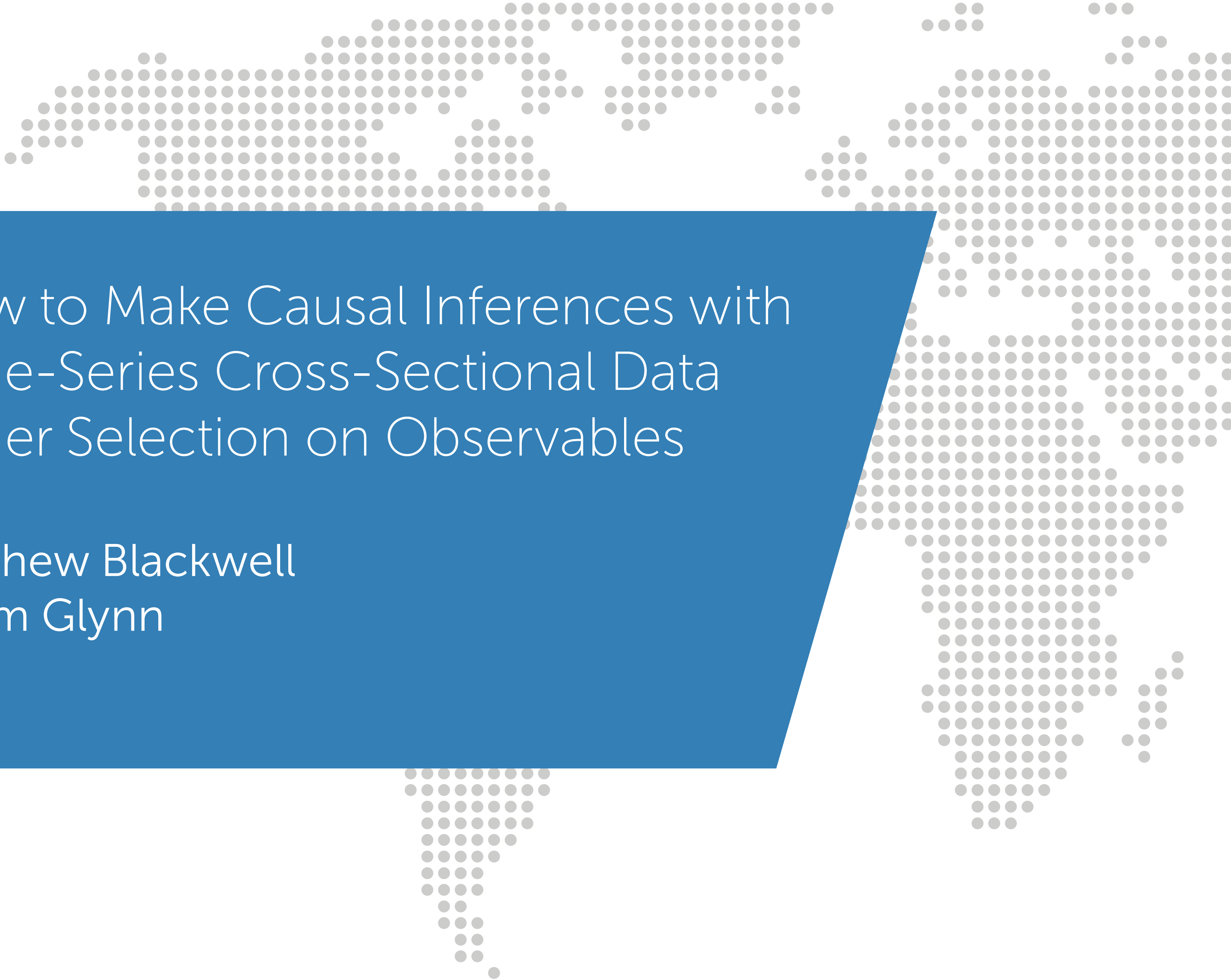
V-Dem
INSTITUTE



How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables

Matthew Blackwell
Adam Glynn

May 2018

Working Paper
SERIES 2018:67
THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Please address comments and/or queries for information to:

V–Dem Institute
Department of Political Science
University of Gothenburg
Sprängkullsgatan 19, PO Box 711
SE 40530 Gothenburg
Sweden
E-mail: contact@v-dem.net

How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables[*]

Matthew Blackwell[†]
Adam Glynn[‡]

April 26, 2018

[*] We are grateful to Neal Beck, Jake Bowers, Patrick Brandt, Simo Goshev, and Cyrus Samii for helpful advice and feedback and Elisha Cohen for research support. Any remaining errors are our own. This research project was supported by Riksbankens Jubileumsfond, Grant M13-0559:1, PI: Staffan I. Lindberg, V-Dem Institute, University of Gothenburg, Sweden; by Knut and Alice Wallenberg Foundation to Wallenberg Academy Fellow Staffan I. Lindberg, Grant 2013.0166, V-Dem Institute, University of Gothenburg, Sweden; by European Research Council, Grant 724191, PI: Staffan I. Lindberg, V-Dem Institute, University of Gothenburg, Sweden; as well as by internal grants from the Vice-Chancellor's office, the Dean of the College of Social Sciences, and the Department of Political Science at University of Gothenburg.

[†] Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St, ma 02138. web: http://www.mattblackwell.org email: mblackwell@gov.harvard.edu

[‡] Department of Political Science, Emory University, 327 Tarbutton Hall, 1555 Dickey Drive, Atlanta, ga 30322 email: aglynn@emory.edu

Abstract

Repeated measurements of the same countries, people, or groups over time are vital to many fields of political science. These measurements, sometimes called time-series cross-sectional (TSCS) data, allow researchers to estimate a broad set of causal quantities, including contemporaneous and lagged treatment effects. Unfortunately, popular methods for TSCS data can only produce valid inferences for lagged effects under very strong assumptions. In this paper, we use potential outcomes to define causal quantities of interest in this settings and clarify how standard models like the autoregressive distributed lag model can produce biased estimates of these quantities due to post-treatment conditioning. We then describe two estimation strategies that avoid these post-treatment biases—inverse probability weighting and structural nested mean models—and show via simulations that they can outperform standard approaches in small sample settings. We illustrate these methods in a study of how welfare spending affects terrorism.

1 Introduction

Many inquiries in political science involve the study of repeated measurements of the same countries, people, or groups at several points in time. This type of data, sometimes called time-series cross-sectional (TSCS) data, allows researchers to draw on a larger pool of information when estimating causal effects. TSCS data also give researchers the power to ask a richer set of questions than data with a single measurement for each unit (for example, see Beck and Katz, 2011). Using this data, researchers can move past the narrowest contemporaneous questions—what are the effects of a single event—and instead ask how the *history* of a process affects the political world. Unfortunately, the most common approaches to modeling TSCS data require strict assumptions to estimate the effect of treatment histories without bias and make it difficult to understand the nature of the counterfactual comparisons.

This paper makes three contributions to the study of TSCS data. Our first contribution is to define counterfactual causal effects and discuss the assumptions needed to identify them nonparametrically. We relate these quantities of interest to common quantities in the TSCS literature, like impulse responses, and show how to derive them from the parameters of a common TSCS model, the autoregressive distributed lag (ADL) model. These treatment effects can be nonparametrically identified under a key selection-on-observables assumption called sequential ignorability; unfortunately, however, many common TSCS approaches rely on more stringent assumptions, including a lack of causal feedback between the treatment and time-varying covariates. This feedback, for example, might involve a country's level of welfare spending affecting domestic terrorism, which in turn might affect future levels of spending. We argue that this type of feedback is common in TSCS settings. While we focus on a selection-on-observables assumption in this paper, we discuss the tradeoffs with this choice compared to standard fixed-effects methods, noting that the latter may also rule out this type of dynamic feedback.

Our second contribution is to provide an introduction to two methods from biostatistics that can estimate the effect of treatment histories without bias and under weaker assumptions than common TSCS models. We focus on two methods: (1) *structural nested mean models* or SNMMs (Robins, 1997) and (2) *marginal structural models with inverse probability of treatment weighting* or MSMs with IPTWs (Robins, Hernán and Brumback, 2000). These models allow for consistent estimation of lagged effects of treatment by paying careful attention to the causal ordering of the treatment, the outcome, and the time-varying covariates. The SNMM approach generalizes the standard regression modeling of ADLs and often imply very simple and intuitive multi-step estimators. The MSM approach focuses on modeling the treatment process to develop weights that adjust for confounding in simple weighted regression models. Both of these approaches have the ability to incorporate weaker modeling assumptions than traditional TSCS models. We describe the modeling choices involved and provide guidance on how to implement these methods.

Our third contribution is to show how traditional models like the ADL are biased for lagged treatment effects in common TSCS settings, while MSMs and SNMMs are not. This bias arises from the time-varying covariates—researchers must control for them to accurately estimate contemporaneous effects, but they induce post-treatment bias for lagged effects. Thus, ADL models can only consistently estimate lagged effects when time-varying covariates are unaffected by past treatment. SNMMs and MSMs, on the other hand, can estimate these effects even when such feedback exists. We provide simulation evidence that this type of feedback can lead to significant bias in ADL models compared to the SNMM and MSM approaches. Overall, these latter methods could be promising for TSCS scholars, especially those who are interested longer-term effects.

This paper proceeds as follows. Section 2 clarifies the causal quantities of interest available with TSCS data and shows how they relate to parameters from traditional TSCS models. Causal assumptions are a key part of any TSCS analysis and we discuss them in Section 3. Section 4 discusses post-treatment bias stemming from traditional TSCS approaches, and Section 5 introduces the SNMM and MSM approaches which avoid this post-treatment bias and shows how to estimate causal effects using these methodologies. We present simulation evidence of how these methods outperform traditional TSCS models in small samples in Section 6. In Section 7, we present an empirical illustration of each approach, based on Burgoon (2006), investigating the connection between welfare spending and terrorism. Finally, Section 8 concludes with thoughts on both the limitations of these approaches and avenues for future research.

2 Causal quantities of interest in TSCS data

At their most basic, TSCS data consists of a treatment (or main independent variable of interest), an outcome, and some covariates all measured for the same units at various points in time. In our empirical setting below, we focus on a dataset of countries with the number of terrorist incidents as an outcome and domestic welfare spending as a binary treatment. With one time period, only one causal comparison exists: a country has either high or low levels of welfare spending. As we gather data on these countries over time, there are more counterfactual comparisons to investigate. How does the history of welfare spending affect the incidence of terrorism? Does the spending regime *today* only affect terrorism today or does the recent history matter as well? The variation over time provides the opportunity and the challenge of answering these complex questions.

To fix ideas, let X_{it} be the treatment for unit i in time period t. For simplicity, we focus first on the case of a binary treatment so that $X_{it} = 1$ if the unit is treated in period t and $X_{it} = 0$ if the unit is untreated in period t (it is straightforward to generalize them to arbitrary treatment types). In our running example, $X_{it} = 1$ would represent a country that had high welfare spending in year t and $X_{it} = 0$ would be a country with low welfare spending. We collect all of the treatments for a given unit into a *treatment history*, $X_i = (X_{i1}, \dots, X_{iT})$, where T is the number of time periods in the study. For example, we might have a country that always had high spending, $(1, 1, \dots, 1)$, or a country that always had low spending, $(0, 0, \dots, 0)$. We refer to the partial treatment history up to t as $X_{i,1:t} = (X_{i1}, \dots, X_{it})$, with $x_{1:t}$ as a possible particular realization of this random vector. We define Z_{it}, $Z_{i,1:t}$, and $z_{1:t}$ similarly for a set of time-varying covariates that are causally prior to the treatment at time t such as the government capability, population size, and whether or not the country is in a conflict.

The goal is to estimate causal effects of the treatment on an outcome, Y_{it}, that also varies over time. In our running example, Y_{it} is the number of terrorist incidents in a given country in a given year. We take a counterfactual approach and define potential outcomes for each time period, $Y_{it}(x_{1:t})$ (Rubin, 1978; Robins, 1986).[1] This potential outcome represents the incidence of terrorism that would occur in country i in year t if i had followed history of welfare spending equal to $x_{1:t}$. Obviously, for any country in any year, we only observe one of these potential outcomes since a country cannot follow multiple histories of welfare spending over the same time window. To connect the potential outcomes to the observed outcomes, we make the standard *consistency asssumption*. Namely, we assume that the observed outcome and the potential outcome are the same for the observed history: $Y_{it} = Y_{it}(x_{1:t})$ when $X_{i,1:t} = x_{1:t}$.

To create a common playing field for all the methods we evaluate, we limit ourselves to making causal inferences about the time window observed in the data—that is, we want to study the effect of welfare spending on terrorism for the years in our data set. Under certain assumptions like stationarity of the covariates and error terms, many TSCS methods can make inferences about the long-term effects beyond the end of the study. This extrapolation is, of course, required with a single time series, but with the multiple units we have in TSCS data, we have the ability to focus our inferences on a particular window and avoid

[1] The definition of potential outcomes in this manner implicitly assumes the usual stable unit treatment value assumption (SUTVA) (Rubin, 1978). This assumption is questionable for the many comparative politics and international relations applications, but we avoid discussing this complication in this paper in order to focus on the issues regarding TSCS data. Implicit in our definition of the potential outcomes is that outcomes at time t only depend on past values of treatment, not future values (Abbring and van den Berg, 2003).

these assumptions about the time-series processes. We view this as a conservative approach because all methods for handling TSCS should be able to generate sensible estimates of causal effects in the period under study. Of course, there is a tradeoff with this approach: we cannot study some common TSCS estimands like the long-run multiplier that are based on time-series analysis. We discuss this estimand in particular in the supplemental materials.

Given our focus on a fixed time window, we will define expectations over cross-sectional units and consider asymptotic properties of the estimators as the number of these units grows (rather than the length of the time series). Of course, asymptotics are only useful in how they guide our analyses in the real world of finite samples, and we may worry that "large-N, fixed-T" asymptotic results may not provide a reliable approximation when N and T are roughly the same size, as is often the case for TSCS data. Fortunately, as we show in the simulation studies of Section 6, our analysis of the various TSCS estimators holds even when N and T are small and close in size. Thus, we do not see the choices of "fixed time-window" versus "time-series analysis" or large-N versus large-T asymptotics to be consequential to the conclusions we draw.

2.1 The effect of a treatment history

For an individual country, the causal effect of a particular history of welfare spending, $x_{1:t}$, relative to some other history of spending, $x'_{1:t}$, is the difference $Y_{it}(x_{1:t}) - Y_{it}(x'_{1:t})$. That is, it is the difference in the potential or counterfactual level of terrorism when the country follows history $x_{1:t}$ minus the counterfactual outcome when it follows history $x'_{1:t}$. Given the number of possible treatment histories, there can be numerous causal effects to investigate, even with a simple binary treatment. As the length of time under study grows, so does the number of possible comparisons. In fact, with a binary treatment there are 2^t different potential outcomes for the outcome in period t. This large number of potential outcomes allows for a very large number of comparisons and a host of causal questions: does the stability of spending over time matter for the impact on the incidence of terrorism? Is there a cumulative impact of welfare spending or is it only the current level that matters?

These individual-level causal effects are difficult to identify without strong assumptions, so we often focus on estimating the *average causal effect* of a treatment history (Robins, Greenland and Hu, 1999; Hernán, Brumback and Robins, 2001):

$$\tau(x_{1:t}, x'_{1:t}) = E[Y_{it}(x_{1:t}) - Y_{it}(x'_{1:t})]. \tag{1}$$

Here, the expectations are over the units so that this quantity is the average difference in outcomes between the world where all units had history $x_{1:t}$ and the world where all units

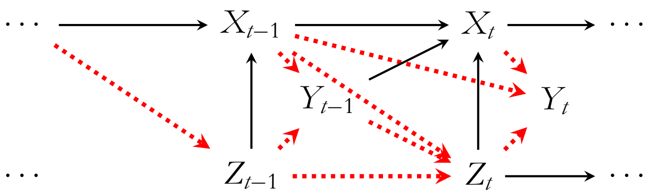


Figure 1: Directed acyclic graph (DAG) of a typical TSCS data. Dotted red lines are the causal pathways that constitute the average causal effect of a treatment history at time t.

had history $x'_{1:t}$. For example, we might be interested in the effect of a country having always high welfare spending versus a country always having low spending levels. Thus, this quantity considers the effect of treatment at time t, but also the effect of all lagged values of the treatment as well. A graphical depiction of the pathways contained in $\tau(x_{1:t}, x'_{1:t})$ is presented in Figure 1, where the red arrows correspond to components of the effect. These arrows represent all of the effects of X_{it}, $X_{i,t-1}$, $X_{i,t-2}$, and so on, that end up at Y_{it}. Note that many of these effects flow through the time-varying covariates, Z_{it}. This point complicates the estimation of causal effects in this setting and we return to it below.

2.2 Marginal effects of recent treatments

As mentioned above, there are numerous possible treatment histories to compare when estimating causal effects. This can be daunting for applied researchers who may only be interested in the effects of the first few lags of welfare spending. Furthermore, any particular treatment history may not be well-represented in the data if the number of time periods is moderate. To avoid these problems, we introduce causal quantities that focus on recent values of treatment and average over more distant lags. We define the potential outcomes just intervening on treatment the last j periods as $Y_{it}(x_{t-j:t}) = Y_{it}(X_{i,1:t-j-1}, x_{t-j:t})$. This "marginal" potential outcome represents the potential or counterfactual level of terrorism in country i if we let welfare spending run its natural course up to $t-j-1$ and just set the last j lags of spending to $x_{t-j:t}$.[2]

With this definition in hand, we can define one important quantity of interest, the *contemporaneous effect of treatment* (CET) of X_{it} on Y_{it}:

$$\begin{aligned}\tau_c(t) &= E[Y_{it}(X_{i,1:t-1}, 1) - Y_{it}(X_{i,1:t-1}, 0)],\\ &= E[Y_{it}(1) - Y_{it}(0)],\end{aligned}$$

[2] See Shephard and Bojinov (2017) for a similar approach to defining recent effects in time-series data.

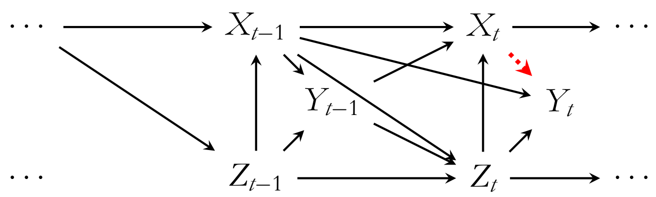


Figure 2: DAG of a TSCS setting where the dotted red line represents the contemporaneous effect of treatment at time t.

Here we have switched from potential outcomes that depend on the entire history to potential outcomes that only depend on treatment in time t. The CET reflects the effect of treatment in period t on the outcome in period t, averaging across all of the treatment histories up to period t. Thus, it would be the expected effect of switching a random country from low levels of welfare spending to high levels in period t. A graphical depiction of a CET is presented in Figure 2, where the red arrow corresponds to component of the effect. It is common in pooled TSCS analyses to assume that this effect is constant over time so that $\tau_c(t) = \tau_c$.

Researchers are also often interested in how more distant changes to treatment affect the outcome. Thus, we define the lagged effect of treatment, which is the marginal effect of treatment in time $t-1$ on the outcome in time t, holding treatment at time t fixed: $E[Y_{it}(1,0) - Y_{it}(0,0)]$. More generally, the j-step lagged effect is defined as follows:

$$\begin{aligned} \tau_l(t,j) &= E[Y_{it}(X_{i,1:t-j-1}, 1, 0_j) - Y_{it}(X_{i,1:t-j-1}, 0, 0_j)], \\ &= E[Y_{it}(1, 0_j) - Y_{it}(0_{j+1})], \end{aligned} \tag{2}$$

where 0_s is a vector of s zero values. For example, the two-step lagged effect would be $E[Y_{it}(1,0,0) - Y_{it}(0,0,0)]$ and represents the effect of welfare spending two years ago on terrorism today holding the intervening welfare spending fixed at low levels. A graphical depiction of the one-step lagged effect is presented in Figure 3, where again the red arrows correspond to component of the effect. These effects are similar to a common quantity of interest in both time-series and TSCS applications called the *impulse response* (Box, Jenkins and Reinsel, 2013).

Another common quantity of interest in the TSCS literature is the *step response*, which is the culmulative effect of a permanent shift in treatment status on some future outcome (Box, Jenkins and Reinsel, 2013; Beck and Katz, 2011). The step response function, or SRF, describes how this effect varies by time period and distance between the shift and the

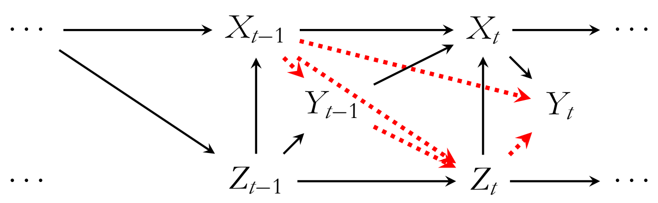


Figure 3: DAG of a panel setting where the dotted red lines represent the paths that constitute the lagged effect of treatment at time $t-1$ on the outcome at time t.

outcome:

$$\tau_s(t,j) = E[Y_{it}(1_j) - Y_{it}(0_j)], \tag{3}$$

where 1_s has a similar definition to 0_s. Thus, $\tau_s(t,j)$ is the effect of a j periods of treatment at time $t-j$ on the outcome at time t. Without further assumptions, there are separate lagged effects and step responses for each pair of periods. As we discuss next, traditional modeling of TSCS data imposes restrictions on the data-generating processes in part to summarize this large number of effects with a few parameters.

2.3 Relationship to traditional TSCS models

The potential outcomes and causal effects defined above are completely nonparametric in the sense that they impose no restrictions on the distribution of Y_{it}. To situate these quantities in the TSCS literature, it is helpful to see how they are parameterized in a particular TSCS model. One general model that encompasses many different possible specifications is called an autoregressive distributed lag (ADL) model:[3]

$$Y_{it} = \beta_0 + \alpha Y_{i,t-1} + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \varepsilon_{it}, \tag{4}$$

where ε_{it} are i.i.d. errors, independent of X_{is} for all t and s. The key features of such a model are the presence of lagged independent and dependent variables and the exogeneity of the independent variables. This model for the outcome would imply the following form for the potential outcomes:

$$Y_{it}(x_{1:t}) = \beta_0 + \alpha Y_{i,t-1}(x_{1:t-1}) + \beta_1 x_t + \beta_2 x_{t-1} + \varepsilon_{it}. \tag{5}$$

In this form, it is clear to see what TSCS scholars have long pointed out: causal effects are complicated with lagged dependent variables since a change in x_{t-1} can have both a direct

[3] For introductions to modeling choices for TSCS data in political science, see De Boef and Keele (2008) and Beck and Katz (2011).

effect on Y_{it} and an indirect effect through $Y_{i,t-1}$. This is why even seemingly simple TSCS models such as the ADL imply quite complicated expressions for long-run effects.

The ADL model also has implications for the various causal quantities, both short-term and long-term. The coefficient on the contemporaneous treatment, β_1, is constant over time and does not depend on past values of the treatment, so it is equal to the CET, $\tau_c(t) = \beta_1$. One can derive the lagged effects from different combinations of α, β_1, and β_2:

$$\tau_l(t, 0) = \beta_1, \tag{6}$$

$$\tau_l(t, 1) = \alpha\beta_1 + \beta_2, \tag{7}$$

$$\tau_l(t, 2) = \alpha^2\beta_1 + \alpha\beta_2. \tag{8}$$

Note that these lagged effects are constant across t. The step response, on the other hand, has a stronger impact because it accumulates the impulse responses over time:

$$\tau_s(t, 0) = \beta_1, \tag{9}$$

$$\tau_s(t, 1) = \beta_1 + \alpha\beta_1 + \beta_2, \tag{10}$$

$$\tau_s(t, 2) = \beta_1 + \alpha\beta_1 + \beta_2 + \alpha^2\beta_1 + \alpha\beta_2. \tag{11}$$

Note that the step response here is just the sum of all previous lagged effects. It is clear that one benefit of such a TSCS model is to summarize a broad set of estimands with just a few parameters. This helps to simplify the complexity of the TSCS setting while introducing the possibility of bias if this model is incorrect or misspecified.

3 Causal assumptions and designs in TSCS data

Under what assumptions are the above causal quantities identified? When we have repeated measurements on the outcome-treatment relationship, there are a number of assumptions we could invoke in order to identify causal effects. In this section we discuss several of these assumptions. We focus on cross-sectional assumptions given our fixed time-window approach. That is, we make no assumptions on the time-series processes such as stationarity even though imposing these types of assumptions will not materially affect our conclusions about the bias of traditional TSCS methods. This result is confirmed in the simulations of Section 6, where the data generating process is stationary and the biases we describe below still occur.

3.1 Baseline randomized treatments

A powerful, if rare, research design for TSCS data is one that randomly assigns the entire history of treatment, $X_{1:T}$, at time $t = 0$. Under this assumption, treatment at time t cannot be affected by, say, previous values of the outcome or time-varying covariates. In terms of potential outcomes, the baseline randomized treatment history assumption is:

$$\{Y_{it}(x_{1:t}) : t = 1, \ldots, T\} \perp\!\!\!\perp X_{i,1:T} | Z_{i0}, \tag{12}$$

where $A \perp\!\!\!\perp B|C$ is defined as "A is independent of B conditional on C." This assumes that the entire history of welfare spending is independent of all potential levels of terrorism, possibly conditional on baseline (that is, time-invariant) covariates. Hernán, Brumback and Robins (2001) called $X_{i,1:T}$ *causally exogeneous* under this assumption. The lack of time-varying covariates or past values of Y_{it} on the right-hand side of the conditioning bar in (12) implies that these variables do not confound the relationship between the treatment and the outcome. For example, this assumes there are no time-varying covariates that affect both welfare spending and the number of terrorist incidents. Thus, baseline randomization relies on strong assumptions that are rarely satisfied outside of randomized experiments and is unsuitable for most observational TSCS studies.[4]

Baseline randomization is closely related to exogeneity assumptions in linear TSCS models. For example, suppose we had the following distributed lag model with no autoregressive component:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \eta_{it} \tag{13}$$

Here, baseline randomization of the treatment history implies the usual identifying assumption in linear TSCS models, *strict exogeneity* of the errors:

$$E[\eta_{it}|X_{i,1:T}] = E[\eta_{it}] = 0. \tag{14}$$

This is a mean independence assumption about the relationship between the errors, η_{it}, and the treatment history, $X_{i,1:T}$.

3.2 Sequentially randomized treatments

Beginning with Robins (1986), scholars in epidemiology have expanded the potential outcomes framework to handle weaker identifying assumptions than baseline randomization.

[4] A notable exception are experiments with a panel design that randomize rollout of a treatment (e.g., Gerber et al., 2011).

These innovations centered on sequentially randomized experiments, where at each period, X_{it} was randomized conditional on the past values of the treatment *and* time-varying covariates (including past values of the outcome). Under this *sequential ignorability* assumption, the treatment is randomly assigned not at the beginning of the process, but at each point in time and can be affected by the past values of the covariates and the outcome.

At its core, sequential ignorability assumes there is some function or subset of the observed history up to time t, $V_{it} = g(X_{i,1:t-1}, Y_{i,1:t-1}, Z_{i,1:t})$, that is sufficient to satisfy no unmeasured confounders for the effect of X_{it} on future outcomes. Formally, the assumption states that, conditional on this set of variables, V_{it}, the treatment at time t is independent of the potential outcomes at time t:

Assumption 1 (Sequential Ignorability)**.** *For every treatment history $x_{1:T}$ and periods t,*

$$\{Y_{is}(x_{1:s}) : s = t, , \ldots , T\} \perp\!\!\!\perp X_{it} | V_{it}. \tag{15}$$

For example, a researcher might assume that sequential ignorability for current welfare spending holds conditional on lagged levels of terrorism, lagged welfare spending, and some contemporaenous covariates, so that $V_{it} = \{Y_{i,t-1}, X_{i,t-1}, Z_{it}\}$. Unlike baseline randomization and strict exogeneity, it allows for observed time-varying covariates like conflict status and lagged values of terrorism to confound the relationship between welfare spending and current terrorism levels, so long as we have measures of these confounders. Furthermore, these time-varying covariates can be affected by past values of welfare spending.

In the context of traditional TSCS models such as (4), sequential ignorability implies the *sequential exogeneity* assumption:

$$E[\varepsilon_{it}|X_{i,1:t}, Z_{i,1:t}, Y_{i,1:t-1}] = E[\varepsilon_{it}|X_{it}, V_{it}] = 0. \tag{16}$$

According to the model in (4), the time-varying covariates here would include the lagged dependent variable. This assumption states that the errors of the TSCS model are mean independent of welfare spending at time t given the conditioning set that depends on the history of the data up to t. Thus, this allows the errors for levels of terrorism to be related to future values of welfare spending.

Sequential ignorability weakens baseline randomization to allow for feedback between the treatment status and the time-varying covariates, including lagged outcomes. For instance, sequential ignorability allows for the welfare spending of a country to impact future levels of terrorism and for this terrorism to affect future welfare spending. Thus, in this dynamic case, treatments can affect the covariates and so the covariates also have potential responses: $Z_{it}(x_{1:t-1})$. This dynamic feedback implies that the lagged treatment may have both a direct effect on the outcome and an indirect effect through this covariate. For example, welfare spending might directly affect terrorism by reducing resentment among potential terrorists, but it might also have an indirect effect if it helps to increase levels of state capacity which could, in turn, help combat future terrorism.

In TSCS models, the lagged dependent variable, or LDV, is often included in the above time-varying conditioning set, V_{it}, to assess the dynamics of the time-series process or to capture the effects of longer lags of treatment in a simple manner.[5] In either case, sequential ignorability would allow the LDV to have an effect on the treatment history as well, but baseline randomization would not. For instance, welfare spending may have a strong effect on terrorism levels which, in turn, affect future welfare spending. Under this type of feedback, a lagged dependent variable must be in the conditioning set V_{it} and strict exogeneity will be violated.

3.3 Unmeasured confounding and fixed effects assumptions

Sequential ignorability is a selection-on-observables assumption—the researcher must be able to choose a (time-varying) conditioning set to eliminate any confounding. A oft-cited benefit of having repeated observations is that it allows scholars to estimate causal effects in spite of time-constant unmeasured confounders. Linear fixed effects models, for instance, have the benefit of adjusting for all time-constant covariates, measured or unmeasured. This would be very helpful if, for instance, each country had its own baseline level of welfare spending that was determined by factors correlated with terrorist attacks but the year-to-year variation in spending within a country was exogeneous. At first glance, this ability to avoid time-constant omitted variable bias appears to be a huge benefit.

Unfortunately, these fixed effects estimation strategies require within-unit baseline randomization to identify any quantity other than the contemporaneous effect of treatment (Sobel, 2012; Imai and Kim, 2016). Specifically, standard fixed effects models assume that previous values of covariates like GDP growth or lagged terrorist attacks (that is, the LDV) have no impact on the current value of welfare spending. Thus, to estimate any effects of lagged treatment, fixed effects models would allow for time-constant unmeasured confounding but would also rule out a large number of TSCS applications where there is feedback between the covariates and the treatment. Furthermore, the assumptions of fixed-effects-style models in nonlinear settings can impose strong restrictions on over-time variation in the treatment

[5] In certain parametric models, the LDV can be interpreted as summarizing the effects of the entire history of treatment. More generally, the LDV may effectively block confounding for contemporaneous treatment even if it has no causal effect on the current outcome.

and outcome (Chernozhukov et al., 2013). For these reasons, and because there is a large TSCS literature in political science that relies on selection-on-observables assumptions, we focus on situations where sequential ignorability holds. We return to the avenues for future research on fixed effects models in this setting in the conclusion.

4 The post-treatment bias of traditional TSCS models

Under sequential ignorability, standard TSCS models like the ADL model in Section 2.3 can become biased for common TSCS estimands. The basic problem with these models is that sequential ignorability allows for the possibility of post-treatment bias when estimating lagged effects in the ADL model. While this problem is well known in statistics (Rosenbaum, 1984; Robins, 1997; Robins, Greenland and Hu, 1999), we review it here in the context of TSCS models to highlight the potential for biased and inconsistent estimators.

The root of the bias in the ADL approach is the nature of time-varying covariates, Z_{it}. Under the assumption of baseline randomization, there is no need to control or adjust for these covariates beyond the baseline covariates, Z_{i0}, because treatment is assigned at baseline—future covariates cannot confound past treatment assignment. The ADL approach thrives in this setting. But when baseline randomization is implausible, as we argue is true in most TSCS settings, we will typically require conditioning on these covariates to obtain credible causal estimates. And this conditioning on Z_{it} is what can create large biases in the ADL approach.

To demonstrate the potential for bias, we focus on a simple case where we are only interested in the first two lags of treatment and sequential ignorability assumption holds with $V_{it} = \{Y_{i,t-1}, Z_{it}, X_{i,t-1}\}$. This means that treatment is randomly assigned conditional on the contemporaneous value of the time-varying covariate and the lagged values of the outcome and the treatment. Given this setting, the ADL approach would model the outcome as follows:

$$Y_{it} = \beta_0 + \alpha Y_{i,t-1} + \beta_1 X_{it} + \beta_2 X_{i,t-1} + Z_{it}'\delta + \varepsilon_{it}. \tag{17}$$

Assuming this functional form is correct and assuming that ε_{it} are independent and identically distributed, this model would consistently estimate the contemporaneous effect of treatment, X_{it}, given the sequential ignorability assumption. But what about the effect of lagged treatment? In the ADL approach, one would combine the coefficients as $\widehat{\alpha}\widehat{\beta}_1 + \widehat{\beta}_2$. The problem with this approach is that, if Z_{it} is affected by $X_{i,t-1}$, then Z_{it} will be post-treatment and in many cases induce bias in the estimation of $\widehat{\beta}_2$ (Rosenbaum, 1984; Acharya, Blackwell and Sen, 2016). Why not simply omit Z_{it} from our model? Because this would bias the estimates of the contemporary treatment effect, $\widehat{\beta}_1$ due to omitted variable bias.[6]

In this setting, there is no way to estimate the direct effect of lagged treatment without bias with a single ADL model. Unfortunately, even weakening the parametric modeling assumptions via matching or generalized additive models will fail to overcome this problem—it is inherent to the data generating process (Robins, 1997). These biases exist even in favorable settings for the ADL, such as when the outcome is stationary and treatment effects are constant over time. Furthermore, as discussed above, standard fixed effects models cannot eliminate this bias because it involves time-dependent causal feedback. Traditional approaches can only avoid the bias under special circumstances such as when treatment is randomly assigned at baseline or when the time-varying covariates are completely unaffected by treatment. Both of these assumptions lack plausibility in TSCS settings, which is why many TSCS studies control for time-varying covariates. Below we demonstrate this bias in simulations, but we first turn to two methods from biostatistics that can avoid these biases.

5 Two methods for estimating the effect of treatment histories

If the traditional ADL model is biased in the presence of time-varying covariates, how can we proceed with estimating both contemporaneous and lagged effect of treatment in the TSCS setting? In this section, we show how to estimate the causal quantities of interest in Section 2 under sequential ignorability using two approaches developed in biostatistics to specifically address this potential for bias in this type of setting. The first approach is based on structural nested mean models (SNMMs), which, in their simplest form, represent an extension of the ADL approach to avoid the post-treatment bias described above. The second class of estimators, based on marginal structural models (MSM) and inverse probability of treatment weighting (IPTW), is *semiparametric* the sense that it models the treatment history, but leaves the relationship between the outcome and the time-varying covariates unspecified. Because of this, MSMs have the advantage of being robust to our ability or inability to model the outcome. We focus our attention on these two broad classes of models because they are commonly used approaches that both (a) avoid post-treatment bias in this setting and (b) do not require the parametric modeling of time-varying covariates.

[6] A second issue is that ADL models often only include conditioning variables to identify the contemporaneous effect, not any lagged effects of treatment. Thus, the effect of $X_{i,t-1}$ might also suffer from omitted variable bias. This issue can be more easily corrected by including the proper condition set, $V_{i,t-1}$, in the model.

One modeling choice that is common to all of these approaches, including the ADL, is the choice of causal lag length. Should we attempt to estimate the effect of the entire history of welfare spending on terrorist incidents with potential outcome $Y_{it}(x_{1:t})$? Or should we only investigate the contemporaneous and first lagged effects with potential outcome $Y_{it}(x_{t-1}, x_t)$? As we discussed above, we can always focus on effects that marginalize over lags of treatment beyond the scope of our investigation. Thus, this choice of lag length is less about the "correct" specification and more about choosing what question the researcher wants to answer. A separate question is what variables and their lags need to be included in the various models in order for our answers to be correct. We discuss the details of what needs to be controlled for and when in our discussion of each estimator.

5.1 Structural nested mean models

Our first class of models, called structural nested mean models, can be seen as an extension of the ADL approach that allows for estimation of lagged effects in a relatively straightforward manner (Robins, 1986, 1997). At their most general, these models focus on parameterizing a conditional version of the lagged effects (that is, the impulse response function):[7]

$$b_t(x_{1:t}, j) = E[Y_{it}(x_{1:t-j}, 0_j) - Y_{it}(x_{1:t-j-1}, 0_{j+1}) | X_{1:t-j} = x_{1:t-j}]. \tag{18}$$

Robins (1997) refers to these impulse responses as "blip-down functions." This function gives the effect of a change from 0 to x_{t-j} in terms of welfare spending on levels of terrorism at time t, conditional on the treatment history up to time $t-j$. Inference in SNMMs focuses on estimating the causal parameters of this function. The conditional mean of the outcome given the covariates needs to be estimated as part of this approach, but this is seen as a nuisance function rather than the object of direct interest.

Given the chosen lag length to study, a researcher must only specify the parameters of the impulse response up to that many lags. If we chose a lag length of 1, for example, then we might parameterize the impulse response function as:

$$b_t(x_{1:t}, j; \gamma) = \gamma_j x_{t-j}, \qquad j \in \{0, 1\}. \tag{19}$$

Here, γ_j is the impulse effect of a one-unit change of welfare spending at lag j on levels of terrorism which does not depend on the past treatment history, $x_{1:t-1}$ or the time period t.

[7] Because of focus on being faithful to the ADL setup, we assume that the lagged effects are constant across levels of the time-varying confounders as is standard in ADL models. One can include interactions with these variables, though SNMMs then require additional models for Z_{it}. See Robins (1997, section 8.3) for more details.

Keeping the desired lag length, we could generalize this specification and have an impulse response that depended on past values of the treatment:

$$b_t(x_{1:t}, j; \gamma) = \gamma_{1j} x_{t-j} + \gamma_{2j} x_{t-j} x_{t-j-1}, \qquad j \in \{0, 1\}, \tag{20}$$

where γ_{2j} captures the interaction between contemporaneous and lagged values of welfare spending. Note that, given the definition of the impulse response, if $x_t = 0$, then $b_t = 0$ since this would be comparing the average effect of a change from 0 to 0. Choosing this function is similar to modeling $X_{i,t-j}$ in a regression—it requires the analyst to decide what nonlinearities or interactions are important to include for the effect of treatment. If Y_{it} is not continuous, it is possible to choose an alternative functional form (such as one that uses a log link) that restricts the effects to the proper scale (Vansteelandt and Joffe, 2014).

Note that the non-interactive impulse response function in (19) can be seen as an alternative parameterization of the ADL (1,1) in (4). When $j = 0$ in (19) and an ADL (1,1) model holds, then the contemporaneous effect of γ_0 corresponds to the β_1 parameter from the ADL model. When $j = 1$ in (19) and an ADL (1,1) model holds, then the impulse response effect of γ_1 corresponds to the $\alpha\beta_1 + \beta_2$ combination of parameters from the ADL model. We derive this connection in more detail below, but one important difference can be seen in this example. The SNMM approach directly models the impulse response effects while the ADL model recreates the impulse response effects from all constituent path effects.

The key to the SNMM identification approach is that problems of post-treatment bias can be avoided by using a transformation of the outcome that leads to easy estimation of each conditional impulse responses (γ_j). This transformation is

$$\widetilde{Y}^j_{it} = Y_{it} - \sum_{s=0}^{j-1} b_t(X_{i,1:t}, s), \tag{21}$$

which, under the modeling assumptions of equation (19), would be

$$\widetilde{Y}^j_{it} = Y_{it} - \sum_{s=0}^{j-1} \gamma_s X_{i,t-s}. \tag{22}$$

These transformed outcomes are called the *blipped-down* or *demediated* outcomes. For example, the first blipped-down outcome, which we will use to estimate first lagged effect, subtracts the contemporaneous effect for each unit off of the outcome, $\widetilde{Y}^1_{it} = Y_{it} - \gamma_0 X_{it}$. Intuitively, this transformation subtracts off the effect of j lags of treatment, creating an estimate of the counterfactual level of terrorism at time t if welfare spending had been set to 0 for j

periods before t. Robins (1994) and Robins (1997) show that, under sequential ignorability, the transformed outcome, $\widetilde{Y}^j_{it}$, has the same expectation as this counterfactual, $Y_{it}(x_{1:t-j}, 0_j)$, conditional on the past. Thus, we can use the relationship between $\widetilde{Y}^j_{it}$ and $X_{i,t-j}$ as an estimate of the j-step lagged effect of treatment, which can be used to create $\widetilde{Y}^{j+1}_{it}$ and estimate the lagged effect for $j+1$. This recursive structure of the modeling is what gives SNMM the "nested" moniker.

We focus on one approach to estimating the parameters called *sequential g-estimation* in the biostatistics literature (Vansteelandt, 2009).[8] This approach is similar to an extension of the standard ADL model in the sense that it requires modeling the conditional mean of the (transformed) outcome to estimate the effect of each lag under study. In particular, for lag j the researcher must specify a linear regression of $\widetilde{Y}^j_{it}$ on the variables in the assumed impulse response function, $b_t(x_{1:t}, j; \gamma)$ and whatever covariates are needed to satisfy sequential ignorability.

For example, suppose we focused on the contemporaneous effect and the first lagged effect of welfare spending and we adopted the simple impulse response $b_t(x_{1:t}, j; \gamma) = \gamma_j x_{t-j}$ for both of these effects. As in Section 4, we assume that sequential ignorability held conditional on $V_{it} = \{X_{i,t-1}, Y_{i,t-1}, Z_{it}\}$. Sequential g-estimation involves the following steps:

1. For $j = 0$, we would regress the untransformed outcome on $\{X_{it}, X_{i,t-1}, Y_{i,t-1}, Z_{it}\}$, just as we would for the ADL model. If the modeling is correctly specified (as we would assume with the ADL approach), the coefficient on X_{it} in this regression will provide an estimate of the blip-down parameter, γ_0 (the contemporaneous effect).

2. We would use $\widehat{\gamma}_0$ to construct the one-lag blipped-down outcome, $\widetilde{Y}^1_{i,t} = Y_{it} - \widehat{\gamma}_0 X_{it}$.

3. This blipped-down outcome would be regressed on $\{X_{i,t-1}, X_{i,t-2}, Y_{i,t-2}, Z_{i,t-1}\}$ to estimate the next blip-down parameter, γ_1.

If more than two lags are desired, we could use $\widehat{\gamma}_1$ to construct the second set of blipped-down outcomes, $\widetilde{Y}^2_{i,t} = \widetilde{Y}^1_{i,t} - \widehat{\gamma}_1 X_{i,t-1}$, which could then be regressed on $\{X_{i,t-2}, X_{i,t-3}, Y_{i,t-3}, Z_{i,t-2}\}$ to estimate γ_2. This iteration can continue for as many lags as desired. What this approach avoids is ever estimating a causal effect while including a post-treatment covariate for that effect. That is, when estimating the effect of welfare spending at lag j, only variables causally prior to welfare spending at that point are included in the regression. Standard errors for all of the estimated effects can be estimated using a consistent variance estimator presented in the Supplemental Materials or via a block bootstrap.

[8] See Acharya, Blackwell and Sen (2016) for an introduction to this method in political science.

This sequential g-estimation approach requires the correct specification of the relationship between the (transformed) outcome and the covariate and treatment histories. It thus requires a similar regression model to the ADL approach described above. More complicated SNMM estimators can incorporate a model for the treatment process, providing some robustness to the modeling choices for the outcome. These estimators are consistent for the parameters of the SNMM when either the model for the (transformed) outcome or the model for the treatment process is correctly specified. This property is called *double robustness* because there are "two shots" to achieve consistency. Vansteelandt and Joffe (2014) provides a review of these methods for SNMMs.

Relationship to the ADL model

As we mentioned in Section 4, the ADL approach and the sequential g-estimation version of SNMM presented above are very similar when the time-varying covariates, Z_{it}, are not affected by treatment. One intuition for this result is that the ADL model and the SNMM with linear model are equivalent when there are no covariates aside from the LDV. To see this, suppose that the ADL model in (4) is correct and perform the first transformation from step 2 above, noting, as above, that the contemporaneous effect is the same for both models $\gamma_0 = \beta_1$:

$$
\begin{aligned}
Y_{it} - \gamma_0 X_{it} &= Y_{it} - \beta_1 X_{it} && (23)\\
&= \beta_0 + \alpha Y_{i,t-1} + \beta_2 X_{i,t-1} + \varepsilon_{it} && (24)\\
&= \beta_0 + \alpha(\beta_0 + \alpha Y_{i,t-2} + \beta_1 X_{i,t-1} + \beta_2 X_{i,t-2} + \varepsilon_{i,t-1}) + \beta_2 X_{i,t-1} + \varepsilon_{it} && (25)\\
&= (\beta_0 + \alpha\beta_0) + \alpha^2 Y_{i,t-2} + \underbrace{(\alpha\beta_1 + \beta_2)}_{\gamma_1} X_{i,t-1} + \alpha\beta_2 X_{i,t-2} + (\alpha\varepsilon_{i,t-1} + \varepsilon_{it}) && (26)
\end{aligned}
$$

From this, we can see that the coefficient on $X_{i,t-1}$ for this transformed outcome is simply the impulse response at lag 1, which is exactly the quantity that the SNMM targets. Given the ADL and SNMM assumptions above, this quantity will be $\alpha\beta_1 + \beta_2$ for the ADL model and γ_1 for the SNMM. Of course, this correspondence will continue for all lagged effects and Table 1 shows how the two sets of quantities relate for various lags.

Furthermore, in the Supplemental Materials we show that the sequential g-estimation estimator with no covariates except a lagged dependent variable is nearly mechanically equivalent to a traditional ADL estimator with one lag. The difference is that the traditional ADL model relies on an assumption that the contemporaneous effect is constant over time, whereas sequential g-estimation relaxes this assumption. This provides an useful interpretation of the ADL model in terms of counterfactual causal effects. It is important to note,

Lag	ADL	SNMM
0	β_1	γ_0
1	$\alpha\beta_1 + \beta_2$	γ_1
2	$\alpha^2\beta_1 + \alpha\beta_2$	γ_2
3	$\alpha^3\beta_1 + \alpha^2\beta_2$	γ_3
4	$\alpha^4\beta_1 + \alpha^3\beta_2$	γ_4

Table 1: The lagged effects, or impulse responses, under the ADL (1,1) in (4) and SNMM in (19).

however, that this equivalence also relies on the form of the ADL model, which uses only three parameters regardless of the number of lags, while the SNMM in this version uses a new parameter for every lag. Additionally, the equivalence disappears once there is an additional time-varying covariate (Z_{it}) in the model.

5.2 Marginal structural models

One potential downside of the SNMM approach is that it requires the analyst to correctly model the relationship between the time-varying covariates and the outcome. This can be difficult when the outcome is a complicated process and there is little theoretical guidance for specifying the outcome-covariate relationships. An alternative that relies instead on modeling the treatment-covariate relationship is called a marginal structural model or MSM (Robins, Hernán and Brumback, 2000).[9] To specify an MSM, we first choose a potential outcome lag length to study and write a model for the marginal mean of those potential outcomes in terms of the treatment history. At the most general, then, an MSM would be the following:

$$E[Y_{it}(x_{1:t})] = g(x_{1:t}; \beta), \tag{27}$$

where the function g operates similarly to a link function in a generalized linear model.[10] These models are similar to the impulse response functions in the SNMM approach, b_t, because they provide structure for the treatment-outcome relationship. For instance, suppose that we were focused on the contemporaneous effect and the effect of the first two lags and so we had to model $E[Y_{it}(x_{t-2:t})] = g(x_{t-2:t}; \beta)$, marginalizing over further lags and other covariates. If Y_{it} were continuous, as in the case of the number of terrorist incidents, we

[9] For a detailed introduction to and application of MSMs in political science, see Blackwell (2013).

[10] These marginal structural models are similar in spirit to *transfer functions* the context of pure time-series data (Box, Jenkins and Reinsel, 2013).

might take g to be linear and focus on the additive effects of each period of treatment:[11]

$$g(x_{t-2:t}; \beta) = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2}. \tag{28}$$

If Y_{it} were binary, we might instead assume g to have a logistic form:

$$g(x_{t-2:t}; \beta) = \frac{\exp(\beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2})}{1 + \exp(\beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2})}. \tag{29}$$

In both of these cases, we have restricted our attention to the last three periods of treatment and so we cannot answer questions about longer-term effects with these models. On the other hand, as we increase the number of lags under study, the number of parameters needed to summarize the effects grows and the model can become unwieldy. Thus, we may consider focusing on the effect of the cumulative number of treated periods, $\sum_{s=1}^{t} x_{is}$. This allows for the entire history of treatment to affect the outcome in a structured, low-dimensional way. Under any of these models, the average causal effect becomes:

$$\tau(x_{1:t}, x'_{1:t}) = g(x_{1:t}; \beta) - g(x'_{1:t}; \beta). \tag{30}$$

Of course, the MSM specification will place restrictions on the average casual effects. A MSM that is a function of only the cumulative treatment, for instance, implies that $\tau(x_{1:t}, x'_{1:t}) = 0$ if $x_{1:t}$ and $x'_{1:t}$ have the same number of treated periods, even if their sequence differs.

How can a researcher estimate an MSM? If one blindly follows model (28) and regresses Y_{it} on $\{X_{i,t-2}, X_{i,t-1}, X_{it}\}$ using ordinary least squares, there will be omitted variable bias in the estimated coefficients. But as we have seen above, simply including time-varying covariates in these models can lead to post-treatment bias. Fortunately, the causal parameters of these models are estimable using an inverse probability of treatment weighting (IPTW) approach where we adjust for time-varying covariates using the propensity score weights, not the outcome model itself, avoiding post-treatment bias (Robins, Hernán and Brumback, 2000). The weighting balances the distribution of the time-varying covariates across values of the treatments, so that omitting these variables in the reweighted data produces no omitted variable bias.

To use IPTW, a researcher must develop a model for the probability of treatment in period t given the variables that satisfy sequential ignorability. For example, suppose that sequential ignorability holds conditional on some conditioning set V_{it}. If X_{it} is binary, then we must obtain a consistent estimate of $\pi_t(v) = \Pr[X_{it} = 1 \mid V_{it} = v]$. This might be a pooled

[11] When the treatment is binary and the chosen lag length is short, we can relax the linearity assumption here by saturating the modeling with all interactions between the periods under study.

logit, a generalized additive model with a flexible functional form, a boosted regression (McCaffrey, Ridgeway and Morral, 2004), or a covariate-balancing propensity score (CBPS) model (Imai and Ratkovic, 2015). The IPTW approach requires this model to provide consistent estimates of the conditional predicted probability of treatment.[12] In spite of this requirement, some methods for propensity score estimation such as CBPS have good finite-sample properties in the face of model misspecification (Imai and Ratkovic, 2015).

We use the predicted probabilities from this treatment model to construct weights for each country-year. For example, suppose that V_{it} included lagged levels of terrorism, $Y_{i,t-1}$, lagged welfare spending, $X_{i,t-1}$, and a set of time-varying covariates, Z_{it}. Then, for a binary treatment, we would construct the weights as:

$$\widehat{SW}_{it} = \prod_{t=1}^{t} \frac{\widehat{\Pr}[X_{it} \mid X_{i,t-1}; \widehat{\gamma}]}{\widehat{\Pr}[X_{it} \mid Z_{it}, Y_{i,t-1}, X_{i,t-1}; \widehat{\alpha}]}. \tag{31}$$

The denominator of each term in the product is the predicted probability of observing unit i's observed treatment status in time t (X_{it}), conditional on the covariates that satisfy sequential ignorability.[13] When we multiply this over time, it is the probability of seeing this unit's treatment history conditional on the past. The numerators here are the marginal probability of the observed treatment history and stabilize the weights to make sure they are not too variable which can lead to poor finite sample performance (Cole and Hernán, 2008). For instance, to construct this numerator we might run a pooled logistic regression of welfare spending in year t on welfare spending in year $t-1$, omitting any time-varying covariates or lagged dependent variables. While this choice of numerator is not required for consistency of the estimator (it can be replaced with 1, for instance), it can help to stabilized weights that are highly variable and thus increase efficiency.

Under these assumptions, the expectation of Y_{it} conditional on $X_{i,1:t}$ in the reweighted data is equal to the MSM:

$$E_{SW}[Y_{it} | X_{i,1:t} = x_{1:t}] = E[Y_{it}(x_{1:t})]. \tag{32}$$

Here $E_{SW}[\cdot]$ is the expectation in the reweighted data. For example, if we used the linear MSM in (28), then we can estimate the causal parameters of MSM by running a weighted least squares regression of the outcome, Y_{it} on $\{X_{i,t-2}, X_{i,t-1}, X_{it}\}$ with $\widehat{SW}_{it}$ as the weights.

[12] This requirement makes it difficult to apply IPTW to fixed-effects settings with binary treatments since estimating the unit-specific models would face an incidental parameters problem, at least for a fixed time window.

[13] To ensure the weights are well-defined, the conditional probability of treatment given the past must be bounded away from 0 and 1. In the biostatistics literature, this assumption is called *positivity* and is similar to the overlap condition in the matching literature.

If sequential ignorability holds, the coefficients on the components of $X_{i,1:t}$ from this regression will have a causal interpretation, though they may depend on the particular modeling choices of the MSM (Robins, Hernán and Brumback, 2000). Standard errors can be estimated via a block bootstrap of units. Note that, unlike the ADL and SNMMs, this approach does not require a model for the relationship between the time-varying covariates and the outcome.

Finally, when the conditional probability of treatment is close to 0 or 1, the IPTW approach can have large and unstable weights, leading to high variance and sometimes small sample biases (Imai and Ratkovic, 2015). SNMMs, on the other hand, tend to be more stable in this setting. And while MSMs and SNMMs can accommodate general types of covariates, SNMMs also tend to be more stable when the treatment is continuous since weighting by a continuous density (as would be required with IPTW) is sensitive to small perturbations in the data (Goetgeluk, Vansteelandt and Goetghebeur, 2008).

5.3 Modeling checklist

In this section, we review the key modeling choices required to implement these methods.

Causal lag length: First, one must choose the lag length to study. At the most general, one can investigate the effect of an entire treatment history, but these are usually too highly dimensional to study without further assumptions. In MSMs, one can reduce this dimensionality by assuming that treatment history only affects the outcome through the average level of treatment or the cumulative amount of treatment up to time t. Alternatively, a researcher can focus on the marginal effects of the last j lags of treatment.

Conditioning set for sequential ignorability: Separate from the question of what to study, is the question of what covariates to choose so that the question can be answered. Sequential ignorability is an assumption about conditional independence: welfare spending is independent of the potential outcomes conditional on past treatment and some set of baseline and time-varying covariates. Thus, scholars must choose a set of covariates for each time period that blocks all confounding for the treatment-outcome relationship—that is, there must be no omitted variables after controlling for that conditioning set. In the context of welfare spending, these covariates might include: lagged welfare spending ($X_{i,t-1}$) and the lagged terrorist activity ($Y_{i,t-1}$), time-varying economic factors like GDP growth unemployment (Z_{it}), and baseline characteristics such as region of the world (Z_{i0}). This conditioning set of variables will be included in the models for the outcome in the SNMM approach or the in the models for the treatment in the MSM approach.

Modeling treatment or outcome: In all of the methods described in this paper, the analyst must specify the functional form of how the treatment history and outcome relate. In the SNMM approach, this is done through the IRF or blip-down functions, while in the MSM, this is done through the specification of the MSM itself. To actually estimate these models, however, a researcher must additionally model either the relationship between the outcome and the covariates in the conditioning set (in the ADL or SNMM approaches) or the relationship between the treatment and the covariates in the conditioning set (in the MSM approach). Because the quality of the casual estimates depends on this modeling, we encourage researchers to choose the approach for which more substantive knowledge can be mustered to help with the modeling task. For example, suppose we were estimating the effect of central bank interest rate changes on support for incumbent candidates. It may be easier to model the central bank interest rate changes if we have detailed information on central bank deliberations about changes that help us specify the model. In other cases, there may be more substantive information about the outcome model.

Functional form assumptions: Finally, in either model that is chosen, the analyst must correctly specify the model in the sense that the functional form assumed for the variables in the conditioning set is correct. This may require, for instance, taking the natural log of population, including a squared term for GDP growth to allow for a nonlinear relationship, or including an interaction between two important covariates. This task is common to all modeling strategies and is not unique to the current setting. A researcher can weaken these modeling assumptions by replacing linear or generalized linear models with generalized additive models that allow the functional forms of chosen covariates to be estimated along with the other model parameters (Beck and Jackman, 1998).[14] Finally, we note that all of these models, including the ADL, assume the correct periodization and causal structure in the data. For instance, we must know that Z_{it} is in fact causally prior to X_{it} even though they could be measured in the same time period. Thus, a significant amount of subject-matter expertise may be required to ensure these specifications are correct. For the purposes of our discussion, however, these issues are common to all TSCS methods and will not affect the comparison between methods.

[14] A growing literature has developed several approaches to flexibly estimating linear (and sometimes generalized linear) models that would reduce the modeling burden on the researcher even further. These models include sparse additive models (Ravikumar et al., 2009), kernal regularized least squares (Hainmueller and Hazlett, 2014), and generalized boosted models (McCaffrey, Ridgeway and Morral, 2004).

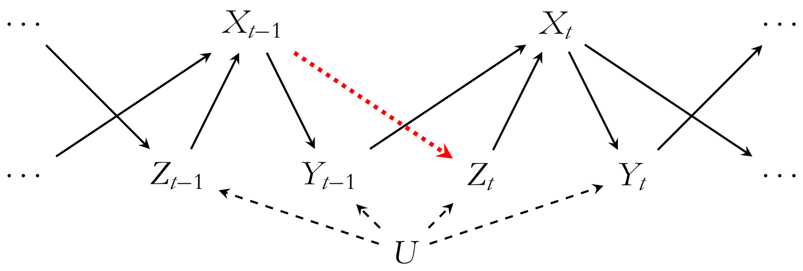


Figure 4: Direct acyclic graph of the simulation study. Dotted red line represents the key causal quantity varied in the simulations, whether past treatment affects future covariates. Dashed lines represent unmeasured confounding.

6 Simulation evidence

To investigate the small sample properties of the various estimators, we conducted a simulation study of a TSCS setting with a treatment, an outcome, and a single covariate, all time-varying. We describe the simulation in more detail in the Supplemental Materials, but the main causal relationships in the design are displayed in Figure 4. Here, the treatment history only has a contemporaneous effects— lagged treatments, $X_{i,1:t-1}$, and outcomes, $Y_{i,1:t-1}$, have no direct or indirect effect on current outcomes, Y_{it}, that don't go through current treatment, X_{it}. The treatment-outcome relationship is confounded due to a time-constant unmeasured confounder, U_i, but conditioning on $\{Y_{i,t-1}, Z_{it}\}$ can block this confounding and ensure sequential ignorability for X_{it}. Finally, the distribution of $\{Y_{it}, X_{it}, Z_{it}\}$ is Markovian and stationary within each unit, which should be an ideal setting for the ADL approach.

To show how the causal structure can affect the performance of the estimators, we consider two scenarios that vary the feedback between the treatment and the time-varying covariate. In the first, we allow for lagged treatment to affect future covariates so that $X_{i,t-1} \to Z_{it}$, and in the second, we close this path. These two scenarios represent when the time-varying counfounder, Z_{it}, is post-treatment to lagged treatment and when it is not. Unfortunately, when Z_{it} is post-treatment, conditioning on it will induce post-treatment bias for the effect of lagged treatment, because conditioning will open a backdoor path from Z_{it} through U_i to Y_{it}. However, we must condition on Z_{it} to remove the omitted variable bias for contemporaneous treatment. This is the dilemma that traditional TSCS models like the ADL model cannot solve, because a single model cannot simultaneously control for Z_{it} and not control for Z_{it}.

We generate data from this model varying numbers of time periods and units and focus on the lagged effect of treatment, $E[Y_{it}(1,0) - Y_{it}(0,0)]$, which in this case is 0. We compared several methods for estimating this quantity: (1) an ADL as in Section 4 with estimate $\widehat{\alpha}\widehat{\beta}_1 + \widehat{\beta}_2$; (2) an SNMM sequential g-estimation with additive linear models for the outcome for each lag; (3) a linear, additive MSM with $g(x_{t-1}, x_t; \beta) = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1}$; and (4) a raw model with no controls that only includes X_{it} and $X_{i,t-1}$.[15] For reference, we also compare these estimators to the infeasible estimator that simply takes the sample average of $Y_{it}(0,1) - Y_{it}(0,0)$ across all unit-periods.

Figure 5 shows the results of these simulation. The left column shows the root mean squared error (RMSE) of the various estimators when the time-varying confounder, Z_{it}, is affected by past treatment. While the SNMM and MSM approaches have roughly similar estimation error across different sample sizes, they vastly outperform the ADL approach. The high RMSE of the ADL approach that persists across sample sizes is due to a large degree of post-treatment bias on the coefficient on $X_{i,t-1}$ due to conditioning on Z_{it}. This bias propagates to the ADL computation of the total effect of lagged treatment. The ADL model even performs worse that a model that has significant omitted variable bias due to excluding all time-varying covariates from the model (labelled "Raw" in the figure). These results hold even though the DGP here is stationary and the sample size and the number of time periods are small and similar in size, meaning that they are unlikely to depend on a "large-N, small-T" setting.

The right column of Figure 5 shows the results when the time-varying confounder is not affected by treatment. Here, the ADL has lower estimation error than any of the other methods, slightly beating out SNMM. The ADL model performs well in this setting since the lagged dependent variable is the only variable affected by past treatment. As we show in the Supplemental Materials, in this case the ADL model is essentially a correctly specified SNMM. This correct specification breaks down when time-varying covariates are affected by treatment. Given the robustness of SNMM to this feature of the casual process and given the similarity in modeling choices for the SNMM and ADL approaches, we recommend using the SNMM as a working replacement for the ADL model whenever lagged effects are of interest.[16]

7 Empirical illustration: Welfare spending and terrorism

Burgoon (2006) studied the effect of domestic welfare spending on terrorist activity within

[15] For each of these approaches except the last, we include the relevant covariates, correctly specified in terms of their functional form. In the Supplemental Materials, we weaken use misspecified functional forms for all models and the results are qualitatively similar.

[16] The ADL approach is also biased when omitting Z_{it}, but including $Y_{i,t-1}$ (results not reported here). There is no permutation of controls that eliminate the bias of ADL when Z_{it} is affected by $X_{i,t-1}$.

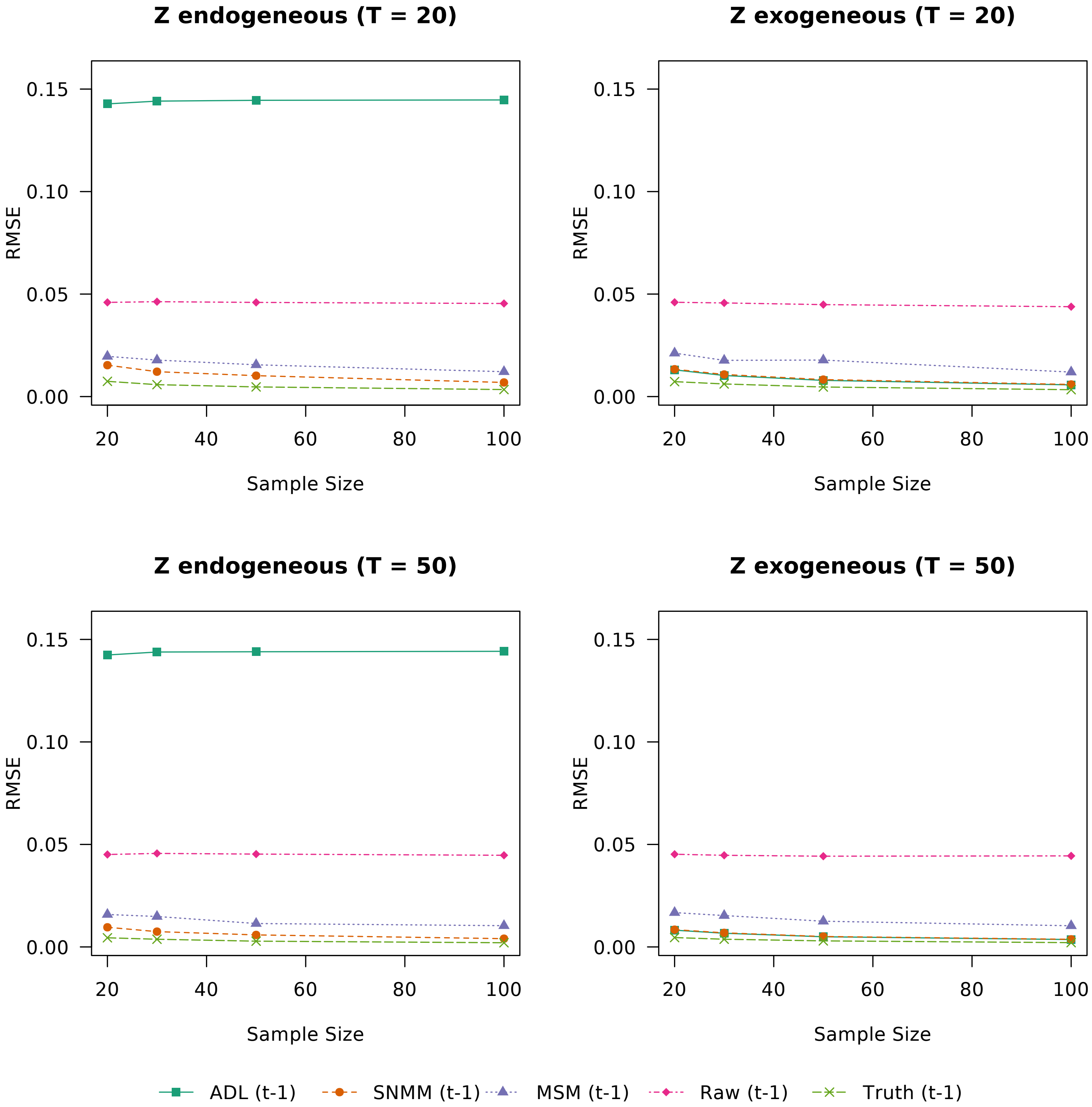


Figure 5: Simulation results when the time-varying confounder is post-treatment (left column) and when the time-varying confounder is not post-treatment (right column). Points represent the root mean squared error (RMSE) of each estimator for the lagged effect of treatment.

countries and used TSCS data to show that increasing spending leads to lower levels of terrorist activity within a country. But how does the timing of this spending matter? Can we assess the effects of lagged government spending on future values of terrorist activity? We apply the models of this paper to show how they differ from traditional approaches to answering these questions.

To do this, we closely follow the specification of Burgoon (2006). The dependent variable is the number of transnational terrorist incidents occurring in a country, omitting purely domestic terrorism such as the Oklahoma City bombing in the United States. Burgoon (2006) uses a negative binomial regression model to estimate the effect of contemporaneous spending, whereas we use a linear model. To account for overdispersion, we use the square root of the number of transnational terrorist incidents as our dependent variable. This approach recovers very similar substantive results as that of Burgoon (2006).

A first step for any of the methods we describe in the paper is to choose a conditioning set of covariates that can satisfy sequential ignorability. Given that Burgoon (2006) interprets the effect of spending in a causal fashion, we follow this selection-on-observables approach and assume that the control variables in the paper's models are sufficient to satisfy sequential ignorability. These include a set of regional and year dummies as baseline covariates and the following time-varying covariates: a lagged dependent variable, left-party control of government, Polity score and its lag, log population, a measure of government capability, whether the country is in a conflict, and the amount of trade logged. In this context, the sequential ignorability assumption states that welfare spending is exogeneous with respect to terrorism conditional on previous terrorist incidents, the time-varying covariates, and region and year fixed effects. Note that if there were unmeasured confounding beyond these controls, the estimates of causal effects in this application could be biased. One could, however, perform a sensitivity analysis to determine how much of the estimated effect disappears under various departures from sequential ignorability (Blackwell, 2014).

To begin, we compare how the ADL and the SNMM approaches differ in terms of their estimates in this context. For the SNMM, we assume each lag has a simple additive effect as in (19), $\gamma_j x_{t-j}$, with no interactions between treatment and lagged treatment. For the ADL model, we use the specification described above while including a lag of treatment in the model to allow for some flexibility in the lag structure. We use the formulas for calculating lagged effects from an ADL model, as described in Section 2.3. This ADL regression is also the first-stage regression for our sequential g-estimation approach, since under sequential ignorability, it can estimate the contemporaneous effect of treatment. Note that the functional form assumptions for the ADL and the SNMM approach are the same, so that any differences between these methods are likely due to the above biases of the ADL model. We

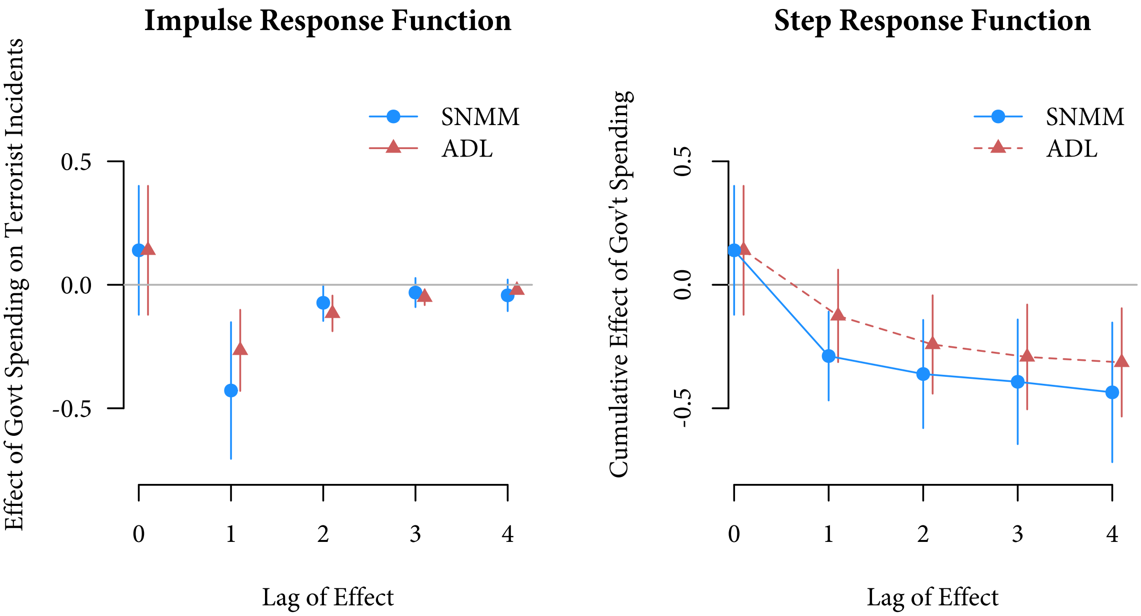


Figure 6: Left: Estimated effect of government spending on the terrorist incidents at various lags, along with 95% confidence intervals. Right: implied step response function at various lags. Data from Burgoon (2006).

focus on a lag length of four years for comparing the SNMM and ADL approaches. Finally, we use the consistent and cluster-robust variance estimator in the Supplemental Materials for the SNMM and a standard cluster-robust variance estimator for the ADL.

Figure 6 shows the estimated contemporaneous and lagged effect of welfare spending on terrorist activity. For instance, one-year lag has $\widehat{\gamma}_1$ for the SNMM, estimated from a regression of the blipped-down outcome on lagged treatment and its conditioning set, and $\widehat{\alpha}\widehat{\beta}_1 + \widehat{\beta}_2$ for the ADL approach. The two approaches are equivalent for the contemporaneous effect but differ in their estimates of the lagged effect. Both methods show a significant and negative effect on lagged spending, but the coefficient from the SNMM approach is about 60% larger in magnitude than the ADL approach. These differences continue with the lags—the effect of the second and third lags are 60% greater in the ADL approach, whereas the effect of the fourth lag is almost double the magnitude for SNMMs. These differences lead to large differences in the estimated cumulative effect of the step response function at the end of four years, with the SNMM estimate almost 40% larger in magnitude.

Why do these differences between the SNMM and the ADL occur? Differences in assumptions about functional form of the covariates are ruled out since the SNMM and ADL models handle these covariates in the exact same way. Furthermore, each of the two approaches rely on a similar assumption about no unmeasured confounding. We believe that the difference between these two approaches is in the post-treatment bias induced by conditioning on the time-varying controls in the ADL approach. In this case, it is highly unlikely that the time-varying covariates are exogenous to welfare spending. For example, one time-varying covariate is the proportion of the government held by left-wing parties. It would be unreasonable to assume that past values of welfare spending are unrelated to future electoral prospects of leftist parties, as would have to be the case for the ADL model to be correct in this case. Indeed, if we regress proportion of the government controlled by leftist parties on lagged welfare spending and the conditioning set for lagged spending, there is a statistically significant and positive coefficient on lagged welfare spending. Thus, it does appear that post-treatment bias could loom large in the estimated effects of the ADL approach.

In the above analysis, we focused on a lag length of 4, even though the data run from 1978 until 1995. Can we learn more about the effects of the history of welfare spending on terrorism? To do this, we turn to marginal structural models where we can develop models that summarize the effects of the entire treatment history in low dimensions. We have seen that lagged welfare spending appears to have an effect and so we may want to know if having a long history of spending also decreases incidences of terrorism. To implement this, we first create a binary measure of welfare spending, X^*_{it}, that is 1 if the country-year had spending (as a function of its GDP) above the global average and 0 if the spending was below the average. We then specify the following MSM:

$$E[Y_{it}(x^*_{i,1:t})] = \beta_0 + \beta_1 x^*_{it} + \beta_2 \left(\sum_{s=1}^{t-1} x^*_{is} \right). \tag{33}$$

Here, the mean of the potential outcomes is a function of the contemporaneous level of spending and the number of lagged periods that have above-average spending. We focus on this simple model, though it is possible to include further lags or interactions between different parts of the history.

We use three approaches to estimating the parameters of the MSM. First, we take the standard ADL-like approach of including the entire set of baseline and time-varying covariates in a regression model. Second, we run the same regression with only the baseline covariates. As we have discussed above, the first of these approaches is likely to produce post-treatment bias and the second is likely to produce omitted variable bias. We compare these to a third approach that uses the IPTW method described above. To create the weights, we fit a logistic regression of the binary treatment on the first two lags of treatment, the cumulative sum of treatment through $t-3$, and the baseline and time-varying covariates described above. We use predicted probabilities from this model to create the weights as

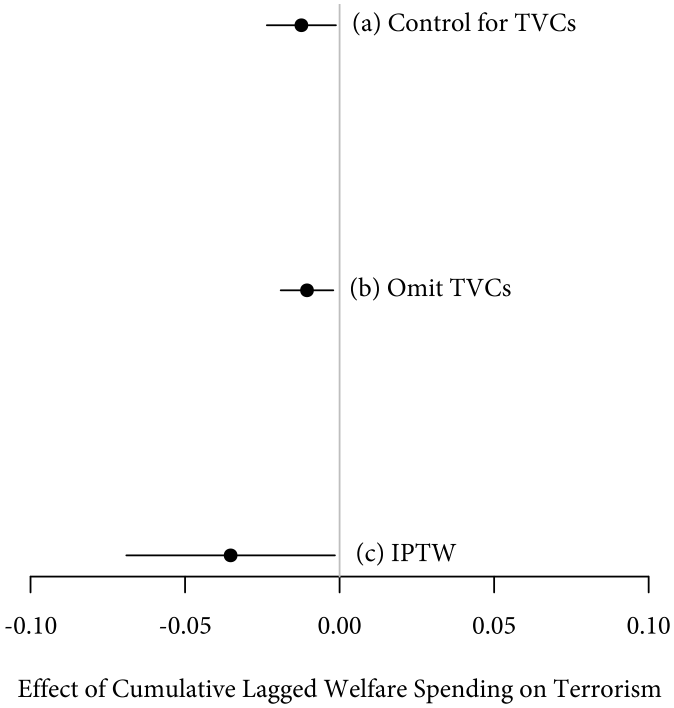


Figure 7: Estimated effect of the cumulative number of high welfare spending years through $t - 1$ on terrorism incidents in year t, fixing welfare spending in year t. The three approaches are (a) when controlling for time-varying covariates, (b) when omitting those variables, and (c) using IPTW. Lines are 95% confidence intervals based on a block bootstrap with 1,000 replications.

in (31), which we use a in WLS regression of the above MSM. We use a block bootstrap to estimate standard errors and trim the weights at 10 to help guard against highly unstable weights (Cole and Hernán, 2008).

Figure 7 shows the results of these models. Both of the traditional approaches estimate a relatively small negative effect of lagged welfare spending on terrorism. The IPTW approach, on the other hand, shows a much larger negative and statistically significant impact, which is consistent with the results of the analysis from both the SNMM and ADL models above. It is interesting to note that the implied post-treatment and omitted variable biases in the first and second models, respectively, are in the same direction. This agreement tempts us to confirm the approximate validity of their results; after all, a natural intuition would be that the true effect must be between these two estimates. Unfortunately, this intuition, while natural, is incorrect. The biases of both approaches can be in the same direction, negating their usefulness as bounds (Blackwell, 2013). Finally, we note that the rather large increase the standard errors in the IPTW approach is driven in part by large weights due to predicted probabilities being close to 0 or 1. This can happen with slowly-changing treatments and is one reason to prefer an SNMM approach in this setting.

8 Conclusions, drawbacks, and future research

Repeated measurements over time of countries, people, or governments expand the scope of causal inference methods. TSCS data allow us to estimate both contemporaneous effects and the effects of more distant lags of treatment. But with an expanded scope comes complications. The usual TSCS regression methods break down for lagged effects. Nevertheless, we have shown that two approaches developed in biostatistics can overcome these difficulties and recover effect estimates across a wide variety of settings.

Both SNMMs and MSMs have their own drawbacks, of course. Even though sequential ignorability nonparametrically identifies any average causal effect of a treatment history, both approaches will almost always depend on modeling to estimate these effects since the covariates needed to justify such an assumption will be highly dimensional. While these modeling assumptions can be weakened to some extent through generalized additive models or other semiparametric techniques, there will always be some degree of model dependence that follows from these approaches. Another problem is that sequential ignorability is a strong, untestable assumption that might be violated. One approach to mitigating this problem is to conduct a formal sensitivity analysis using the methods of either Blackwell (2014) or relying the bias formulas presented in Acharya, Blackwell and Sen (2016). These sensitivity analyses can give researchers a sense of how reliant their results are on sequential ignorability holding.

In this paper, we focused on the usual sequential ignorability assumption as commonly invoked in epidemiology. Many TSCS applications in political science rely on a "fixed effects" assumption that there is time-constant, unmeasured heterogeneity in units. Linear models can easily handle these types of assumptions, though nonlinear fixed effects models pose greater difficulties. Estimating the above causal quantities with these models, however, remains elusive except under strong assumptions like baseline randomization (Chernozhukov et al., 2013; Sobel, 2012). A valuable direction for future work would be to develop fixed effects methods that could estimate causal effects under a within-unit version of sequential ignorability.

Bibliography

Abbring, J. H. and G. J. van den Berg. 2003. "The Nonparametric Identification of Treatment Effects in Duration Models." *Econometrica* 71(5):1491–1517.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–529.

Beck, Nathaniel and Jonathan N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6(1):1–36.

Beck, Nathaniel and Jonathan N. Katz. 2011. "Modeling Dynamics in Time-Series–Cross-Section Political Economy Data." *Annual Review of Political Science* 14(1):331–352.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2):596–627.

Blackwell, Matthew. 2013. "A Framwork for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2):504–520.

Blackwell, Matthew. 2014. "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." *Political Analysis* 22(2):169–182.

Box, George EP, Gwilym M Jenkins and Gregory C Reinsel. 2013. *Time series analysis: forecasting and control*. Wiley.

Burgoon, Brian. 2006. "On Welfare and Terror Social Welfare Policies and Political-Economic Roots of Terrorism." *Journal of Conflict Resolution* 50(2):176–203.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81(2):535–580.

Cole, Stephen R. and Miguel A. Hernán. 2008. "Constructing inverse probability weights for marginal structural models." *American Journal of Epidemiolology* 168(6):656–64.

De Boef, Suzanna and Luke Keele. 2008. "Taking Time Seriously." *American Journal of Political Science* 52(1):185–200.

Gerber, Alan S, James G. Gimpel, Donald P. Green and Daron R Shaw. 2011. "How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105(01):135–150.

Goetgeluk, Sylvie, Sijn Vansteelandt and Els Goetghebeur. 2008. "Estimation of Controlled Direct Effects." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70(5):1049–1066.

Greene, William H. 2012. *Econometric analysis*. 7 ed. Printice Hall.

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.

Hernán, Miguel A., Babette A. Brumback and James M. Robins. 2001. "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments." *Journal of the American Statistical Association* 96(454):440–448.

Imai, Kosuke and In Song Kim. 2016. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Working Paper.
URL: *http://imai.princeton.edu/research/files/FEmatch.pdf*

Imai, Kosuke and Marc Ratkovic. 2015. "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association* 110(511):1013–1023.

Liang, Kung-Yee and Scott L Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73(1):13–22.

McCaffrey, Daniel F, Greg Ridgeway and Andrew R Morral. 2004. "Propensity score estimation with boosted regression for evaluating causal effects in observational studies." *Psychol Methods* 9(4):403–425.

Ravikumar, Pradeep, John Lafferty, Han Liu and Larry Wasserman. 2009. "Sparse additive models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(5):1009–1030.

Robins, James M. 1986. "A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect." *Mathematical Modelling* 7(9-12):1393–1512.

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics* 23(8):2379–2412.

Robins, James M. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane. Vol. 120 of *Lecture Notes in Statistics* New York: Springer-Verlag pp. 69–117.

Robins, James M., Miguel A. Hernán and Babette A. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.

Robins, James M., Sander Greenland and Fu-Chang Hu. 1999. "Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome." *Journal of the American Statistical Association* 94(447):–687.

Rosenbaum, Paul R. 1984. "The consquences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society. Series A (General)* 147(5):656–666.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6(1):34–58.

Shephard, Neil and Iavor Bojinov. 2017. "Time series experiments and causal estimands: exact randomization tests and trading." Working Paper.
URL: *https://scholar.harvard.edu/files/shephard/files/cause20170718.pdf*

Sobel, Michael E. 2012. "Does Marriage Boost Men's Wages?: Identification of Treatment Effects in Fixed Effects Regression Models for Panel Data." *Journal of the American Statistical Association* 107(498):521–529.

Swank, Duane and Sven Steinmo. 2002. "The New Political Economy of Taxation in Advanced Capitalist Democracies." *American Journal of Political Science* 46(3):642–655.

Vansteelandt, Sijn. 2009. "Estimating Direct Effects in Cohort and Case–Control Studies." *Epidemiology* 20(6):851–860.

Vansteelandt, Stijn and Marshall Joffe. 2014. "Structural Nested Models and G-estimation: The Partially Realized Promise." *Statistical Sciecence* 29(4):707–731.

A Long Run Multiplier

Another quantity of interest in traditional TSCS models is the long-run multiplier (LRM), which is the effect of a one-unit change the equilibrium level of X_t on the equilibrium level of Y_t (Greene, 2012, pp. 422, De Boef and Keele, 2008).

We do not fully consider this quantity in this paper because its definition requires additional assumptions that, while relatively easy to discuss within the context of standard parametric TSCS models, are more complicated within the nonparametric approach. Most simply, our fixed time-window approach essentially precludes assessment of this quantity. However, in the interest in clarifying the differences between our approach and the econometric TSCS traditions, we provide a short discussion here. Equilibrium in the potential outcomes framework would be the long-run averages of the potential outcomes under a constant treatment history, if they exist. For instance, the equilibrium level of Y_{it} under treatment would be:

$$\lim_{t \to \infty} E[Y_{it}(1_t)].$$

The LRM, then, is the average causal effect with a comparison between always treated, $(1, 1, \dots)$, and never treated, $(0, 0, \dots)$ as we let t go to infinity:

$$LRM = \lim_{t \to \infty} E[Y_{it}(1_t) - Y_{it}(0_t)]. \tag{34}$$

Identification of the LRM suffers from a few challenges. First, there is no guarantee that the limit in (34) exists. One of the principal reasons the time series literature focuses on the dynamics of the outcome is to ensure that the empirical processes are stable (or stationary) and that such limits exist. Identification, then, will depend on *some* assumptions about the distribution of the dependent variable. Second, even if the limit exists, the LRM cannot be nonparametrically identified without further restrictions since it depends on estimating the mean potential outcome after an infinite number of time periods.

B Consistent variance estimation

In this section we present a consistent estimator for the variance of the SNMM approach with linear models, a no time-varying interactions assumption, and time-constant impulse response. Let w^j_{it} be a $1 \times k_j$ vector of unit i covariates for estimating the IRF at lag j. In general, this vector will be some function of the treatment and the time-varying covariates $w^j_{it} = f(z_{i,1}, x_{i,1}, \dots, z_{i,t-j}, x_{i,t-j})$. Some of these covariates, $\tilde{x}^j_{it}$, are those in the impulse response function and will be used to transform the outcome for the next lag. The remaining covariates, $\tilde{z}^j_{it}$, are covariates used to satisfy sequential ignorability. These two sets of covariates partition the vector, $w^j_{it} = (\tilde{x}^j_{it}, \tilde{z}^j_{it})$.

We collect these vectors into a $T_j \times k_j$ matrix of covariates for unit i at lag j, W_{ij}, where the number of observations per unit, T_j, will depend on the covariates chosen. For instance, certain lagged covariates might be missing in earlier time periods since they would have occurred before baseline measurements. We define the matrices $\tilde{X}_{ij}$ and $\tilde{Z}_{ij}$ similarly. Let $V_i = (y_i, W_{i0}, \ldots, W_{iJ})$ be the observed data for unit i.

Let γ_j be a $k_j \times 1$ vector of coefficients for w_{it}^j and let β_j be the subvector of γ_j associated with the IRF covariates, $\tilde{x}_{it}^j$. The vector $\gamma = (\gamma_0', \gamma_1'''', \ldots, \gamma_J')'$ is the target of inference. Under sequential ignorability and a linear model with time-constant effects for $y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})$, the system of equations must satisfy the following moment conditions:

$$E[W_{i0}'''(y_i - W_{i0}\gamma_0)] = 0 \tag{35}$$

$$E[W_{i1}'(y_i - \tilde{X}_{i0}\beta_0 - W_{i1}\gamma_1)] = 0 \tag{36}$$

$$E[W_{i2}'(y_i - \tilde{X}_{i0}\beta_0 - \tilde{X}_{i1}\beta_1 - W_{i2}\gamma_2)] = 0 \tag{37}$$

$$\vdots \qquad = 0$$

$$E[W_{iJ}'(y_i - (\sum_{j=0}^{J-1} \tilde{X}_{ij}\beta_j) - W_{iJ}\gamma_J)] = 0 \tag{38}$$

To simplify notation, we assume that y_i and $\tilde{X}_{ij}$ are properly truncated whenever appropriate so that they are conformable with the other matrices.

Let $g(V_i, \gamma)$ be the $K \times 1$ vector of estimating equations defined above, where $K = \sum_{j=1}^{J} k_j$ is the dimensionality of γ. Thus, we can compactly write the moment conditions as $E[g(V_i, \gamma^*)] = 0$, where γ^* is the true value of the parameters. The usual GMM approach here is to find $\hat{\gamma}$ such that $(1/n)\sum_i g(V_i, \hat{\gamma}) = 0$. Here we have as many moment conditions as parameters to estimate so there is an exact solution, which can easily be found with standard software by iterating through the lags, estimating $\hat{\gamma}_j$ and using it to transform y_i to estimate $\hat{\gamma}_{j+1}$. The point estimate from that approach will be identical to one from estimating all parameters jointly. The standard errors on $\hat{\gamma}$, though, will be incorrect because they ignore the fact that estimates for one period depend on estimates from previous periods.

Standard theory on GMM estimators can help us derive asymptotically correct standard errors. Let γ^* be the true value of Define the $K \times K$ matrices $G \equiv E[\nabla_\gamma g(V_i, \gamma^*)]$ and $B \equiv E[g(V_i, \gamma^*)g(V_i, \gamma^*)']$. Then, under regularity conditions, $\hat{\gamma}$ will be asymptotically Normal with asymptotic variance,

$$\text{Avar}(\hat{\gamma}) = (G'G)^{-1}G'BG(G'G)^{-1}/N.$$

Let $\tilde{W}_{ij} = [\tilde{X}_{ij}\ 0]$ be the matrix of covariates at lag j with zeros replacing any covariates not included in the IRF. Then it is easy to show that with the above moment conditions, G

will have the following form:

$$G = E \begin{bmatrix} W'_{i0}W_{i0} & 0 & 0 & 0 & \cdots & 0 \\ W'_{i1}\tilde{W}_{i0} & W'_{i1}W_{i1} & 0 & 0 & \cdots & 0 \\ W'_{i2}\tilde{W}_{i0} & W'_{i2}\tilde{W}_{i1} & W'_{i2}W_{i2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ W'_{iJ}\tilde{W}_{i0} & W'_{iJ}\tilde{W}_{i1} & W'_{iJ}\tilde{W}_{i2} & W'_{iJ}\tilde{W}_{i3} & \cdots & W'_{iJ}W_{iJ} \end{bmatrix}. \tag{39}$$

Let W_j be the stacked $NT_j \times k_j$ matrix of all W_{ij} and define $\tilde{W}_j$ similarly. Then, under the appropriate regularity conditions, a consistent estimator of G will be:

$$\widehat{G} = N^{-1} \begin{bmatrix} W'_0W_0 & 0 & 0 & 0 & \cdots & 0 \\ W'_1\tilde{W}_0 & W'_1W_1 & 0 & 0 & \cdots & 0 \\ W'_2\tilde{W}_0 & W'_2\tilde{W}_1 & W'_2W_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ W'_J\tilde{W}_0 & W'_J\tilde{W}_1 & W'_J\tilde{W}_2 & W'_J\tilde{W}_3 & \cdots & W'_JW_J \end{bmatrix}. \tag{40}$$

To estimate B it is useful to derive it for this specific context. Let $u_{ij}(\gamma) = y_i - \sum_{s=0}^{j} \tilde{X}_{is}\beta_s - W_{ij}\gamma_j$ be errors associated with lag j. Then we can write B in the following form:

$$B = E \begin{bmatrix} W'_{i0}u_{i0}(\gamma)u_{i0}(\gamma)'W_{i0} & W'_{i0}u_{i0}(\gamma)u_{i1}(\gamma)'W_{i1} & \cdots & W'_{i0}u_{i0}(\gamma)u_{iJ}(\gamma)'W_{iJ} \\ W'_{i1}u_{i1}(\gamma)u_{i0}(\gamma)'W_{i0} & W'_{i1}u_{i1}(\gamma)u_{i1}(\gamma)'W_{i1} & \cdots & W'_{i1}u_{i1}(\gamma)u_{iJ}(\gamma)'W_{iJ} \\ \vdots & \vdots & \ddots & \vdots \\ W'_{iJ}u_{iJ}(\gamma)u_{i0}(\gamma)'W_{i0} & W'_{iJ}u_{iJ}(\gamma)u_{i1}(\gamma)'W_{i1} & \cdots & W'_{iJ}u_{iJ}(\gamma)u_{iJ}(\gamma)'W_{iJ} \end{bmatrix}. \tag{41}$$

Letting $\widehat{u}_{ij} = u_{ij}(\widehat{\gamma})$ be the residuals from lag j, we can consistently estimate B with:

$$\widehat{B} = N^{-1} \sum_{i=1}^{N} \begin{bmatrix} W'_{i0}\widehat{u}_{i0}\widehat{u}'_{i0}W_{i0} & W'_{i0}\widehat{u}_{i0}\widehat{u}'_{i1}W_{i1} & \cdots & W'_{i0}\widehat{u}_{i0}\widehat{u}'_{iJ}W_{iJ} \\ W'_{i1}\widehat{u}_{i1}\widehat{u}'_{i0}W_{i0} & W'_{i1}\widehat{u}_{i1}\widehat{u}'_{i1}W_{i1} & \cdots & W'_{i1}\widehat{u}_{i1}\widehat{u}'_{iJ}W_{iJ} \\ \vdots & \vdots & \ddots & \vdots \\ W'_{iJ}\widehat{u}_{iJ}\widehat{u}'_{i0}W_{i0} & W'_{iJ}\widehat{u}_{iJ}\widehat{u}'_{i1}W_{i1} & \cdots & W'_{iJ}\widehat{u}_{iJ}\widehat{u}'_{iJ}W_{iJ} \end{bmatrix}. \tag{42}$$

Given these two consistent estimators, we can apply standard asymptotic theory to derive the following estimator which is consistent for $\text{Avar}(\widehat{\gamma})$:

$$\widehat{\text{Var}}[\widehat{\gamma}] = (\widehat{G}'\widehat{G})^{-1}\widehat{G}'\widehat{B}\widehat{G}(\widehat{G}'\widehat{G})^{-1}. \tag{43}$$

Note that this estimator is robust to heteroskedasticity and serial correlation. The asymptotic properties hold as $N \to \infty$ with both T and J fixed, so this estimator is likely to perform best if N is large relative to T and J. One could impose a system homoskedasticity assumption and estimate the variance under a feasible GLS approach, which might be more efficient if T and N are closer in size. Alternatively, there are several finite-sample corrections that can improve inference with T is large.

C Proof of Sequential g-estimation/ADL equivalence

Suppose the vectors Y_t, Y_{t-1}, X_t and X_{t-1} have been centered, and define the X matrix $X = [X_{t-1} \quad X_t \quad Y_{t-1}]$ to be the combination of these column vectors. Let $\hat{\beta}$ be the coefficient vector from the regression of Y_t on X so that $\hat{\beta} = (X'X)^{-1}X'Y_t$ and has entries, $\hat{\beta} = (\hat{\beta}_2, \widehat{\beta_1}, \hat{\alpha})'$. Note the lack of an intercept due to centering of all variables.

The SNMM approach can be accomplished by blipping down and regressing on X_{t-1}. This can also be re-written as the difference between the coefficient on X_{t-1} from the simple regression of Y_t on X_{t-1} and the coefficient on X_{t-1} from the simple regression of X_t on X_{t-1} times the coefficient on X_{t-1} from the multiple regression.

$$\begin{aligned}
\tilde{Y}_t &= Y_t - X_t\hat{\beta}_1 \\
\hat{\psi}_1 &= (X'_{t-1}X_{t-1})^{-1}X'_{t-1}\tilde{Y}_t \\
&= (X'_{t-1}X_{t-1})^{-1}X'_{t-1}(Y_t - X_t\hat{\beta}_1) \\
&= (X'_{t-1}X_{t-1})^{-1}X'_{t-1}Y_t - (X'_{t-1}X_{t-1})^{-1}X'_{t-1}X_t\hat{\beta}_1
\end{aligned}$$

We also know from the normal equations of the full multivariate regression that

$$\begin{aligned}
(X'_{t-1}X_{t-1})\hat{\beta}_2 + (X'_{t-1}X_t)\hat{\beta}_1 + (X'_{t-1}Y_{t-1})\hat{\alpha} &= (X'_{t-1}Y_t) \\
\hat{\beta}_2 &= (X'_{t-1}X_{t-1})^{-1}X'_{t-1}Y_t \\
&\quad - (X'_{t-1}X_{t-1})^{-1}(X'_{t-1}X_t)\hat{\beta}_1 - (X'_{t-1}X_{t-1})^{-1}(X'_{t-1}Y_{t-1})\hat{\alpha} \\
\hat{\beta}_2 + (X'_{t-1}X_{t-1})^{-1}(X'_{t-1}Y_{t-1})\hat{\alpha} &= (X'_{t-1}X_{t-1})^{-1}X'_{t-1}Y_t - (X'_{t-1}X_{t-1})^{-1}(X'_{t-1}X_t)\hat{\beta}_1 \\
&= \hat{\psi}_1
\end{aligned}$$

Note that $\hat{\psi}_1 = \hat{\beta}_2 + (X'_{t-1}X_{t-1})^{-1}(X'_{t-1}Y_{t-1})\hat{\alpha}$ is close to the estimated impulse response from the ADL approach $(\hat{\beta}_2 + \hat{\beta}_1\hat{\alpha})$. The difference is that the ADL approach uses the contemporaneous effect $\widehat{\beta}_1$ (the estimate of the effect of X_t on Y_t) while the sequential g-estimation approach uses $(X'_{t-1}X_{t-1})^{-1}(X'_{t-1}Y_{t-1})$ (the estimate of the effect of X_{t-1} on Y_{t-1}). Therefore, note that the approaches will only be equivalent when the effects of X on Y are constant across time.

D Simulation Details

For the simulations, we generated the baseline covariate as $Z_{i1} \sim \mathcal{N}(0.4, 0.1^2)$ and a time-constant omitted variable as $U_i \sim \mathcal{N}(0, 0.1^2)$. Then, in each period, we generated the data with the following specification:

$$\begin{aligned}
Y_{it}(1,1) &= 0.8 + \mu_{1.1} + \mu_{2.11} + 0.9 \cdot U_i + \mathcal{N}(0, 0.1^2) \\
Y_{it}(1,0) &= 0.8 + \mu_{1.1} + 0.9 \cdot U_i + \mathcal{N}(0, 0.1^2) \\
Y_{it}(0,1) &= 0.8 + \mu_{2.01} + 0.9 \cdot U_i + \mathcal{N}(0, 0.1^2) \\
Y_{it}(0,0) &= 0.8 + 0.9 \cdot U_i + \mathcal{N}(0, 0.1^2) \\
Z_{it}(1) &= \gamma_0 + \gamma_1 + 0.7 \cdot U_i + \mathcal{N}(0, 0.1^2) \\
Z_{it}(0) &= \gamma_0 + 0.1 \cdot U_i + \mathcal{N}(0, 0.1^2) \\
Z_{it} &= X_{i,t-1}Z_{it}(1) + (1 - X_{i,t-1})Z_{it}(0) \\
\Pr[X_{it} = 1 | Z_{it}, Y_{i,t-1}] &= \mathsf{inv.logit}(\alpha_0 + \alpha_1 \cdot Z_{it} + \alpha_2 \cdot Y_{i,t-1})
\end{aligned}$$

In the simulations in the paper, we set $\mu_{1.1} = \mu_{2.01} = \mu_{2.11} = -0.1$, $\alpha = (-1.3, 1.5, 2.5)$, and $\gamma_0 = 0.5$. In the two settings discussed in the paper, γ_1 was set to -0.5 when the time-varying confounder is affected by treatment and 0 when it is not. Note that for each time-series, i, the DGP does not depend on t and the DGP for period t only relies on data for periods t and $t-1$. Conditional on $X_{i,t-1:t}$, the vector $\{Y_{it}, Z_{it}\}$ is clearly stationary because its only remaining time-varying variation comes from i.i.d. errors. It is easy to show that after marginalizing over these vectors, the process X_{it} forms a time-homogeneous Markov chain, implying the overall DGP is stationary. We checked this via simulation by simulating 1000 time-series of length $T = 1000$ and found that the means and autocovariances of each process were constant over time. Furthermore, all process clearly rejected the unit-root null hypothesis of an augmented Dickey-Fuller test.

In addition to the results in the paper, we also conducted a second simulation study with misspecification in the time-varying covariates. In particular, we assume that Z_{it} was not directly observable, and instead we observe a non-linear transformation of that covariate, $Z^*_{it} = \exp(0.25 * (Z_{it} - 0.5)^3)$. If an analyst knew this deterministic transformation, then she could correctly specify the functional form as $\log(Z_{it})^{1/3}$. Theoretically, this misspecification two possible effects. First, it can increase omitted variable bias for the contemporaneous effect of treatment because we are not conditioning on the correct confounders. Second, it could actually reduce post-treatment bias for the lagged effect since we now conditioning on a noisier version of the post-treatment variable. Thus, we show four scenarios in figure 8 that vary whether or not we have the correct Z_{it} and investigating the RMSE of the contemporaneous and lagged effects. In these results, the Z_{it} is endogenous.

From these results, we can see that even under misspecification, the ADL model has significant bias for the lagged effect. The ADL and MSM models have worse performance in that setting, but they still outperform the ADL model. Likewise, the SNMM and MSM approaches also see increases in RMSE for the contemporaneous effect due to omitted variable bias and model misspecification. These results show that post-treatment bias of the ADL model can easily overwhelm any problems with misspecification of the proposed modeling strategies in this paper.

E Additional illustration: The effect of trade on taxation in OECD countries

In this section, we describe another empirical illustration that shows how the MSM/IPTW approach gives strikingly different results from the conventional TSCS approach when we apply each to the data from Swank and Steinmo (2002). These scholars estimate the effects of domestic economic policies on tax rates in advanced industrialized democracies. Here we focus on one of their explanatory variables, trade openness, and its effect on one of their outcomes, the effective tax rate on labor. In their models, Swank and Steinmo find trade openness to have no statistically significant effect on these tax rates, but they only considered the effect of trade openness in the previous year. While Swank and Steinmo discuss the long-run effects of economic policies, they only estimate the contemporaneous effect of this trade policy, leaving aside any effects of history.

Swank and Steinmo adhere to the guidance of previous methodological research on TSCS data (Beck and Katz, 1996). The authors regress the tax rate in a given year on economic and political features of each country from the previous year. In addition to trade openness ($X_{i,t-1}$), these attributes include liberalization of capital controls, unemployment, leftist share of the government, and importantly, a lagged measure of the dependent variable. We refer to the lagged dependent variable as $Y_{i,t-1}$ and the set of attributes (excluding trade openness) as $Z_{i,t-1}$. Thus we can write their main estimating equation as:

$$Y_{it} = \beta_0 + \beta_1 X_{i,t-1} + \beta_2 Y_{i,t-1} + \beta_3 Z_{i,t-1} + \varepsilon_{it}. \tag{44}$$

Keep in mind that β_1 only has a causal interpretation as the CET when sequential ignorability holds and when the effect of $X_{i,t-1}$ is constant across the covariates, $Y_{i,t-1}$ and $Z_{i,t-1}$, and across time.

To uncover any historical effects of trade openness on the labor tax rate, we expand the model of Swank and Steinmo beyond a single lag. We instead take the cumulative years of trade openness as our main independent variable:[17]

$$Y_{it} = \beta_0 + \beta_1 \left(\sum_{k=1}^{t-1} X_{i,k} \right) + \beta_2 Y_{i,t-1} + \beta_3 Z_{i,t-1} + \nu_{it}. \tag{45}$$

Unfortunately, post-treatment bias ruins the causal interpretation of the coefficient on our new measure, β_1. Earlier values of trade openness, such as $X_{i,t-2}$, might affect the lagged tax rate, for instance. To avoid this difficulty, we can take a second approach—omitting the time-varying confounders, $Y_{i,t-1}$ and $Z_{i,t-1}$, from our model. Here we would estimate the effect of trade openness only conditioning on a time trend:

$$Y_{it} = \widetilde{\beta}_0 + \widetilde{\beta}_1 \left(\sum_{k=1}^{t-1} X_{i,k} \right) + \widetilde{\beta}_2 t + \eta_{it}. \tag{46}$$

While this method avoids the issue of post-treatment bias entirely, it admits the possibility of omitted variable bias. If past values of the tax rate affect future trade openness, for instance, then excluding these lags of the dependent variable will produce bias in our estimated effects. Each approach has its drawbacks, but we can learn a great deal by comparing their results to our preferred weighting method.

What do these approaches discover about the effects of trade openness? As Figure 9 shows, both methods—omitting and controlling for time-varying confounders—lead to the same basic conclusion: there is no statistically significant effect of trade openness on tax policy.[18] These results are consistent with the findings of Swank and Steinmo (2002). An alternative to both of these approaches is the above weighting method. To implement

[17] Here we need trade openness as a binary treatment, so we create a new trade openness variable which is 1 if the county-year had a score at or above the median of the entire sample. The results are substantively unchanged if we use continuous measures, though, as noted above, IPTW in those situations has much poorer properties (Goetgeluk, Vansteelandt and Goetghebeur, 2008).

[18] We estimate both of these models using a generalized estimating equations approach with robust standard errors, allowing for arbitrary correlation of observations within a country (Liang and Zeger, 1986).

IPTW in this case, we omit the time-varying confounders from the tax rate model and instead include those in a propensity score model to create weights as in (31). We then use those weights in a weighted GEE model. Instead of controlling for the time-varying confounders in our regression model, these weights adjust for the confounding in the time-varying covariates without inducing post-treatment bias. Figure 9 shows the IPTW estimates are not only significant and positive, but also far larger in magnitude than either of the other approaches.

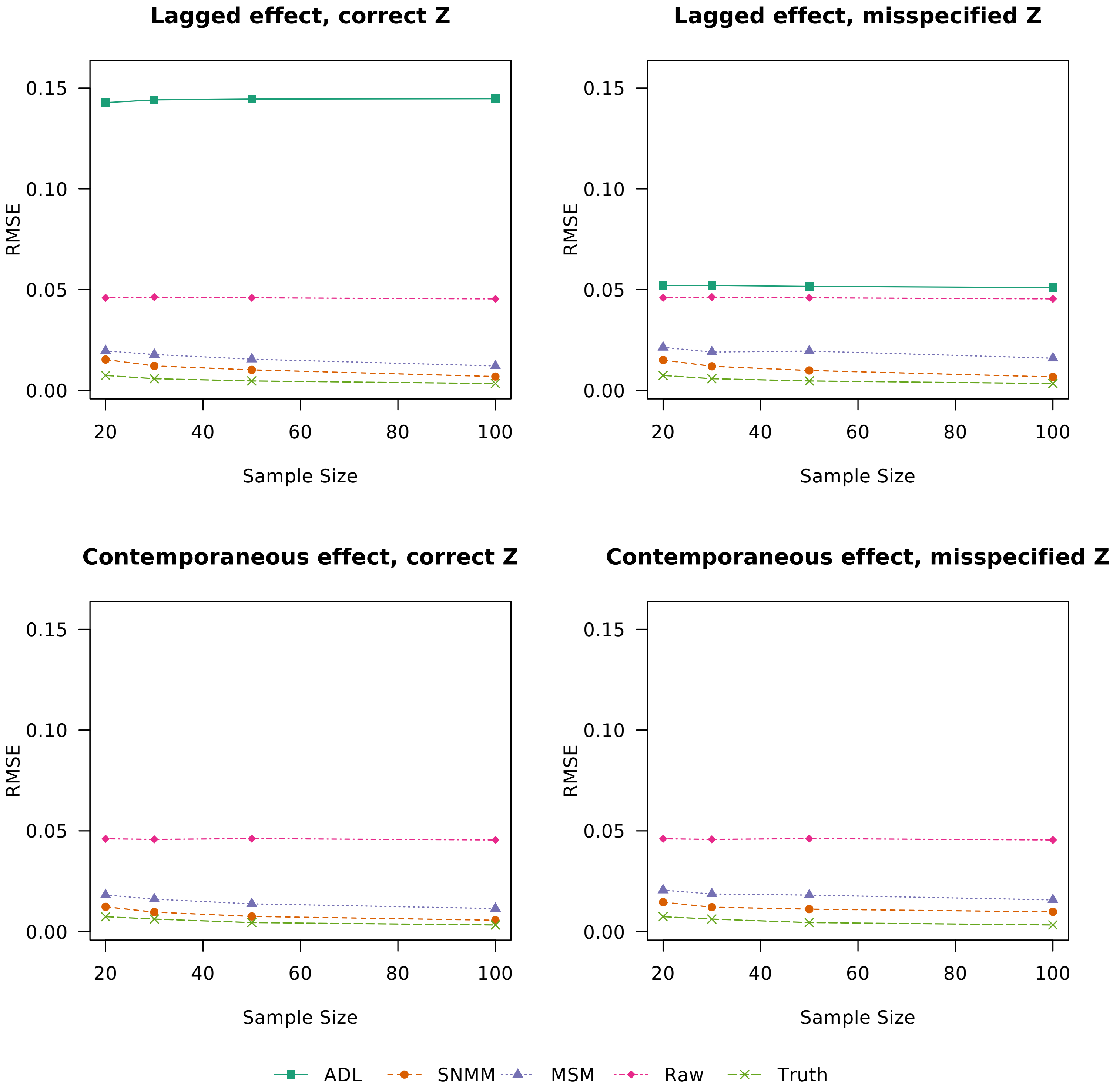


Figure 8: Simulation results when the time-varying confounder is correctly specified in all models (left column) and when it is incorrectly specified in all models (right column). Top row is the RMSE for the lagged effect of treatment and the bottom row is the contemporaneous effect of treatment. In the bottom row, the ADL results are identical to the SNMM. In these simulations, $T = 20$.

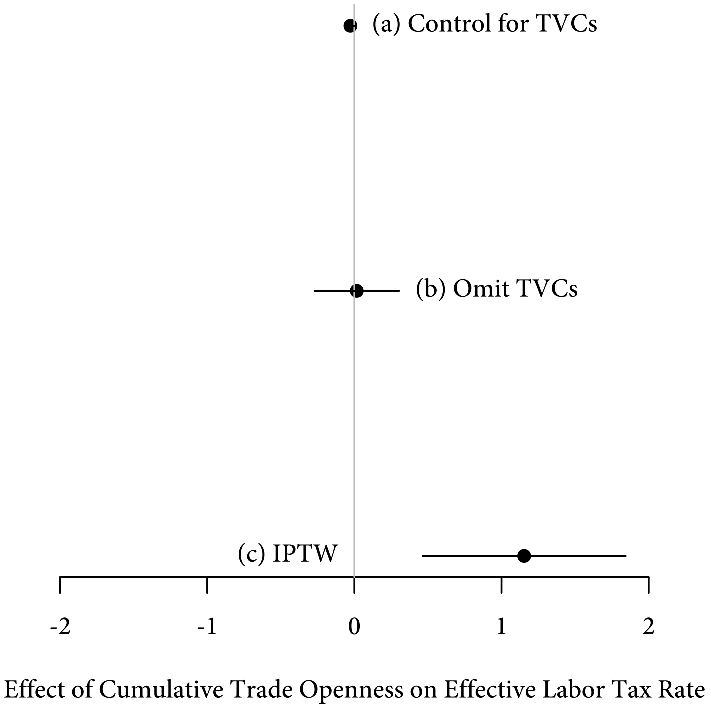


Figure 9: Estimated effect on labor tax rates of cumulative trade openness using three models. They represent the estimated effect and 95% confidence interval (a) when controlling for variables that trade openness affects, (b) when omitting those variables from the model, and (c) using the recommended IPTW approach.