency deliverables. These strategic adjustments were designed to create the image that GAO could "do more with less."

These adjustments allowed GAO to present data to Congress in its annual reports that made production appear steady, or even increasing depending on the metric. Importantly, this evidence of manipulation does not extend to aggregate performance metrics the agency *did not report to Congress*, and that Congress had no easy way of tabulating. I derive one such measure, the volume of policy recommendations to the executive branch. On this metric, I find large effects: the volume of GAO recommendations for executive action decreased by an estimated 600 recommendations in the fiscal years closely following the Republican Revolution. For context, in FY1994 GAO issued approximately 1,300 unique policy recommendations, in FY1995 that figure fell to around 700.

Police Patrols, Fire Alarms, and Savvy Bureaucrats

How does Congress oversee the administrative state? The literature on Congressional oversight presents three useful categorizations: police patrol versus fire alarm oversight (McCubbins and Schwartz 1984), formal versus informal, and public versus private. "McNollgast" (McCubbins, Noll, and Weingast) argue in a famous series of articles that Congress has an understandable preference for "fire alarm" style oversight (McCubbins and Schwartz 1984). In this view, Congress relies on interest groups, outside actors, or merely events to pull the fire alarm that triggers Congressional scrutiny. This is in contrast with "police patrol" oversight, where Congress undertakes constant monitoring of the executive branch expending its own valuable time and attention. While this view has been subjected to a wide variety of theoretical and empirical criticism (some of which I describe momentarily), the basic framework is nonetheless useful for conceptualizing Congressional oversight of the administrative state (Gailmard 2014).

Despite this, many scholars and observers of the modern Congress continue to worry the institution is unable to effectively oversee the executive branch (LaPira, Drutman, and Kosar 2020). Despite expansive authority and a slew of constitutional, as well as statutory, mechanisms in the Congressional oversight arsenal, many argue Congress provides insufficient oversight of executive

policymaking. Critics argue that although Congress appears to frequently employ its oversight authority, for example, through numerous oversight hearings (Lewallen 2020), this authority is wielded ineffectively, and is often thwarted by savvy bureaucrats. In this telling, executive bureaucrats, armed with superior information and understanding of the process of policymaking, outmaneuver their Congressional overseers and "bend the rules" to drag policy closer to their preferred ends (Potter 2019).

Other research focuses on understudied aspects of oversight. For example, Lowande (2018) argues, while models of Congressional oversight focus largely on committee hearings, individual members of Congress often engage with executive agencies through direct correspondence. His findings suggest oversight is driven more by institutional roles and parochial district interests rather than member ideology. Other work interrogates member correspondence as a mechanism for oversight, finding descriptive representation translates to improved substantive representation in the rulemaking process (Lowande, Ritchie, and Lauterbach 2019). Subsequent work by Lowande and Augustine Potter (2021) emphasizes the importance of ex-post Congressional review through the notice and comment process for agency rulemaking. One measure of Congressional oversight efficacy is the extent of improper payments made by executive agencies. Recent work suggests that Congressional hearings appear to have no measurable effect on improper payments made by agencies (Ban and Hill 2023).

Theories of Congressional oversight largely ignore the Government Accountability Office, which functions as a centralized, formal, "police force" that oversees the entire executive branch on behalf of Congress.[1] For scale, the agency employs more than 3,000 full-time staff. This exceeds the total staff of the House and Senate committee and leadership offices combined (Reynolds 2023). It also dwarfs staffing levels at the other support agencies by an order of magnitude, for example the Congressional Research Service (CRS) employs roughly 600, and the Congressional Budget Office merely 250. While Congressional committees have turned toward high-profile controversies and investigations, the GAO continues to produce reams of policy recommendations for federal programs of all shapes and sizes. Agencies will often implement GAO recommendations either entirely, or in-part, and the lion's share of their proposals receive at least some response from the executive agency in question (roughly 75%) (GAO 2023).

[1]In the Appendix, I further detail the relatively scant existing literature on the GAO.

The Evolution of GAO's Oversight and Policymaking Role

GAO's Current Mission and Role in Policymaking The stated mission of the Government Accountability Office, previously the Government Accounting Office, is to be the "congressional watchdog." In the agency's own language:

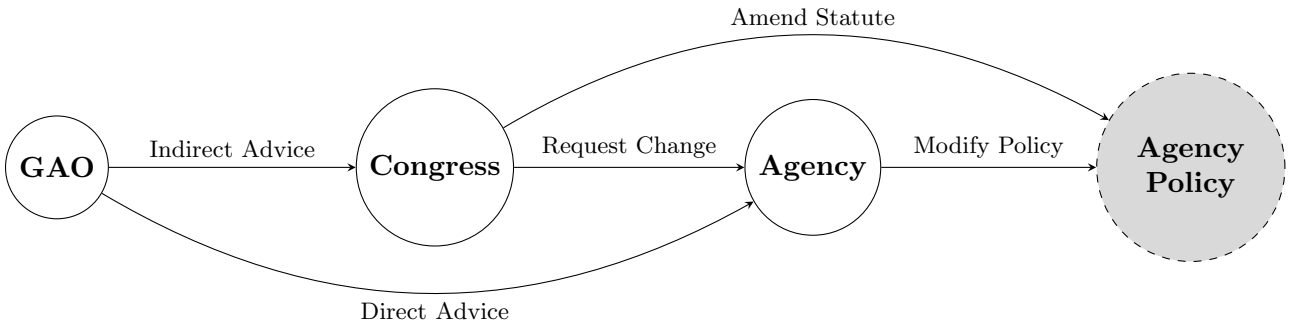
> "GAO provides Congress, the heads of executive agencies, and the public with timely, fact-based, non-partisan information that can be used to improve government and save taxpayers billions of dollars (GAO 2023)."

In the modern era, the agency is tasked with overseeing executive outlays and program implementation, uncovering cost and efficiency savings where possible. This mission is notably broad; the GAO releases reports and recommendations across the range of executive functions. The recommendations range wildly in scope and importance. Some reports make modest recommendations to, for example, suggesting the HHS secretary improve the public reporting of autism research (Dicken 2024). Other reports make recommendations with more significant implications, for example, recommending the Department of Energy safeguard against insider leaks of nuclear information (Bawden 2023). These recommendations also sometimes serve as the basis for statutory interventions by the Congress, or changes in administrative policy. For example, GAO closely monitored the outlay of Federal COVID dollars during the pandemic. Congress later revoked unspent authority based in part on this reporting (Kociolek 2023). In recent years, GAO's spotlight has been cast with increasing frequency on large spending legislation (e.g., the American Recovery and Reinvestment Act). The agency has also faced recent scrutiny from the Trump administration and Congressional Republicans for resisting executive interference with the agency through the Department of Government Efficiency (DOGE).

Most GAO policy recommendations constitute a form of ex post review. An agency implements a new policy program, or continues to implement an old program, and GAO then makes recommendations regarding the implementation of said program once it is in place. The agency can choose to implement or ignore these recommendations. If ignored, the recommendation stays "open" and listed publicly by GAO. These open recommendations can stay active for decades, the oldest open recommendation as of March, 2024 was released in the year 2000, and recommends changes to U.S. Navy accounting practices (Commons 2000).

GAO makes policy recommendations for executive policymaking both *indirectly* via communications with various Congressional entities and *directly* by communicating independently with executive agencies. These communications can be public or private, and reports can begin with Congressional request or by GAO's own volition. Figure 1.2 summarizes the pathways for GAO recommendations into the executive policymaking process.

FIGURE 1.2. GAO Policy Advice and Information in Executive Policymaking



Changes in GAO's Role over Time The agency's mission has evolved from its earlier history in the first half of the 20th century. Before and during the Second World War, the agency employed thousands of auditors to oversee federal expenditures (see Figure 1.1). Following this period, the agency drew down its staff, corresponding with end of the war, stabilizing its workforce to 4-6,000 full time employees (Reynolds 2023). Like the other support agencies, the GAO is facially nonpartisan and unlike Congressional committees, has no partisan division of staff. This relative neutrality makes GAO an important information reservoir for members of Congress, who are faced with an increasingly polarized information environment. Indeed, Congressional staffers perceive information from GAO, and the other support agencies, to be more trustworthy than information from committee staff, party leadership, the administration, etc. (LaPira, Drutman, and Kosar 2020).

During the 1970s and 80s, the role of the agency expanded beyond that of a simple audit agency. GAO began to engage in more expansive policy and program analyses of executive programs, expanding past simple accounting and auditing (Bowsher 2021). This expansion of GAO's oversight role coincided with the expanding size and scope of the federal domestic policy programs in the post-war period. Alongside this expansion, GAO started to issue public recommendations for executive

action. These recommendations propose changes to the implementation of federal programs, often with a focus on reducing the program's cost. The agency also began to publicly issue "matters" for Congressional attention, statutory policy recommendations for Congressional consideration.

The mid-century period of expansion and stability in GAO's operations came to a close in the early 1990s around 1992. Even before the Republican Revolution, GAO faced increasing bipartisan scrutiny, particularly in the Senate (Rubin 2002). This scrutiny came as a result of GAO's expanded role as a policy shop, which put the agency's occasionally controversial reports and policy recommendations into the limelight. For example, GAO published a widely cited analysis regarding the Savings and Loan Crisis (Bowsher 2021). Moreover, some members of Congress found GAO to be overly sclerotic and slow in responding to Congressional requests, and there was a broader concern among some members about agency drift (i.e., GAO's outputs and role had moved beyond Congress' intentions). During the 80s and 90s, much of GAO's published analysis was self-directed as opposed to pursuant to Congressional request. Beyond efficiency, another Congressional complaint against the support agencies was the perception of partisan bias.

In 1992, a provision of a Senate appropriations bill attempted to mandate GAO staff work only on assignments that stemmed directly from Congressional requests; this provision was removed in the House (Rubin 2002). Critics of the agency did successfully cut GAO's staffing by roughly 10% from 1992-1994 (Reynolds 2023). The early resource reductions at GAO led the organization to layoff some junior staff, particularly administrative staff, and freeze hiring for empty positions (Commons 2000). This led to a "top-heavy" staffing regime, but most of the organization's senior policy staff were shielded.

GAO during the Republican Revolution Following 1994 wave election, Republicans regained majority positions in both chambers for the first time in 40 years, and enacted sweeping changes across the legislative branch. Among other things, they abruptly slashed GAO's staff by 20% over a single fiscal year, and another 11% the next fiscal year. The new Republican Chairman of the Senate committee on Government Affairs said about GAO in March 1995: "How can GAO be made more efficient? How can they emulate the private sector to do more with less?"[2] This second round of cuts ate further into GAO's junior ranks, but this time senior staff were not fully

[2]Quote retrieved from committee hearing notes available via ProQuest Congressional, hearing ID: HRG-1995-SGA-0038.

insulated. GAO was also forced to close several of its field offices around the United States and internationally, which it did to protect the headquarters office to the fullest extent possible. According to interviews with GAO staff from the period, cuts resulted in a reduction in the quality of the agency's published material, and a decline in primary source audits and data collection. Rather than sending agents to do independent data collection regarding the implementation of a federal program out in the field, GAO would instead rely on secondary data sources when possible (Rubin 2002).

Did GAO Bureaucrats See The 1994 Republican Wave Coming?

Is it fair to treat the Republican wave in the 1994 election as quasi-exogenous? Scholars writing at the time certainly thought so. For example, Ornstein and Schenkenberg (1995) writes:"...no one would have expected the massive Republican takeover of both houses of Congress in the 1994 election. Indeed, politicians, pundits, and strategists on both sides of the political spectrum were caught off guard. In the Washington Post's final election predictions, only three of the fourteen analysts surveyed picked the Republicans to win both Houses, and not one picked them to muster the full 230 seats in the House that they achieved." Even forecasters who did correctly predict the surprise 1994 election outcome, for example Alesina, Londregan, and Rosenthal (1993, 1996), admit: "The last two American elections, the presidential election of 1992 and the mid-term congressional election of 1994 have been widely regarded as surprises, relative to forecasts."

Importantly, what is necessary for identification is not that *no one* saw a Republican wave coming, but rather, that GAO officials had no special ability to foresee it. If these bureaucrats did predict a Republican wave, they may have attempted taken steps to strategically adjust agency outputs in advance of the election. Of course, as with all identifying assumptions, this exogeneity assumption is not directly testable, but there is a lack of quantitative or qualitative evidence to the contrary. Informally, one staffer present in the legislative support agencies during this period suggests reports published in the lame-duck period of 1994 were unaltered and largely drafted prior to the election result. Given this, I set the threshold cutpoint for the analysis in the following sections as January 4, 1995 at the start of the new Congress following the Republican wave.

This assumption is only plausible for aggregate outcome measures that GAO did not report to Congress during this period. GAO's incentive, and ability, to distort aggregate metrics makes

any figure it was required to report to Congress, such as total reports or total testimonies (or the aggregate of both), suspect. On the other hand, this assumption is much more plausible for the number of executive policy recommendations made by GAO year to year. This statistic was not tallied for the Congress, and any attempt to do so would have been extremely tedious and costly at the time. Moreover, a central complaint of the Republican members of Congress against GAO was that it had strayed too far into making controversial, public recommendations. Given these qualitative facts, it seems unlikely that GAO had any clear incentive to manipulate the count of policy recommendations in the same manner that it manipulated total reports. If anything, GAO may have reduced public recommendations for executive action to avoid scrutiny over its recommendations. It is worth mentioning that GAO did begin to report recommendation counts in FY2001, following an overhaul of its annual reporting (Walker 2000). This fiscal year rests outside the bandwidth over which treatment effects are calculated in the following steps.

Why Attack Legislative Capacity? Partisan Competition, Theories, Ideological (Self?) Sabotage, and Fenno's Paradox

Reliance on Legislative Subsidies: Where does the Budget Constraint Come From? Formal models of lobbying often assume legislators have a budget constraint to acquire informational and policymaking capacity (e.g., learning about a policy area and crafting legislation in that area) (Hall and Deardorff 2006). While this is typically a reasonable assumption, it is not obvious why there is a resource constraint on legislative capacity in budgetary terms in the United States. Spending on legislative branch appropriations has not consumed a significant portion of the federal budget in the modern era. Moreover, it is not as if spending on capacity somehow comes at the legislator's personal financial expense, they are merely appropriating taxpayer dollars or borrowed funds.

So why doesn't Congress appropriate itself more resources to craft policy, develop expertise, and collect policy information? I argue, following Fong, Lowande, and Rauh (2025), the true constraint legislators face isn't fiscal, but rather political. Outside interest groups often provide substitutes for Congressional staff and expertise, and populist policies to "rein in" the legislative branch are generally popular in the mass public. This public opposition to increasing spending on Congress itself, or otherwise expanding capacity, makes legislative subsidies from lobbying groups tempting

to members of Congress. These members can receive informational and policy development subsidies from like-minded lobbyists without expending political capital to increase Congress' internal capacity.

Partisan Competition or Ideological Sabotage? Crosson et al. (2021) suggest two divergent theories for the declines in Congressional capacity, declines that begin in the Gingrich era. One explanation is "ideological sabotage." This straightforward theory posits that Republicans hold an ideological commitment to reducing the size and scope of government generally, and thus, attempts to reduce the size of the legislative branch can be viewed merely as a component part of this broader ideological agenda. Reducing legislative capacity in this theory serves both as a signal to the electorate that members are serious about shrinking government and also as a policy designed to hobble the Congress' ability to expand existing programs or create new ones due to a simple lack of policy expertise. This perspective focuses on the role of the ascendant Republican Revolutionaries with little role for Congressional Democrats. Per Crosson et al. 2021, "Similar rhetoric is unheard of among Democrats, and certainly cannot be attributed to them, since the Contract was a Republican initiative." [3]

While this explanation has intuitive appeal, several conceptual problems are apparent. First, while Republicans were the primary proponents for changes to the legislative state, many of the changes were enacted on a bipartisan basis. For example, nearly half of the Democratic caucus in the House, as well as Democratic President Clinton, supported the Legislative Branch Appropriations Act of FY 1995 that eliminated the Office of Technology Assessment. The same measure passed easily in the Senate with a voice vote and no objections. Moreover, prior to the RR, a Democratically controlled Congress had already authorized a smaller set of changes to the legislative branch that included a reduction in funding for the legislative support agencies.

Second, as is discussed above, it is not clear that reducing legislative capacity contributes to the project of reducing the size and scope of government. Most obviously, the legislative branch is a vanishingly small fraction of federal spending. What is indeed more likely is that capacity reductions alter the nature of enacted policy, with an increasing share originally generated by outside interest groups relative to in-house staff (McGee 2022; Fong, Lowande, and Rauh 2025).

[3]Moreover, several famous, elite conservatives such as Milton Friedman championed the "citizen legislator" model in the mid-20th century, providing an intellectual underpinnings for this realized worldview.

This relative substitution of expertise has no obvious connection to reducing federal spending, or the scope of federal programs, but instead, changes the process by which original legislation is crafted and what policy is contained therein.

On the other hand, regardless of the substantive argument for reducing legislative capacity, it's possible Republicans pursued cuts to the legislative state in order to signal a commitment to small government philosophy to voters (Crosson et al. 2021). This somewhat cynical flavor of the ideological sabotage theory posits: "members of the Republican Party viewed the costs in legislative matters to be far below the electoral advantage gained from cultivating an image of fiscal frugality (Crosson et al. 2021)."

A second explanation is that increases in partisan competition and the centralization of the policymaking process drove reductions in capacity. As majority margins narrowed and chamber control flipped more frequently, the expected payoff to individual policy entrepreneurship fell while the value of party-coordinated messaging and agenda control rose. Members, in this account, rationally shifted staff toward communications and constituency service, and away from policy development, as leaders increasingly wrote key bills and constrained amendment opportunities. This theory also explains Democratic capitulation to Republican capacity proposals. When chamber control is uncertain, both parties anticipate being out of power soon. Symmetric capacity reductions are a form of institutional "mutual disarmament": each side limits the other's ability to develop and oversee policy when it is in the majority, and members on both sides reallocate their scarce attention to matters that can increase their electoral prospects.

Explaining GAO's Response to the Republican Revolution Regardless of why Republicans chose to go after the legislative state generally and the GAO specifically, how did the GAO shift its oversight output in response to the change in majority and subsequent spending cuts? In short, they put forward an image to Congress that it could do more with less. In their annual reports, officials reported improving performance statistics during this period. These claimed increases in the number of GAO reports suggest deep cuts to the agency somehow improved its performance. Can it really be true the agency was doing more, and higher quality, oversight work in the face of steep cuts and elevated scrutiny?

Qualitative evidence suggests the answer is no (Rubin 2002). Agency officials instead employed a variety of mechanisms to maintain, or elevate, performance statistics during this period. These

mechanisms include: splitting large reports into multiple components to be released separately; strategically timing report releases around aggregation cutoffs; shortening reports; cannibalizing language from prior material; reducing the internal review period for each report; relying on secondary rather than primary data sources, among other mechanisms. The agency had strong incentives to avoid the *appearance* of quality reductions, and took whatever actions necessary to shield reductions from direct Congressional view. This is summarized nicely by Rubin (2002), "The GAO was unable to document reductions in the quality or quantity of its services, because such declines might be used as an excuse for further cuts or even agency termination."

The motivation for strategic manipulation of the timing and quantity of GAO products by agency officials is straightforward. Staff had clear incentives to preserve their jobs and existing authority to the extent possible in the face of a hostile Congress. These incentives were no doubt foremost in the minds of support agency leadership and staff following the total dissolution of the Office of Technology Assessment in FY1995. Following the dissolution of OTA, GAO faced a credible threat of disbandment from Congressional Republicans and attempted to adapt to avoid this worst case outcome. This suggests two general mechanisms: avoidance and acquiescence. First, GAO officials took actions, where possible, to avoid potentially hostile Congressional attention. As such, we should expect GAO to avoid issuing reports that might "make a splash" in the news media or on the hill. We should also expect GAO reports to contain less specific policy recommendations, for the executive branch or for Congress, instead favoring ambiguous language that is more difficult to scrutinize. Second, GAO officials took steps to acquiesce to the demands of Congressional Republicans by focusing on substantive policy topics favored by Republican chairman and rank-and-file members. In this chapter, I test the first mechanism, in chapter two I explore the second mechanism.

Importantly, the qualitative literature suggests GAO manipulated performance statistics *that were reported to Congress*. There is no evidence that other measures of GAO's productivity that neither the agency, nor Congress, compiled during the period were manipulated by the agency. This important nuance motivates the measurement of oversight productivity using metrics that were not public reported in the period of study. To examine GAO's response, I quantify the agency's outputs in terms of both seen and unseen statistics. Most simply, I measure the simple count of GAO reports, a metric that was reported to Congress. Second, I measure the number of

recommendations for executive policymaking that GAO generated over the entire period, a metric not tabulated for Congress. Third, I measure the number of matters for Congress GAO produced over the period, another metric not tabulated for Congress at the time. For recommendations and matters I operationalize these concepts both as raw counts and also as a share of reports containing these items.

My expectations are summarized in Table 1.1.

TABLE 1.1. Summary of Productivity Measures & Expectations

DV Measure	**Description**	**Report to Congress?**	**Expectation**
Total Reports	All unique GAO reports.	Yes	No Effect
Total Recommendations	All unique GAO recommendations for Executive Policymaking.	No	Decrease ↓
Recommendation Share	Share of Reports with recommendations.	No	Decrease ↓
Total Matters	All unique GAO matters for Congress.	No	Decrease ↓
Matters Share	Share of Reports with matters.	No	Decrease ↓

Data and Methods

Dataset The dataset employed in this paper is the universe of published, and publicly available,[4] GAO documents spanning back to the creation of the agency. To assemble this dataset, I built a web-scraping and feature extraction package[5] that iteratively accesses and downloads every unique content page through the primary source, GAO's online report database. The dataset contains all unique page entries as well as each full report, approximately 55,000 entries spanning from 1922-2022. Figure 1.3 displays the raw counts of all public entries by fiscal year. It is important to note, this figure aggregates over all document types (e.g., testimonies, reports without recommendations, reports with recommendations, etc.). Additionally, public report entries are relatively sparse prior to the 1970s. Given this sparsity, all analyses trim the pre-1980 period out of an abundance of caution that only years with complete report entries (i.e., all actual reports are included)

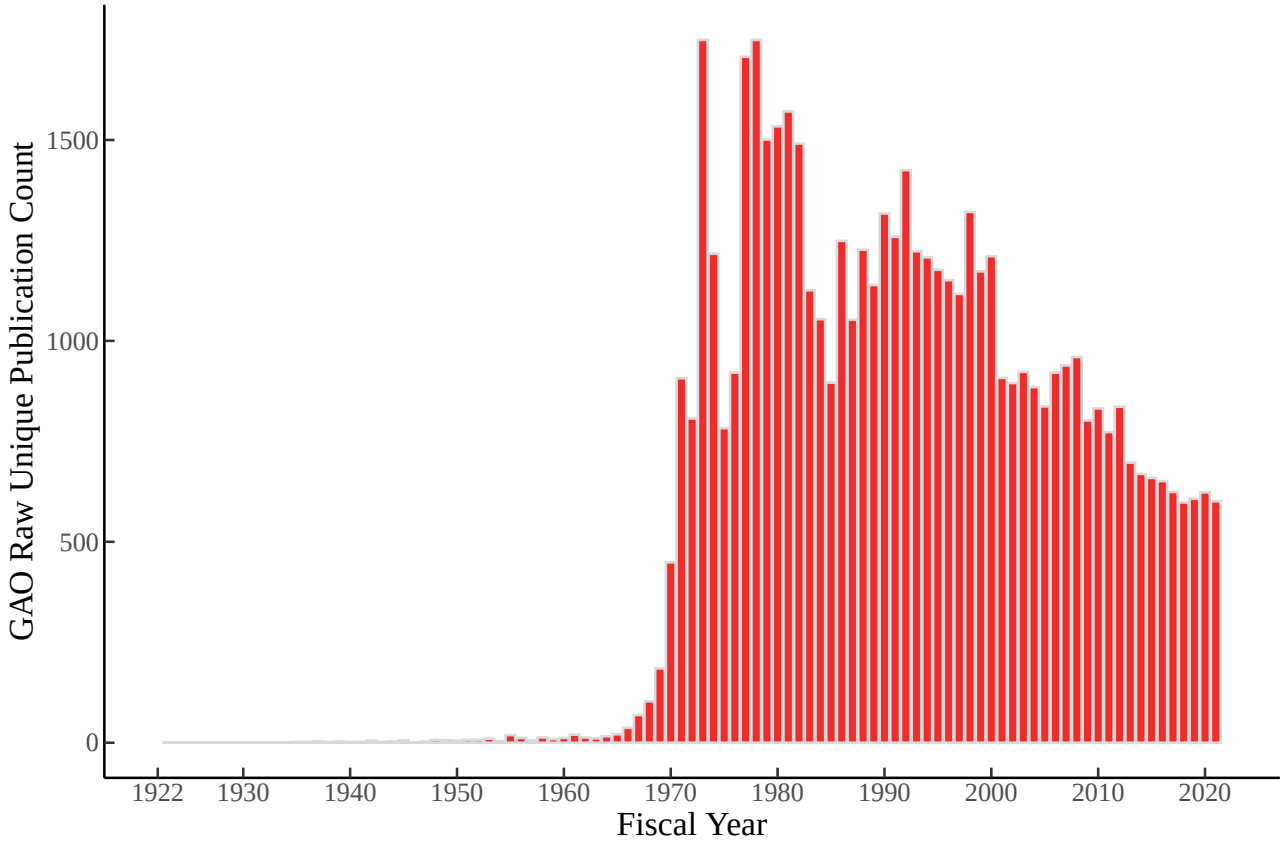
[4]Some subset of reports contain classified information and are not available to the public. A full listing of these reports for fiscal years 2015-2023 is available in the public archive here. Given the limited information available on these reports, and their small number, they are omitted from this analysis.

[5]The `R` package, `gao`, capable of generating an up-to-date version of the dataset employed here will be made available upon publication on Github here (not currently public).

are considered. Moreover, even with this aggressive manual trimming, none of the data prior to approximately 1985 is selected in the automatic bandwidth procedure used to calculate main effects throughout the rest of the chapter.

As such, the covered period in the following sections is 1980 through 2022, and the total entry count is trimmed to approximately 43,000. More generally, there are two broad categories of entries: reports and testimonies. I focus mainly on reports, as Congressional testimonies fluctuate near exclusively based on direct Congressional requests for testimony (i.e., agency officials do not propose to testify, they must attend). Within reports, some GAO products contain explicit recommendations for agency action (e.g., the Department of Defense should audit the F-35 program), while others make no explicit recommendation and are merely descriptive (e.g., the Small Business Association is implementing a new anti-fraud program). Of my sample, about a third (14,000) contain explicit recommendations and the remaining two-thirds (29,000) do not.

FIGURE 1.3. Unique Count of GAO Reports and Testimonies by Fiscal Year



While the `gao` package retrieves both full report files (.pdf) and webpages (.html), the remaining analysis relies exclusively on the webpage data. While the report files obviously contain the full raw text of a given GAO report, the web page entries contain a wealth of hand-coded metadata tags not contained in the full reports, including report topics (e.g., national defense), relevant executive agencies, the current status of GAO's recommendations if any are made in the report (i.e., implemented by the agency, ignored, etc.), and more. These tags are summarized in Figure 1.2 below.

TABLE 1.2. Feature Sets in GAO Webpages vs. Reports

	Webpage (html)	**Report (pdf)**
Report Title	✓	✓
Report Summary	✓	✓
Report Number	✓	✓
Initial Publication Date	✓	✓
Public Release Date	✓	
Recommendations List	✓	✓
Recommendations Status	✓	
Executive Agency Affected	✓	✓
Report Topics	✓	

The webpage data allows us to easily construct several measures of GAO's productive outputs that the report data alone would miss, for example, the number of GAO recommendations for executive policymaking. These GAO reports with explicit policy recommendations are of particular interest, as GAO includes a hand-coded, and regularly updated table, on each report page listing the recommendations and their current status. For example, see Figure 1.4 below. These recommendations for executive action are counted per report in my dataset. GAO also gives policy advice to Congress in what it calls "matters" for Congressional attention. These policy prescriptions for Congressional action are less prevalent relative to recommendations for executive action, roughly 10-20% of total GAO recommendations provided in a given fiscal year are "matters". The topic of each report is assigned by GAO's internal typology. This typography contains 31 unique categories spanning substantive areas ranging widely from defense to environment. A majority of reports are coded as single topic (53%), another third contain two topics (31%), most of the rest contain 3-4 topics (14%) and a very small number contain either no listed topic (less than 1%) or more than 4

topics (1.6%). These topic codings, as well as other extracted unique content labels, are employed to interrogate the possibility of heterogeneous treatment effects in the results section.[6]

FIGURE 1.4. Example GAO Recommendation Table

Recommendations for Executive Action

Agency Affected	Recommendation	Status
Department of Justice	The Attorney General should ensure that the Assistant Attorney General of the National Security Division and the Director of the Federal Bureau of Investigation, in consultation with the Secretaries of Homeland Security and State, take additional steps to enhance understanding of transnational repression among federal agencies and state and local law enforcement agencies, such as by establishing a formal interagency definition of transnational repression or conducting additional training. (Recommendation 1)	Open When we confirm what actions the agency has taken in response to this recommendation, we will provide updated information.
Department of Justice	The Attorney General should develop and draft a coordinated, department-wide position on any identified gaps in current legislation for addressing transnational repression and, if appropriate, submit a legislative proposal to the Office of Management and Budget in accordance with OMB Circular A-19. (Recommendation 2)	Open When we confirm what actions the agency has taken in response to this recommendation, we will provide updated information.
Department of State	The Secretary of State should ensure that the Assistant Secretaries for Political-Military Affairs and Democracy, Human Rights, and Labor include steps to coordinate with and collect information on transnational repression from other agencies, such as the Departments of Justice and Homeland Security, in its new transnational repression-related tools and guidance to align the arms transfer decisions-making process with the 2023 Conventional Arms Transfer Policy. (Recommendation 3)	Open When we confirm what actions the agency has taken in response to this recommendation, we will provide updated information.

Estimation Procedure To analyze the effect of the Republican Revolution (or RR) on GAO's productive outputs, I employ a Regression Discontinuity in Time (RDiT) design, where the unexpected result of the 1994 election and subsequent party turnout in both the House and Senate is the source of exogenous variation underlying the strategy.

Formally, we have a binary treatment assignment mechanism D_i, a running variable time (daily) X_i, and a cutpoint c where:

[6]See the proceeding Chapter for the investigation of heterogeneous treatment effects.

$$D_i = \begin{cases} D_i = 1 & \text{if } X_i > c \\ D_i = 0 & \text{if } X_i < c \end{cases} \tag{1.1}$$

In this case, D_i corresponds to whether or not the count-day pair (e.g., 10 recommendations-September 1, 1990) is before or after the start of the 104th Congress. Our cutpoint (c) is the beginning of the 104th Congress election. We are interested in estimating the effects of the RR on GAO's various outputs, Y_i. Our estimand of interest is the local average treatment effect at the threshold, τ_{TD} defined as:

$$\tau_{TD} = \lim_{x \to c^+} \mathbb{E}\left[Y_{1i} \mid X_i = x\right] - \lim_{x \to c^-} \mathbb{E}\left[Y_{0i} \mid X_i = x\right] \tag{1.2}$$

In words, we are interested in the difference in expectations for Y_i around the threshold. The crucial identifying assumption is continuity at the threshold. Continuity in this setting implies there is no time-varying, confounding variable(s) Z that influences both GAO productivity and treatment assignment.

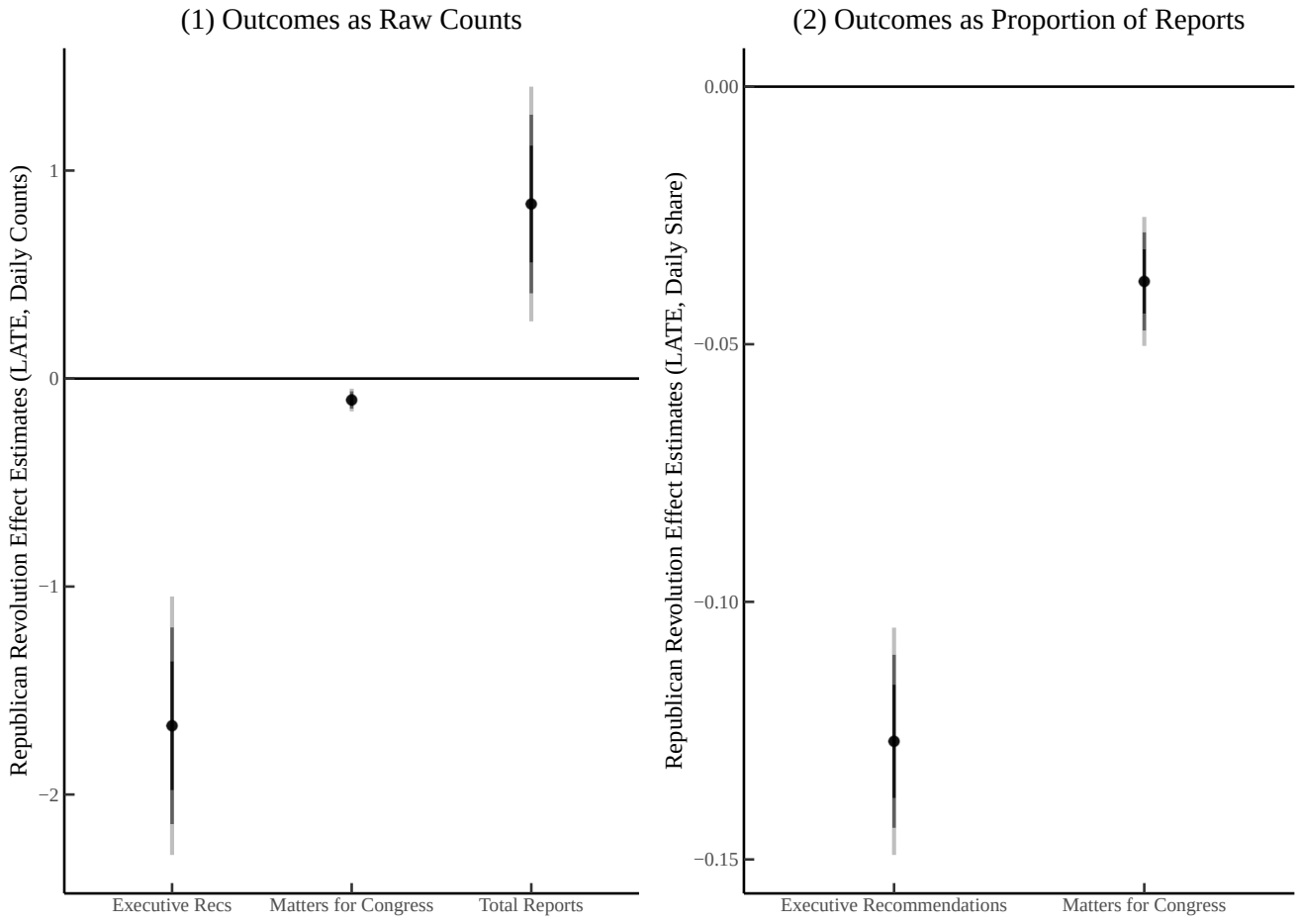
To estimate $\hat{\tau_{TD}}$, I follow Calonico, Cattaneo, and Titiunik (2014)'s estimation procedure. First, the procedure selects an optimal bandwidth using a standard algorithm (Calonico, Cattaneo, and Titiunik 2014).[7] Then, I estimate separate local linear regression models on either side of the cutoff using triangular kernels for density estimation following the recommendation of Lee and Lemieux (2010). A triangular kernel increases the weight of observations close to the threshold, and linearly decreases the weight of observations further from the threshold reaching zero weight at the bandwidth cutoff. To incorporate possible serial correlation between days of report production, I cluster standard errors by month. Finally, I estimate $\hat{\tau_{TD}}$ by taking the simple difference in the fitted local linear regression predictions at the threshold. I repeat the process across all dependent variable measures. For credible inference, this procedure follows the default settings recommended by Calonico, Cattaneo, and Titiunik (2014), the most popular and widely cited method for estimating RDD effects.

[7] I also run a sensitivity analysis over bandwidth in the Appendix.

Results

Main Effects To begin with, I present the main RDD results across both simple count and share dependent variable measures in Figure 1.5. For all results, I present the bias-corrected estimates following Calonico, Cattaneo, and Titiunik (2014). [8]

FIGURE 1.5. Republican Revolution Effects: Reports, Recommendations, Matters



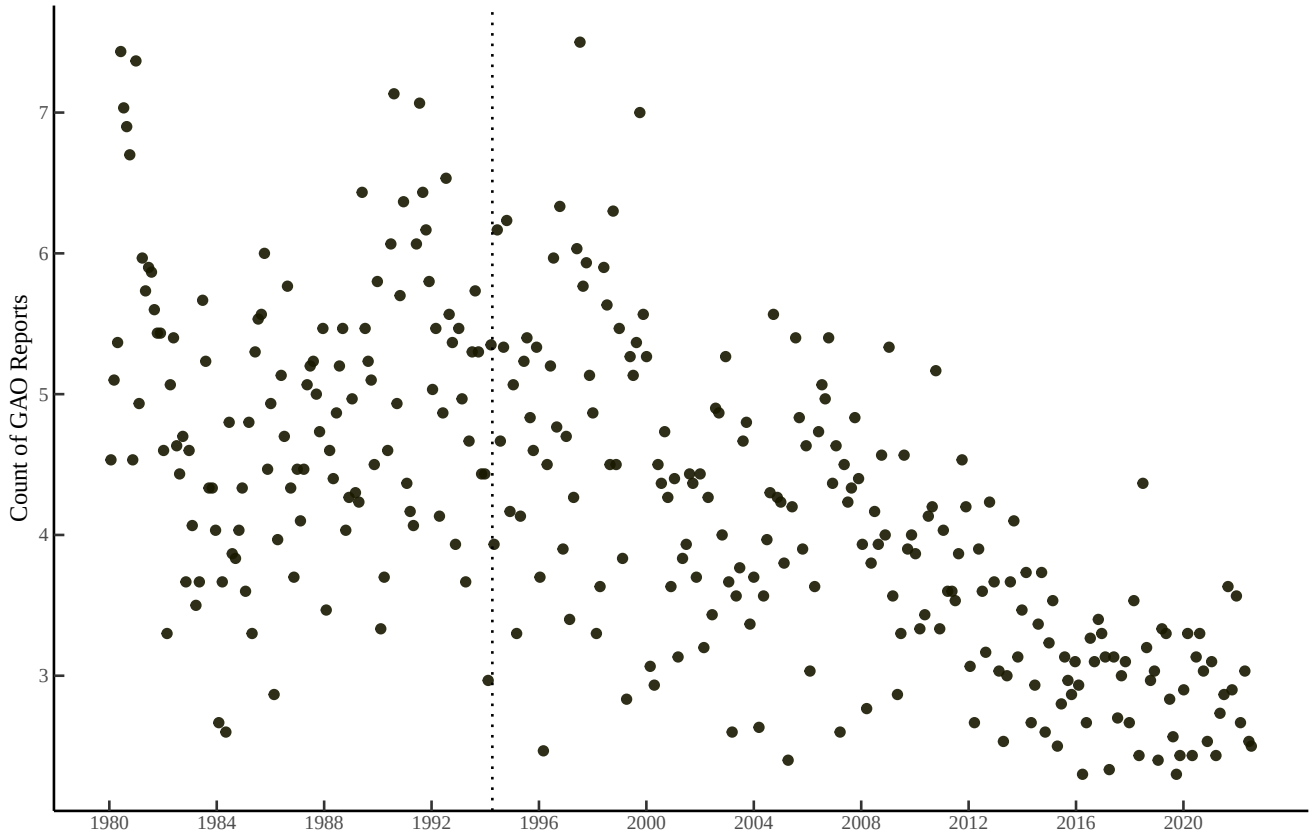
Note: Black points represents bias-corrected, local average treatment effect point estimates from the RDD models with corresponding 80%, 95%, and 99% confidence intervals. Estimates produced using default settings using `rdrobust`.

Beginning with total report volume, there is an estimated increase in raw report counts of .84 [.43,1.2] reports per day around the threshold. Extrapolating, this yields approximately extra

[8] Results following a conventional RDD procedure, as well as a procedure that adds an additional robustness step, are available in the Appendix. Across all main specifications the results across the three procedures are virtually substantively identical with only minor numerical differences in the estimates.

GAO 25 reports over a 30 day period, or 300 extra annually. This result, however, is not robust to reasonable alternative model specifications (see Appendix). Depending on the choice of bandwidth, polynomial order, and kernel weight the RR effect on total report volume is estimated to be positive, null, or even negative. Raw report volume over time is displayed in Figure 1.6 as well.[9]

FIGURE 1.6. GAO 30-Day Average Report Counts over Time



Note: Black points represent average report counts within 30-day bins. The black dotted line is the threshold, January 4, 1995 when the new Republican majority Congress was seated. Figures generated using a modified function from `RDHonest` (Armstrong and Kolesár 2020).

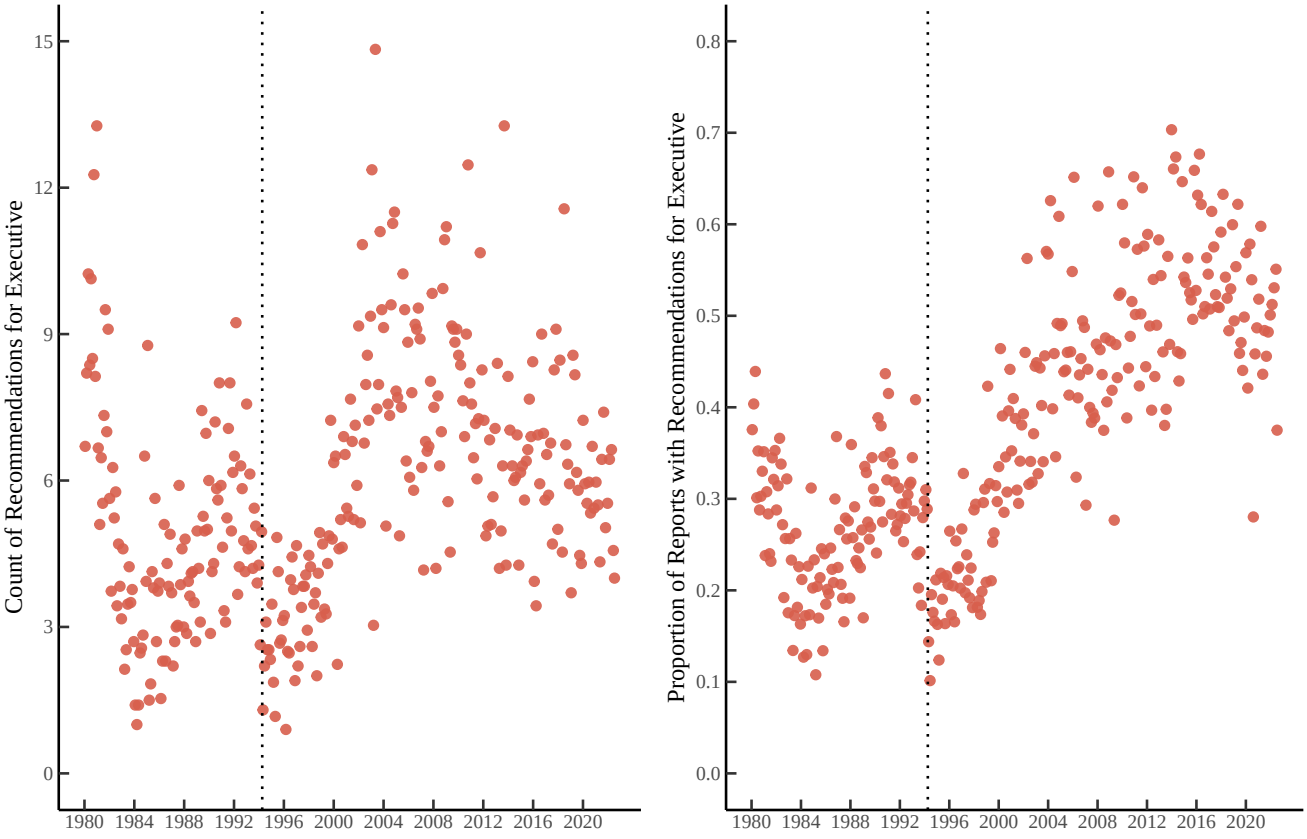
Interestingly, raw report volume secularly declines in the more recent period following roughly FY2001, but no clear discontinuity around the threshold is present. This secular decline continues through the rest of the observed period. Perhaps coincidentally, FY2001 is the same year GAO

[9]Figures 1.6,1.7, and 1.8 display the raw data used to calculate the RR effects over the time range employed in the analysis. To avoid over-interpretation, these simple figures omit both the fitted local linear regression lines as well as corresponding intervals following the recommendation of both Calonico, Cattaneo, and Titiunik (2014) and Armstrong and Kolesár (2020). The pointwise confidence intervals sometimes displayed in plots of this variety are not employed for inference and can give readers a false sense for the nature of the results and estimation procedure.

stopped reporting total report counts in their annual reports, switching to other performance metrics.

Next, for recommendations for executive policymaking, the RR caused an estimated decrease in recommendation count of -1.7 [-2.1,-1.3] per day. Extrapolating over 30 days, this yields a decrease of roughly 51 recommendations, or 612 less annually. In terms of the share of reports that contained recommendations for the executive, the RR caused an estimated 12.7% [-14.3%,-11.2%] reduction in the share of reports that contained any recommendation. The raw data underlying the RDD models is displayed in Figure 1.7. Results for recommendation counts are strongly robust to all alternative model specifications (360 unique combinations of kernel weights, bandwidth, and polynomial order, see Appendix). Results for the share of reports with recommendations are nearly as robust, with exclusively negative point estimates for the RR effect, and confidence bands that mostly do not overlap zero.

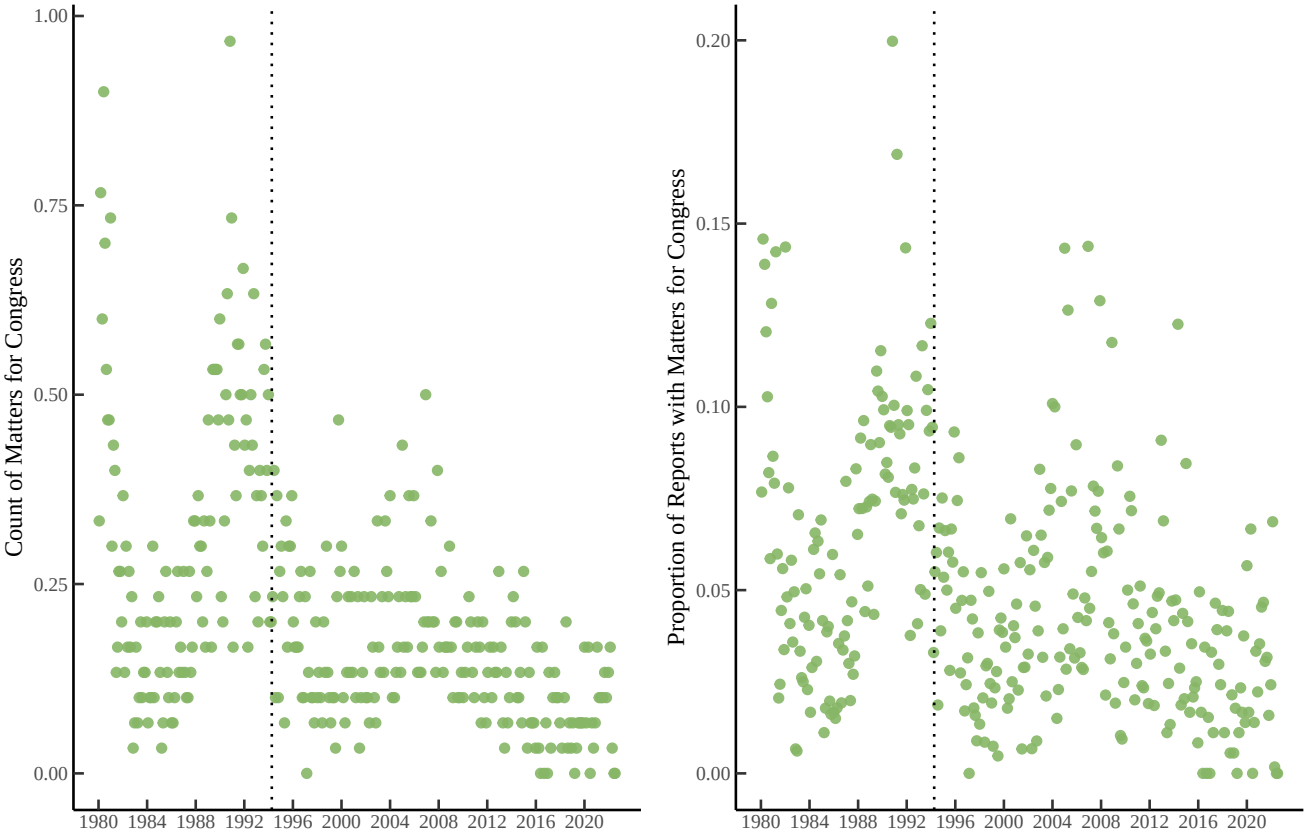
FIGURE 1.7. GAO 30-Day Average Recommendations for the Executive over Time



Note: Points represent average recommendation count and average recommendation share of total reports within 30-day bins. The black dotted line is the threshold, January 4, 1995 when the new Republican majority Congress was seated. Figures generated using a modified function from `RDHonest` (Armstrong and Kolesár 2020).

Finally, for matters for Congress, the RR caused an estimated decrease in matters count of -.10 [-.141,-.066] per day. Extrapolating over 30 days yields approximately three less matters for Congress, or roughly 36 less annually. For the share of reports that contained matters for Congress, the estimated RR effect is -3.8% [-4.7%,-2.9%]. These results suggest GAO restricted matters for Congress in a similar fashion to executive recommendations, however for Congressional matters the effects appear to have lingered permanently. Figure 1.8 shows the raw data used in these calculations, where we observe another secular decline in the later periods of the time series.

FIGURE 1.8. GAO 30-Day Average Matters for Congress over Time



Note: Points represent average matters for Congress count and average matters share of total reports within 30-day bins. The black dotted line is the threshold, January 4, 1995 when the new Republican majority Congress was seated. Figures generated using a modified function from `RDHonest` (Armstrong and Kolesár 2020).

The results for matters for Congress, both in terms of counts and shares of reports, are sensitive to alternative model specifications, in particular the choice of bandwidth. Under alternative bandwidths the sign of the estimated RR effect generally changes from positive to null or negative.

Discussion

Taken together, these results suggest the Republican Revolution chilled GAO's production of public recommendations for executive policymaking, and likely also matters for Congressional consideration. Moreover, I confirm earlier qualitative work that claims GAO strategically manipulated its aggregate statistics to avoid further scrutiny in the post-'94 period. The differential trends across these three measures of productivity around the threshold of interest are explainable in part by

variation over which statistics were reported to Congress. GAO's clear incentive to avoid making controversial recommendations in public following the RR is also likely a factor underlying the difference in these trends.

These results beg the question: why would Congressional Republicans enact cuts to GAO in the first place? GAO is an agent of Congress, and one of its most significant tools to oversee the executive branch. Many Republicans rightly or wrongly viewed GAO as an institution biased towards Democratic members of Congress, however, it was certainly possible to redirect the agency without reducing its overall capacity as an oversight body. One can imagine several possible explanations. In the view of GAO's director at the time, Chuck Bowsher, Republicans faced a trade-off: to pursue cuts at GAO, or somewhere else in the legislative branch, or fail to follow through on their promise to reduce the size of Congress (Bowsher 2021). In this view, while cuts to GAO were substantively meritless, the alternative would be to further cut committee or personal staff for members, or renege on the Contract with America promise altogether. By extensively cutting GAO, Republicans were able to successfully fulfill their campaign promise while avoiding even deeper cuts to partisan legislative staff. Taken together, this would suggest the political value of enacting cuts to the legislative branch outweighed the harm to Congressional capacity in the eyes of Republicans at the time. An added possibility is GAO's reaction to increased scrutiny—an overall decrease in the volume of recommendations for the executive branch—was an unintended consequence of efforts by Republicans to improve efficiency in the legislative branch. In other words, Republican members may have believed GAO could truly do "more with less."

So what portion of the estimated effects are attributable to majority party turnover relative to resource reductions? These data are unable to fully address this question, as both events occur near simultaneously. On one hand, qualitative evidence from the period is suggestive that party turnover is the major factor. The resource reductions at GAO also began in the pre-trend period, and no significant reduction in recommendation volume is noticeable in FY93 or FY94. It is nonetheless possible that the sharper resource reductions that occurred post-'94 did reduce the agency's "real" capacity to produce the same volume of public recommendations for executive policy. It is also possible the resource reductions had nonlinear effects, increasing in magnitude following FY94.

Additionally, these results cannot untangle the extent to which the observed reduction in GAO's public recommendations for executive action, or matters for Congressional consideration, were replaced/substituted with informal, private communications. Anecdotal reports suggest GAO transmission of private, informal policy advice may have increased following the Republican Revolution (Rubin 2002), but this suggestion is unfortunately not quantitatively verifiable.

Conclusion

Can Congress wrangle the administrative state and exert political control over agencies' actions? Scholars have long debated the mechanisms and effectiveness of Congressional oversight, often reaching divergent conclusions (McCubbins and Schwartz 1984; Potter 2019). This chapter focuses on an understudied tool of centralized, "police patrol" oversight through the Government Accountability Office. The GAO is the largest single Congressional oversight body, and the only that employs a nonpartisan staffing regime. I detail overlooked pathways for GAO's recommendations to enter the executive policymaking process, both through the Congress and directly to executive agencies. Moreover, I examine the agency's response to severe scrutiny and budget cuts during the Republican Revolution of 1994.

With a novel dataset, I quantify the volume of GAO's productive outputs using three different measures. With these data, I implement a regression discontinuity in time design that exploits the 1994 election as a source of exogenous variation. In so doing, I am able to estimate the effects of the Republican Revolution on GAO's productivity. I find strong evidence the Republican Revolution and corresponding changes at GAO chilled the agency's willingness to make public recommendations for executive and Congressional action. This chilling effect significantly reduced the volume of GAO's recommendations to the executive branch in latter half of the 1990s. These findings should be taken in the context of possible substitution effects; GAO officials may have increased informal, private communication of recommendations for executive action during this period. This analysis cannot speak to this possibility. It is nonetheless possible that the reductions estimated here occurred without a corresponding change in private communication; if true, Congress effectively hamstrung a crucial mechanism for executive oversight in this period. It is also plausible the chilling effect on GAO had lasting cultural repercussions for the agency, possibly reducing officials willingness to engage on controversial policy matters in the future.

Beyond the empirical results, this chapter illuminates a clear trade-off in legislative oversight, whether of the executive branch, or of Congress' own support agencies. Attempts to gather information about an agency, or force it to disclose information, can result in obfuscation where the agency replies with false information. This finding adds to the existing theoretical literature on the usefulness of ex post review (Patty and Turner 2021). Practically, this difficulty casts a shadow on simple count measures of agency priority or productivity, particularly measures that are reported to an agency's principal. Unfortunately this dynamic complicates efforts to oversee bureaucratic agencies as well as scholarly efforts to interrogate and quantify agency production.

Finally, this chapter contributes to a growing literature that details the complex web of oversight institutions and mechanisms employed by Congress beyond simple oversight hearings. This multifaceted regime provides avenues for Congressional oversight, informal and formal, public and private, police patrol and fire alarm, suited to the heterogeneous needs and desires of members. Whether or not this regime prevents policy drift in the administrative state remains to be seen.

Bibliography

Alesina, Alberto, John Londregan, and Howard Rosenthal. 1993. "A model of the political economy of the United States." *American Political Science Review* 87 (1): 12–33.

———. 1996. "The 1992, 1994 and 1996 elections: A comment and a forecast." *Public Choice* 88 (1): 115–125.

Armstrong, Timothy B, and Michal Kolesár. 2020. "Simple and honest confidence intervals in nonparametric regression." *Quantitative Economics* 11 (1): 1–39.

Ban, Pamela, and Seth J Hill. 2023. "Efficacy of Congressional Oversight." *Working Paper.*

Bawden, Allison. 2023. "Nuclear Security: DOE Should Take Actions to Fully Implement Insider Threat Program," https://www.gao.gov/products/gao-23-105576.

Bimber, Bruce Allen, et al. 1996. *The politics of expertise in Congress: The rise and fall of the Office of Technology Assessment.* SUNY Press.

Bowsher, Charles. 2021. "Interview with Chuck Bowsher, Comptroller General of the United States, 1981-1996." https://www.gao.gov/assets/2021-12/Bowsher_transcript_0.txt.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82 (6): 2295–2326.

Clarke, Andrew J. 2020. "Congressional capacity and the abolition of legislative service organizations." *Journal of Public Policy* 40 (2): 214–235.

Commons, Gladys J. 2000. "Financial Management: Improvements Needed in the Navy's Reporting of General Fund Inventory," https://www.gao.gov/products/gao-01-37r.

Crosson, Jesse M, Alexander C Furnas, Timothy Lapira, and Casey Burgat. 2021. "Partisan competition and the decline in legislative capacity among congressional offices." *Legislative Studies Quarterly* 46 (3): 745–789.

Dicken, John E. 2024. "Autism Research and Support Services: Federal Interagency Coordination and Monitoring Efforts Could Be Further Strengthened," https://www.gao.gov/products/gao-24-106446.

Fong, Christian, Kenneth Lowande, and Adam Rauh. 2025. "Expertise acquisition in Congress." *American Journal of Political Science* 69 (1): 5–18.

Gailmard, Sean. 2014. "McNollgast's" Administrative Procedures as Instruments of Political Control"." *Oxford Handbook of Classics in Public Policy,* 1–31.

GAO. 2023. "2023 Annual Report: Additional Opportunities to Reduce Fragmentation, Overlap, and Duplication and Achieve Billions of Dollars in Financial Benefits."

Gingrich, Newt. 1994. "A Contract with America." *House of Representatives Archive.*

Hall, Richard L, and Alan V Deardorff. 2006. "Lobbying as legislative subsidy." *American Political Science Review* 100 (1): 69–84.

Kociolek, Kristen. 2023. "COVID-19 Relief: Funding and Spending as of Jan. 31, 2023."

LaPira, Timothy M, Lee Drutman, and Kevin R Kosar. 2020. *Congress Overwhelmed: The Decline in Congressional Capacity and Prospects for Reform.* University of Chicago Press.

Lee, David S, and Thomas Lemieux. 2010. "Regression discontinuity designs in economics." *Journal of economic literature* 48 (2): 281–355.

Lewallen, Jonathan. 2020. *Committees and the Decline of Lawmaking in Congress.* University of Michigan Press.

Lowande, Kenneth. 2018. "Who polices the administrative state?" *American Political Science Review* 112 (4): 874–890.

Lowande, Kenneth, and Rachel Augustine Potter. 2021. "Congressional oversight revisited: Politics and procedure in agency rulemaking." *The Journal of Politics* 83 (1): 401–408.

Lowande, Kenneth, Melinda Ritchie, and Erinn Lauterbach. 2019. "Descriptive and substantive representation in congress: Evidence from 80,000 congressional inquiries." *American Journal of Political Science* 63 (3): 644–659.

McCubbins, Mathew D, and Thomas Schwartz. 1984. "Congressional oversight overlooked: Police patrols versus fire alarms." *American journal of political science,* 165–179.

McGee, Zachary A. 2022. "Problem solving and the demand for expert information in congress." *Legislative Studies Quarterly* 47 (1): 53–77.

Ornstein, Norman J, Thomas E Mann, and Michael J Malbin. 2002. *Vital Statistics on Congress: 2001-2002.* American Enterprise Institute.

Ornstein, Norman J, and Amy L Schenkenberg. 1995. "The 1995 Congress: The first hundred days and beyond." *Political Science Quarterly* 110 (2): 183–206.

Patty, John W, and Ian R Turner. 2021. "Ex Post Review and Expert Policy Making: When Does Oversight Reduce Accountability?" *The journal of politics* 83 (1): 23–39.

Potter, Rachel Augustine. 2019. *Bending the rules: Procedural politicking in the bureaucracy.* University of Chicago Press.

Reynolds, Molly E. 2023. "Vital Statistics."

Rubin, Irene S. 2002. *Balancing the federal budget: Trimming the herds or eating the seed corn?* CQ Press.

Walker, David M. 2000. "GAO Performance and Accountability Report, 2000." https://www.gao.gov/assets/gao-01-626sp.pdf.

CHAPTER 2

The Republican Revolution in High Resolution
The Heterogeneous Effects of the Republican Revolution on Oversight at the Government Accountability Office

This chapter investigates heterogeneity in the Republican Revolution effects identified in the previous chapter. Specifically, I study whether the Republican Revolution's impact on GAO outputs varied across policy domains and agencies. To explore the possible treatment effect heterogeneity in the main RR effects, I employ a novel causal machine learning approach tailored to this high dimensional, regression discontinuity design setting that employs sample splitting and further restricts the sample the bandwidths employed for main effects. Utilizing this estimation procedure, I present significant evidence for treatment effect heterogeneity in the RR effects estimated in Chapter 1 over substantive topics and agencies. My results elevate National Defense along with Natural Resources and Environment and the Environmental Protection Agency as key moderators. I estimate sizable declines in the probability a report makes recommendations or matters over most topics, but find conversely, an increase in the probability that reports related to national defense contain recommendations or matters for Congress. These results are consistent with GAO's strategic accommodation of the new majority's oversight priorities. Finally, I detail the relevance of these results for the mechanisms outlined in the first Chapter.

Introduction

As is established in the previous chapter, the Republican Revolution caused GAO to issue fewer recommendations for executive action and matters for Congress while holding the raw count of reports equivalent or even slightly increasing. *Were these effects constant over substantive policy topics and executive departments?* In other words, did the GAO issue less recommendations for executive action overall, or concentrated among certain types of reports? While it isn't possible to derive theoretical expectations for every subcategory of reports (e.g., how did GAO change its output for reports dealing with the National Park Service?), I argue based on the theoretical mechanisms described in the first chapter GAO would have been prudent to shift its visible policy recommendations for either the executive or Congress to topics in line with the demands of its new principal, Republican Committee chairs.

In the remainder of this chapter, I proceed to interrogate the possibility of heterogeneity in the main effects identified previously. In order to do so, I craft a novel estimation procedure that applies causal machine learning, in particular a variant of Wager and Athey (2018)'s causal forest algorithm, in the context of a regression discontinuity design. The approach has a variety of benefits, for example, allowing for estimation and inference over a high dimensional set of potential moderators without manually specifying the single moderator of interest. This is especially useful here given the sheer volume of potential moderating variables and the difficulty of deriving clean theoretical predictions for each individual covariate. I find, using a modified causal forest approach, that the RR effects identified in Chapter 1 vary significantly over substantive topics and agencies. For example, reports focusing on National Defense have a higher probability of containing explicit policy recommendations for the executive following the Republican Revolution in contrast to other report types. These results suggest GAO made an explicit attempt to acquiesce to the oversight interests of the new Republican majority while largely eschewing explicit recommendations for other topics and agencies. These results add important nuance to the main effects identified in Chapter 1, and further clarify the potential mechanisms at play in these results.

Background and Theoretical Expectations

Following the direct attack on GAO in 1995, Comptroller General Bowsher's term ended shortly thereafter in September of 1996. It took over two years for a new Presidentially appointed and

Senate confirmed Comptroller General to take office in November of 1998. Upon his confirmation, newly minted GAO director David M. Walker championed several pieces of reform legislation that, among other things, rebranded the agency from the Government Accounting Office to the Government Accountability Office, which better reflects the nature of GAO's work products which extend well beyond mere accounting. Both in the interim period and with a new comptroller general, I argue GAO officials as in the previous Chapter, these officials attempted to pivot towards the policy interests of its new principal (i.e., acquiescing), attempting to cater to the interests of Republican committee chairs (who exerted pressure on GAO's annual appropriations) and Republican rank-and-file members. The rhetorical basis of Republican attacks on GAO was the claim GAO was a mere client of Democratic committee chairs and an unduly partisan agency. In this political environment, anytime GAO issued a public recommendation for executive action, or matter for Congress, they created an opportunity for scrutiny and even scandal. Given the credible possibility of disbandment, it would be logical for GAO officials to avoid publishing explicit policy recommendations that could garner the hostile attention of Republican members. Further, officials would have been wise to cater to the specific requests and interests of Republican members.

Given these mechanisms, we should expect GAO to shift public recommendations and matters for Congress to issue areas favored by by Republican members and away from other issues that likely to attract public attention needlessly. In this period, Republican members expressed interest in GAO auditing the Pentagon and investigating National Defense issues generally [1]. It is more difficult to derive clear theoretical expectations for other topics given the scope of both topic and agency labels that could serve as potential moderators. This difficulty further necessitates a principled and relatively agnostic approach to estimating conditional effects in this setting as outlined in the subsequent section.

Methods, Identification, & Data

Unlike in the previous chapter, the unit of analysis for the heterogeneous treatment effect models is the individual GAO report. This change makes the results in a substantial increase in the statistical power available to estimate the conditional effects.[2] As is widely discussed in the literature (see e.g.,

[1] These requests were made in public Congressional hearings following the RR.

[2] For main effects, aggregating at the daily level yields sufficient power and also allows for ease of interpretability, and construction of proportional outcome measures. Here, the outcome measures are reduced to increase available power.

Ratkovic, Druckman, and Green (2021) and Wager (2024) on conditional average treatment effect estimation, power is the most significant constraint in confidently exploring potentially complex heterogeneous treatment effect functions. I further include all available reports in the analysis to account for possible substitution between report types (e.g., reports that might have previously contained executive recommendations now published merely as reports without recommendations). Finally, for the sake of simplicity I reduce the dependent variables for this chapter to simple binary: whether or not each individual report contains 1) recommendations for executive action and 2) matters for Congress.

Identification Formally, this Chapter targets the conditional average treatment effect at the threshold, where we are interested in moderators of the main RDD effect in Chapter 1. Moreover, in addition to the identification assumptions needed to identify main effects in Chapter 1, it is now necessary to make additional assumptions to estimate conditional effects in the RDD setting. First, we must assume there are no omitted time-varying confounders Z that cause both our outcome of interest and a moderator of interest around the threshold. We further need common support on both sides of the threshold for moderating covariates. The first identifying assumption is quite strong and in these high dimensional observational data it's difficult to assuage concerns of potential omitted interaction bias (Blackwell and Olson 2022). Given this difficulty, the HTE results presented throughout should be viewed as somewhat exploratory, and taken with less confidence relative to the main effects.

Dataset The dataset for this chapter is the same as the previous chapter, but in this case we make use of the extensive handcoded topic and agency labels included on GAO's website. These 332 unique report level covariate features run the gambit from substantive topics (e.g., National Defense) to agency labels (e.g., the Department of Energy). This is a very high dimensional and sparse matrix of covariates to assess heterogeneous effects which motivates the use of a machine learning approach better suited to this type of data.

Methods To identify heterogeneous treatment effects (HTEs), political scientists often turn to simple subset regressions executed on the entire available sample (Hainmueller, Mummolo, and Xu 2019). This approach is fraught for a number of reasons, among them the issue of multiple testing (Ratkovic, Druckman, and Green 2021), possible omitted interaction bias (Blackwell and

Olson 2022), and often low statistical power. Moreover, given the large and sparse covariate space available to investigate conditionality in the main effects, saturated regressions would succumb to the curse of dimensionality. In addition to the concerns that plague any HTE analysis, the RDD setting involves further challenges. For example, how should kernel weights, bandwidth trimming, and natural clusters in the data be incorporated into the estimation procedure?

Instead of the traditional regression approach, I adapt Causal Forest (CF)(Wager and Athey 2018)—a causally oriented machine learning method built on random forest (Breiman 2001)—to estimate HTEs in the Regression Discontinuity in Time (RDiT) setting.[3] This approach yields doubly robust conditional average treatment effect estimates (CATEs) and employs powerful forest-based curvefitting algorithms to detect possible heterogeneity. Moreover, I follow a principled estimation procedure that involves sample-splitting and state-of-the-art tests for the existence of treatment effect heterogeneity as well as causal variable importance metrics(Bénard and Josse 2023).[4]

To estimate the CF model, for each dependent variable I first restrict the sample to the automatically selected RDD bandwidth employed in the main effects and treat reports following the RR as treated and those prior as untreated. For simplicity, I employ simple uniform weights for the HTE analysis so each observation is equally weighted regardless of distance from the RR threshold.[5] Next, I train the CF model using all available topic and label covariates, tuning the model and reducing the feature space using a best linear fit test on out-of-sample observations. For each model, I recursively retrain the model dropping any covariates with feature/covariate importance of zero for estimating the CATE function. While this might seem like an aggressive trimming procedure, in this case all this procedure does is recursively drop covariates that are never chosen in the heterogeneous treatment effect model despite being available to the model at many opportunities over thousands of honest decision trees.

Causal Variable Importance However, before summarizing the CATEs, one must first determine over which marginal dimensions[6] to calculate conditional effects.

[3]Chapter 3 contains a more expansive explanation of the inner workings of CF, this is a cursory overview.

[4]In the Appendix, I provide a suite of evidence that this estimation procedure is well powered and suited for this particular task. I demonstrate via Monte Carlo simulations the performance of my approach versus a state-of-the-art regression based approach that underperforms in this high dimensional setting.

[5]In the Appendix, I demonstrate robustness to alternative kernel weights for the two models.

[6]In the Appendix, I investigate the possibility of interactive, or multi-dimensional heterogeneity, in the modeled RR effects by employing a partial dependence framework (Hastie et al. 2009). Specifically, I calculate overall and

To avoid multiple testing problems (i.e., simply generating marginal CATEs for every available covariate), I instead select covariates based on causal variable importance scores using a modified leave-one-covariate-out procedure (Bénard and Josse 2023). This procedure involves retraining the CF model while dropping one, or a group of covariates, and then reassessing the models performance over R-loss with that variable(s) excluded. Intuitively, R-loss is the change in model fitness for the CATE model. An advantage of this approach is highly correlated covariates that likely represent the same latent dimension can be considered together. The variant of this method applied here groups covariates if their simple Pearson correlation is above .5. For instance, in this case the labels of "Energy" and the "Department of Energy" are highly correlated and are thus considered jointly. Normally, variable importance metrics come with no statistical guarantees, but this specific procedure is guaranteed to generate consistent estimates assuming only SUTVA and no unobserved confounders. The precise process for the C-VI calculation is shown in the following steps:[7]

(1) **Local centering** Compute out-of-bag nuisance estimates $\hat{\mu}(X_i) = \mathbb{E}[Y_i \mid X_i]$ & $\hat{\pi}(X_i) = \Pr(W_i{=}1 \mid X_i)$, then:

$$\tilde{Y}_i = Y_i - \hat{\mu}(X_i), \quad \tilde{W}_i = W_i - \hat{\pi}(X_i).$$

(2) **Full model** Fit a causal forest on $(\tilde{Y}_i, \tilde{W}_i, X_i)$ & obtain the CATE predictor $\hat{\tau}(X)$. Its empirical R-loss:

$$\widehat{R}_n(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{Y}_i - \hat{\tau}(X_i)\tilde{W}_i\}^2$$

(3) **Drop one covariate** Re-train the forest *without* the j^{th} covariate.

(4) **Variable–importance score**

$$\text{VI}_j = \widehat{R}_n\big(\hat{\tau}_{(-j)}\big) - \widehat{R}_n\big(\hat{\tau}\big) \quad \geq 0.$$

Larger VI_j corresponds with a greater degradation in model performance.

pairwise Friedman's H-Statistics to investigate the existence of strong interactions in the modeled CATE function. I find little evidence of important interactions, which is logical considering the high fraction of reports that deal with either single topics or clusters of related topics.

[7]For more detail on Nie and Wager (2021)'s R-learner and associated R-loss see Chapter 3.

Results

To begin with, I recalculate main effects for both dependent variables as a sanity check. Using the CF models, I estimate the RR reduced the probability an individual GAO report contained any recommendation by -6.8% [-8.2,-5.4%]. Further, for matters for Congress, the model estimates a -4.9% [-5.6%, -4.0%] reduction in the probability reports contained matters for Congress following the RR. Subsequently, I employ an omnibus suite of tests to determine whether or not heterogeneity is present based on out-of-sample testing: a best linear fit test and a test that employs rank-weighted average treatments (Yadlowsky et al. 2024). In brief, these tests interrogate whether or not treatment effect heterogeneity is present by assessing the CF model's performance on different subsamples of the available data.[8] Results are presented in Table 2.1 below for each test for recommendations for executive action. All tests indicate heterogeneity is present at common thresholds for statistical significance.

TABLE 2.1. Omnibus Heterogeneity Test Results for Executive Recommendations

Heterogeneity test	**Estimate**	**p-value**	**HTEs detected?**
Best Linear Fit Test (Chernozhukov et al., 2024)	1.25576057174549	5.047605×10^{-12}	TRUE
High vs. Low Test (Athey et al., 2017)	0.0832717824861555	1.606988×10^{-9}	TRUE
Sequential RATE Test (Wager, 2024)	Not applicable	6.271183×10^{-7}	TRUE
RATE OOB Test (Two-Sided, Wager, 2024)	0.0164396413332923	4.880977×10^{-2}	TRUE
RATE OOB Test (One-Sided, Wager, 2024)	0.0164396413332923	2.440489×10^{-2}	TRUE

The results for Matters for Congress are presented in Table 2.2 below. In this case, four out of five HTE tests pass, the only failing test is a heuristic measure of heterogeneity that compares the average effect above and below the median in the model CF cates.

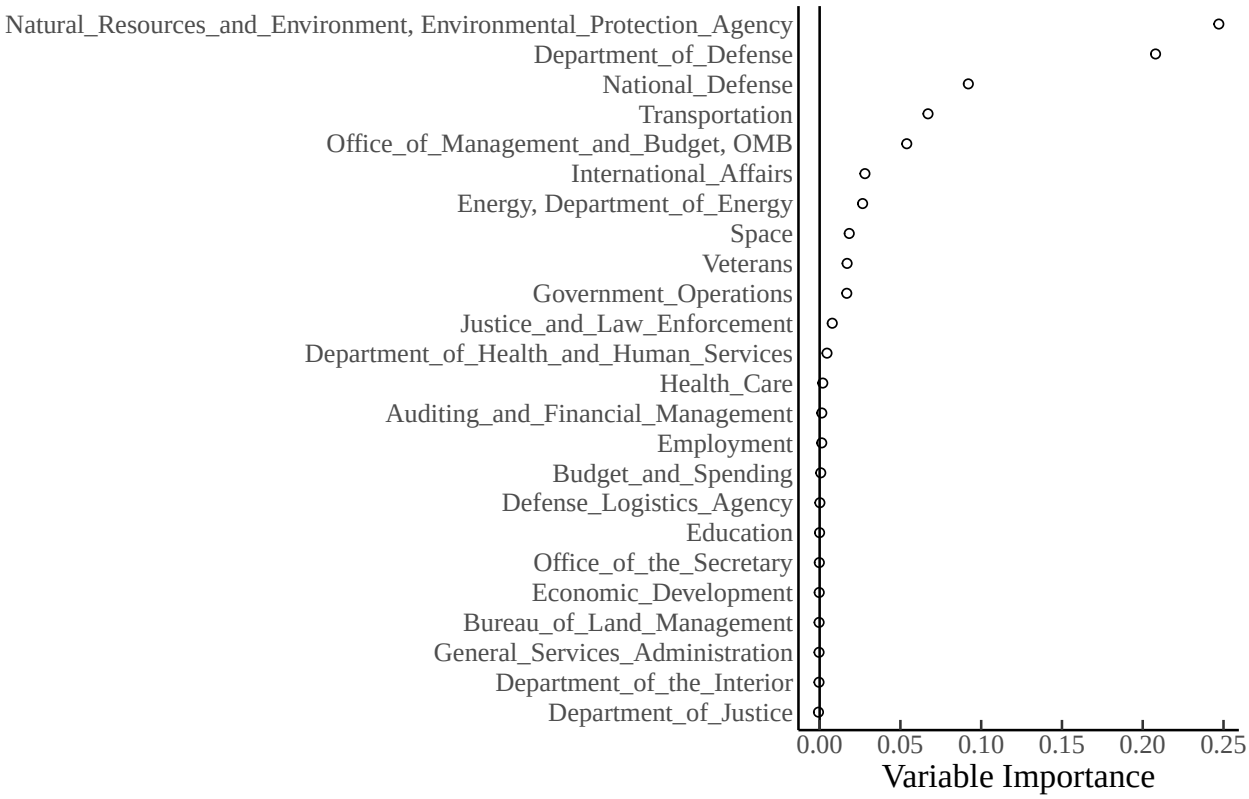
TABLE 2.2. Omnibus Heterogeneity Test Results for Matters for Congress

Heterogeneity test	**Estimate**	**p-value**	**HTEs detected?**
Best Linear Fit Test (Chernozhukov et al., 2024)	0.667129215283121	4.595399×10^{-6}	TRUE
High vs. Low Test (Athey et al., 2017)	0.0142187627282021	9.228546×10^{-2}	FALSE
Sequential RATE Test (Wager, 2024)	Not applicable	4.562997×10^{-3}	TRUE
RATE OOB Test (Two-Sided, Wager, 2024)	0.0136604691212965	1.876293×10^{-3}	TRUE
RATE OOB Test (One-Sided, Wager, 2024)	0.0136604691212965	9.381463×10^{-4}	TRUE

[8]Further detail in the Appendix for this chapter. I also present a simple distribution of the CATE predictions ranked by magnitude for each dependent variable in the appendix.

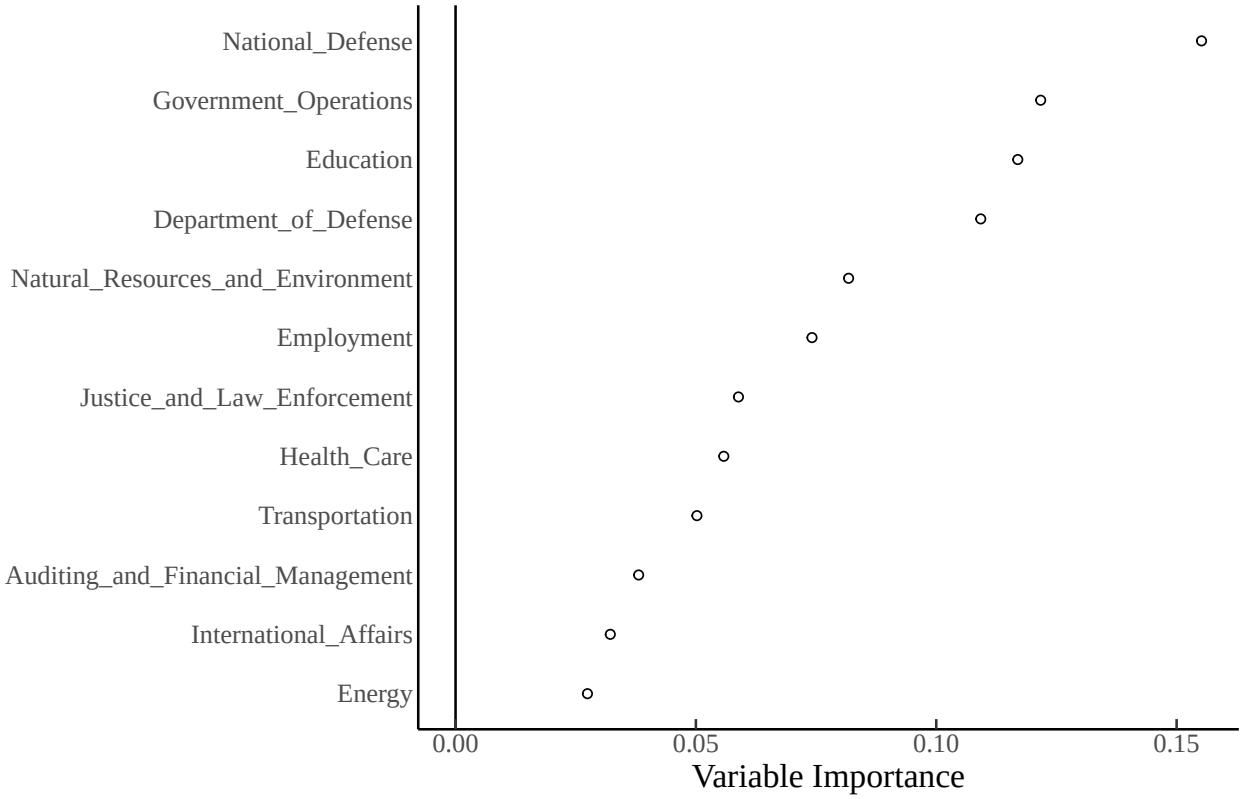
I estimate these causal variable importance (C-VI) scores, and the results are presented in Figures 2.1 and 2.2 below.

FIGURE 2.1. Causal Variable Importance Scores for HTEs: Executive Recommendations



Note: Scores on nominal scale, higher values indicate greater importance for determining treatment effect heterogeneity. Scores calculated through a modified, leave-one-covariate-out procedure developed by Bénard and Josse (2023). Variables are grouped by correlation, with covariates correlated above .5 grouped together. Defense and Department of Defense are considered separately given a correlation slightly below the threshold of .5.

FIGURE 2.2. Causal Variable Importance Scores for HTEs: Matters for Congress



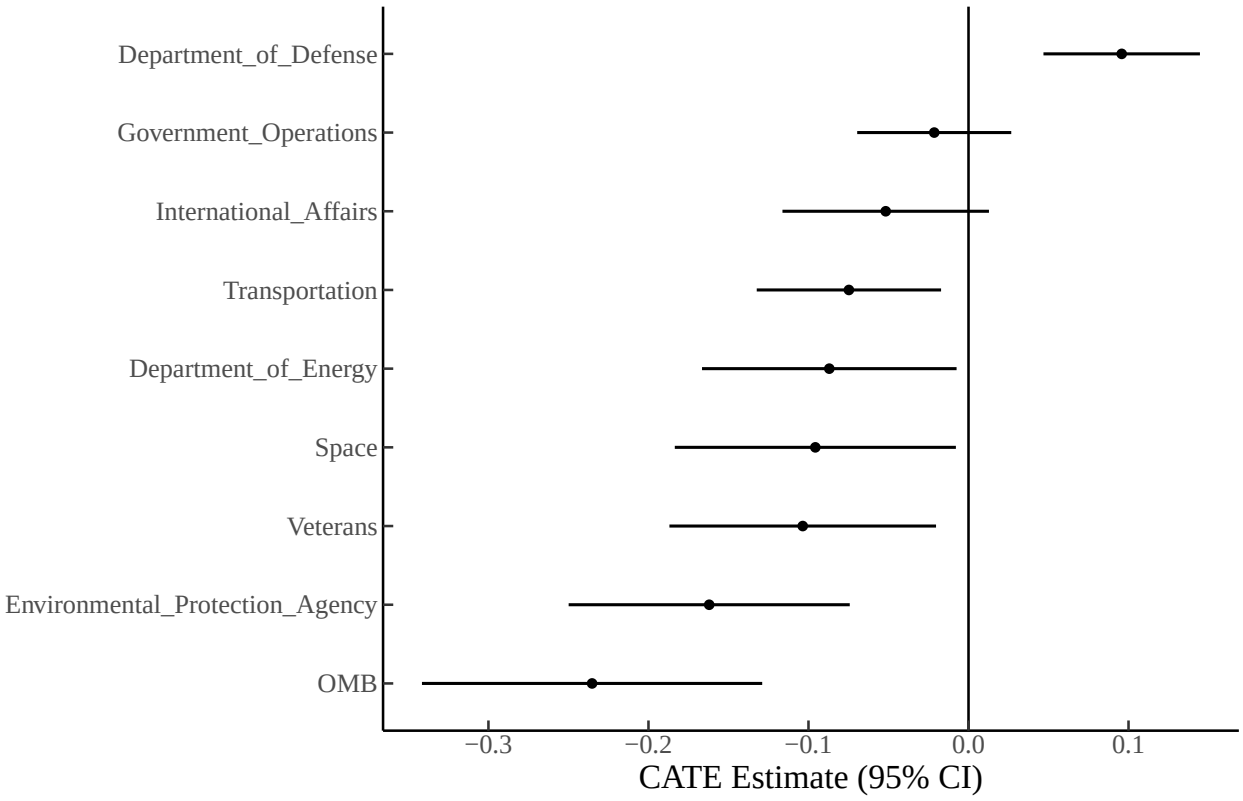
Note: Scores on nominal scale, higher values indicate greater importance for determining treatment effect heterogeneity. Scores calculated through a modified, leave-one-covariate-out procedure developed by Bénard and Josse (2023). Variables are grouped by correlation, with covariates correlated above .5 grouped together. Defense and Department of Defense are considered separately given a correlation slightly below the threshold of .5.

The C-VI scores for executive recommendations indicate the most important variables for determining treatment effect heterogeneity in the CF model are National Resources and Environment/Environmental Protection Agency as well as Department of Defense and National Defense, followed by several other somewhat less important categories. For matters for Congress, the C-VI scores indicate the most important variables are again National Defense and the Department of Defense, followed by Government Operations and Education, Natural Resources and Environment as well as other features.

To determine the magnitude and direction of these effects, I now proceed to estimate CATE summaries over the top 10 dimensions identified by the C-VI scores. There are several approaches for summarizing the CATEs, in this instance I opt to employ a best linear projection approach

(Wager and Athey 2018). This method projects the CF results into a linear space, allowing for the interpretation of the results as one would a simple linear model (in this case, a linear probability model). Moreover, the best linear projection procedure yields doubly-robust estimates with robust standard errors. These CATE results are displayed in Figures 2.3 and 2.4.

FIGURE 2.3. Change in Probability GAO Report Makes Recommendation for Executive

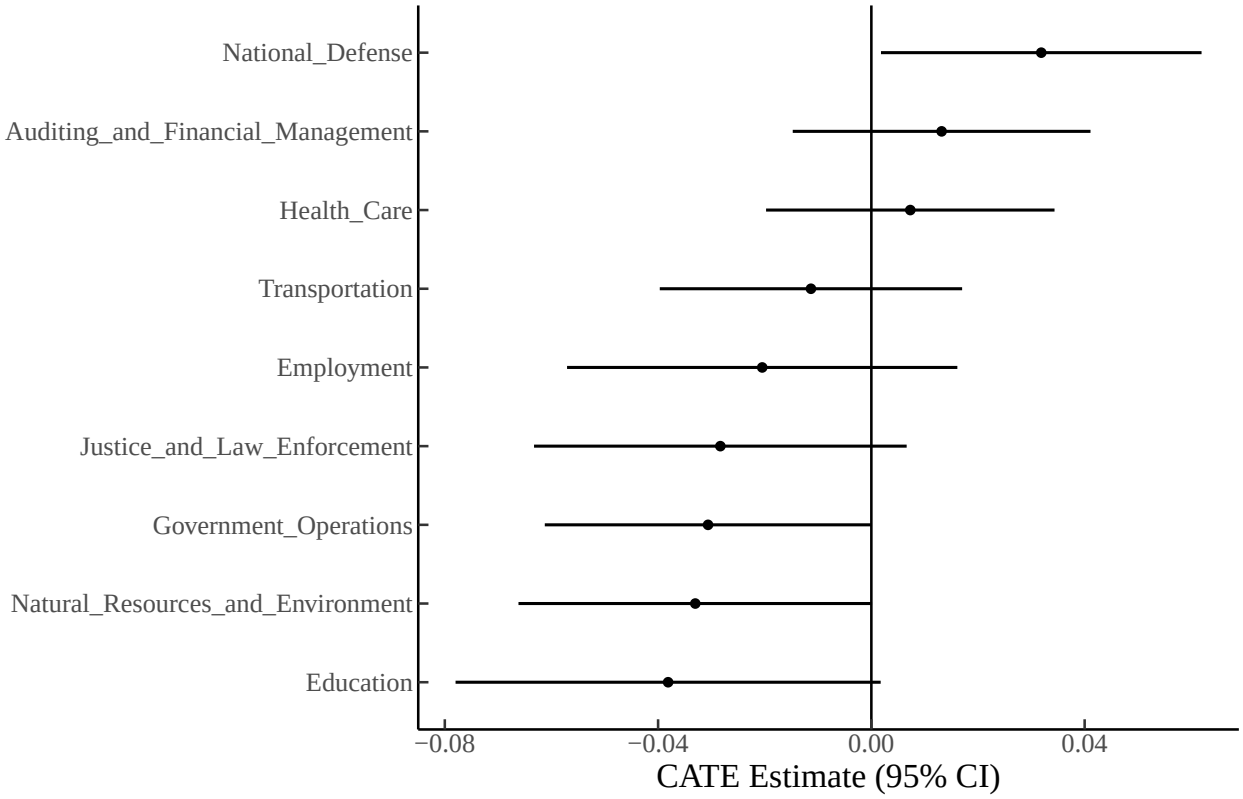


Note: Doubly-robust conditional average treatment effect estimates with 95% confidence intervals. Estimates derived from a tuned causal forest model and a best linear projection approach, allowing for interpretation of the effects similar to a linear probability model. Covariates selected on the basis of variable importance for determining treatment effect heterogeneity. Models tuned for out-of-bag fit, and variables trimmed based on feature importance. For collinear covariates, only one is included (e.g., only Department of Defense and not also National Defense).

These estimates indicate divergent results over report topics. For example, for the Office of Management and Budget as well as the Environmental Protection Agency, there are large estimated reductions in the probability a GAO report contains recommendations for executive action following the RR. On the other hand, the probability that reports related to the Department of Defense contain recommendations for executive action increases following the RR. Other categories

identified by the C-VI scores have negative estimated effects of smaller magnitude, or are statistically insignificant.

FIGURE 2.4. Change in Probability GAO Report Contains Matter for Congress



Note: Doubly-robust conditional average treatment effect estimates with 95% confidence intervals. Estimates derived from a tuned causal forest model and a best linear projection approach, allowing for interpretation of the effects similar to a linear probability model. Covariates selected on the basis of variable importance for determining treatment effect heterogeneity. Models tuned for out-of-bag fit, and variables trimmed based on feature importance. For collinear covariates, only one is included (e.g., only Department of Defense and not also National Defense).

The results for matters for Congress are directionally similar to the results for recommendations, but in this case most of the 95% confidence bands overlap zero or sit on the knife-edge of statistical significance. Overall, these results do not inspire confidence, but directionally the categories fall as expected, with an increase in the probability reports contain matters for defense in particular.

Discussion

The heterogeneous treatment effect results of this chapter suggest GAO both avoided Congressional attention over some policy domains and acquiesced to Congressional Republicans by focusing on issues such as National Defense. Agents at GAO rationally redirected the agency's attention in ways that (at least on paper) were designed to appeal to the new majority party. These results also provide an additional robustness check for the main results in Chapter 1, by employing different measures of the outcomes of interest and taking a different, causal machine learning estimation approach, I am able to valid again the main RR effects. Moreover, GAO clearly avoided issuing recommendations for policy areas that could attract attention to the agency, like the Office of Management and Budget, or agencies more closely associated with the Democratic party, like the Environmental Protection Agency.

Conclusion

While the results in this chapter present suggestive evidence for heterogeneity in the RR effects over both recommendations for the executive and matters for Congress, these results are somewhat exploratory and require strong identifying assumptions relative to the main effects. Nonetheless, the results presented here provide some evidence for heterogeneity in the RR effects presented in Chapter 1.

Bibliography

Bénard, Clément, and Julie Josse. 2023. "Variable importance for causal forests: breaking down the heterogeneity of treatment effects." *arXiv preprint arXiv:2308.03369.*

Blackwell, Matthew, and Michael P Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30 (4): 495–514.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45:5–32.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* 27 (2): 163–192.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Vol. 2. Springer.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika* 108 (2): 299–319.

Ratkovic, Marc, James N Druckman, and Donald P Green. 2021. "Subgroup analysis: pitfalls, promise, and honesty." *Advances in Experimental Political Science,* 271–88.

Wager, Stefan. 2024. "Sequential Validation of Treatment Heterogeneity." arXiv: 2405.05534 [econ.EM]. https://arxiv.org/abs/2405.05534.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2024. "Evaluating treatment prioritization rules via rank-weighted average treatment effects." *Journal of the American Statistical Association,* no. just-accepted, 1–25.

CHAPTER 3

Leaving No Variance on the Table: Causal Machine Learning for Average and Conditional Effects in Cross-Sectional Data

Sam Fuller[†] Jack T. Rametta[*]

[*]Harvard University [*]University of California, Davis

Machine learning methods in general and tree-based methods in particular are on the rise in the social sciences. While ML is widely valued for predictive accuracy, new causally-oriented machine learning methods, like causal forest, leverage this power for statistical inference and treatment effect estimation. Previous work has established the theoretical benefits of these methods, and compared their relative empirical performance for different causal tasks, but no work to date has quantified the performance of CF compared to standard methods across various research contexts. This is an especially glaring omission given recent debates regarding the utility of machine learning for tabular data analysis in the social sciences in the first place. In this chapter, we employ a diverse set of over two million Monte Carlo simulations reflecting both experimental and observational data in cross-sectional settings to evaluate the performance of a popular CML approach, causal forest, relative to standard regression approaches. We find causal forest outperforms standard methods for average and conditional effects across a variety of evaluative criteria including statistical power, precision, bias, and the accurate detection and consequent estimation of heterogeneous treatment effects. Conversely, we find that standard approaches have vastly larger false-positive rates for conditional treatment effects: 9–66 times larger than rates for causal forest. These results suggest researchers should consider leveraging causal machine learning to analyze their cross-sectional data.

Introduction

Research on causally-oriented machine learning (CML) has accelerated in recent years. Beginning with the development of causal trees and causal forest (CF) (Athey and Imbens 2016; Athey, Tibshirani, and Wager 2019; Athey and Wager 2019), new estimation techniques are now coming forward at a rapid clip (e.g., Dorie et al. 2022; Chernozhukov et al. 2023; Chen et al. 2024). These new approaches aim to integrate the predictive power of blackbox machine learning algorithms into the potential outcomes framework, reorienting predictive algorithms for the task of causal inference. The CML literature has matured in recent years and applications of this approach outside of industry are now increasingly common in the social sciences. However, while there is significant research on the performance of causal machine learning estimators, studies tend to compare different CML approaches to one another while neglecting standard parametric approaches for the same causal task. For example, Wager and Athey (2018) run simulations to compare the performance of causal forest with k-nearest neighbors for conditional average treatment effect estimation, but not to OLS which is still widely used to estimate heterogeneous effects in political science and other disciplines. In practice, most social science research relies on tried and true parametric approaches and no study has explored quantitatively the costs and benefits of CML methods as compared to these standard methods like OLS. Consequently, we argue, it is exceedingly difficult for researchers to know when to use CF, or other related CML method like double machine learning(Chernozhukov et al. 2023), *and* to justify its use—particularly to skeptical readers or reviewers.

This paper aims to remedy this gap in research by leveraging the power of Monte-Carlo simulations to evaluate the performance of causal forest versus OLS in estimating both average treatment effects (ATEs) and conditional average treatment effects (CATEs) across a variety of common research scenarios in a cross-sectional setting. We choose to focus on causal forest because while there are now many CML approaches to choose from, causal forest is by far the most widely adopted approach. It is furthermore likely the easiest for social scientists with little background in machine learning to adopt "off-the-shelf." To compare CF with traditional methods, we build a conceptual model that represents a broad spectrum of common scenarios researchers may find themselves in. This model is designed to accurately represent actual research and is not designed to privilege any particular model over one-another. Overall, our framework motivates an extensive set of over

two-million simulations that seeks to enable researchers to choose when and where to use causal forest, specifically, and causal machine learning, generally, for analyzing their data.

We find across our simulations that causal forest outperforms common parametric approaches like OLS in terms of statistical power, coverage, precision, bias, and the accurate detection and consequent estimation of heterogeneous treatment effects. Importantly, our simulations are designed to be a tough-test for CF over OLS, finding that even in situations where OLS should perform admirably (consistent, linear effects, normally distributed errors) CF still matches or surpasses its performance. For example, in one set of simulations we find, while maintaining the same coverage rate, that CF standard errors are up to *40% shorter* than those recovered from OLS. Moreover, we find that CF has higher powered across a wide variety of simulation settings, in some cases requiring only *half* the sample size to achieve the same power as standard estimators. Perhaps counterintuitively, these improvements in inferential performance stem from increases in *predictive performance* achieved by the forest-based models.

For estimating heterogeneous treatment effects, we find CF improves in performance over simple multiplicative interaction models in terms of both bias, false positives, and false negatives. For our false positive simulations, OLS models have anywhere between a 43.1% and 53.5% false positive rate (FPR), whereas CF's FPR range is approximately 9 to upwards of 66 times smaller, ranging from 0.75% to only 5.00%. This finding alone suggests that researchers should take caution when attempting to apply OLS for detecting and estimating heterogeneous treatment effects. It is both easier and more robust to estimate effects using CF—or another CML model—for this challenging estimation task.

Our results, however, not only support the use of CF and CML, but also highlight significant issues with standard approaches. We argue that our results highlight three interrelated concerns with standard and even improved approaches (like Interflex from Hainmueller, Mummolo, and Xu 2019): 1) that standard approaches, particularly for the estimation of CATEs, are highly susceptible to false positives *and* false negatives (even with adjustments for multiple-testing like a Bonferroni correction); and 2) current approaches, even when they are not susceptible to these problems, are leaving money on the table by failing to exploit important information in pretreatment covariates; and 3) even when pretreatment covariates are included in standard models, their inability to flexibly

model the relationship between these covariates and the outcome without manual specification may also leave explainable variance on the table.

Our paper proceeds as follows. First, we summarize the previous research on treatment effect estimation, including previous work on causal forest. Second, we introduce a general conceptual framework for understanding research design in a cross-sectional setting and then introduce causal forest and doubly robust machine learning. We then demonstrate the utility of DRML and CF by leveraging a diverse and extensive set of simulations that directly reflect scenarios in our conceptual framework for cross-sectional analyses. Finally, we end with a discussion of the implications of our findings and point towards directions for future research.

Treatment Effect Estimation

Guidelines for the estimation of treatment effects in cross-sectional data, particularly experiments, have stayed relatively stable over the past twenty years. This is in large part due to the beauty of random assignment: even a simple difference-in-means can recover unbiased estimates of the average treatment effect (ATE). However, research questions in political science are rarely constrained to the ATE: heterogeneous effects are political scientists' bread and butter. The most common approach for estimating conditional average treatment effects (CATEs) is to pre-register conditional expectations and calculate these effects using an interaction (conditioning variable $\times$ treatment) in a regression model. While there is some previous research that cautions against this approach, particularly Hainmueller, Mummolo, and Xu (2019) and Ratkovic (2021a), it is still unclear when the standard interaction approach is likely to produce inaccurate estimates—or even false positives and negatives.

Even when we solely focus on the ATE in an experimental context, treatment effects may be small and/or difficult to recover without large sample sizes. Additionally, given political science's interest in complex—and idiosyncratic—relationships, pretreatment covariates are almost always significantly correlated with our outcomes of interest. This, in combination with potentially small treatment effects, has led many researchers to include control variables for ATE regression estimation. There is significant debate about whether to include control variables, and if so, how they can be included without inducing bias through post-hoc regression adjustments—what Mutz, Pemantle, and Pham (2019) pejoratively calls the "balance test and adjust" procedure. More recent research

defends the inclusion of controls and the use of regression-based estimators for ATE estimation in experiments (Lin 2013; Negi and Wooldridge 2021) with the important caveat that included covariates are pre-registered.

Furthermore, previous research in political science has demonstrated the challenging nature of heterogeneous treatment effect estimation. Hainmueller, Mummolo, and Xu (2019) outline several estimation concerns involved in the estimation of multiplicative interaction models, in particular the assumption of a constant linear moderator and potential lack of common support over the moderating variable of interest. They propose a useful suite of diagnostic tests and alternative estimators[1] to assess the validity of these assumptions in any particular study. Notably, however, the authors choose to explicitly avoid the application of machine learning models for CATE estimation. Also, their approach relies on the researcher to provide a single, prespecified moderator-of-interest. While their approach decreases the probability of false positive results, and serves as a robustness check against omitted interaction bias (Blackwell and Olson 2022) among other estimation concerns, it is nonetheless fundamentally rooted in traditional approaches to conditional average treatment effect estimation as opposed to newer approaches that allow for a more agnostic exploration of heterogeneous effects.

Previous Research on the Performance of Causal Machine Learning and Causal Forest Research on the benefits of causal machine learning methods, like causal forest, for social science data is limited. There is some work on how these methods should be used and incorporated into existing social science practice (e.g., analysis plans and pre-registrations), but is limited to two book chapters, Ratkovic (2021a) and Green and White (2023). First, Ratkovic (2021a) provides guidelines on how to use CF to detect treatment effect heterogeneity and then to estimate conditional average treatment effects (CATEs), focusing on issues like multiple testing and sample splitting. Green and White (2023), on the other hand, provides a detailed and rigorous introduction to these methods and also provides examples of implementations. However, neither explores nor determines what

Advanced methods—particularly those that leverage ML—have no clear guidelines for when and how they should be used and little-to-no understanding of the performance of these methods as compared to standard OLS estimation. While there is significant evidence of the performance

[1]These tools are compiled in the `interflex` `R` package available on CRAN.

of causal forest in the abstract, to date, no study has established quantitatively CF's costs and benefits as compared to standard parametric methods like OLS.[2] Consequently, researchers do not know when and how to use these methods over standard approaches for ATEs or conditional effects in both observational and experimental data.[3] Given this, it is far less confusing why these methods have not proliferated in the discipline.

Prior simulation studies outside of political science do compare the efficacy of causal machine learning approaches relative to one another, but they do not compare their performance relative to traditional estimation approaches. For example, Knaus, Lechner, and Strittmatter (2021) compares the relative performance of a wide variety of tree-based and regularization-based algorithms for individual average and conditional average treatment effects. They find that tree-based methods generally outperform regularization methods across most of their data generating processes. Similarly, Hahn, Murray, and Carvalho (2020) provide a suite of simulation studies to compare their Bayesian regression tree approach to causal forest and other algorithms. In contrast to prior work, we compare causal machine learning with standard, traditional estimators, focusing on ordinary least squares in particular. Our objective with this comparison is to demonstrate the benefits of causal machine learning even in research contexts where we may expect limited benefits (e.g., random treatment assignment, low-dimensional covariates, small sample sizes, etc.). It is obviously useful to understand the performance of these methods vis-a-vis one another, but it is critically important to explore the performance of these methods in common research scenarios against methods that are standard for these analyses.

Outlining Cross-Sectional Research Contexts

In this section we describe a stylized framework that outlines the scenarios researchers commonly face when analyzing cross-sectional data in experimental or observational contexts. These scenarios are broken down by the relationship of covariates to both the outcome and to treatment assignment. We summarize this space of possibilities using the poles of the two dimensions: covariates are either highly related or completely unrelated to the outcome and the same for treatment assignment. This dynamic, resulting in four main categories, is summarized below in Figure 3.3.

[2]For example, Wager and Athey (2018) run simulations to compare CF's performance versus k-nearest neighbors, but do not interrogate its performance as compared to OLS or other parametric methods.

[3]Athey and Wager (2019) do illustrate *how* one may analyze observational data using CF, but do not make recommendations on when one should.

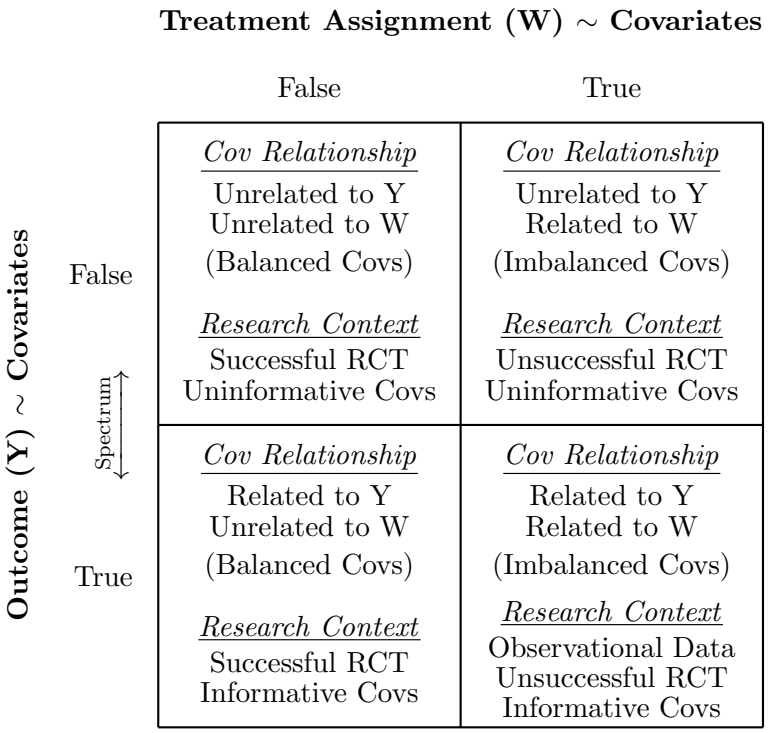


FIGURE 3.1. Stylized Research Context Framework

While there are numerous considerations that inform one's research design and analytical process, we argue that these dimensions capture the most important considerations when analyzing cross-sectional data. First, we look at the top row where covariates are unrelated to the outcome. While these situations could manifest in observational data, they are most common in experimental contexts, with the top-left including situations where a RCT is successful and the top-right where a RCT is unsuccessful (there is some sort of randomization issue leading to covariate imbalance between treatment and control groups). Regardless of the imbalance, however, there is no threat to recovering the true treatment effect because in either scenario covariates are unrelated to the outcome. The top-right example may be a scenario where, for example, an online-experiment looking at the influence of partisan policy endorsements on policy support was randomized by the respondent's IP address (e.g., all even-numbered addresses were assigned to treatment, all odd-numbered to control). In this context, we can safely ignore the imbalanced covariate(s), like IP address.

However, in the bottom row, covariates are related to the outcome. Continuing with our example above, the bottom-left cell would be a scenario where randomization was successful, but there are covariates like strength of partisan identity and self-reported ideology that are related to policy

support. Here, while these covariates are related to the outcome—and we very well may want to also explore their potential conditioning effect on the treatment—randomization is still successful, so we have an unbiased estimate of the ATE using specified parametric estimators (SPEs).[4] By specified parametric estimators, we refer to estimators that require the manual specification of covariates' functional forms and that make parametric assumptions with regard to either the relationship between covariates and the outcome variable and/or the standard errors (e.g., ordinary least squares regression, difference-in-means in a regression framework, etc.). However, when we are in the bottom-right cell and covariates are related to *both* the outcome and treatment assignment, we will have biased estimates of the ATE when we have no adjustments. In fact, this cell represents most observational data analyses, where we have no evidence of random assignment or other source of exogenous variation to underlie a design-based inferential approach, and in cases where a RCT has randomization issues with covariates that are related to the outcome.

Put simply, in the top row we do not need to care about covariates and in the bottom row we should be highly concerned about covariates, especially in the bottom-right cell. Again, these cells are the ends of the spectra here, so researchers will most often find themselves somewhere in between the poles. We will note here, however, that other considerations are of course still important for researchers to consider. These include research/experimental design, choices of covariates and their measurement, unobserved confounders, and potential included colliders—just to name a few. Additionally, as we mentioned above, this framework is focused on cross-sectional data. And while the methods we discuss in this paper *can* be employed in panel or time-series contexts (e.g., see Arkhangelsky and Imbens 2022), we do not focus on them here.

Conditional Effects In political science research, conditional effects are an essential tool to evaluate hypotheses, especially with experimental data. Often we are interested in how treatments differentially affect subgroups or how some latent dimension conditions a treatment. Unfortunately, estimating conditional effects is, fundamentally, a difficult statistical task that requires researchers to carefully consider a wide variety of factors, especially in the design phase of surveys and experiments.

[4]While some causal machine learning methods use semi-parametric estimators, like causal forest with AIPW, the modeling of covariates' relationship to both the outcome and treatment is fully non-parametric. Other estimators, such as flavors of Chernozhukov et al. (2023)'s double machine learning estimator, are fully non-parametric.

These range from what covariates should be measured and how, to modeling choices (e.g., interactions or subsetting), to what criteria to use to determine if heterogeneity/conditionality is present, choosing an estimand (e.g., CATEs or individual treatment effects, ITEs), to finally choosing an estimator for conditional effects (e.g., causal forest, interactive regression models, etc.). Common threats to the estimation of these effects include omitted interaction bias (Blackwell and Olson 2022); a lack of common support over the moderating variable (Hainmueller, Mummolo, and Xu 2019); insufficient statistical power (Arel-Bundock et al. 2025); multiple testing bias when testing multiple conditioning variables, and more (Ratkovic 2021b). These concerns and threats to validity are important to consider for both observational and experimental work: even with experimental data, CATEs do not have the same design-based guarantees as ATEs, and so we should be cautious when estimating them.

Causal Forest & Doubly Robust Machine Learning

Causal Forest Causal forests build upon the general random forest concept in two major ways.[5] First, it combines the standard random forest process with orthogonalization through residual-on-residual regression introduced by Robinson (1988) which, as we mentioned earlier, relies on the Frisch-Waugh-Lovell theorem. In the initial "building" step, this entails estimating two separate random forest models: one of treatment propensity $\hat{\gamma}^{(-i)}(X_i)$ and one of the outcome model $\hat{\omega}^{(-i)}(X_i)$, where the superscript $-i$ denotes that the model was estimated on observations not including i (that it was cross-fit, or the effect is estimated "out-of-bag"). Importantly, causal forest modifies the loss function of classic random forests to instead maximize the squared difference in subgroup treatment effects: $n_L n_R(\hat{\tau}_L - \hat{\tau}_R)^2$. This average subgroup treatment effect ($\hat{\tau}$) is estimated by residual-on-residual regression where the residuals originate from the propensity and outcome forests. Subsequently forest weights $\alpha(x)$—which simply predict a weighted average of outcomes—are used to estimate the conditional average treatment effect (CATE) function:

$$\tau(x) := lm\left(Y_i - \hat{\omega}^{(-i)}(X_i) \sim W_i - \hat{\gamma}^{(-i)}(X_i), \text{weights} = \alpha_i(x)\right) \tag{3.1}$$

Second, in order to avoid overfitting and regularization bias, CF adds a new tree-building parameter: honest subsample splitting. Essentially, the honesty parameter dictates what portion

[5]For a description of decision trees and random forest, see the Appendix.

of the sample is used to determine splits in each tree and what portion of the sample is used to estimate the effects of each bin/leaf. This is in contrast to classical random forest, which uses a random subsample to build each entire tree. This feature leads to numerous benefits, enumerated in Athey and Imbens (2016), but most importantly enables the construction of valid confidence intervals and improved coverage versus "dishonest" models.

A subsequent article by Wager and Athey (2018) shows that CF provides valid statistical inferences for treatment effects: given the usual selection on observables assumptions (i.e., unconfoundedness), CF estimates converge on the truth (i.e., are consistent) and are asymptotically normal. Similarly, Athey, Tibshirani, and Wager (2019) establish the process for calculating valid confidence intervals for CF effect estimates using a form of infinitesimal jackknife bootstrapping (in truth, for any generalized random forest variant). By leveraging the subsample splitting process of CF, the variance estimation procedure is able to calculate variance and thus CIs. To quote the `grf` package material:

> *In each training pass, we sample the full dataset to create a subsample of half its size. Then, a small group of trees is trained on this half-sample. In particular, for each tree we draw a subsample of the half-sample, and grow the tree using these examples... When predicting, a variance estimate is also computed by comparing the variance in predictions within groups to the total variance.*

Taken together, these adaptations and additions position CF as a model that leverages the predictive power of machine learning to answer causal questions and estimate causal quantities.[6] Importantly, while the causal forest algorithm and its estimation of individual treatment effects (ITEs) is fully non-parametric, in order to perform inference with conditional effects it is necessary to summarize the inherently noisy ITE estimates with a semi-parametric estimator such as AIPW (using the "best linear projection" from Semenova and Chernozhukov 2021).

An example of the power of this method is illustrated in Figure 3.2, where we plot the predicted CATE values from a CF model and from a standard OLS in comparison to the true DGP. Specifically, the DGP is a simple stepwise function:

[6]For more details on why CF provides valid statistical inferences and confidence intervals, see Wager and Athey (2018) and Athey, Tibshirani, and Wager (2019), respectively.

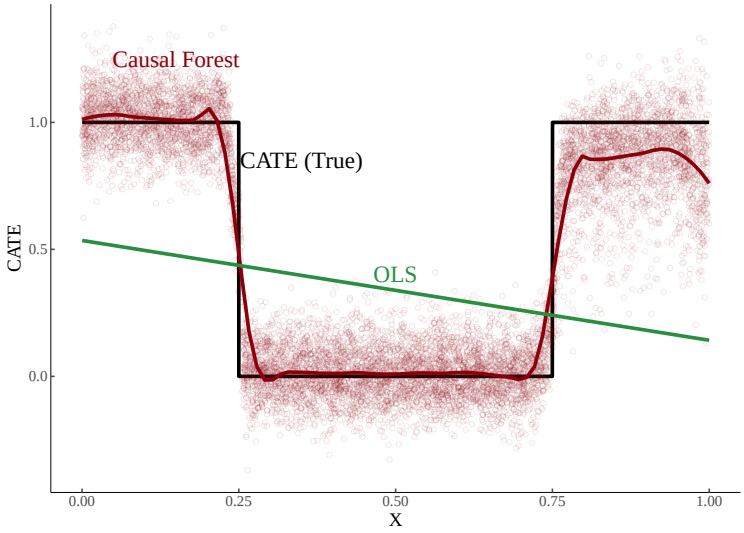


FIGURE 3.2. Comparison of Causal Forest and OLS in Recovering a Stepwise Functional Form

$$y = \begin{cases} 0 + \mathcal{N}(0,1) & \text{if } 0.25 < x < 0.75 \\ 1 + \mathcal{N}(0,1) & \text{otherwise} \end{cases} \tag{3.2}$$

Here we see that CF can handle complex, nonlinear functional forms with ease, whereas OLS is forced to provide a highly inaccurate approximation. This flexibility and highly accurate predictions are core to the benefits that CF provides. We outline these benefits in further detail in the doubly robust estimation framework section to follow.

Doubly Robust Machine Learning To understand why CF provides significant benefits it helps for us to first establish the basic structure and logic of doubly-robust estimation.

Here we specify a partially linear model following Robins and Rotnitzky (2001), which introduces this formalization.

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W,X) = 0,$$

$$W = \gamma(X) + V, \qquad \mathbb{E}(V|X) = 0,$$

where Y is our outcome variable and $W = \{0, 1\}$ is an indicator variable of treatment.[7] The vector $X = (X_1, ..., X_p)$ is composed of all other covariates, and ϵ and V are random errors for each equation.

In plain language, we have two equations we are working with. The first, $Y = \theta W + \omega(X) + \epsilon$, models the outcome variable as a function of treatment (W) and covariates (X) as well as random error (ϵ) that, in expectation, is equal to zero (and can be ignored) when controlling for treatment and covariates: $\mathbb{E}(\epsilon|W, X) = 0$. Here, θ is the causal effect of treatment (the coefficient for W) and $\omega(\cdot)$ represents the functional relationship between covariates (X) and the outcome (Y). This first equation, however, is just a simple regression framework for modeling the effect of some treatment. The "D" or "Doubly" of DR, and why this approach enables us to recover causal estimates, arises from the use of the second equation for treatment effect estimation.

This second equation, $W = \gamma(X) + V$, models treatment as a function of covariates ($\gamma(X)$) and random error (V). Again, the random error in expectation is zero when controlling for covariates, $\mathbb{E}(V|X) = 0$, so we can ignore that term. In a RCT context where randomization was successful, we would essentially be able to ignore this equation because treatment assignment would, ideally, be orthogonal (unrelated) to any covariates, even if they are related to our outcome of interest. Specifically, we attempt to estimate $\gamma(\cdot)$ which is the functional relationship between covariates (X) and treatment assignment (W).

DR estimation also has important desirable properties: 1) it is semi-parametrically efficient and, most importantly, 2) it is robust to model misspecification in either the outcome or treatment propensity models (i.e., we only need to correctly specify one of these models to recover consistent, causal effects). The first property relates to the fact that the model is efficient (recovers the real treatment effect with a limited number of observations) even though the functional form of some of the components of the model (namely $\omega(\cdot)$ and $\gamma(\cdot)$) are unknown. The second property is that DR estimation will provide robust treatment effects as long as we model the outcome and/or treatment assignment correctly. This component can be evaluated theoretically—are we including variables that we have good reason to believe are related to the outcome and/or treatment assignment?—and empirically—based on model statistics, are we accurately specifying these models between training (data used to build the models) and testing sets (data held out from the training set)?

[7] Here we focus on a binary treatment for simplicity, multi-arm and continuous treatments also work in this framework.

As we mentioned above, these two equations can be and have been modeled using parametric estimators, like OLS for the outcome model and MLE for the propensity model. However, while parametric estimators are easy to interpret, correctly specifying one of these model is *difficult* and is subject to numerous considerations.[8] These include the choice and number of regressors; any variable transformations, like squaring a term; and the choice of interactions. Above and beyond these considerations, there are fundamental questions about the underlying functional relationship between the covariates and the outcome or propensity: is it purely additive? Is it nonlinear? How should missingness in covariates be incorporated? As Robins and Rotnitzky (2001) note, a flexible, nonparametric estimator would be optimal in this context where predictive accuracy is paramount. In the following section, we discuss the applicability of nonparametric machine learning algorithms for these tasks.

Referring back to Equation **??**:

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W, X) = 0,$$

$$W = \gamma(X) + V, \qquad \mathbb{E}(V|X) = 0,$$

We now see where machine learning fits into this equation (literally and figuratively). Specifically, we can use it to model the relationship between the outcome and pretreatment covariates, $\omega(X)$, and treatment assignment and pretreatment covariates, $\gamma(X)$. With these models in hand, particularly the treatment assignment model, we can then estimate the treatment effect, θ, in the first equation $Y = \theta W + \omega(X) + \epsilon$.

This is the ML component of DRML, and there are vast benefits of this approach over both specified parametric estimators and standard DR estimation (e.g., a logistic regression for treatment propensity estimation). Broadly, the benefit of DRML is the combination of DR estimation with the predictive power of flexible machine learning algorithms. We are able to much more confidently model both the outcome, $Y = \omega(X)$, and the propensity to be treated, $W = \gamma(X)$. Additionally, we illustrate below across all of our simulations the significant performance gains of DRML over SPE approaches.

[8] We note, however, that maximum-likelihood estimators, like logistic regressions, are neither easily interpretable *nor* particularly powerful or accurate.

The specific implementation of DRML with CF, however, is more nuanced than simply plugging in the ML model for $\omega(X)$ and $\gamma(X)$. By default, CF employs the augmented inverse propensity weighting estimator to calculate CATEs and ATEs:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\underbrace{\frac{W_i Y_i}{\hat{\gamma}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{\gamma}(X_i)}}_{\text{IPW Component}} + \underbrace{\hat{\omega}_{(1)}(X_i) - \hat{\omega}_{(0)}(X_i)}_{\text{Outcome Component}}\right] \tag{3.3}$$

Like before, W_i is a binary treatment indicator, Y_i is the observed outcome, X_i are the pre-treatment covariates, $\hat{\gamma}(X_i)$ is the estimated propensity score, $\hat{\omega}_{(0)}(X_i)$ is the estimated outcome regression model for the given observation i assuming they were treated, and $\hat{\omega}_{(0)}(X_i)$ is the estimated outcome regression model for the same observation assuming they were untreated. In this setup, we can use ML to model the nuisance functions $\hat{\gamma}(X)$ and $\hat{\omega}(X)$. Taken together, this composite estimator outputs semi-parametrically efficient average treatment effect estimates in very general settings (Chernozhukov et al. 2023). This method, as we show with our simulations, is often a first-best approach for calculating ATEs.

To understand how AIPW has doubly robust properties[9] , we decompose Equation 3.3 into two components, the regression adjustment component D and the residual IPW estimator R as follows:

$$\begin{aligned}
\hat{\tau}_{AIPW} &= D + R \\
D &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\omega}_{(1)}\left(X_i\right) - \hat{\omega}_{(0)}\left(X_i\right)\right) \\
R &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{\gamma}\left(X_i\right)}\left(Y_i - \hat{\omega}_{(1)}\left(X_i\right)\right) - \frac{1-W_i}{1-\hat{\gamma}\left(X_i\right)}\left(Y_i - \hat{\omega}_{(0)}\left(X_i\right)\right)\right)
\end{aligned} \tag{3.4}$$

While it is true that this estimator is "weakly" doubly robust and guaranteed consistent if either the outcome or propensity model is well estimated, the property of "strong" double robustness is only possible when both models are at least somewhat well specified. As we have hinted at above, the ML component of DRML is much better at correctly modeling *both* the treatment and outcome than standard parametric approaches. Consequently, we can be far more confident that we are able to accurately model both components such that we reach strong double robustness.

[9]Notation follows Wager 2022.

To be specific, strong double robustness would mean our estimated $\hat{\tau}_{\text{AIPW}}$ is a first order equivalent of the true, "oracle" function of the treatment effect: τ_{AIPW}. Of course, this is not verifiable in real-world data analyses. However, as the model fitness of the outcome and propensity models improve, we can be increasingly confident our estimator is performing well. This requirement for strong double robustness motivates cross validation and tuning for both models. It also justifies attempts to model both the response surface and treatment with different algorithms that might better fit the idiosyncratic data-generating process at hand. For example, it may be optimal in some circumstances to choose different algorithms, or even an ensemble of algorithms, for the different nuisance functions.

Testing for Heterogeneity and Estimating Conditional Effects Traditional methods for estimating conditional effects, namely simple interactive regression models (conditioning variable $\times$ treatment), are often ill-suited for first determining if treatment effect heterogeneity is present and second estimating conditional effects. This is due to: 1) standard parametric methods restrict the functional form of the conditioning effect of variables; 2) standard parametric methods can lead to both false negatives and positives due to nonlinear CATE DGPs (see simulations below), multicollinearity, a lack of statistical power (Arel-Bundock et al. 2025), overfitting (Ratkovic 2021b), and omitted interaction bias (Blackwell and Olson 2022); and 3) excessive researcher degrees of freedom in specifying these models (Hainmueller, Mummolo, and Xu 2019).

DRML, in contrast, offers a robust approach for subgroup analysis/estimating conditional effects that deals with these issues in a systematic way, improving on traditional methods across the board. In particular, one can estimate a DRML CATE model that includes all relevant pretreatment covariates that addresses the possibility of omitted interaction bias while not falling prey to the curse of dimensionality (Altman and Krzywinski 2018). Moreover, DRML methods natively involve forms of sample splitting or cross-fitting to avoid overfitting and the powerful predictive capabilities of ML curve fitting algorithms can improve on the power requirements needed to confidently estimate conditional effects by improving efficiency and precision (see our simulation evidence). Finally, the saturated nature of DRML CATE estimation actually restricts researcher degrees of freedom, so long as the estimator is paired with a detailed pre-registration procedure (as we discuss in further detail later).

DRML With Uninformative Covariates An important concern which we hope to address with this paper is identifying when, how, and to what extent DRML fails, or introduces bias, when estimating treatment effects. Two obvious situations where this may happen are represented by the top row of Figure 3.3: covariates are uninformative and treatment assignment is either random or non-random. Here in these situations covariates have no relationship with the outcome (either directly or through a conditioning effect with the treatment), and thus it would be reasonable to assume that estimating a DRML model using covariates to predict said outcome (and treatment propensity) could potentially induce significant bias. However, mathematically this is highly improbable and our simulations confirm this.[10]

Referring back to the earlier question:

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W, X) = 0,$$
$$W = \gamma(X) + V, \qquad \mathbb{E}(V|X) = 0,$$

in the context where covariates are completely uninformative for both the outcome (YX) and treatment propensity models (WX), in expectation the relationship of the covariates on the outcome would be zero, $\mathbb{E}(\omega(X)) = 0$, and the treatment propensity would simply be the proportion of units treated, $\mathbb{E}(\gamma(X)) = n_1/n$, where n_1 is the number of observations in the treatment group and n the total number of observations. Consequently, this equation can be simplified to:

$$Y = \theta W + \epsilon, \quad \mathbb{E}(\epsilon|W) = 0, \tag{3.5}$$
$$W = n_1/n + V, \qquad \mathbb{E}(V) = 0,$$

Equation C.1 illustrates that, in the extreme cases where covariates are wholly unrelated to both the outcome and treatment, we are essentially recovering a difference-in-means estimate. Even more so, our simulations confirm this across a wide variety of scenarios. This is all to say that the DRML framework essentially encapsulates the standard difference-in-means calculation used in RCTs when covariates are wholly uninformative. Critically, as long as one component of DR estimation is correctly specified, either the treatment propensity or outcome model, then we are guaranteed unbiased effect estimates. In contexts where covariates are uninformative, nonparametric ML estimators generate predicted treatment probabilities centered around n_1/n. Thus, DRML

[10] We will note here that there are some scenarios, approximately less than 1% of all cases, *and only* in sample sizes of 50 and 100, that DRML estimation introduces marginal bias, as compared to difference-in-means.

treatment effect estimates are unbiased and generally share the properties of the difference-in-means estimator.[11]

Evaluating the Performance of DRML with Monte Carlo Simulations

In this section we report the results of our extensive set of simulations. These simulations are intended to a) be a tough test of the ability of CF and DRML to outperform standard estimation procedures and b) be illustrative of the most likely scenarios researchers will face when analyzing cross-sectional data. Critical to note is that, due to time and computational constraints, we use default, untuned CF models across all of our simulations. Consequently, the significant gains that we report here should be interpreted as a lower-bound, in that well-tuned models should perform even better than untuned models at ATE estimation and particularly CATE detection.

Specifically, we run three series of simulations. The first focuses on ATE estimation across a variety of common research scenarios. The second series examines the statistical power achieved by CF as compared to standard estimation approaches. We examine both a relatively simple and a relatively complex DGP, where we would expect CF to provide more benefits in the latter—and fewer in the former—scenario. And finally, the final series of simulations covers CATE estimation across four categories of DGPs, including scenarios where OLS should perform admirably (e.g., one-dimensional, linear CATEs) and where it is more likely to have issues (e.g., two-dimensional, nonlinear CATEs).

In total we compare five estimators across these simulations: AIPW, overlap-weighting (ATO/OW), difference in means, regression adjustment based on balance statistics, and a regression-adjustment approach often called the Lin estimator (Lin 2013).[12] For each set of simulations we note what estimators are used/compared as well as how they are evaluated (e.g., power and bias). Finally, we also note the the general parameters for the data-generating-processes (DGPs) in each subsection. For interested readers, we detail the specific algorithms for the simulations in the Appendix.

[11] Importantly, as we note below, researchers can test whether treatment assignment is independent of covariates (WX). If estimated treatment propensities are centered—and especially if they are normally distributed—around n_1/n, then one can be reasonably confident that randomization was accomplished.

[12] At a high-level the Lin adjustment works by 1) centering/de-meaning all covariates, then 2) interacting all covariates with the treatment variable, and finally 3) regressing the outcome variable on the treatment, these centered covariates, and all of the covariate-treatment interaction terms.

Average Treatment Effects For the first series of simulations, we focus on ATE estimation and thus motivate these simulations by adapting Figure 3.3 into Figure **??**. Specifically, Figure **??** reports the four research contexts and a) their three sets of corresponding simulations (two for the bottom cells and another for the top row) and b) how DRML performs in said simulations, as compared to SPEs.

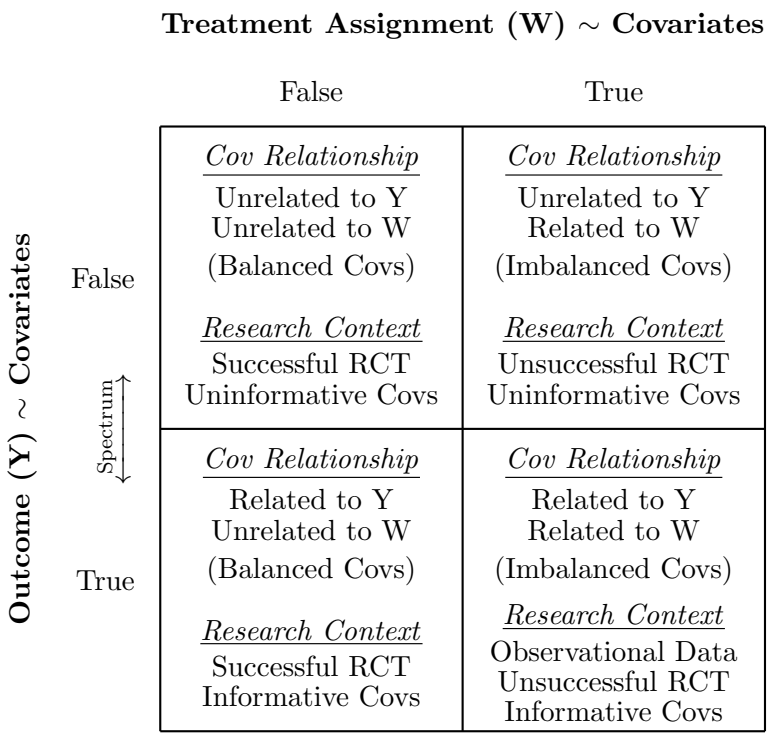


FIGURE 3.3. Stylized Research Context Framework

In the top row, even if covariates are related to treatment assignment (top-right cell), simple bivariate regressions or even difference-in-means between treatment groups (assuming a binary treatment) are sufficient to recover the true ATE, assuming power requirements are met. In general, because the covariates are unrelated to the outcome, there can be no gains from including said covariates in any analyses, including regression adjustments.

In the lower row, however, where the outcome is related to covariates, there are significant potential benefits from adjustments. Even when covariates are not related to treatment assignment (or imbalanced between treatment and control groups in an experiment), if you can accurately model the relationship between covariates and the outcome there are significant potential gains

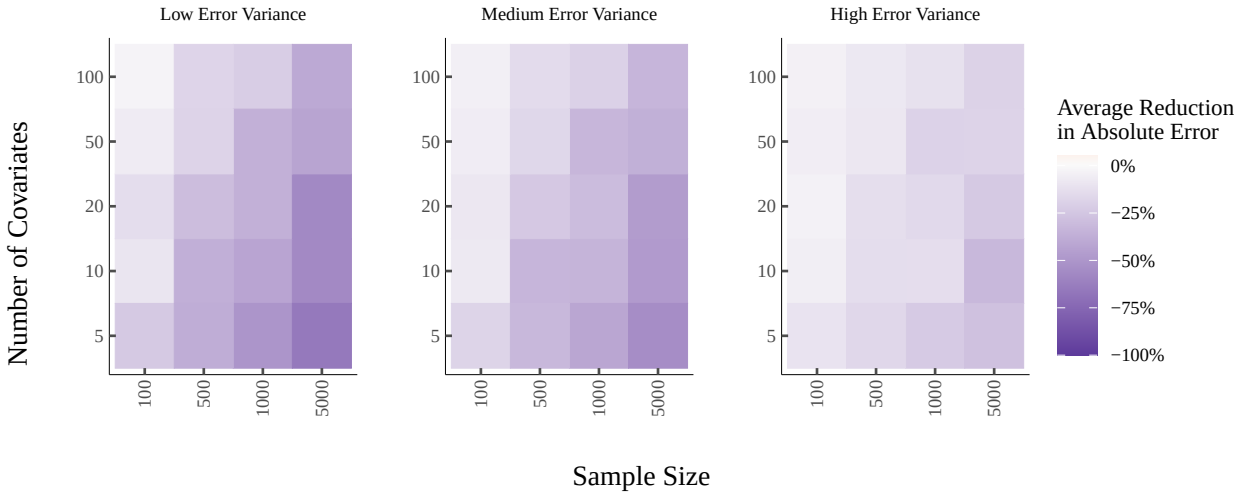


FIGURE 3.4. ATE Results: Standard Experimental Simulations
Note: Cells represent the average reduction in absolute error of the DRML estimated ATEs versus unadjusted difference-in-means, regression-adjusted, and Lin-adjusted effects.

to both precision and coverage. These benefits arise from the reduction in variance achieved by explicitly modeling the outcome variable.

Standard Experimental Context Here, our simulations attempt to approximate a standard experimental context where treatment assignment was successfully randomized but covariates are informative for predicting the outcome (the bottom left panel of Figure **??**). The vast majority of social science experiments will be similar to this scenario. Specifically, we simulate 480 sets of 2,000 simulations (960,000 total tests) over a wide range of simulation parameters, listed in Table 3.1. The effect size for all simulations is 1 and treatment assignment is simple complete randomization. The specific DGP is explained by Algorithm 1 in the Appendix, but note that there is no contamination in these simulations (treatment assignment is purely random, so `contamination = FALSE`).

TABLE 3.1. DRML Performance Test Simulation Parameters

Parameter	Possible Settings
Sample Size (n)	100, 500, 1000, 5000
Number of Covariates (k)	5, 10, 20, 50, 100
Heterogeneous Effects (`heterogeneity`)	True or False
Error Variance (σ^2)	Low (1), Medium (5), High (10)
Linear Effects (`linear`)	True or False
Interactions (`interaction`)	True or False

We evaluate performance using two metrics, reduction in absolute and standard error, averaging each metric across all simulations within each set. This results in 1,440 comparisons for each metric: 480 per method ($480 \times 3 = 1{,}440$). Specifically, our results are summarized in Figure 3.4 by the number of covariates (Y-axis), sample size (X-axis), and error variance (panel). Each grid location represents the average absolute error reduction of all our comparisons across all other parameter combinations.

These results are broken down even further in Figure 3.5, which displays the performance of all four estimators in terms of coverage (Panel 1), average absolute bias (AAB; Panel 2), standard error length (Panel 3), and AIPW's relative percentage reduction average absolute error (Panel 4).[13] In terms of coverage, all four estimators achieve advertised, or close-to advertised, coverage rates across sample sizes with the exception of the Lin estimator, which struggles in the smallest sample size (100 observations).[14] DRML (AIPW) performs best in all settings and reduces the average length of standard errors while maintaining coverage. For example, in simulations with $5,000$ observations, DRML standard errors are on average 40% shorter relative to difference-in-means across all other simulation settings. Additionally, we find that AIPW's relative reduction in AAB is large and increasing with sample size.[15]

Overall, we find that the DRML recovers vastly more fine-grained, accurate effect estimates across nearly every single set of simulations while maintaining proper coverage. DRML is only outperformed in 19 out of 1,440 comparisons, merely $\sim$1% of the time (the specific instances are reported in the Appendix.). Importantly, these 19 instances are largely idiosyncratic parameter combinations and exist exclusively in extremely small samples (100 observations).

Systematic Imbalance In this section, we present simulations to demonstrate the performance of DRML estimators in circumstances where systematic imbalance, and resulting confounding, is present (the bottom right panel of Figure **??**). To be clear, these are circumstances where treatment assignment is contaminated by covariates and those same covariates simultaneously affect the

[13] We calculate the relative percentage reduction within each sample size by subtracting each other estimator's AAB from AIPW's AAB and dividing this by AIPW's AAB.

[14] This difficulty arises from degrees of freedom issues when the number of covariates is large relative to the sample size, for further detail on the coverage results see the Appendix.

[15] The slight exception to this is Lin's very poor performance in the 100 observation simulations. This performance is low for Lin because of issues with very low degrees of freedom, particularly when there are many covariates. In fact, Lin cannot be used when there are 50 or 100 covariates in sample sizes of 100 because there are too few observations.

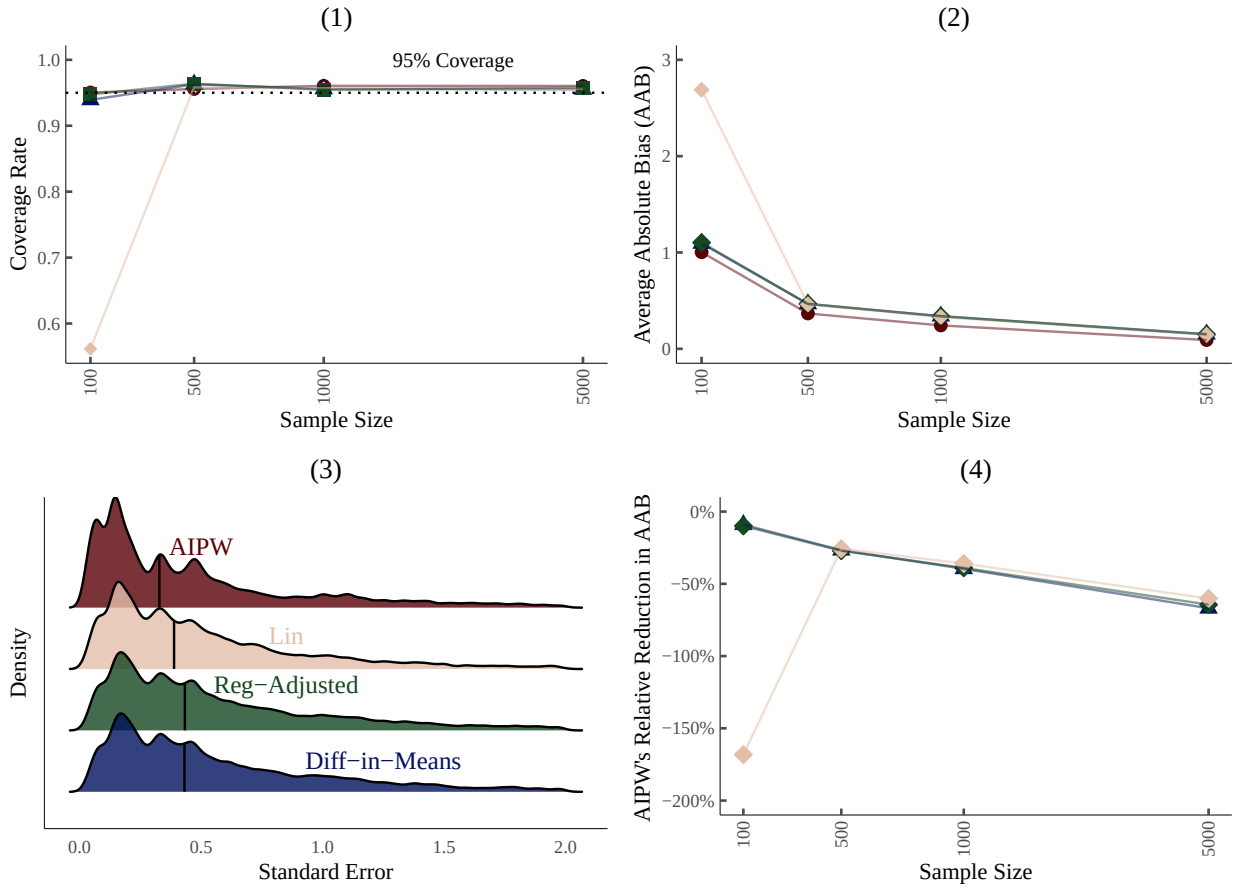


FIGURE 3.5. DRML Performance Test Results

outcome, inducing confounding. The simulation parameter settings follow those above in Table 3.1, but treatment assignment is also determined by the same covariates that also predict the outcome. More information about this nonrandom assignment (where `contamination = TRUE`) can be found in Step 5 of Algorithm 1 in the Appendix.

In all our simulations in this section there is systematic imbalance over either a single dimension or more complex, multidimensional and interactive imbalance. As above, we compare the performance between difference-in-means, regression adjustment via balance statistics, the Lin estimator, and AIPW (DRML) in Figure 3.6.[16] In panel (1) we report coverage rates and in panel (2) average absolute bias, each across sample sizes. In panel (3) reports the distributions of standard

[16]For DRML we opt to report only the results for AIPW here rather than also including OW given the OW results are essentially identical and difficult to visually differentiate. OW results, however, are available in our replication materials for interested readers.

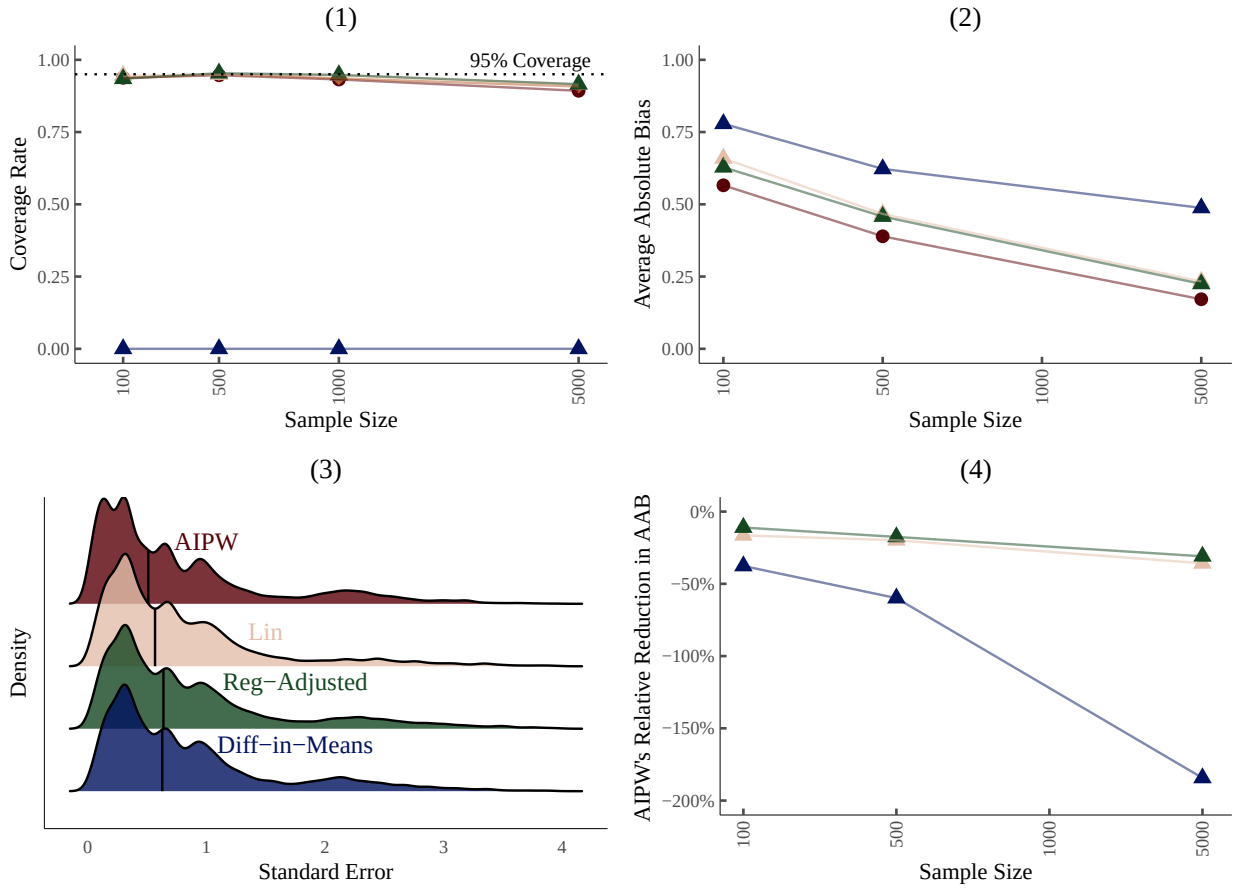


FIGURE 3.6. ATE Results: Systematic Imbalance Simulations

error lengths, and panel (4) reports AIPW's relative percentage reduction in average absolute bias compared to each of the three other estimators.

Throughout all of these simulations, with all simulation parameters, DRML outperforms other estimators when systematic imbalance is present. Average absolute bias is dramatically lower, especially compared to unadjusted difference-in-means that fails to account for the imbalance. Standard error lengths are slightly shorter for DRML as well, and coverage rates are close to advertised rates. However, all three estimators fail to fully reach 95% coverage in some simulation settings and of course the simple difference-in-means estimator is unable to reach desired coverage rates given systematic imbalance.

Uninformative Covariates One reasonable concern is that employing any estimator, including DRML, that attempts to make adjustments based on covariates may introduce bias through these adjustments. Consequently, in this section we compare the performance of three estimators (AIPW,

OW, and difference-in-means) for average treatment effects in the unlikely circumstance where *all measured covariates are unrelated to the outcome* (the upper row in Figure **??**).

The simulation settings are presented in Table 3.2. We add two new parameters in these simulations, first to vary the fraction of treated units and second the magnitude of the treatment effect. In all cases, the treatment effect is homogeneous and the number of pretreatment covariates is five. Each unique simulation setting is repeated 100 times, resulting in a total of 27,000 simulations.

TABLE 3.2. Uninformative Covariates Simulation Parameters

Parameter	Possible Settings
Sample Size (n)	50, 100, 200, 500, 1000, 2000
Fraction Treated (`treatfrac`)	10%, 30%, 50%, 70%, 90%
Treatment Effect Magnitude (τ)	Small (0.5), Medium (1), Large (2)
Signal-to-Noise Ratio (`SNR`)	Low (1), Medium (2), High (5)

The comparison between DRML (both AIPW and overlap weighting estimators) and difference-in-means is presented in Figure 3.7. Panel (1) reports the coverage rates across sample size, panel (2) the average absolute bias across sample sizes, panel (3) reports the distributions of standard error lengths, and panel (4) reports difference-in-means' relative reduction in average absolute bias compared to AIPW and overlap weighting, respectively.[17]

As we suggested above, when covariates are unrelated to both the outcome and treatment assignment, DRML estimators essentially reduce to difference-in-means as the sample size increases. However, in small samples (50–200), AIPW does increase bias relative to difference-in-means. This is a reflection of two factors: 1) we are forcing DRML to estimate a propensity model when the propensity score is known by design, and 2) the fact that AIPW divides by extreme propensity scores in situations with extremely few treated or untreated units (e.g., most treatment propensities are close to zero or one). As panel (4) demonstrates, this problem can easily be addressed in small samples by employing the ATO/OW estimator that avoids this step, or alternatively inputting a known propensity score if possible. Despite the increase in bias, AIPW maintains proper coverage across all sample sizes, whereas difference-in-means (and overlap weighting) only reach proper coverage in sample sizes of 2,000. These differences in coverage are likely a result from differences

[17] We calculate the relative percentage reduction within each sample size by subtracting each other estimator's AAB from difference-in-mean's AAB and dividing this by difference-in-mean's AAB.

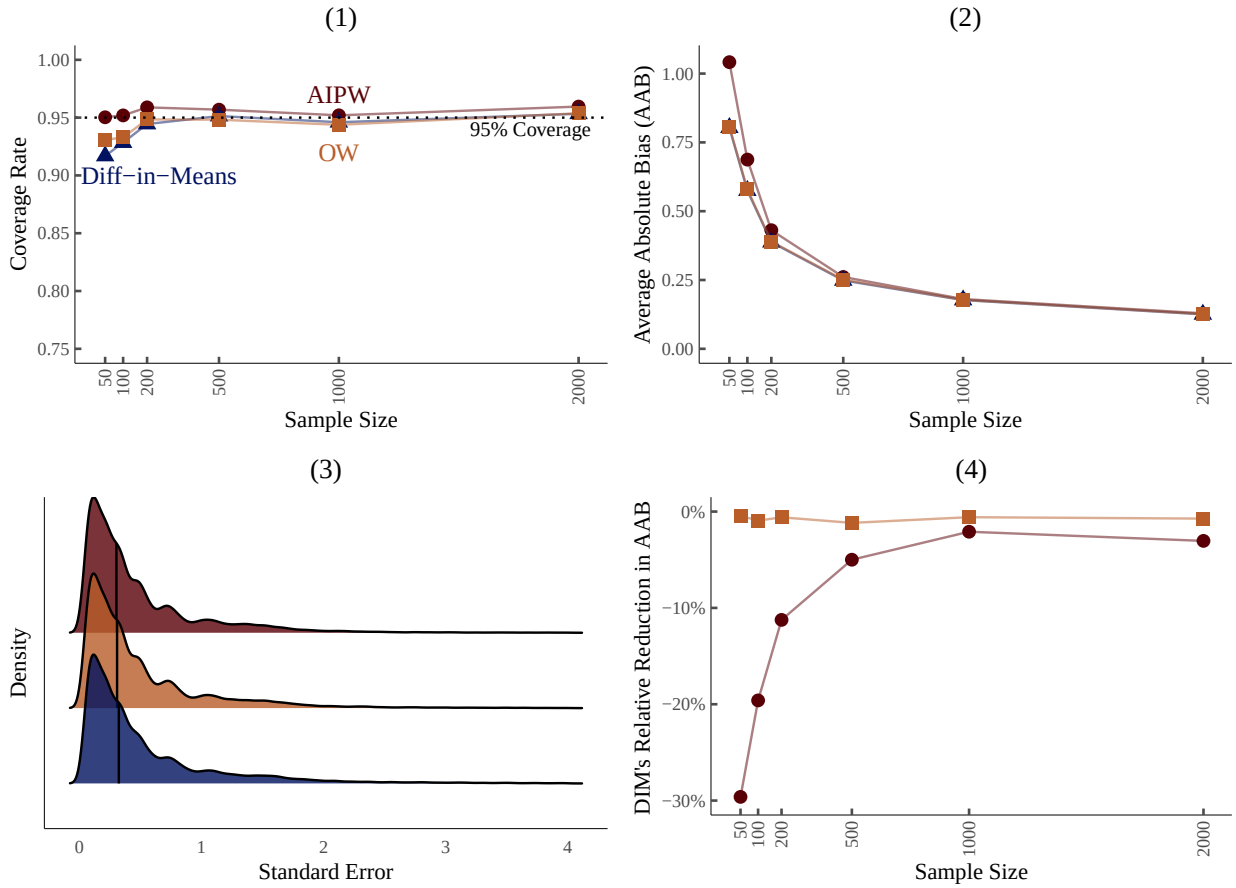


FIGURE 3.7. ATE Results: Uninformative Covariates Simulations

in the construction of the standard errors for AIPW/CF relative to ATO/OW and difference-in-means.

Again, it should be stressed that this simulation is an exceedingly tough test of the performance of AIPW/DRML since *no* covariates are related to either the treatment or outcome **and** we are forcing DRML to model a known propensity distribution. In realistic settings the probability that all covariates are completely uninformative is essentially zero (given that we would hope researchers would be including at least some amount of established or theoretically motivated covariates). Additionally, if a researcher was faced with this situation (i.e., they are analyzing a small-n RCT) they could, and should, specify the treatment propensity for all observations (e.g., 0.5 in a simple binary treatment). Specifying the known treatment propensity would reduce the bias in AIPW to essentially zero. Overall, despite the difficulty of these tests for AIPW, these results suggest that

DRML performs at essentially the same level as difference-in-means in contexts in which we would be most concerned about its performance.

Statistical Power In this section we investigate the power requirements for DRML estimators versus other popular choices for average treatment effects. We do this both in the standard experimental context, with a relatively simple linear additive DGP, and also in a high dimensional, complex setting where the benefits of an ML-powered outcome model are likely to be stark. In the low dimensional setting, we estimate the outcome models using simple honest random forests and in the high dimensional setting we switch to a boosted variant of random forest.[18]

The high dimensional Monte Carlo simulation settings are listed in Table 3.3. The effect size (1) is constant across parameter settings. We repeat each unique combination setting 2,000 times, which yields a total of 32,000 simulations.

TABLE 3.3. High Dimensional Simulation Parameters

Parameter	Possible Settings
Treatment Effect Magnitude (τ)	Small (0.5), Medium (1), Large (2)
Signal-to-Noise Ratio (`SNR`)	Minuscule (7), Low (5), Medium (3), High (1)
Sample Size (n)	25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400, 500, 750, 1000, 2000, 5000

We report the comparisons for the low-dimensional simulations, the same that are reported in Figure 3.4 and generated by the parameters in Table 3.1, in Figure 3.8 and the high-dimensional simulations in Figure 3.9. In each figure, panels reflect a given signal-to-noise ratio (higher meaning that the effect in relation to random noise is larger).

In the simple data-generating processes (Figure 3.8), DRML outperforms SPEs across all sample sizes and signal-to-noise ratios. For the more complicated simulations (Figure 3.9), DRML vastly outperforms difference-in-means across all sample sizes (though performance converges at sample sizes of 2,000 and 5,000, depending on the signal-to-noise ratio). When compared to the Lin estimator, DRML is either superior (in high and medium ratios) or equivalent (in low and minuscule ratios). Again, these comparisons suggest that DRML provides lower, or at worst equal, power requirements compared to standard approaches.

[18] As in all our simulations, both algorithms and splitting procedures are implemented by `grf`, and default settings and tuning procedures are used given computational restraints.

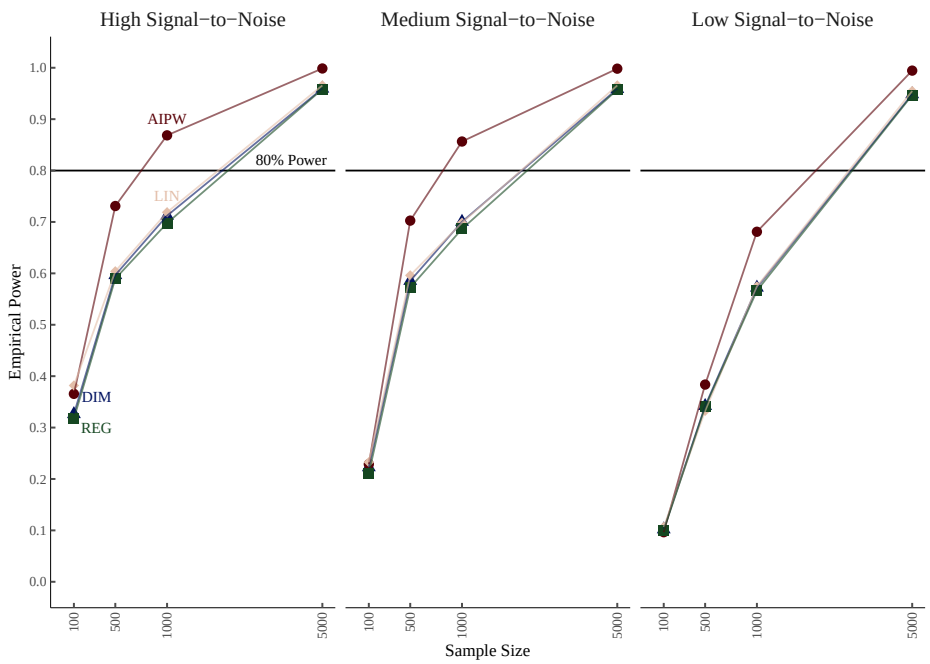


FIGURE 3.8. Power Results: Low-Dimensional Simulations

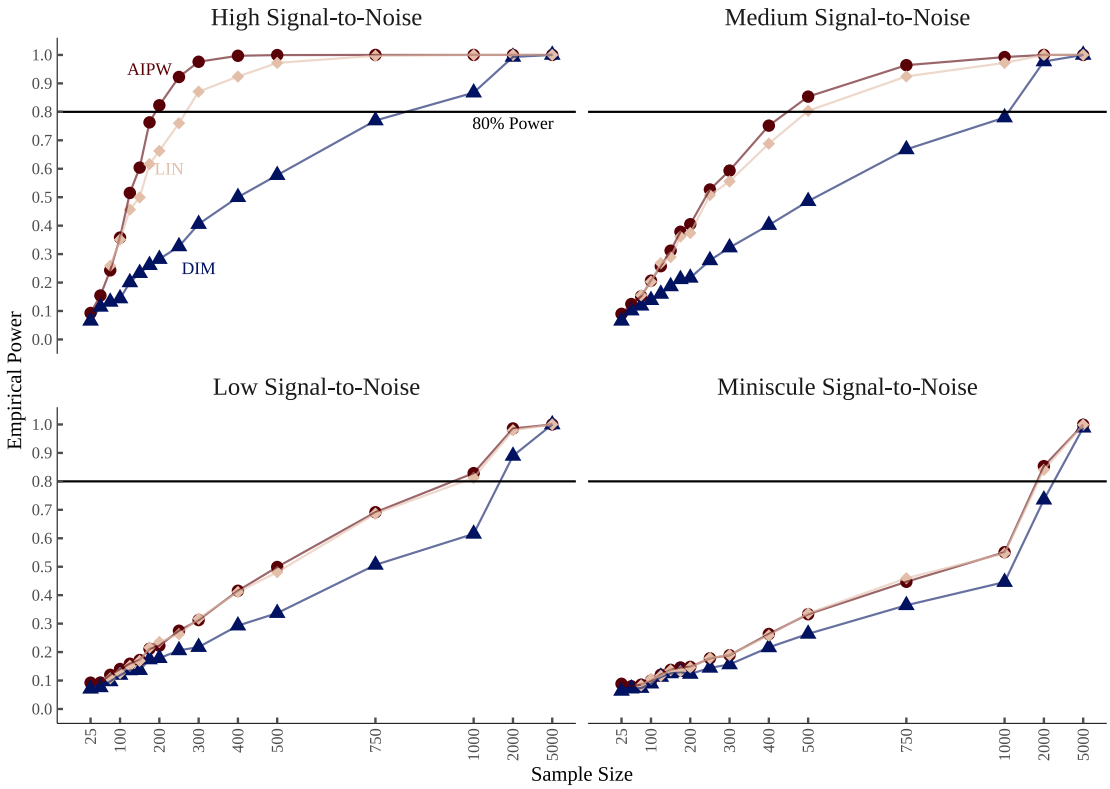


FIGURE 3.9. Power Results: High-Dimensional Simulations

Conditional Average Treatment Effects In this section we execute another set of simulations to compare the performance of CF and saturated interactive OLS regression models for recovering CATEs. The purpose of these simulations is to compare CF to the simple interactive regression approach that is common in applied research. Looking at Table 3.4, the DGPs for these simulations vary over sample size, number of predictors, signal-to-noise ratio, the presence or lack of treatment effect heterogeneity, linearity of treatment effects, and the dimensionality of the CATE function. If heterogeneity is present, it can be linear and simple such that it would be well represented by a simple linear regression, or multidimensional and nonlinear, such that a simple linear regression will be misspecified. Each unique setting is repeated 100 times resulting in a total set of 51,200 total simulations.

TABLE 3.4. CATE Simulation Parameters

Parameter	Possible Settings
Signal-to-Noise Ratio (`SNR`)	Minuscule (1), Low (3), Medium (5), High (7)
Sample Size (n)	500, 1000, 2000, 5000
Number of Predictors (p)	5, 10, 15, 20
Heterogeneity Present? (`heterogeneity`)	True, False
Dimension of CATE (`dimension`)	1D, 2D
Nonlinear Effects (`linear`)	True, False

For each simulation, we execute both a CF and a saturated OLS model. We then proceed by assessing whether or not heterogeneity is detected. For CF, we use the default check, the best linear fit test (described above), and for the saturated regression, we check if any of the treatment interaction terms are statistically significant. First, we examine the results for simulation settings where no treatment effect heterogeneity is present in the true DGP. Table 3.5 reports the false positive rates (FPR) across the two estimators for these simulation runs.

A few observations stem from these results. First, the FPR for saturated OLS is remarkably high across all settings, indicating that at least one interaction term is statistically significant at the standard threshold ($p \leq .05$). This result underscores the need for some form of multiple testing adjustment, or out-of-sample validation, in CATE modeling. Second, the FPR for CF is quite low and always at or below 5%. This illustrates the importance of omnibus tests for treatment effect heterogeneity that eschew simple checks for statistical significance.

	Sample Size							
	500		1000		2000		5000	
S-N Ratio	CF	OLS	CF	OLS	CF	OLS	CF	OLS
High	0.75%	50.0%	1.94%	51.7%	2.69%	53.4%	4.12%	53.5%
Medium	2.94%	45.6%	1.69%	45.0%	4.44%	48.4%	4.75%	46.4%
Low	2.25%	43.9%	1.81%	45.1%	4.31%	46.7%	4.62%	46.8%
Miniscule	2.00%	43.1%	2.69%	43.2%	4.62%	45.6%	5.00%	45.0%

TABLE 3.5. Comparison of False Positive Rates (FPR): No Heterogeneity Present

Finally, we report the comparative performance, as measured by mean absolute error, of DRML (using CF) and a saturated OLS in Figure 3.10. These results are broken down by the data-generating process (panel), sample size (columns in each panel), signal-to-noise ratio (rows in each panel), and method (CF in red, OLS in blue). Both estimators perform better as sample size and SNR increase, but CF dominates the saturated models across the parameter space and is strongest at detecting nonlinear and multidimensional heterogeneity. Even more impressive, this *untuned* model detects and estimates these complexities without the manual specification of squared, or even three-way, interaction terms.

Overall, these simulations demonstrate the superiority of the DRML two-step approach: 1) assess whether treatment effect heterogeneity is present whatsoever and *if and only if* heterogeneity is detected 2) estimate CATEs for important variables. The alternative, estimating a simple interaction regression model on the entire sample and simply checking the statistical significance of interaction terms, is an unsound approach for identifying and estimating conditional effects.

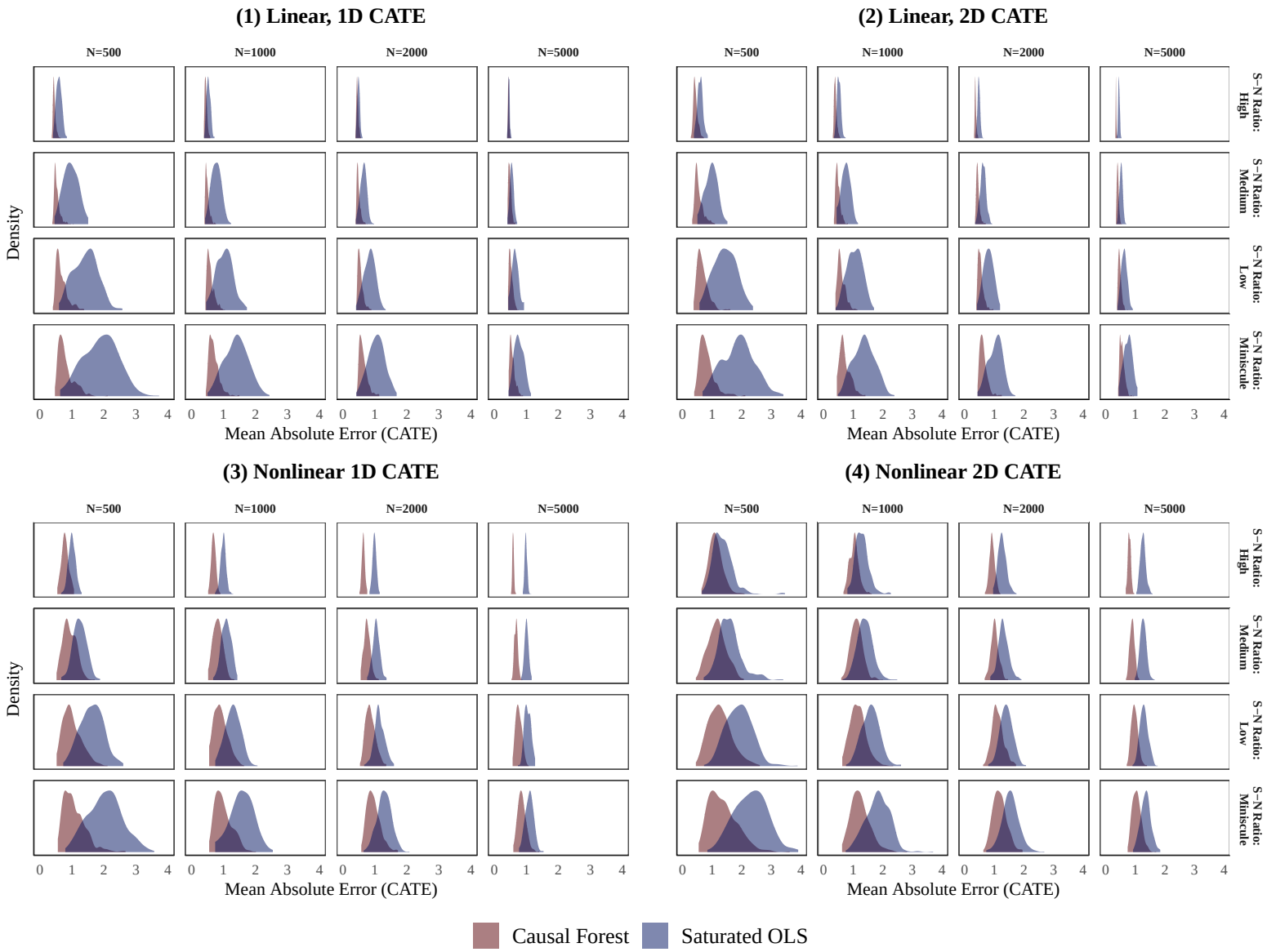


FIGURE 3.10. CATE Recovery Results: Simulations Across Sample Size, DGP, and Signal-to-Noise Ratio

Discussion and Future Work

Parametric methods—like linear and logistic regressions—have long been standard tools for social science research. While there has been significant growth in the use of causal inference designs over correlational regressions, these approaches are limited to research questions where there is some source of exogenous variation or shock that enables causal identification. Furthermore, parametric methods are frequently employed to analyze observational data where causal identification is murky, or even impossible, and even more often to analyze experimental data where identification is guaranteed. While this latter group of methods perform admirably on many types of data, they have well-founded and well-known issues like the curse of dimensionality, high researcher's degrees of freedom, and inflexible functional form assumptions (among others).

To address these issues, many researchers have turned to semi- and non-parametric methods which in turn have led to the steady growth of causally-oriented machine learning methods, like causal forest or Bayesian causal forest (Hahn, Murray, and Carvalho 2020). Critically, these methods have focused on providing statistical guarantees to these machine learning methods, such that researchers can be much more confident in their results. However, the novelty and nonparametric nature of these approaches—along with the intuitive tension between causal inference and black box ML—has limited the spread and use of causal ML. Furthermore, to date there has been no systematic, simulation-based test of causal forest in comparison to standard parametric methods.

This paper provides the most complete evidence to date of causal forest's significant utility over standard methods. To do this, we first introduce the concept of doubly robust machine learning and provide a clear explanation of its logic and how causal forest fits into this framework. And second, we use nearly two-million simulations to test CF's performance compared to difference-in-means, regression adjustment, and the Lin estimator. These simulations tested the comparative performance of these methods across: 1) ATE accuracy when covariates are either informative or uninformative; 2) ATE accuracy under nonrandom treatment assignment (systematic imbalance); 3) statistical power for recovering ATEs; and 4) CATE accuracy including false positive and negative rates.

Conceptually, DRML is intuitive, powerful, and—critically—has strong statistical guarantees for recovering accurate and consistent treatment effects. Additionally, DRML, generally, and causal forest, specifically, is able to test for and model treatment effect heterogeneity in a far more holistic

and principled way than parametric approaches. This, along with DRML's performance in our simulations, has significant implications for social science research.

First, causal forest should be used in almost every experimental analysis, particularly when it is likely that there are informative covariates and especially when estimating conditional effects. With informative covariates, CF is more accurate, has higher coverage rates, and requires fewer observations for the same level of statistical power than standard methods. This fact should weigh even more heavily in light of the widespread issue of low statistical power in social science experiments (Arel-Bundock et al. 2025). Even when there are no informative covariates, DRML performs almost identically to difference-in-means. Again, as we mention above, this scenario is highly unlikely in social science research. Between these two scenarios CF either provides significant benefits for estimating ATEs, or performs just as well as standard methods.

Furthermore, CF is overwhelmingly superior to standard methods at estimating CATEs: CF is far more accurate, particularly for simulations with nonlinear and/or multidimensional CATEs and those with low signal-to-noise ratios. Even more importantly, CF has a false positive rate between *9 to 66 times* smaller than the standard, interactive OLS approach. These results have serious implications for any CATEs estimated using standard approaches. Simply put, researchers *should not* use standard methods for CATE estimation, as the threat of false positives is severe.

Second, DRML and CF represent a marked improvement over standard approaches for analyzing observational data without clear causal identification. Across all of our simulations with nonrandom assignment (systematic imbalance) CF performs better than difference-in-means, regression adjustment, and the Lin estimator in terms of absolute bias and standard error lengths and has nearly equivalent coverage.

While in this paper we focus on forest-based DRML approaches (and particularly CF), this framework easily incorporates other approaches for modeling either the outcome variable or treatment propensity. As such, one could imagine employing, for example, a kernel-regularized least squares regression (Hainmueller and Hazlett 2014) for the outcome and covariate balancing (Imai and Ratkovic 2014) for the propensity scores. Our recommendation here then is for using the DRML framework generally, rather than CF in particular, when faced with observational data that does not fit clearly into a causal identification strategy.

There are also a wide variety of promising avenues for future work as the causal machine learning literature evolves inside and outside of political science. Most obviously, the DRML approach could be expanded to address panel data, considering issues such as systematic attrition. Another outstanding question is exactly how researchers should conceptualize and implement model tuning and algorithmic selection for DRML. In other words, how exactly should researchers a) choose which models to employ for their particular research design and b) tune the selected model in a systematic and principled way.

Moreover, there are many competing measures of variable importance (e.g., permutation importance, Shapley values, etc.) that may be useful for causal tasks, the validity and usefulness of these competing approaches is unclear. For example, recent work on variable importance has focused on addressing multicollinearity (Bénard and Josse 2023; Verdinelli and Wasserman 2024). Future research should also work to create methods that are effective at detecting important interactions. Finally, while the causal ML literature has focused deeply on supervised machine learning (predicting an outcome), there has been little to no work on unsupervised ML (also known as scaling or dimensional analysis). This concern regarding measurement is critically underappreciated, as no prediction method can fully address poor measures; the old adage holds true: "garbage in, garbage out."

Bibliography

Altman, Naomi, and Martin Krzywinski. 2018. "The curse(s) of dimensionality." *Nature Methods* 15 (6): 399–400.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D Stanley. 2025. "Quantitative political science research is greatly underpowered." *Journal of Politics,* https://doi.org/https://doi.org/10.1086/734279.

Arkhangelsky, Dmitry, and Guido W Imbens. 2022. "Doubly robust identification for causal panel data models." *The Econometrics Journal* 25 (3): 649–674.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–1178. https://doi.org/10.1214/18-AOS1709. https://doi.org/10.1214/18-AOS1709.

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." *Observational Studies* 5 (2): 37–51.

Bénard, Clément, and Julie Josse. 2023. "Variable importance for causal forests: breaking down the heterogeneity of treatment effects." *arXiv preprint arXiv:2308.03369.*

Blackwell, Matthew, and Michael P Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30 (4): 495–514.

Chen, Xiaohong, Ying Liu, Shujie Ma, and Zheng Zhang. 2024. "Causal inference of general treatment effects using neural networks with a diverging number of confounders." *Journal of Econometrics* 238 (1): 105555.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2023. *Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.* Working Paper, Working Paper Series 24678. National Bureau of Economic Research. https://doi.org/10.3386/w24678. http://www.nber.org/papers/w24678.

Dorie, Vincent, George Perrett, Jennifer L Hill, and Benjamin Goodrich. 2022. "Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning." *Entropy* 24 (12): 1782.

Green, Jon, and Mark H. White II. 2023. "Machine Learning for Experiments in the Social Sciences." In *Elements in Experimental Political Science.* Cambridge University Press. https://doi.org/10.1017/9781009168236.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho. 2020. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." *Bayesian Analysis* 15 (3): 965–1056.

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach." *Political Analysis* 22 (2): 143–168.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* 27 (2): 163–192.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76 (1): 243–263.

Knaus, Michael C, Michael Lechner, and Anthony Strittmatter. 2021. "Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence." *The Econometrics Journal* 24 (1): 134–161.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7 (1): 295–318. https://doi.org/10.1214/12-AOAS583. https://doi.org/10.1214/12-AOAS583.

Mutz, Diana C, Robin Pemantle, and Philip Pham. 2019. "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data." *The American Statistician* 73 (1): 32–42.

Negi, Akanksha, and Jeffrey M Wooldridge. 2021. "Revisiting regression adjustment in experiments with heterogeneous treatment effects." *Econometric Reviews* 40 (5): 504–534.

Ratkovic, Marc. 2021a. "Subgroup Analysis: Pitfalls, Promise, and Honesty." In *Advances in Experimental Political Science,* edited by James N. Druckman and Donald P. Green, 271–288. Cambridge University Press. https://doi.org/10.1017/9781108777919.020.

———. 2021b. "Subgroup Analysis: Pitfalls, Promise, and Honesty." In *Advances in Experimental Political Science,* edited by James N. Druckman and Donald P.Editors Green, 271–288. Cambridge University Press. https://doi.org/10.1017/9781108777919.020.

Robins, James M, and Andrea Rotnitzky. 2001. "Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'." *Statistica Sinica* 11 (January): 920–936.

Robinson, Peter M. 1988. "Root-N-consistent semiparametric regression." *Econometrica: Journal of the Econometric Society,* 931–954.

Semenova, Vira, and Victor Chernozhukov. 2021. "Debiased machine learning of conditional average treatment effects and other causal functions." *The Econometrics Journal* 24 (2): 264–289.

Verdinelli, Isabella, and Larry Wasserman. 2024. "Feature importance: A closer look at shapley values and loco." *Statistical Science* 39 (4): 623–636.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Discussion

Discussion & Future Work

These three chapters have painted a picture of the Government Accountability's response to the Republican Revolution of 1994 and subsequent majority party turnover in the House and Senate, as well as outlining the potential utility of causal machine learning methods for social scientists. Substantively, one obvious unanswered question this dissertation raises: why is attacking Congressional capacity popular in the mass public in the first place? What are the behavioral roots of capacity attacks? The previous literature, including this dissertation, explores institutional explanations for attacks of capacity (i.e., the supply side), but often belies these behavioral questions, merely taking for granted the unpopularity of capacity increases and the institution more generally (i.e., the demand side). Understanding the political economy of capacity attacks is an issue that should concern both scholars and reformers, as future efforts to buttress the Legislative State with additional staff, funding, etc. are likely to run headlong into members incentives to attack capacity. Moreover, existing efforts to reform Congress' institutional regime tend to focus on continuing to *restrict* the behavior of members and their staff, for example limiting their ability to hold assets, pursue related jobs after office, etc. Putting aside the merit of these policies, they continue to push in a restrictionist direction at the same time the macro-political environment has worsened, with some members increasingly fearful for their personal safety in office.

Methodologically, this dissertation presents a variety of causal machine learning methods that lie near the edge of current state-of-the-art approaches for average and conditional effect estimation. The applicability of these methods and best practices for their use are areas of active research that is rapidly evolving. While some researchers are increasingly convinced of the utility of these new methods, and it is now possible to see causal forest or double machine learning applied in top political science journals, there is still widespread disagreement in the social sciences over the utility of these approaches. For example, Morucci and Spirling (2024) argues simple parametric

models are typically superior for social scientists given the limited "intrinsic dimension" of many social science datasets. They argue, in brief, that the noise floor, or explainable variation, in most social science datasets is quite high, limiting the utility of complex models (e.g., neural nets) over simple ones (e.g., ordinary least squares regression). Moreover, obviously simple models do yield relative ease of interpretation relative to approaches that employ blackbox algorithms in some part of the estimation procedure. Of course, this argument is not new. Modern debates regarding the applicability of machine learning for scientific studies follow conceptually from debates in statistics (see, e.g., Breiman (2001)'s two cultures of data modeling).

Future work could explore the veracity of the "low dimension" thesis: is it truly the case that most social science datasets have a high noise floor? One might be skeptical of this claim given recent work that applies machine learning to widely analyzed tabular datasets, extracting increases in predictive performance and subsequent resolution/interpretability of the data generating process (Hare and Kutsuris 2023). Second, one might ask if model complexity is the conceptual frame for understanding new causal machine learning methods? These approaches often feature novel ways to incorporate the predictions of machine learning methods into estimating equations for inference, but the machine learning methods themselves are not necessarily highly complex and bespoke, potentially as simple as a ridge regression for instance. Returns to model complexity differ over curve-fitting algorithms with highly divergent approaches. For example, to extract maximum performance, it is typically necessary to extensively tune a gradient boosted decision tree model and allow the model to learn (i.e., become more complex) over a large number of iterations. On the other hand, some approaches based on simple decision trees see a much more rapid decrease in returns to complexity. For example, in a simple random forest model there are typically fast diminishing returns to additional decision trees, and extensive hyperparameter tuning typically yields only minor performance benefits. The returns to increasing model complexity are simply different between the two approaches. This is also true for neural networks and other methods. One might also wonder whether out-of-the-box neural networks, as employed in Morucci and Spirling (2024), are truly the relevant comparison for simple parametric methods applied to tabular data. This is an especially questionable approach given voluminous Monte Carlo evidence that networks tend to underperform on tabular datasets (Shwartz-Ziv and Armon 2022). [19]

[19]This is true with the exception of the more recent tabular foundation models adapted from Bayesian statistics (see e.g., Hollmann et al. 2022; Robertson et al. 2025.

To summarize, this dissertation asks more questions than it answers and provides a variety of avenues for future substantive and methodological work.

Bibliography

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45:5–32.

Hare, Christopher, and Mikayla Kutsuris. 2023. "Measuring swing voters with a supervised machine learning ensemble." *Political Analysis* 31 (4): 537–553.

Hollmann, Noah, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. "Tabpfn: A transformer that solves small tabular classification problems in a second." *arXiv preprint arXiv:2207.01848.*

Morucci, Marco, and Arthur Spirling. 2024. "Model Complexity for Supervised Learning: Why Simple Models Almost Always Work Best, And Why It Matters for Applied Research." https://arthurspirling.org/documents/MorucciSpirling_JustDoOLS.pdf.

Robertson, Jake, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. 2025. "Do-PFN: In-context learning for causal effect estimation." *arXiv preprint arXiv:2506.06039.*

Shwartz-Ziv, Ravid, and Amitai Armon. 2022. "Tabular data: Deep learning is not all you need." *Information Fusion* 81:84–90.

APPENDIX A

Supplemental Material for Did the Republican Revolution Hamstring Congressional Oversight? A Case Study of the Government Accountability Office

GAO in the existing literature

A small sample of prior work interrogates GAO for a variety of research purposes. For example, the volume of GAO reports has been used to test theories of interbranch politics. Warren (2012) argues that increasing conflict between Congress and the executive results in increased oversight from the GAO (Warren 2012), as operationalized by raw report counts for 33 federal agencies. Unfortunately, as described in the prior section, raw reports counts are a biased and unreliable measure of GAO's productivity, at least in the 1990s. Further, Warren's analysis lumps report types together, including reports that do not make explicit policy recommendations, and ignores changes in staffing and budget at GAO over the period of study, making the results difficult to substantively interpret.

Subsequent work examines the correlates of GAO report volume over time from 1994-2022 using data aggregated by executive agency, suggesting more reports are produced in times of unified government (Kennedy 2023), relying in part on logic from Lee (2016). While the authors do control for staffing levels in their analysis, they also unfortunately employ raw report counts as a measure of agency oversight attention. Selin and Moore (2023) recover direct committee requests for GAO action over time, and suggest committees appear to be requesting from GAO with less frequency. However, committee staff and member offices often make informal information requests, or simply collect material from the support agencies through internal and publicly available means (i.e., by accessing GAO's voluminous archive). These other avenues to support agency material are unobservable, but do represent some percentage of all usage Congressional usage of support agency information. In other words, it might merely be the case that Congressional staff are more likely to request, or access, GAO information through their website or by picking up the phone.

Additional Empirical Results for the RR Effects

Detailed RDD results from main text The tables below present the full RDD results from `rdrobust` presented in the main text figures. For all results, default parameter settings are employed, this includes triangular kernel weights, bandwidth selection using mean squared error (MSE) optimal bandwidth selection, and cluster-robust standard errors via nearest-neighbor (NN) matching. Notably, the bandwidth selection procedure differs between conventional and bias-corrected, yielding (typically) slight numerical differences between the estimates in this case. In the

tables below, p refers to the polynomial order, q refers to the order of the bias-corrected estimator, h refers to the conventional bandwidth using MSE, b refers to the bias-corrected bandwidth that targets coverage over MSE.

TABLE A.1. RDD estimate: Count of GAO Reports

Method	Coef.	Std. Err.	z	$P > \|z\|$	[95% C.I.]
Conventional	0.735	0.209	3.521	0.000	[0.326 , 1.143]
Bias-Corrected	0.839	0.209	4.024	0.000	[0.431 , 1.248]
Robust	0.839	0.219	3.826	0.000	[0.409 , 1.269]

Notes: N = 9,712. Left / Right obs = 3,560 / 6,152. Eff. N = 444 / 439. Order est. $p = 1$, order bias $q = 2$. BW est. $h = 656.024$, BW bias $b = 1761.878$, $\rho = h/b = 0.372$. BW type = mserd. Kernel = Triangular. VCE = NN.

TABLE A.2. RDD estimate: Count of Executive Recommendations

Method	Coef.	Std. Err.	z	$P > \|z\|$	[95% C.I.]
Conventional	-1.770	0.220	-8.029	0.000	[-2.202 , -1.338]
Bias-Corrected	-1.678	0.220	-7.612	0.000	[-2.110 , -1.246]
Robust	-1.678	0.240	-7.000	0.000	[-2.148 , -1.208]

Notes: N = 9,712. Left / Right obs = 3,560 / 6,152. Eff. N = 420 / 418. Order est. $p = 1$, order bias $q = 2$. BW est. $h = 621.253$, BW bias $b = 1441.774$, $\rho = 0.431$. BW type = mserd. Kernel = Triangular. VCE = NN.

TABLE A.3. RDD estimate: Count of Matters for Congress

Method	Coef.	Std. Err.	z	$P > \|z\|$	[95% C.I.]
Conventional	-0.093	0.019	-4.901	0.000	[-0.131 , -0.056]
Bias-Corrected	-0.103	0.019	-5.425	0.000	[-0.141 , -0.066]
Robust	-0.103	0.021	-4.857	0.000	[-0.145 , -0.062]

Notes: N = 9,712. Left / Right obs = 3,560 / 6,152. Eff. N = 487 / 481. Order est. $p = 1$, order bias $q = 2$. BW est. $h = 722.960$, BW bias $b = 1402.006$, $\rho = 0.516$. BW type = mserd. Kernel = Triangular. VCE = NN.

TABLE A.4. RDD estimate: Share of Reports with Executive Recommendations

Method	Coef.	Std. Err.	z	$P > \|z\|$	[95% C.I.]
Conventional	-0.124	0.008	-15.542	0.000	[-0.139 , -0.108]
Bias-Corrected	-0.127	0.008	-15.965	0.000	[-0.143 , -0.112]
Robust	-0.127	0.009	-14.882	0.000	[-0.144 , -0.110]

Notes: N = 9,712. Left / Right obs = 3,560 / 6,152. Eff. N = 398 / 397. Order est. $p = 1$, order bias $q = 2$. BW est. $h = 590.450$, BW bias $b = 1457.759$, $\rho = 0.405$. BW type = mserd. Kernel = Triangular. VCE = NN.

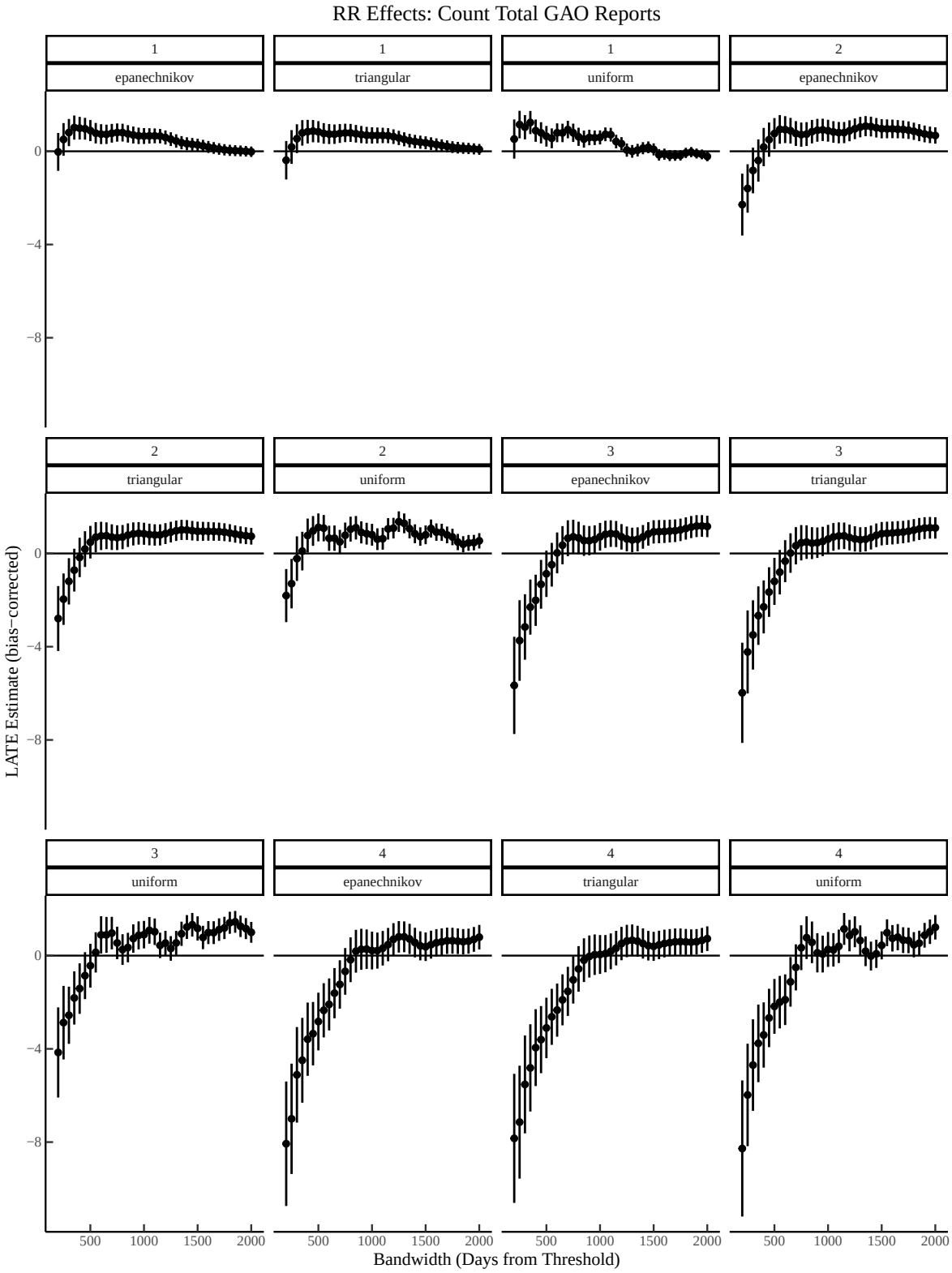
Alternative RDD Model Specifications In this section, the RDD results for each dependent variable and dependent variable measure are displayed over 360 unique combinations of kernels (uniform, triangular, epanechnikov), bandwidths ($bw = seq(100, 3000, 100)$, and polynomial

TABLE A.5. RDD estimate: Share of Reports with Matters for Congress

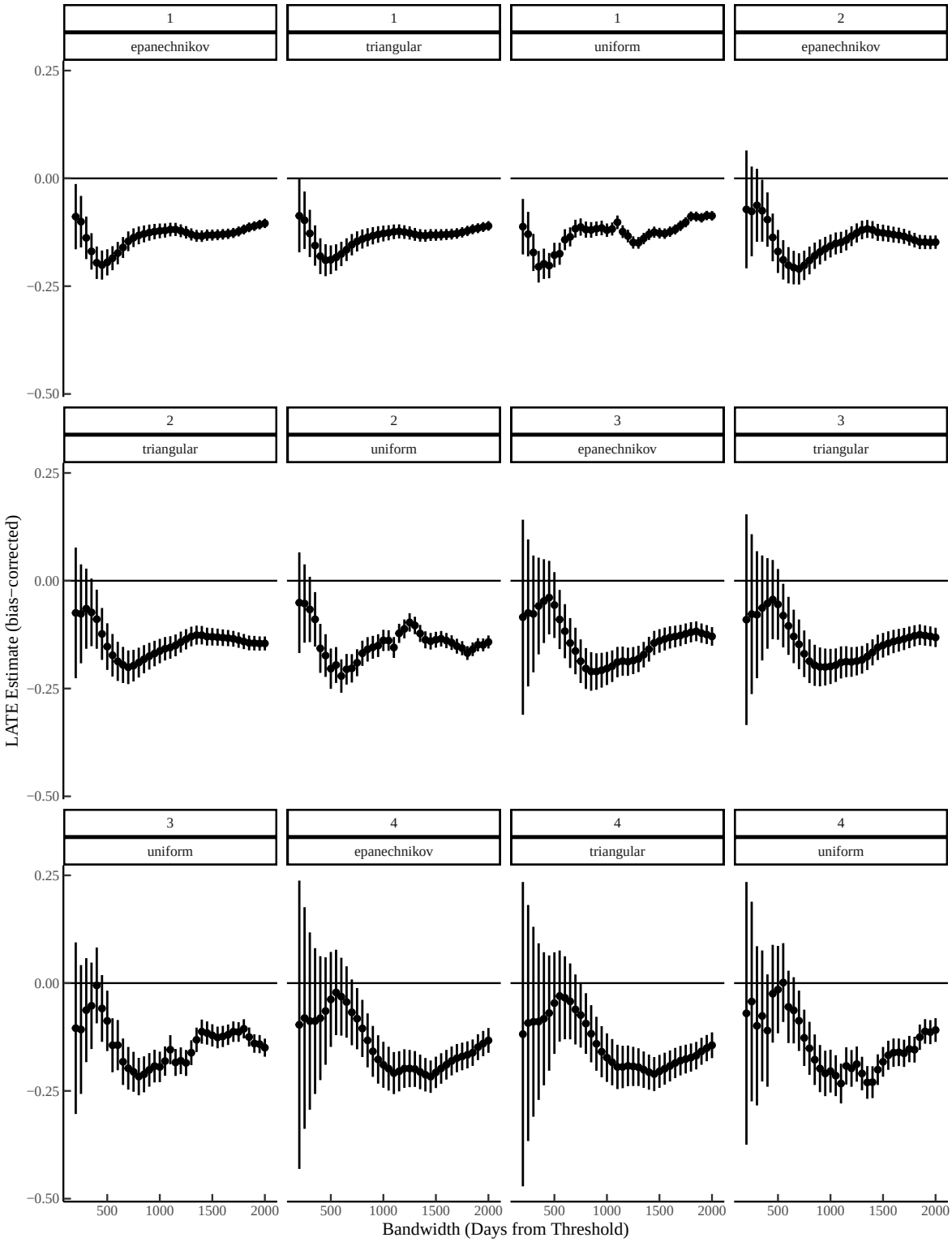
Method	Coef.	Std. Err.	z	$P > \|z\|$	[95% C.I.]
Conventional	-0.035	0.005	-7.694	0.000	[-0.044 , -0.026]
Bias-Corrected	-0.038	0.005	-8.268	0.000	[-0.047 , -0.029]
Robust	-0.038	0.005	-7.785	0.000	[-0.047 , -0.028]

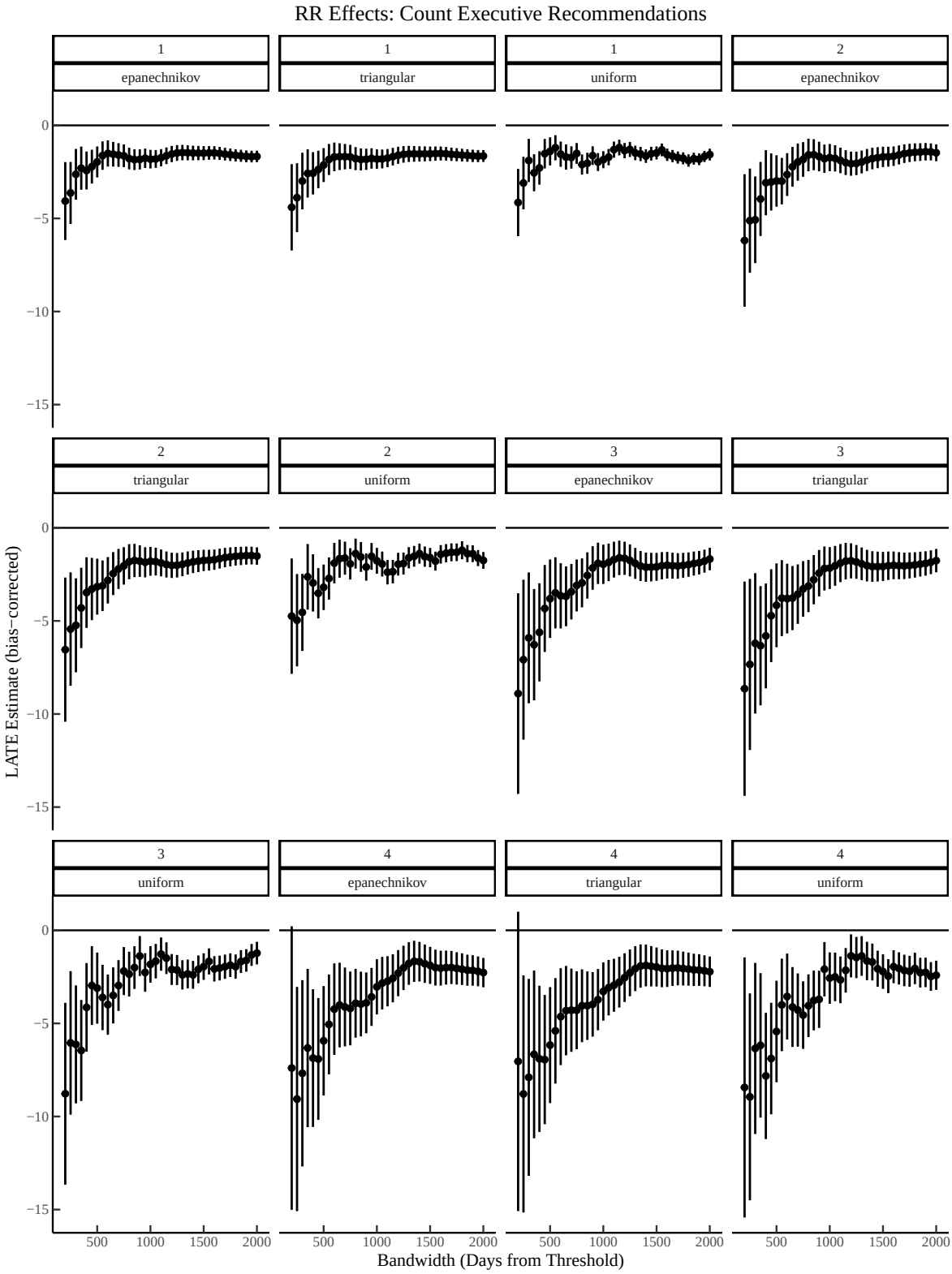
Notes: N = 9,712. Left / Right obs = 3,560 / 6,152. Eff. N = 283 / 281. Order est. $p = 1$, order bias $q = 2$. BW est. $h = 425.826$, BW bias $b = 1122.802$, $\rho = 0.379$. BW type = mserd. Kernel = Triangular. VCE = NN.

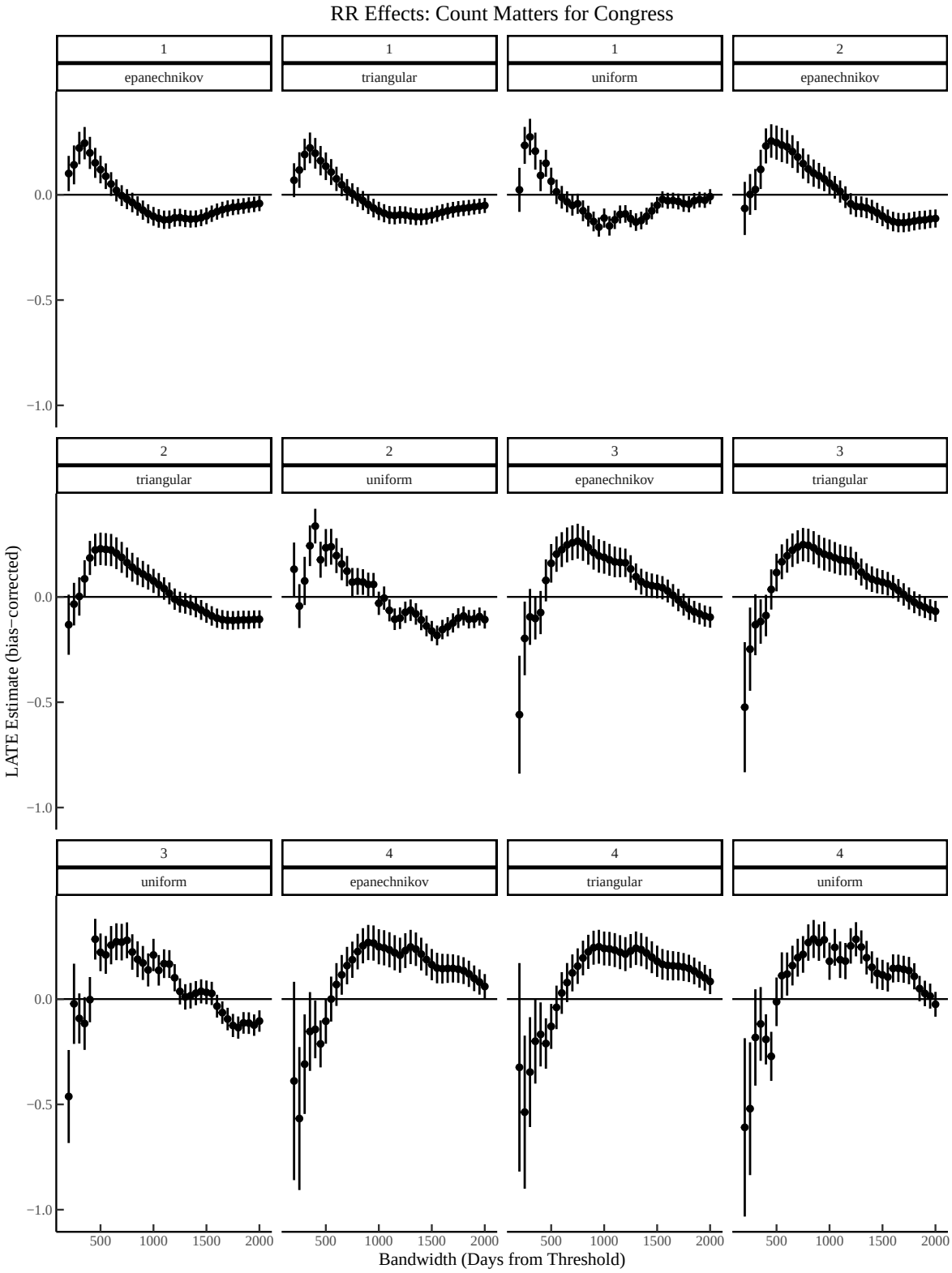
orders (e.g., $p = 1, 2, 3.4$) for the local linear regressions. These combinations are chosen to be reflective a wide-range of differing parameters aside from the defaults employed in the main text. All estimates preserve clustered standard errors at the month level, and only bias-corrected (conservative) confidence intervals and point estimates are presented. These robustness checks represent a fairly exhaustive set of combinations of parameter settings for the local linear regressions. The results are presented in the figures below.

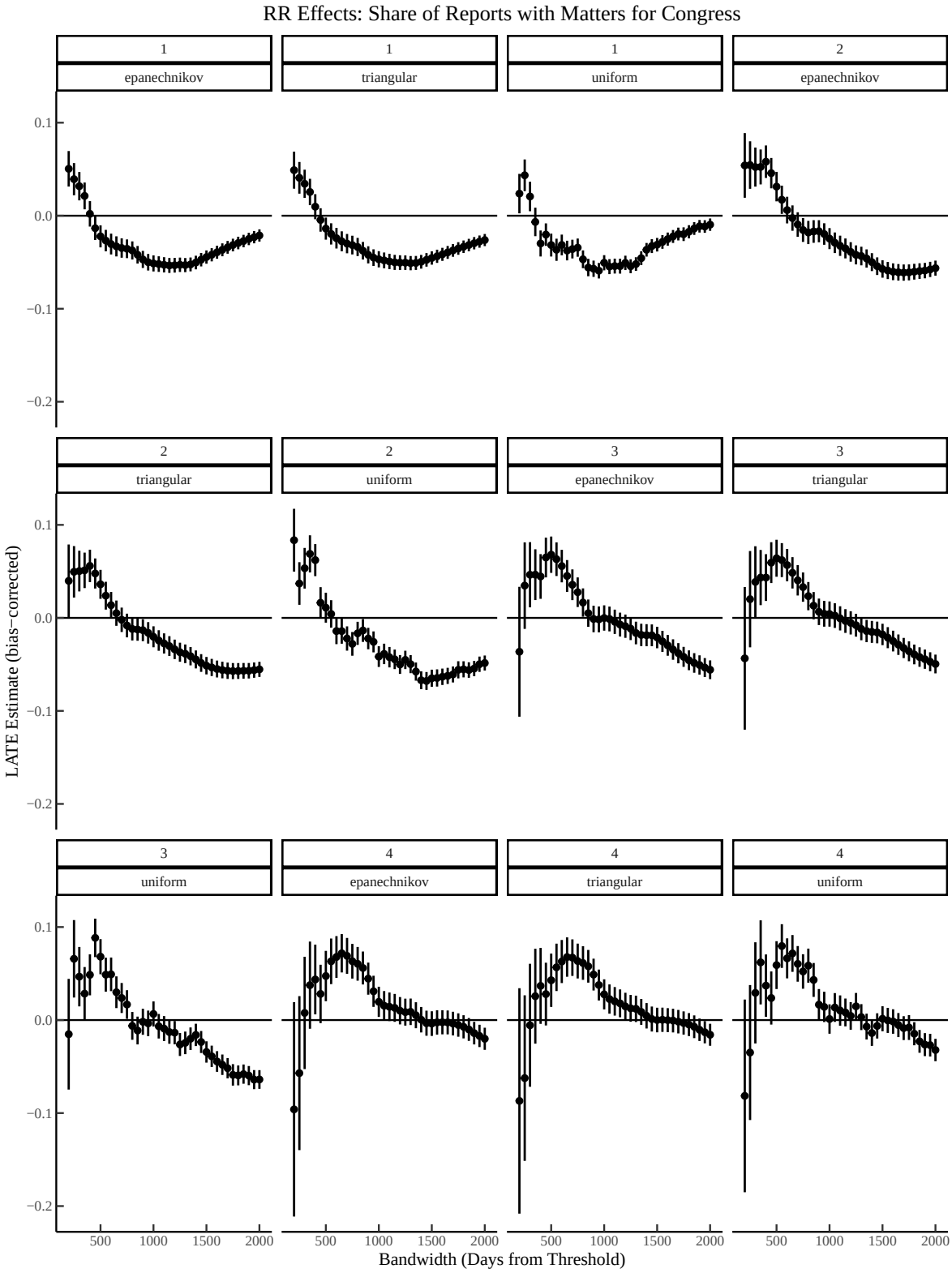

RR Effects: Count Total GAO Reports
1
epanechnikov
1
triangular
1
uniform
2
epanechnikov
2
triangular
2
uniform
3
epanechnikov
3
triangular
3
uniform
4
epanechnikov
4
triangular
4
uniform
LATE Estimate (bias−corrected)
0
−4
−8
500
1000
1500
2000
Bandwidth (Days from Threshold)

RR Effects: Share of Reports with Executive Recommendations




RR Effects: Count Executive Recommendations
1
epanechnikov
1
triangular
1
uniform
2
epanechnikov
2
triangular
2
uniform
3
epanechnikov
3
triangular
3
uniform
4
epanechnikov
4
triangular
4
uniform
LATE Estimate (bias−corrected)
0
−5
−10
−15
500
1000
1500
2000
Bandwidth (Days from Threshold)


RR Effects: Count Matters for Congress
1
epanechnikov
1
triangular
1
uniform
2
epanechnikov
2
triangular
2
uniform
3
epanechnikov
3
triangular
3
uniform
4
epanechnikov
4
triangular
4
uniform
LATE Estimate (bias−corrected)
0.0
−0.5
−1.0
500
1000
1500
2000
Bandwidth (Days from Threshold)


RR Effects: Share of Reports with Matters for Congress
1
epanechnikov
1
triangular
1
uniform
2
epanechnikov
2
triangular
2
uniform
3
epanechnikov
3
triangular
3
uniform
4
epanechnikov
4
triangular
4
uniform
LATE Estimate (bias−corrected)
0.1
0.0
−0.1
−0.2
500
1000
1500
2000
Bandwidth (Days from Threshold)

Bibliography

Kennedy, Joshua B. 2023. "Always Watchful: Political Context and Congressional Oversight through the Government Accountability Office." In *Congress & the Presidency,* 50:342–360. 3. Taylor & Francis.

Lee, Frances E. 2016. *Insecure majorities: Congress and the perpetual campaign.* University of Chicago Press.

Selin, Jennifer L, and Grace Moore. 2023. "Keeping tabs on the executive." *Presidential Studies Quarterly.*

Warren, Patrick L. 2012. "Allies and adversaries: appointees and policymaking under separation of powers." *The Journal of Law, Economics, & Organization* 28 (3): 407–446.

APPENDIX B

Supplemental Material for The Republican Revolution in High Resolution

The Heterogeneous Effects of the Republican Revolution on Oversight at the Government Accountability Office

Additional Heterogeneous Treatment Effect Tests and Checks

Breiman's Permutation Feature Importance To buttress the LOCO feature important results presented in the main text, here I calculate permutation feature importance (Breiman 2001). This measure of importance asks what the degradation of model performance is when a feature is randomly permuted and model fitness, in this case in terms of R-loss, is recalculated for predictions using that permuted feature. This is similar to LOCO but does not involve fully retraining the model for each covariate, and in this form, also does not allow for variable grouping. While this approach is not entirely novel, it more experimental than the LOCO approach and does not come with advertised statistical guarantees Paillard et al. 2024.

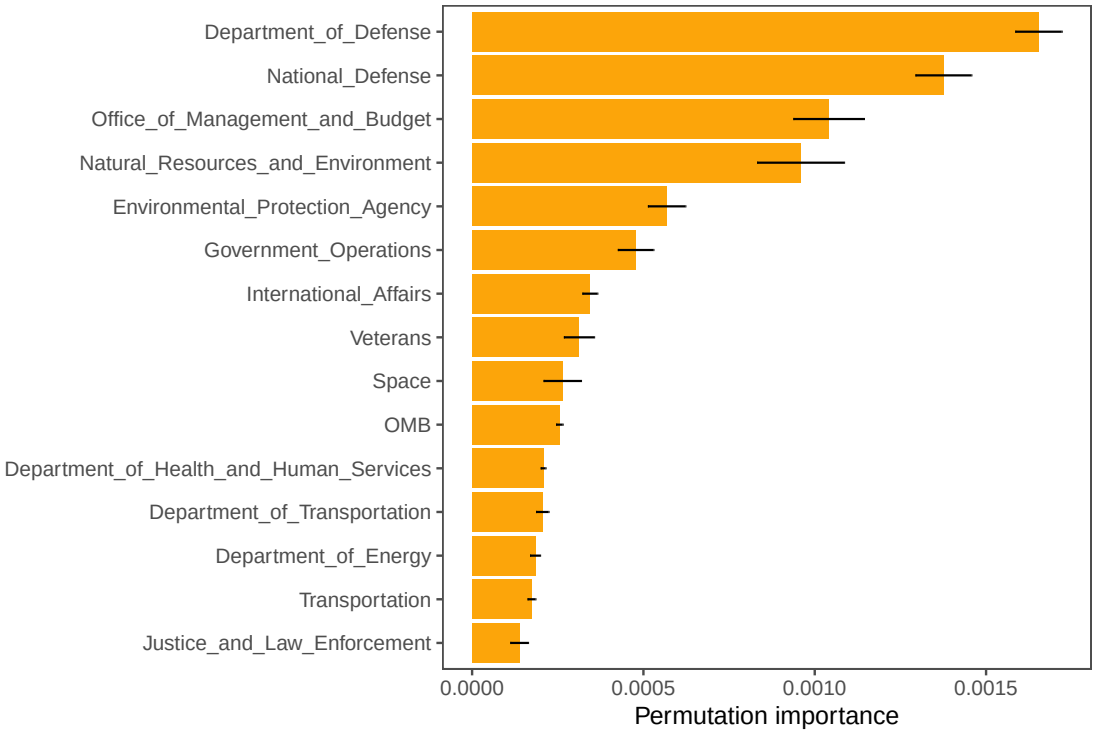


FIGURE B.1. Permutation Feature Importance Alternative to LOCO

In this case the permutation feature importance results, presented in Figure B.1 for executive recommendations only roughly mirror the LOCO results, and would result in a near identical set of covariates to test for marginal effects in the best linear projection results presented in the main text.

Friedman's H Statistic Full Results This section displays the complete Friedman's H-statistic results for the heterogeneous treatment effects analysis. These results include both overall interaction strength and two-way pair interaction strength.

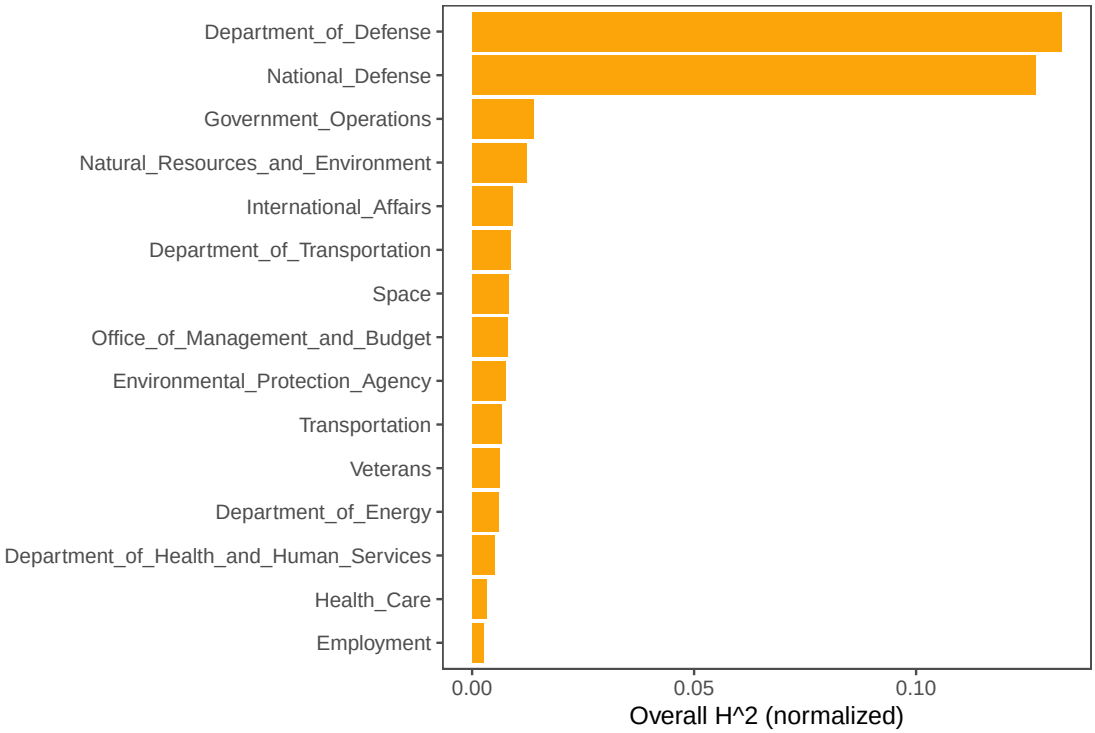


FIGURE B.2. Overall Interaction Strength by Covariate (Friedman's H-Statistic)

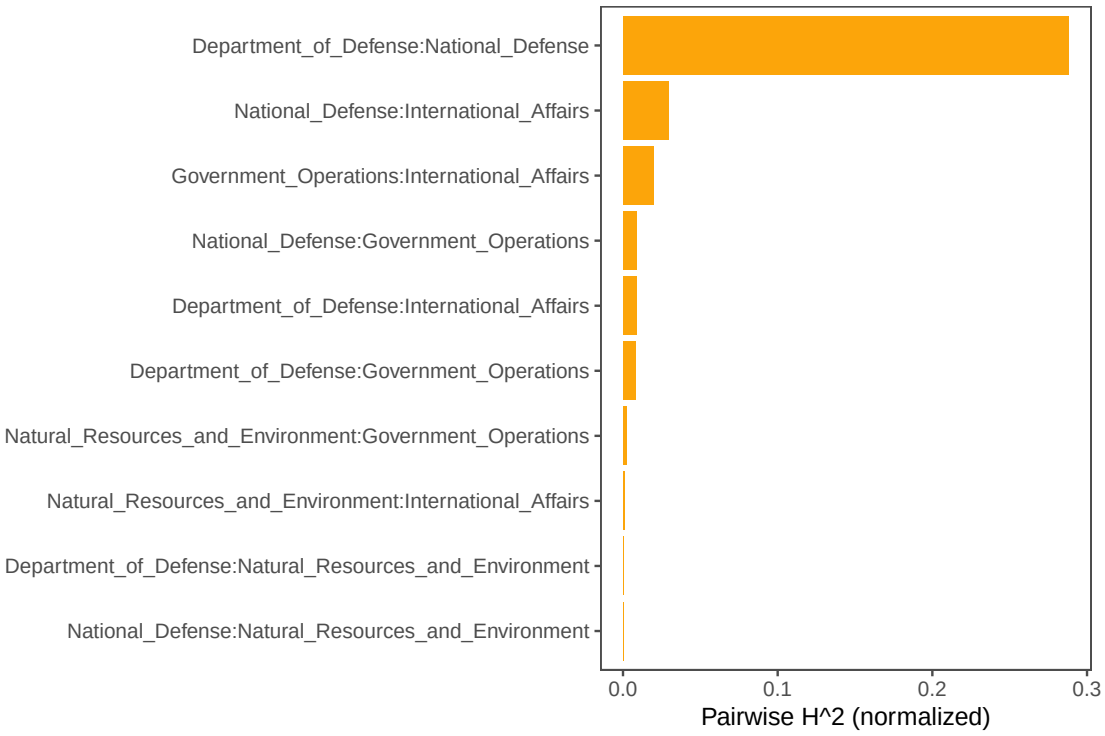
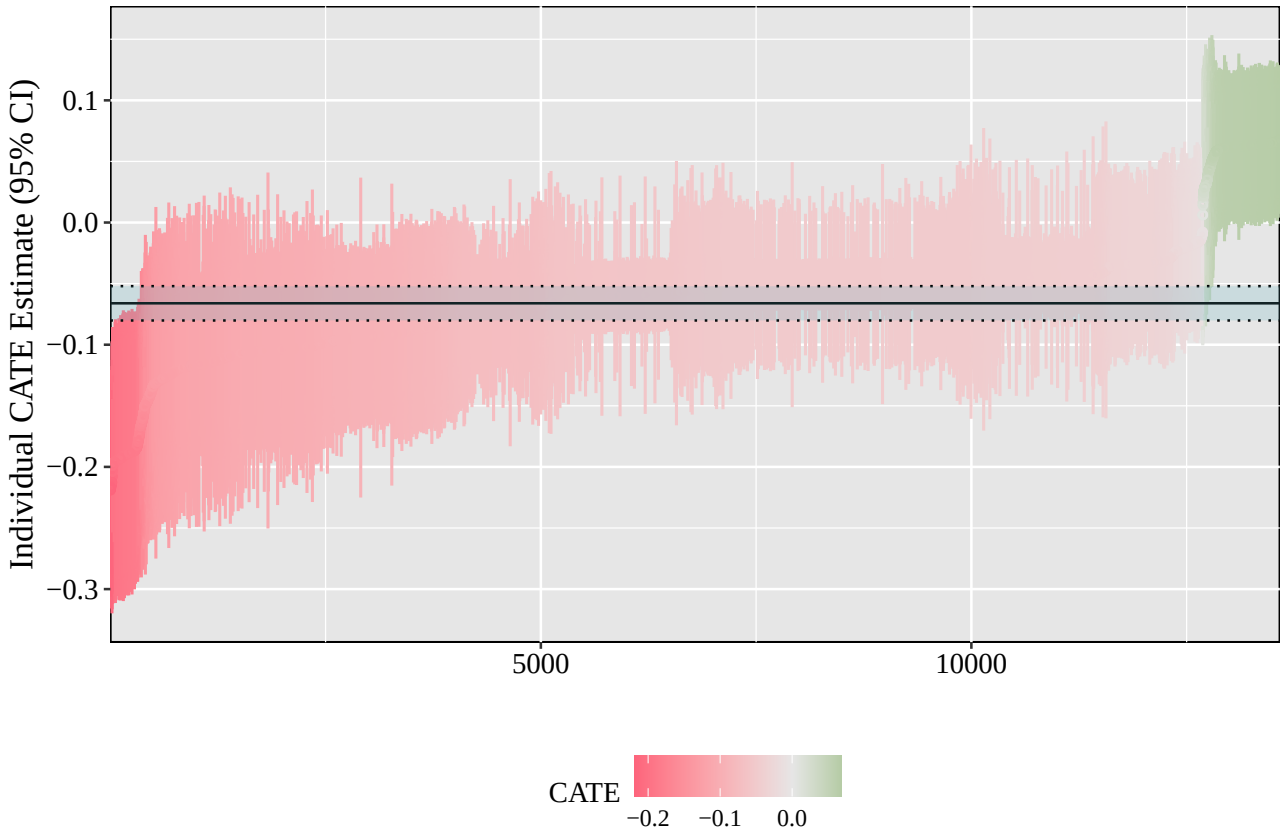


FIGURE B.3. Pairwise Interaction Strength (Friedman's H-Statistic)

Rank HTE Plots by Dependent Variable Aside from the formal tests for treatment effect heterogeneity deployed above, one can also simply examine the distribution of CATE predictions relative to the average treatment effect. Figure B.4 shows the noisy, individual-level CATE estimates with 95% confidence bands rank ordered by effect size. The average treatment effect is shown as a black horizontal line with a 95% confidence region shaded in blue. As the figure makes clear, the CF model is estimating significant heterogeneity in the effects, with individual CATE estimates ranging from -.22 to .07. This indicates for some subset of reports, the RR actually causes an *increase* in the probability the report contains recommendations for the executive.

FIGURE B.4. Rank Ordered CATE Estimates



Note: Individual CATE predictions with 95% confidence intervals. Color corresponds to magnitude of CATE, average treatment effect region shown as black horizontal line with blue 95% confidence region.

While this ranking exercise is useful as a heuristic check, given the inherent difficulty and noisiness of estimating CATEs at the individual level, it is necessary to summarize the CATE estimates at a higher aggregate level to retrieve robust estimates as is done in the main text.

Bibliography

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45:5–32.

Paillard, Joseph, Angel Reyero Lobo, Vitaliy Kolodyazhniy, Bertrand Thirion, and Denis A Engemann. 2024. "Measuring Variable Importance in Heterogeneous Treatment Effects with Confidence." *arXiv preprint arXiv:2408.13002.*

APPENDIX C

Supplemental Material for Leaving No Variance on the Table: Causal Machine Learning for Average and Conditional Effects in Cross-Sectional Data

Adapting Classical Machine Learning Algorithms for Causal Inference Classical machine learning algorithms are prediction machines.[1] To unfamiliar readers, these algorithms may appear exotic—and indeed some of these approaches are quite complex relative to simple regression models—but the general purpose of these approaches is actually very simple. On a fundamental level, ML algorithms—such as multilayer perceptron neural networks, support vector machines, and random forests—take some set of input variables X and build predictive models of some outcome variable Y. Unlike simple regression models, the exact functional form of the relationship between the independent and dependent variables is chosen flexibly by an algorithmic procedure, allowing for the possibility of high-order interactions and non-linearities that would need to be manually specified in a regression.

In political science, researchers have employed classical ML algorithms for a wide variety of tasks (Grimmer, Roberts, and Stewart 2021). For example, Hare and Kutsuris (2023) employs ensembles of machine learning algorithms to predict the vote choice of swing voters. Another piece by Weitzel et al. (2024) attempts to measure subjective bias in democracy scores by comparing expert ratings with predictions from an ML model using objective inputs.

In this article, we focus specifically on methods based on decision trees, also known as forest-based methods. These methods are rightly popular and of particular value to social scientists because of 1) their predictive power when applied to tabular data, 2) their relative interpretability when compared to other ML methods, and 3) their specific adaptations for both causal identification (e.g., causal forest) and common social science data (e.g., CatBoost). First, forest-based methods, broadly, and boosting-methods, specifically, are the most powerful predictors for tabular data (Shwartz-Ziv and Armon 2022), even when compared to computationally expensive deep learning methods (McElfresh et al. 2024). Second, forests are far more intuitive than more complex models like neural networks and also have a well developed ecosystem of interpretability statistics (like variable importance measures) vis-a-vis other models (see Molnar 2020). Finally, there are also a whole host of forest-based methods that satisfy numerous desiderata for social scientists. For example, there are Bayesian versions of forests which can incorporate priors (e.g., Bayesian Additive Regression Trees Chipman, George, and McCulloch 2010), forests that natively handle categorical

[1]For a broad introduction to classical machine learning methods, we recommend Hastie et al. (2009) and Molnar (2020).

data (e.g., CATboost Prokhorenkova et al. 2018), and methods designed to detect treatment effect heterogeneity such as causal forest.

From Simple Decision Trees to Causal Forest In this section we describe the inner workings of causal forest, beginning with the building blocks of a simple decision tree.[2] For motivation, we will proceed with a stylized example of predicting presidential vote-choice using partisanship, age, gender, race, education, and region.

Decision Trees A decision tree is simply a set of if/else rules that generates some prediction, either binary or continuous. They operate quite simply: given a set of variables/predictors $X = (X_1, ..., X_p)$ it will create decision-rules recursively (one after the other) that maximize/minimize some criterion, normally predictive accuracy. Importantly, when we generate a decision tree we do not specify the specific relationship between our predictors and our outcome, rather we use a set of "tuning" parameters for the decision tree algorithm to create a tree. These parameters include, but are not limited to, the minimum number of observations in each "bin" (final prediction) and the maximum number of "splits" (if/else rules). These tuning parameters ensure that the tree does not overfit to the data, since for example, one could create a tree that perfectly predicts all observations using their unique combination of characteristics by setting the minimum node size to one. In actuality, we would test multiple combinations of tuning parameters and use sample-splitting (generating/training the model on a subset of the data, and then testing the model performance on the held-out data) to ensure that we have a predictive, but generalizable model.

This highlights one major concern in ML modeling: the bias-variance tradeoff. Essentially, ML models are "greedy," built with predictive accuracy as their sole goal, so if allowed, they will generate perfect predictions on the data used to build the model. This leads to no bias, as every prediction is completely accurate, but incredibly high variance as the model is not generalizable to new or additional data. For example, we highlight this tradeoff in Figure C.1.

Here, across all panels, we have simulated data generated from a simple Sigmoid function resulting in an S-shape. Within each panel, we report models that reflect different locations along the bias-variance tradeoff: the first (1) being the least complex, with the highest bias and lowest variance, and the final panel (4) having the lowest bias but the highest variance. Specifically, the curve fitting algorithms are of progressive complexity, beginning with a simple linear model, with

[2]The explanations in this section broadly follow Hastie et al. (2009) and Montgomery and Olivella (2018).

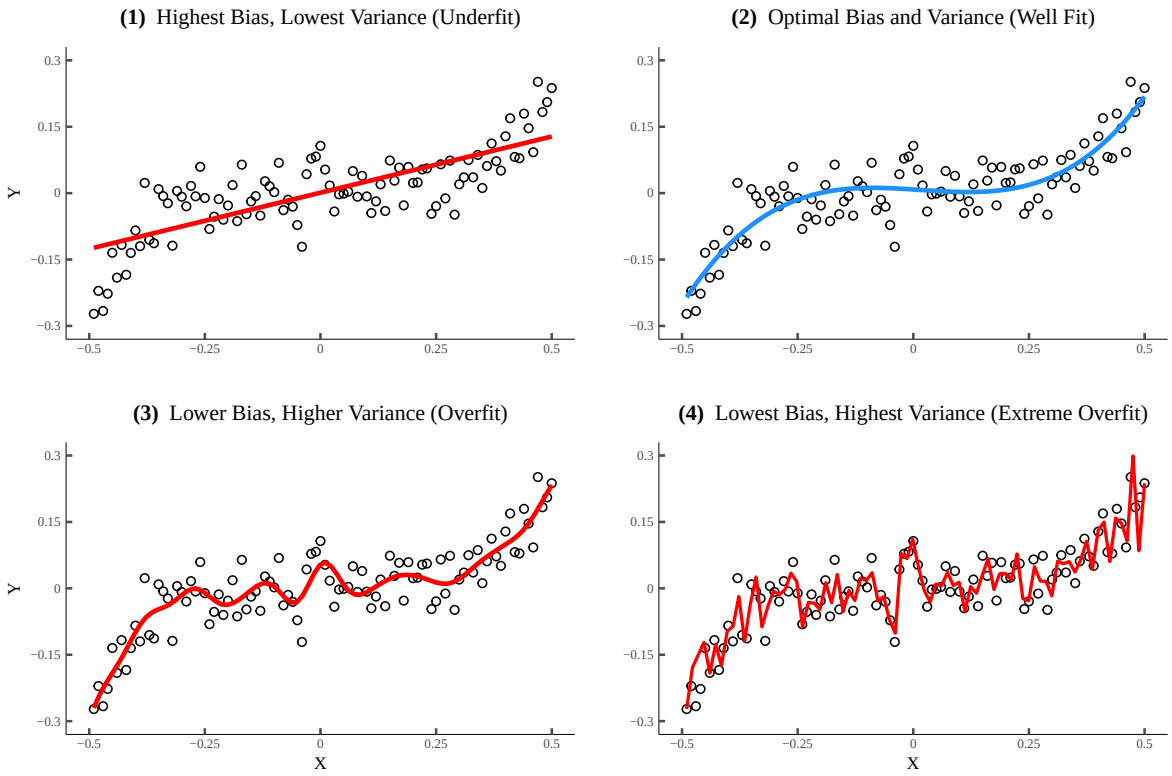
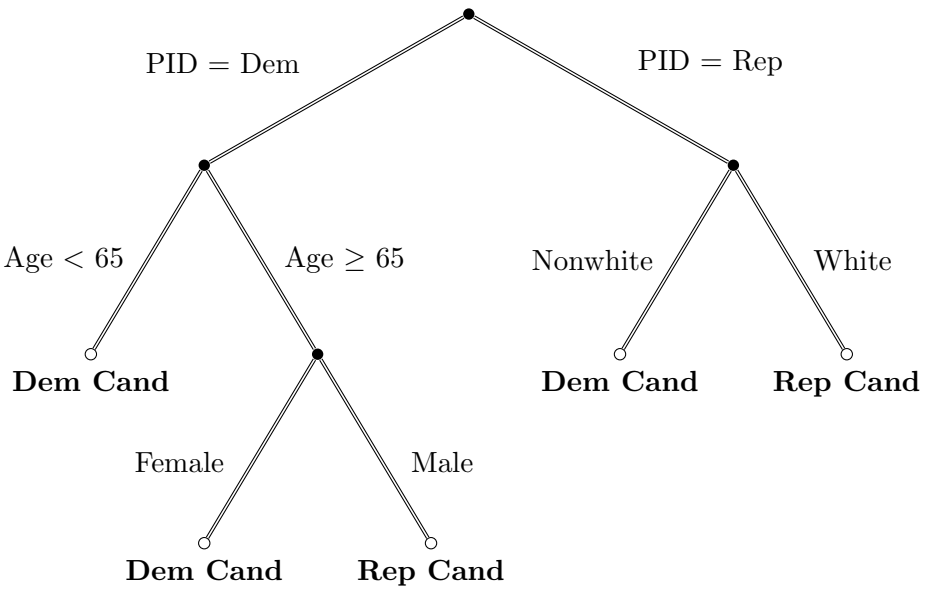


FIGURE C.1. Visualization of the Bias-Variance Tradeoff

each successive panel adding additional splines (1, 3, 20, and 85) to each model. The simple linear model does a decent job at summarizing a positive relationship between X and Y, but clearly misses the non-linearities of the data-generating process (DGP). In contrast, the three spline model (2) does a great job at recovering and representing the true DGP, significantly reducing bias, but only marginally increasing variance in the model. However, adding splines beyond three results in increasingly dramatic overfitting where we can clearly see that the model does not reflect the true, or approximately true, DGP. In the extreme, Panel (4), we have a model that has nearly zero error but incredibly high variance, leading to extremely poor generalizability. This simple example illustrates both the benefits of increasing model complexity but also the possible dangers, we want a highly accurate, but also generalizable model.

Turning back to the motivating example of predicting vote-choice, consider the following decision tree in Figure C.2. This is just one possible tree fit to predict vote-choice, given some set of tuning parameters and subset of data for creating the model. Working top-down, the first and most important split is on PID, with all respondents identifying as Democrats filtering down the left-side

FIGURE C.2. Example Decision Tree Predicting Presidential Vote-Choice



of the tree, and all Republicans down the right. Subsequently, the algorithm uses age (split on 65 years of age) for Democrats, and race (split on white and nonwhite respondents) for Republicans. For Republicans this is the end of the tree, with the tree predicting that nonwhite Republicans vote for the Democratic candidate and white Republicans vote for the Republican. For Democrats, the tree predicts that respondents under the age of 65 vote for the Democratic candidate. However, for those above 65 there is an additional split based on gender, where female Democrats aged 65 and up are predicted to vote for the Democratic candidate whereas male Democrats 65 and up are predicted to vote for the Republican. This exemplifies the concept of recursive partitioning, wherein the data is sequentially split into smaller and smaller bins based on predictively useful covariates, but limited by the specific tuning parameters of the tree (e.g., the minimum bin size could be 10 respondents).

Importantly, within each bin/prediction, there are both correctly and incorrectly classified respondents. Say, for the far left bin (Democrats under 65), 90% of respondents actually voted for the Democrat, but perhaps in the bin with nonwhite Republicans only 55% of respondents voted for Democrats. Regardless, the decision tree model will give a binary prediction that any given respondent with either set of those characteristics will vote for a Democrat. However, the certainty or predicted probabilities for those estimates is different and corresponds to the percentages in the

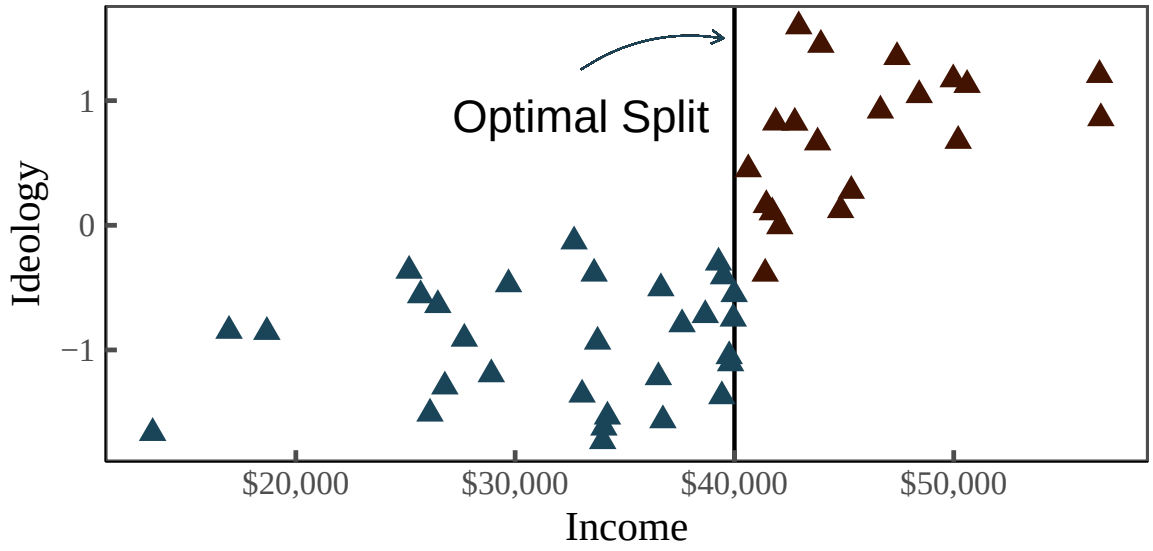


FIGURE C.3. Example Split in Decision (Regression) Tree

bins. Assuming a Democratic vote is 1 and Republican 0, the Democrats under 65 bin would have predicted probabilities of 0.9 and nonwhite Republicans 0.55. These predicted probabilities directly reflect the certainty of the model's predictions for each bin (akin to predicted probabilities for a given combination of predictors in a linear probability model). This is all to say, that while decision trees appear alien when compared to parametric estimators, they in-fact provide the same sorts of estimates (probabilities and classes) *and* their structure and logic is arguably more intuitive.

We can also visualize decision tree splits graphically, which is especially useful in a regression context (continuous outcome). In Figure C.3, we present another stylized example where we are predicting a continuous, granular measure of ideology (−2–2, Y-axis) using respondents' income (X-axis). Here, instead of generating a fitted line through the space like in a regression, we instead see the decision/split rule (the vertical line) that optimally predicts ideology. Respondents to the right of the split rule (approximately located at $40,000) would have higher predicted values, and those to the left of the split would have lower predicted values.

Now importantly, this figure simply represents one split in what could be a more complex decision-tree. Subsequent (or previous) splits may exist that use other variables—or even another value of PID—to split the data for prediction. If this was the only or last split, then the final prediction for points to the left and right of this line would be the average ideology value of both

groups (∼–1 for the left group and ∼1 for the right group). Again, while these are simplified, stylized examples, they should hopefully provide a conceptual basis for understanding trees.

Random Forest Building on these simple decision trees, forest-based methods were first introduced by Breiman (2001), with the invention of random forest. The essential logic of any forest is to assemble numerous trees, by some process/algorithm, and to then aggregate the predictions from the resulting forest. Random forests work through a "wisdom of the crowds" effect, where predictions from a large number of "dumb" decision trees are aggregated, resulting in a "smart" prediction. The random forest algorithm ensures that each tree is, relatively, dumb/underfit through a set of multiple tuning parameters. While there are numerous implementations of random forests, the most common tuning parameters are: 1) limiting the number of variables by randomly selecting some subset available to any given tree (e.g., each tree is created using only 3 out of 10 predictors); 2) requiring a minimum number of observations to be in any given bin (prediction value/category) for each tree; 3) requiring a minimum number of observations to be located down a split in the tree (regardless of the numbers in the eventual prediction bins); and/or 4) limiting the maximum number of splits for any given tree. Regardless of the minutiae of how random forests are constructed, they are incredibly flexible and even without additional features (e.g., boosting) tend to outperform most parametric models in terms of raw predictive power (Hastie et al. 2009).

Mathematically, we can summarize the process of predicting the conditional mean, $\mu(x) = E[Y_i|X_i = x]$, using a random forest in two steps: building the forest and making predictions with the forest.[3] First, each tree generates splits on covariates to maximize the weighted (by the the evenness of the split) squared differences between the resulting groups' means: $n_L n_R(\overline{y}_L - \overline{y}_R)^2$ where, arbitrarily, the L subscript denotes the "left" branch of a given split, and R denotes the "right" branch. The aggregated forest of trees is then used to generate a prediction for each observation. This is done by first calculating the mean outcome for each observation in the final bins within each tree and then averaging the predictions across the entire forest:

$$\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{Y_{i1}\{Xi \in L_b(x)\}}{|L_b(x)|}$$

The Doubly Robust Estimation Framework Having established this conception of research contexts and highlighted the biases introduced by standard estimation approaches in the

[3]We follow much of the notation from the `grf` package implementation page for consistency.

bottom row of Figure 3.3, we proceed by detailing the doubly robust (DR) estimation framework. In simple terms, the DR estimation framework leverages auxiliary information from covariates to 1) adjust for biases in treatment selection (nonrandom assignment; selection bias) and 2) adjust for the relationship between covariates and the outcome, separate from controlling for covariates in a standard multivariable regression.

This class of estimators, and conceptual framework, has its roots in statistical models dating back to 1994, with the first and still widely-used method, augmented inverse propensity weighting (AIPW), introduced by Rotnitzky, Robins, and Scharfstein (1998). Subsequently, Scharfstein, Rotnitzky, and Robins (1999) established the doubly robust (or "doubly protected") properties of the regression representation of AIPW in contexts where data is missing at random or in causal inference contexts where there is assumed to be no unobserved confounding. Importantly, however, the terminology of DR estimation, and the specification of the entire class of estimators, was more concretely established by Robins and Rotnitzky (2001).

Robins and Rotnitzky note the core problem that doubly robust estimation is attempting to address. Assume that we have some binary treatment $W \in \{0, 1\}$ and some vector of covariates $X = (X_1, ..., X_p)$ and a continuous outcome $Y \in \mathbb{R}$.[4] We also define the true relationship between X and Y as $Y = \omega(X)$ where w represents some functional relationship (e.g., linear, nonlinear, stepwise, etc.). Consequently, the "true" relationship can be defined as $Y = \omega(X) + \theta W$ where θ represents our treatment effect. Given our continuous outcome, we may choose to estimate a linear, ordinary least squares (OLS) regression

$$Y = \beta_0 + \beta' X + \theta W$$

In this context, the vector β' represents a linear approximation of the relationship between our outcome and our covariates. Assuming that assignment is random, as in a RCT framework, or that there are no unobserved confounders (and that there is no measurement error) then $\hat{\theta}$ is an accurate estimate of the treatment effect. Importantly though, if there is an interaction between the covariates and the treatment, then $\hat{\theta}$ may be biased. However, as Robins and Rotnitzky note, *even if* there is no interaction between the treatment and covariates, the $\hat{\theta}$ OLS estimate may *still* be severely biased if $\hat{\beta}' X$ does not accurately represent the true, perhaps nonlinear, relationship

[4]Importantly, this framework works with both continuous or binary outcomes as well as multi-arm treatments.

between X and Y: $\omega(X)$. Even more so, if this relationship, $\omega(X)$, is *nonlinear* while being both highly correlated with treatment W and predictive of the outcome Y, then the estimate of the treatment effect, $\hat{\theta}$, will be severely biased.

In response to this concern, alternative estimators were proposed, with the first being propensity score estimation. This adjustment, introduced by Rosenbaum and Rubin (1983), defines the probability of treatment as $P \equiv pr(W = 1|X)$ and the true relationship between P and X as $\gamma(X)$. Ideally, we would estimate this relationship as an "unknown unrestricted function," to quote Robins and Rotnitzky. However, given the (computational) limitations on nonparametric estimation until very recently, these treatment propensities were estimated using a simple logistic regression of the form

$$pr(W = 1|X) = \alpha_0 + \alpha' X$$

where the resulting probabilities $\widehat{P}$ replace the vector of covariates X in our OLS regression leading to the equation

$$Y = \beta_0 + \theta W + \zeta \widehat{P}$$

where ζ represents the effect that treatment propensity $\widehat{P}$ has on our outcome Y. In general, we are able to model our covariates' relationship to treatment assignment (and then its relationship to the outcome) because of the Frisch-Waugh-Lovell theorem. Essentially, this theorem allows us to *separately* estimate the effect of covariates on the outcome and on treatment assignment and still recover the same estimates of the treatment effect θ as if we estimated the entire regression together, like in Equation C.[5]

However, there are issues with both the covariate control and the treatment propensity control regression approaches. Specifically, as noted above, the covariate control approach is biased if the outcome model is misspecified. In contrast, the treatment propensity control approach is biased if the propensity model is misspecified. This is exactly where Robins and Rotnitzky (2001) makes their contribution with doubly robust estimators. Specifically, they show that as long as either the outcome *or* treatment propensity model is correctly specified, then doubly robust estimates are

[5]Technically this is done through residual-on-residual regression, but the core benefit of the FWL theorem is that it allows us to estimate these models separately.

guaranteed to be consistent. This is where doubly robust and doubly protected originates, only one out of the two models need be correctly specified in order to obtain consistent estimates.

This finding by Robins and Rotnitzky is at the heart of the benefits of doubly robust estimation. In the social sciences, we often have a tenuous understanding of the true underlying relationship between some set of covariates and an outcome. Consequently, saturating a regression with covariates, or ignoring them completely, is deeply unsatisfying. This is not to say that non-causal, regression-based research is useless or uninformative, rather it is exceedingly helpful for many considerations, like determining what covariates to measure and how to measure them—even if these models happen to be "wrong" (Little and Pepinsky 2021). However, given the nature of, for example, predicting human behavior, it is critical to have a healthy dose of epistemic humility and to recognize that we are most likely misspecifying our parametric models (for both the outcome and treatment propensities).

Under the standard DR estimation framework in which parametric estimators are used for both the outcome and treatment propensity models—and in which we are forced to specify the specific functional relationship between X and Y or P—we are still likely to misspecify *both* models. However, if we were able to model both $Y = \omega(X)$ and $P = \gamma(X)$ using an "unrestricted function" this significant concern could be greatly ameliorated. This is where the growth and proliferation of flexible, powerful, and nonparametric machine learning models come into play.

As mentioned above, when DR estimation was first established, both processing power to estimate nonparametric models, and the models themselves, were unavailable to researchers. In fact, when Robins and Rotnitzky (2001) was written, Breiman (2001) had not yet introduced his new random forest method, the progenitor of the most powerful nonparametric estimators for tabular data (e.g., *xgBoost* Chen and Guestrin 2016 and *CatBoost* Prokhorenkova et al. 2018). While we detail one ML method, causal forest, that adapts Breiman's random forest for causal inference in the next subsection, suffice to say, it is no longer infeasible to estimate a nonparametric, unrestricted function for both the covariates relationship to the outcome and to treatment propensity. This combination of the doubly robust estimation framework with these nonparametric machine learning models we coin "doubly robust machine learning," or DRML.

Turning back to the formal specification of the doubly robust treatment effect estimation framework, we specify a partially linear model

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W, X) = 0,$$

$$W = \gamma(X) + V, \qquad \mathbb{E}(V|X) = 0,$$

where Y is our outcome variable and $W = \{0, 1\}$ is an indicator variable of treatment.[6] The vector $X = (X_1, ..., X_p)$ is composed of all other covariates, and ϵ and V are random errors for each equation.

In plain language, we have two equations we are working with. The first, $Y = \theta W + \omega(X) + \epsilon$, models the outcome variable as a function of treatment (W) and covariates (X) as well as random error (ϵ) that, in expectation, is equal to zero (and can be ignored) when controlling for treatment and covariates: $\mathbb{E}(\epsilon|W, X) = 0$. Here, θ is the causal effect of treatment (the coefficient for W) and $\omega(\cdot)$ represents the functional relationship between covariates (X) and the outcome (Y). This first equation, however, is just a simple regression framework for modeling the effect of some treatment. The "D" or "Doubly" of DR, and why this approach enables us to recover causal estimates, arises from the use of the second equation for treatment effect estimation.

This second equation, $W = \gamma(X) + V$, models treatment as a function of covariates ($\gamma(X)$) and random error (V). Again, the random error in expectation is zero when controlling for covariates, $\mathbb{E}(V|X) = 0$, so we can ignore that term. In a RCT context where randomization was successful, we would essentially be able to ignore this equation because treatment assignment would, ideally, be orthogonal (unrelated) to any covariates, even if they are related to our outcome of interest. Specifically, we attempt to estimate $\gamma(\cdot)$ which is the functional relationship between covariates (X) and treatment assignment (W).

DR estimation also has important desirable properties: 1) it is semi-parametrically efficient and, most importantly, 2) it is robust to model misspecification in either the outcome or treatment propensity models (i.e., we only need to correctly specify one of these models to recover consistent, causal effects). The first property relates to the fact that the model is efficient (recovers the real treatment effect with a limited number of observations) even though the functional form of some of the components of the model (namely $\omega(\cdot)$ and $\gamma(\cdot)$) are unknown. The second property is that DR estimation will provide robust treatment effects as long as we model the outcome and/or treatment

[6]Here we focus on a binary treatment for simplicity, multi-arm and continuous treatments also work in this framework.

assignment correctly. This component can be evaluated theoretically—are we including variables that we have good reason to believe are related to the outcome and/or treatment assignment?—and empirically—based on model statistics, are we accurately specifying these models between training (data used to build the models) and testing sets (data held out from the training set)?

As we mentioned above, these two equations can be and have been modeled using parametric estimators, like OLS for the outcome model and MLE for the propensity model. However, while parametric estimators are easy to interpret, correctly specifying one of these model is *difficult* and is subject to numerous considerations.[7] These include the choice and number of regressors; any variable transformations, like squaring a term; and the choice of interactions. Above and beyond these considerations, there are fundamental questions about the underlying functional relationship between the covariates and the outcome or propensity: is it purely additive? Is it nonlinear? How should missingness in covariates be incorporated? As Robins and Rotnitzky (2001) note, a flexible, nonparametric estimator would be optimal in this context where predictive accuracy is paramount. In the following section, we discuss the applicability of nonparametric machine learning algorithms for these tasks.

DRML With Uninformative Covariates An important concern which we hope to address with this paper is identifying when, how, and to what extent DRML fails, or introduces bias, when estimating treatment effects. Two obvious situations where this may happen are represented by the top row of Figure 3.3: covariates are uninformative and treatment assignment is either random or non-random. Here in these situations covariates have no relationship with the outcome (either directly or through a conditioning effect with the treatment), and thus it would be reasonable to assume that estimating a DRML model using covariates to predict said outcome (and treatment propensity) could potentially induce significant bias. However, mathematically this is highly improbable and our simulations confirm this.[8]

Looking again at Equation **??**

$$Y = \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W, X) = 0,$$

$$W = \gamma(X) + V, \qquad \mathbb{E}(V|X) = 0,$$

[7] We note, however, that maximum-likelihood estimators, like logistic regressions, are neither easily interpretable *nor* particularly powerful or accurate.

[8] We will note here that there are some scenarios, approximately less than 1% of all cases, *and only* in sample sizes of 50 and 100, that DRML estimation introduces marginal bias, as compared to difference-in-means.

in the context where covariates are completely uninformative for both the outcome (YX) and treatment propensity models (WX), in expectation the relationship of the covariates on the outcome would be zero, $\mathbb{E}(\omega(X)) = 0$, and the treatment propensity would simply be the proportion of units treated, $\mathbb{E}(\gamma(X)) = n_1/n$, where n_1 is the number of observations in the treatment group and n the total number of observations. Consequently, this equation can be simplified to:

$$\begin{aligned} Y &= \theta W + \epsilon, \quad \mathbb{E}(\epsilon|W) = 0, \\ W &= n_1/n + V, \qquad \mathbb{E}(V) = 0, \end{aligned} \tag{C.1}$$

Equation C.1 illustrates that, in the extreme cases where covariates are wholly unrelated to both the outcome and treatment, we are essentially recovering a difference-in-means estimate. Even more so, our simulations confirm this across a wide variety of scenarios. This is all to say that the DRML framework essentially encapsulates the standard difference-in-means calculation used in RCTs when covariates are wholly uninformative. Critically, as long as one component of DR estimation is correctly specified, either the treatment propensity or outcome model, then we are guaranteed unbiased effect estimates. In contexts where covariates are uninformative, nonparametric ML estimators generate predicted treatment probabilities centered around n_1/n. Thus, DRML treatment effect estimates are unbiased and generally share the properties of the difference-in-means estimator.[9]

Details of honest n-step boosted random forests Honest estimation, or sub-sample splitting, is a data splitting technique wherein some fraction of the training data is used to estimate the model (in the case of random forest, estimate the splits in the decision trees) and some held-out fraction of the training data is used to estimate the effects (in the case of random forest, filling in the leaves on the decision trees). Any leaves in a decision tree that are left empty are also pruned in the estimation procedure. The idea of honest splitting is an important debiasing technique popularized by (Athey and Imbens 2016; Wager and Athey 2018).

n-step boosted random forests We employ an n-step Boosted Random (Regression) Forest (Ghosal and Hooker 2020) to estimate treatment propensities. The procedure for estimating the boosted model is as follows: First, estimate a single random forest model. Second, estimate a new, small

[9]Importantly, as we note below, researchers can test whether treatment assignment is independent of covariates (WX). If estimated treatment propensities are centered—and especially if they are normally distributed—around n_1/n, then one can be reasonably confident that randomization was accomplished.

random forest model targeting the out-of-bag error from the first forest. If the new forest can improve on the out-of-bag error from the first forest, a new full random forest model is estimated. This procedure iterates until a new model can no longer explain a large fraction of the error from the previous forest. Thus, the final boosted random forest is an ensemble of n individual forests, or n-steps.

Variable importance For the purpose of this paper, we employ the measure of variable importance introduced by Wager and Athey 2018. To construct this measure, first a matrix is generated that, for each node from 1 to the maximum tree-depth, sums the amount of times each variable appears as a split from any tree in that node. This results in a matrix with rows equal to the number of variables and columns equal to the maximum tree depth and each cell being the number of times that given variable appears as a split in that node depth for all trees. Importantly, we limit this matrix to 4 columns/nodes (the default for this function), excluding subsequent splits. Second, each row is divided by its sum, to calculate the node incidence proportions for every variable (e.g., variable 1 appears 40% of the time in node 1, 20% in node 2, 30% in node 3, and 10% in node 4 resulting in a vector $\vec{V_1} = [.40, .20, .30, .10]$).

The weights for the matrix (vector $\vec{W}$) are determined by a tuning parameter called "decay" which, in our function, is left on the default value of 2 and leads to an exponential discounting of a variable being used for a split at each depth. So variables that split first are fully weighted, those that are split second are weighted by $\frac{1}{4}$, third $\frac{1}{9}$, fourth $\frac{1}{16}$ (i.e., $\vec{W} = [1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}]$). The weights are then divided by the sum of weights (in this instance, divided by $1\frac{61}{144}$) resulting in relative weights being applied to each node column. Then the sum is finally calculated for each variable row vector, leading to a *relative* variable importance score for each variable. There are, of course, alternative measures of variable importance, and these metrics should *not* be interpreted causally in the context of a propensity score model. However, for the application of nonrandom assignment detection and simple identification of predictors that contribute contaminating assignment, this simple variable importance metric is suitable.

Out-of-bag predictions To generate "out-of-bag" predictions, the predictions for each observation are generated using only the trees in the forest that did not use that observation as a training unit. In so doing, we reduce bias in the model predictions stemming from overfitting.

Tuning procedure We employ the cross-validation (CV) procedure developed in the `grf` R package (Tibshirani 2023). In brief, the CV procedure first randomly draws 100 sets of parameter values and for each set, a forest is trained and out-of-bag error estimates are estimated. Specifically, "mini forests" are trained using a small number of trees, for computational speed. However, given these small forests a debiasing process is used through a variance decomposition. Finally, a smoothing function is used to determine optimal parameter values. Also note that CFs use an error measure developed by (Nie and Wager 2021). For more information see the `grf` package materials.

Additional Monte Carlo Results and Detail

Data-Generating Algorithms This section displays the full DGPs for the Monte carlo simulation runs in the main text.

Algorithm 1 Standard Experimental and Systematic Imbalance Simulations Algorithm

Initialization—Set parameter values and randomly draw and set a seed from a discrete uniform distribution: seed $\sim \mathcal{U}\{1, 10000\}$.

Generate covariate values where the set of covariates is $\mathbf{X_j} = \{x_1, ...x_j\}$ and $length(x_j) = n$: $x_j \sim \mathcal{N}(0, 1)$. Randomly select a number of covariates ($c \sim \mathcal{U}\{1, 5\}$) to sample from the generated covariates ($\mathbf{X_j}$) for the outcome DGP: $\mathbf{X_c} \sim \mathbf{X_j}$. For each covariate in $x_c \in \mathbf{X_c}$, we then randomly select a linear coefficient for each covariate from a continuous uniform distribution: $\beta_c \sim \mathcal{U}(-2, 2)$ and $\beta_c = \{\beta_1, ..., \beta_c\}$. Generate a linear effect for each observation (i) and related variable ($\mathbf{X_c}$): $LE_c = \beta_c \times x_c$. And, if `linear = FALSE`, generate a nonlinear effect for each variable in $\mathbf{X_c}$ by squaring each observation's effect: $NE_c = (LE_c)^2$. If `interaction = TRUE` and $c >= 2$, randomly select two covariates from the subset of covariates that are already in the outcome DGP ($\mathbf{X_t} \sim \mathbf{X_c}$) and include their pairwise interaction in the outcome DGP, drawing the coefficient for the interaction from a continuous uniform distribution: $\beta_t \sim \mathcal{U}(-2, 2)$. If `contamination = FALSE`, treatment assignment (W) is orthogonal to x_{ij} and drawn randomly with: $W \sim$ Bernoulli(0.5). If `contamination = TRUE`, then the same random subset of covariates that are in the outcome DGP ($\mathbf{X_c}$) become determinants of treatment assignment. First, we draw $\gamma_c \sim \mathcal{U}(-.5, .5)$ and calculate the probability of assignment using $p = P(W = 1|\mathbf{X_c}) = \text{logistic}(\mathbf{X_c} \times \gamma_c)$ where $\text{logistic}(\cdot) = \frac{1}{1+e^{-z}}$. Treatment assignment is then calculated using: $W \sim \text{Bernoulli}(p)$. If `heterogeneity = FALSE`, the treatment effect is constant: $\tau = 1$. If `heterogeneity = TRUE`, the treatment effect is simply: $\tau_i = 1 + W_i \times \sum_1^c (\mathbf{X_{ic}} \times \gamma_c)$. Finally, the outcome is generated as: $Y_i = W_i \times \tau_i + LE_c + NE_c + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Algorithm 2 Uninformative Covariates Data-Generating Process Algorithm

Initialization—Set parameter values and seeds.

Treatment Assignment: Simulate treatment W as: $W_i \sim \text{Bernoulli}(\texttt{treatfrac}), \quad i = \{1, \ldots, n\}$. **Generate Covariates:** Simulate a covariate matrix where: $x_{ij} \sim \mathcal{N}(0, 1), \quad i = \{1, \ldots, n\}$, $j = \{1, \ldots, 5\}$. **Outcome Generation:** Simulate outcome Y as: $Y = \tau \times W + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \texttt{SNR})$.

Algorithm 3 CATE Simulations Data-Generating Process Algorithm

Initialization—Set parameter values and seed

Generate Treatment Assignment: Simulate treatment assignment W: $W_i \sim \text{Bernoulli}(0.5), \quad i = \{1, \ldots, n\}$.

Generate Correlated Covariates: Simulate p covariates with a pre-specified correlational structure:

- Mean vector: $\boldsymbol{\mu} = \mathbf{0}$.
- Covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \quad \rho = .2$$

Covariates are then drawn as:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Generate Baseline Outcome: The baseline outcome is generated as:

$$\text{Base} = 2x_1 - 3x_2 + 0.5x_4^2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \texttt{SNR}).$$

Generate Conditional Average Treatment Effect (CATE):

- If `linear = TRUE`:

$$\text{CATE} = \begin{cases} W \cdot x_1 \cdot x_2 & \text{if } \texttt{dim = 2 and hetero = TRUE}, \\ W \cdot x_1 & \text{if } \texttt{inter = FALSE and hetero = TRUE}, \\ W & \text{if } \texttt{hetero = FALSE}. \end{cases}$$

- If `linear = FALSE`:

$$\text{CATE} = \begin{cases} W \cdot (x_1 \cdot x_2)^2 & \text{if } \texttt{dim = 2 and hetero = TRUE}, \\ W \cdot (x_1)^2 & \text{if } \texttt{inter = FALSE and hetero = TRUE}, \\ W & \text{if } \texttt{hetero = FALSE}. \end{cases}$$

Generate Final Outcome: The outcome is:

$$Y = \text{CATE} + \text{Base}$$

.

Algorithm 4 High Dimensional Data-Generating Process Algorithm

Initialization—Set parameter values and seeds.

Generate Base Data: Generate 10 covariates for the base outcome model following Friedman (1991):

$$Y_{\text{Friedman}} = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + e$$

where $e \sim \mathcal{N}(0,1)$ and $x_1, x_2, \ldots, x_{10} \sim \text{Uniform}(0,1)$. **Generate Treatment Assignment:** Simulate treatment assignment W as:

$$W_i \sim \text{Bernoulli}(0.5), \quad i = \{1, \ldots, n\}$$

. **Generate Additional Covariates:** Simulate 5 additional covariates:

$$x_{ij} \sim \mathcal{N}(0,1), \quad i = \{1, \ldots, n\},\ j = \{1, \ldots, 5\}$$

. **Outcome Generation:** Generate the final outcome Y as:

$$Y = \tau \times W + Y_{\text{Friedman}} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \texttt{SNR})$.

Additional Coverage Results The coverage rates for each method are presented in Table C.1 below. Results for the Lin estimator are presented in two different ways: one where NAs (i.e., where the estimator failed) are included in the calculation as "uncovered," and the other simply drops the NAs from the calculation. In general, all four methods achieve or closely approach 95% coverage. In practice, one could also implement conformal prediction intervals (Lei and Candès 2021), which guarantee coverage at a pre-specified rate. This approach could be employed for any of the estimators mentioned, one useful software implementation is `cfcausal` (Lei and Candès 2021).

DRML	Unadjusted DIM	Regression-Adjust	Lin (NAs=0)	Lin (Drop NAs)
96%	95%	96%	86%	95%

TABLE C.1. Coverage rates averaged over all simulations

Simulations Where DRML is Outperformed in the Standard Experimental Context Simulations The following list indicates the 19 simulation comparisons where DRML is beaten, either in terms of absolute error or standard error length, relative to one of the other methods.

- n: 100, Covs: 20, Error Var: 5, Linear: True, Interactions: False, Hetero: False
- n: 100, Covs: 20, Error Var: 5, Linear: True, Interactions: False, Hetero: True
- n: 100, Covs: 50, Error Var: 1, Linear: False, Interactions: True, Hetero: False
- n: 100, Covs: 50, Error Var: 1, Linear: True, Interactions: True, Hetero: True
- n: 100, Covs: 100, Error Var: 1, Linear: False, Interactions: False, Hetero: False

- n: 100, Covs: 100, Error Var: 2, Linear: False, Interactions: True, Hetero: False
- n: 100, Covs: 100, Error Var: 2, Linear: False, Interactions: True, Hetero: True
- n: 100, Covs: 100, Error Var: 2, Linear: True, Interactions: True, Hetero: False
- n: 100, Covs: 100, Error Var: 2, Linear: True, Interactions: True, Hetero: True
- n: 100, Covs: 100, Error Var: 5, Linear: False, Interactions: True, Hetero: True
- n: 100, Covs: 20, Error Var: 1, Linear: False, Interactions: False, Hetero: False
- n: 100, Covs: 20, Error Var: 2, Linear: False, Interactions: False, Hetero: False
- n: 100, Covs: 20, Error Var: 5, Linear: False, Interactions: False, Hetero: False
- n: 100, Covs: 20, Error Var: 5, Linear: False, Interactions: False, Hetero: True
- n: 100, Covs: 50, Error Var: 2, Linear: True, Interactions: True, Hetero: True
- n: 100, Covs: 10, Error Var: 2, Linear: True, Interactions: True, Hetero: True
- n: 100, Covs: 10, Error Var: 5, Linear: True, Interactions: True, Hetero: False