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

Matthew Blackwell†
Adam Glynn‡

September 18, 2017

Abstract

Repeated measurements of the same countries, people, or groups over time form the foundation of many fields of quantitative political science. These measurements, sometimes called time-series cross-sectional (TSCS) data, can help researchers answer a variety of causal questions. Repeated measurements, however, can also lead to confusion about what causal question scholars are answering and what methods, data, and assumptions they need to do so. In this paper, we apply the developments in the statistical literature on causal inference to standard TSCS models and clarify how to nonparametrically define and identify certain TSCS quantities of interest within this context. The paper then describes a number of estimation strategies for these quantities, including inverse probability weighting and structural nested mean models. We show that some of these models will, under strong conditions, be equivalent to some traditional econometric models for TSCS data. This result connects two disparate methodological literatures and shows that some traditional TSCS methods can have a valid interpretation in counterfactual/potential outcomes models. We demonstrate these approaches through two empirical examples.

*We are grateful to Neal Beck, Jake Bowers, Patrick Brandt, Simo Goshev, and Cyrus Samii for helpful advice and feedback. Any remaining errors are our own.

†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

1 Introduction

Repeated measurements of the same countries, people, or groups at several points in time form the basis of time-series cross-sectional (TSCS) data. Large swaths of political science collect, use, and even consider the methodological implications of such data. This type of data allows researchers to draw on a larger pool of information when estimating causal effects. Furthermore, TSCS data 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). We 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 our political world. Two common examples of such treatment history effects are impulse responses and unit responses. Often multiple impulse responses (or unit responses), based on differing amounts of history, are presented as impulse response functions (or unit response functions). This variety of options and information, however, can lead to confusion regarding what causal questions we are answering, what methods we need to do so, and what assumptions justify their use.

The definition and estimation of causal effects in traditional TSCS approaches, though, are dependent on the specific statistical model chosen. It is common to write down a parametric model for the outcome and then derive quantities of interest from the parameters of this model. This approach makes it difficult to understand how the parameters of a statistical model relate to causal effects based on comparing counterfactual scenarios. This problem is exacerbated in TSCS models because the quantities of interest are almost always combinations of multiple parameters.

In this paper, we apply recent developments in the statistical literature on causal inference to TSCS data to help clarify the nature of causal reasoning in this setting. We define counterfactual causal effects that can vary over time and over units and relate them to typical TSCS estimands such as impulse response functions and intervention effects. Furthermore, we show how to derive each of these quantities from a common econometric TSCS model, the autoregressive distributed lag model. These connections will help TSCS practitioners understand the hypothetical interventions they are implicitly assuming when using TSCS models.

With these definitions in hand, we then turn to describing and evaluating different assumptions for identifying the various causal quantities of interest. TSCS methods often build off of statistical methods for time-series analysis, where the data are often, though not always, assumed to be stationary, which roughly means that the distribution of data is stable over time. We show how this assumption actually requires that treatment effects are themselves stationary over time, which may be a fairly restrictive assumption in many

settings. Traditional TSCS methods typically sidestep this question entirely by imposing a constant treatment effects assumption over time (and often across units). Stationarity is often combined with an assumption about the exogeneity of the treatment process and we connect these assumptions to randomization assumptions in the causal inference literature. In this paper, we focus on methods suitable for settings in which a selection-on-observables assumption holds and discuss the tradeoffs in the choice between these assumptions and within-unit or fixed-effect assumptions.

To help TSCS scholars estimate the effects of treatment histories under weaker assumptions, we provide an introduction to two methods from biostatistics based on *marginal structural models with inverse probability of treatment weighting* or MSMs with IPTWs (Robins, Hernán and Brumback, 2000) and *structural nested mean models* or SNMMs (Robins, 1997). These models allow researchers to estimate, for example, impulse responses, which are simply the direct effect of past treatment on outcomes for fixed values of future treatment. MSMs with IPTW often imply simple weighted regression estimators but are most stable with binary treatments. The SNMM approach works well with continuous treatments or ordinal treatment with many categories, both of which are very common in political science. Furthermore, for continuous outcomes, SNMM models imply very simple and intuitive multistep estimators and we provide a consistent variance estimator for the estimated effects in the Appendix. In short, these estimators 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.

These approaches to estimating causal effects in TSCS data have two important advantages. First, these models allow for feedback between the time-varying covariates (including the lagged dependent variable) and the treatment, making them appropriate for a wide variety of contexts in political science. For example, these methods allow for control variables to be affected by the explanatory variable of interest in the previous time period. Second, this approach weakens the modeling assumptions on lagged effects that are imposed by traditional TSCS models like the ADL model. Overall, this methodology could be promising for TSCS scholars, especially those who are interested in the effects of the full history of treatment.

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 4.3 introduces the MSM and SNMM approaches and shows how to estimate causal effects using these methodologies. We present an illustration of each approach in Section 5. The first, based on Swank and Steinmo (2002), uses MSMs to investigate the effect of trade

on tax policy in developed countries. The second, based on Burgoon (2006), uses SNMMs to investigate the connection between the welfare state and terrorism. Finally, Section 6 concludes with thoughts on both the limitations of these approaches and avenues for future research.

2 Causal quantities of interest in TSCS data

Repeated measurements greatly expand the range of causal questions available to researchers. At their most basic, TSCS data consists of a treatment, an outcome, and some covariates all measured for the same units at various points in time. By treatment, we mean the main explanatory variable of interest—the cause of the main effect of interest. In this paper we view causal inference as a comparison of various counterfactual outcomes based on changing the value of the treatment. In cross-sectional data with a binary treatment, there are a limited number of counterfactual comparisons to make. Imagine, for instance, a political economy dataset of countries with economic policies as outcomes and internationalization of trade as a binary treatment. With one time period, only one comparison exists: a country has either an open or a closed trade regime. As we gather data on these countries over time, more possibilities arise. How does the history of trade openness in these countries affect tax or budget outcomes? Does their trade regime *today* only affect their policies today or does the recent history matter as well? The variation over time provides the opportunity and the challenge of answering these more complex questions.

To fix ideas, let X_{it} be the treatment or independent variable of interest 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. We collect all of the treatments for a given unit into a *treatment history*, $\underline{X}_i = (X_{i1}, \ldots, X_{iT})$, where T is the number of time periods in the study. For example, we might have an *always treated* unit with history $(1, 1, \ldots, 1)$ or a *never treated* unit with history $(0, 0, \ldots, 0)$ or any combination of these. In addition, we define $\underline{X}_{it} = (X_{i1}, \ldots, X_{it})$ to be the partial treatment history up through time t and $\underline{x}_t$ is a possible particular realization of this random vector. For simplicity, we define most of our quantities of interest in terms of binary treatments, but it is straightforward to generalize them to arbitrary treatment types. We define Z_{it}, $\underline{Z}_{it}$, and $\underline{z}_t$ similarly for a set of time-varying covariates that are causally prior to the treatment at time t.

The goal is to estimate causal effects of the treatment on an outcome, Y_{it}, that also varies over time.[1] We take a counterfactual approach (Rubin, 1978) and define potential

[1] While it is common to estimate the "effect" of the lagged dependent variable (LDV) in TSCS models,

outcomes for each time period, $Y_{it}(\underline{x}_t)$.[2] This potential outcome represents the value that the outcome would take in period t if country i had followed history $\underline{x}_t$. Obviously, for any country in any time period, we only observe one of these potential outcomes since a country cannot follow multiple histories at the same time. 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: $Y_{it} = Y_{it}(\underline{x}_t)$ when $\underline{X}_{it} = \underline{x}_t$. Finally, we make a *no anticipation assumption* so that $Y_{it}(\underline{x}_{t+1}) = Y_{it}(\underline{x}_t)$ when $\underline{x}_{t+1} = (\underline{x}_t, x^*_{t+1})$ for any x^*_{t+1} (Abbring and van den Berg, 2003). This assumption essentially states that future values of the treatment cannot affect past values of the outcome. These definitions are based on the extension of the static potential outcomes framework to the time-varying treatment case by Robins (1986).

2.1 Populations and samples in TSCS data

The causal inference literature distinguishes between two sources of variation in the estimation of causal effects: the selection of units into the sample and the assignment of the units to treatment. This distinction flows into the definition of two common causal quantities of interest: the sample average treatment effect (SATE) and the population average treatment effect (PATE). The SATE is the average treatment effect for units in a fixed sample, so the only variation in the estimation comes from the assignment of units to different treatment levels. The PATE, on the other hand, is the treatment effect for a population of units from which the sample is drawn, so there is an additional source of variation from the selection of units into the sample.

In TSCS data, the definition of the sample from this point of view is fairly easy to define: it is the collection of unit-periods under study. What is sometimes less clear is the population to which one is attempting to make inference. There are four possibilities. First, one can make inference within the sample (as with the SATE) such that the only source of variation comes from the assignment of treatment. Second, one can make inferences to a larger population of units for a fixed time window (this is sometimes referred to as a panel data approach). Third, one can make inferences to a larger population of time periods for a fixed set of units. Fourth, one can make inferences to a larger population of units and

we view this as generally a non-causal question. The goal of including an LDV in a model is typically either to properly account for the "dynamics" of the dependent variable or to make the estimation of a causal effect more plausible. As we show below, both of these goals are accomplished by our approach even if we are not directly interested in the coefficient on the LDV.

[2]The definition of potential outcomes in this manner requires the 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.

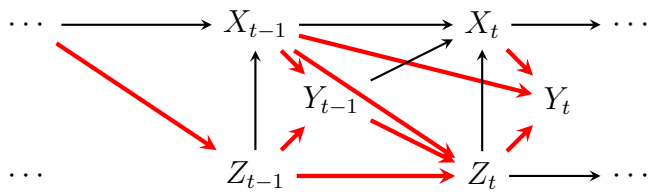

Figure 1: Average Treatment History Effect at Time t

time periods. Much of the literature on TSCS models focuses on the third approach, while here, we keep with the causal inference literature and focus on the second.[3] However, as we discuss below, this does not preclude the assessment of many traditional TSCS questions. Additionally, later in the paper, we show how this choice allows the relaxation of a number of typical TSCS assumptions.

2.2 The effect of a treatment history

With the potential outcomes in hand, we can define the causal quantities of interest available with TSCS data. The most basic quantity is the average treatment history effect, or ATHE (Robins, Greenland and Hu, 1999; Hernán, Brumback and Robins, 2001):

$$\tau(\underline{x}_t, \underline{x}'_t) = E[Y_{it}(\underline{x}_t) - Y_{it}(\underline{x}'_t)]. \tag{1}$$

Here, the expectations are over the units, i, so that this quantity is the average difference in outcomes between the world where all units had history $\underline{x}_t$ and the world where all units had history $\underline{x}'_t$. For example, we might be interested in the effect of a country having always traded openly versus a country always having a closed economy. Thus, the ATHE 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 represented by an ATHE is presented in Figure 1, where the red arrows correspond to components of the effect. These arrows represent all of the effects of X_t, X_{t-1}, X_{t-2}, and so on, that end up at Y_t. Note that many of these effects flow through the time-varying covariates, Z_t. This point complicates the estimation of ATHEs and we return to it below.

While the ATHE is the most basic effect with TSCS data, it allows a dynamic complexity that makes it quite flexible. It is clear from the definition that there are, in fact, many different ATHEs: one for each pair of treatment histories. As the length of time under

[3] It is also typically possible to operate as if one is making inference to a larger population of units and achieve conservative estimates of uncertainty for the SATE. See, for example, Imbens and Rubin (2015), Chapter XX. An extension of this result to TSCS setting is interesting, but beyond the scope of this paper.

study grows, so does the number of possible comparisons. In fact, with a binary treatment there are 2^t different values of the ATHE for the outcome in period t. This large number of comparisons allows for a host of causal questions: does the stability of trade openness over time matter for the impact of trade internationalization on economic policies? Is there a cumulative impact of trade openness or is it only the current institutions that matter?

2.3 Impulse and step effects

While Robins, Greenland and Hu (1999) and Hernán, Brumback and Robins (2001) introduced the definition of treatment history effects to longitudinal studies, their focus was squarely on comparisons useful for medical clinical studies and epidemiology. Here, we extend this general approach to meet the needs of social scientists and define a handful of specific treatment history effects that have been of special interest to TSCS scholars. In order to simplify the notation, we assume a binary treatment, with 1 representing the active treatment condition and 0 representing the control condition.

A very common quantity of interest in both time-series and TSCS applications is the *impulse response*, which is the effect of a single "shock" of treatment in one period on future outcomes (Box, Jenkins and Reinsel, 2013). Of course, this is just an example of a ATHE. Supposing our shock occurred in the first period, $E[Y_{i1}(1) - Y_{i1}(0)]$ would be the zero-step impulse response, $E[Y_{i2}(1,0) - Y_{i2}(0,0)]$ would be the one-step impulse response, $E[Y_{i3}(1,0,0) - Y_{i3}(0,0,0)]$ would be the two-step impulse response, and so on. The impulse response function, or IRF, describes how this response changes as we increase the steps between the shock and the measured outcome. The most general IRF could have its "impulse" at any time period and so we define the IRF as:

$$\tau_r(t,j) = E[Y_{i,t+j}(\underline{0}_{t-1}, 1, \underline{0}_j) - Y_{i,t+j}(\underline{0}_{t+j})], \tag{2}$$

where $\underline{0}_s$ is a vector of s zero values. Thus, $\tau_r(t,j)$ represents the impulse response or effect of a single-period blip of treatment at time t on the outcome at time $t+j$.

Another common quantity of interest is the *step response*, which is the effect of a permanent shift from control to treatment at time t on some future outcome (Box, Jenkins and Reinsel, 2013). This estimand has also been referred to as the effect of an intervention (Abadie, Diamond and Hainmueller, 2010) and unit response function (Beck and Katz, 2011). Here, if the policy is implemented in the second period, the step response for the third period would be $E[Y_{i3}(0,1,1) - Y_{i3}(0,0,0)]$. This type of response is common in policy analysis since it represents the cumulative effect of some policy shift on future outcomes. The step response function, or SRF, describes how this effect varies by time period and

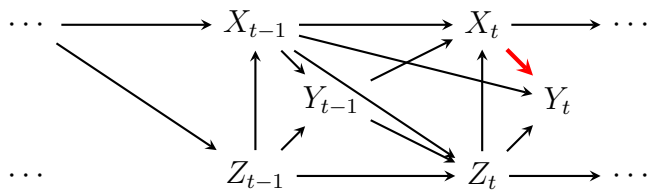

Figure 2: Contemporaneous Effect at Time t

distance between the shift and the outcome:

$$\tau_s(t,j) = E[Y_{i,t+j}(\underline{0}_{t-1}, \underline{1}_{j+1}) - Y_{i,t+j}(\underline{0}_{t+j})], \tag{3}$$

where $\underline{1}_s$ has a similar definition to $\underline{0}_s$. Thus, $\tau_s(t,j)$ is the effect of a j periods of treatment at time t on the outcome at time $t+j$. At its most general there is a separate impulse and step responses for each pair of periods. Econometric modeling of TSCS data imposes restrictions on the data-generating processes in part to limit the number of possible quantities to estimate.

One important special case of the impulse response function is the *contemporaneous effect of treatment* (CET) of X_t on Y_t, regardless of the past history of the treatment. To define this quantity, we write the set of all possible treatment histories at $t-1$ as $\underline{\mathcal{X}}_{t-1}$. Then the CET is:

$$\begin{aligned}
\tau_c(t) &= \sum_{\underline{x}_{t-1} \in \underline{\mathcal{X}}_{t-1}} E[Y_{it}(\underline{x}_{t-1}, 1) - Y_{it}(\underline{x}_{t-1}, 0) | \underline{X}_{i,t-1} = \underline{x}_{t-1}] \Pr[\underline{X}_{i,t-1} = \underline{x}_{t-1}], \\
&= \sum_{\underline{x}_{t-1} \in \underline{\mathcal{X}}_{t-1}} E[Y_{it}(1) - Y_{it}(0) | \underline{X}_{i,t-1} = \underline{x}_{t-1}] \Pr[\underline{X}_{i,t-1} = \underline{x}_{t-1}], \\
&= E[Y_{it}(1) - Y_{it}(0)],
\end{aligned} \tag{4}$$

where $Y_{it}(1)$ is the potential outcome under treatment in time t and the second line holds due to the consistency assumption. 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 a closed to an open trade regime in period t. A graphical depiction of a CET is presented in Figure 2, where the red arrow corresponds to component of the effect. Clearly, this quantity narrows the scope of the effect compared to the ATHE. It is common to assume that this effect is constant over time so that $\tau_c(t) = \tau_c$. For example, this is often assumed in pooled TSCS analyses.

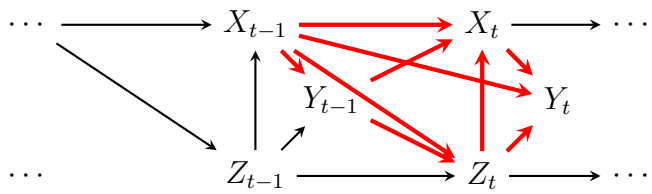

Figure 3: Future Effect of Treatment at Time $t-1$ on Outcome at Time t

A straightforward extension of the CET is the future effect of treatment (FET), which is the effect of treatment in time t on the outcome in time $t+j$. The FET is defined as follows:

$$\begin{aligned}
\tau_f(t,j) &= \sum_{\underline{x}_{t-1} \in \underline{\mathcal{X}}_{t-1}} E[Y_{i,t+j}(\underline{x}_{t-1},1) - Y_{i,t+j}(\underline{x}_{t-1},0) | \underline{X}_{i,t-1} = \underline{x}_{t-1}] \Pr[\underline{X}_{i,t-1} = \underline{x}_{t-1}], \\
&= \sum_{\underline{x}_{t-1} \in \underline{\mathcal{X}}_{t-1}} E[Y_{i,t+j}(1) - Y_{i,t+j}(0) | \underline{X}_{i,t-1} = \underline{x}_{t-1}] \Pr[\underline{X}_{i,t-1} = \underline{x}_{t-1}], \qquad (5) \\
&= E[Y_{i,t+j}(1) - Y_{i,t+j}(0)],
\end{aligned}$$

Note the important difference between a future effect and an impulse response. For an impulse response, a treatment is received in time t, but the unit is forced to receive the control condition from time $t+1$ to time $t+j$. For a future effect, the unit also receives the treatment in time t, but the unit is allowed to receive either the treatment or control conditions from time $t+1$ to time $t+j$ depending on endogenous assignment within those periods. A graphical depiction of an FET is presented in Figure 3, where again the red arrows correspond to component of the effect.

2.4 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 econometric TSCS based on strong assumptions, 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:

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

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

$$LRM = \lim_{t\to\infty} E[Y_{it}(\underline{1}_t) - Y_{it}(\underline{0}_t)]. \tag{6}$$

Identification of the LRM suffers from a few challenges. First, there is no guarantee that the limit in (6) 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.

2.5 Relationship to econometric 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} (with the exception of the LRM). The traditional approach to TSCS data, on the other hand, first assumes a parametric model for Y_{it} and then derives quantities of interest from the parameters of this particular model. It is useful then to show an example of how the two approaches relate to one another. One general model that encompasses many different possible specifications is called an autoregressive distributed lag (ADL) model:[4]

$$Y_{it} = \beta_0 + \alpha Y_{i,t-1} + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \varepsilon_{it}, \tag{7}$$

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:

$$Y_{it}(\underline{x}_t) = \beta_0 + \alpha Y_{i,t-1}(\underline{x}_{t-1}) + \beta_1 x_t + \beta_2 x_{t-1} + \varepsilon_{it}. \tag{8}$$

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

[4]For introductions to econometric 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 the CET and is constant over time. One can derive the long-term effects from a combination of α, β_1, and β_2. The IRF, for instance, will be constant over time because the effects of X_{it} and $X_{i,t-1}$ do not depend on t. In fact, it is easy to show that for the ADL model, the IRF has the following form:

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

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

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

The step response, on the other hand, has a stronger impact because it incorporates the contemporaneous at every lag distance:

$$\tau_s(t,0) = \beta_1, \tag{12}$$

$$\tau_s(t,1) = \beta_1 + \alpha\beta_1 + \beta_2, \tag{13}$$

$$\tau_s(t,2) = \beta_1 + \alpha\beta_1 + \beta_2 + \alpha^2\beta_1 + \alpha\beta_2. \tag{14}$$

Note that the step response here is just the sum of all previous impulse responses. In this model we can even identify the LRM as $\frac{\beta_1+\beta_2}{1-\alpha}$ as long as $|\alpha| < 1$. It is clear that one benefit of such a TSCS econometric model is to restrict the full generality of the data to be able to summarize a broad set of estimands with just a few parameters. But what is the role of the econometric modeling choice here? It is useful to know when these quantities of interest are identified nonparametrically from the data, without the crutch of a specific model.

3 Causal assumptions and designs in TSCS data

Under what assumptions are the above causal quantities identified? When we have repeated measurements outcome-treatment relationships, there are a number of assumptions we could invoke in order to identify causal effects. In this section we discuss several of these assumptions. The first set of assumptions comes out of the time-series literature and has to do with ensuring that the distribution of the data is relatively stable over time.

3.1 Time-series assumptions in TSCS models

TSCS data are often contrasted with what is called panel data in the sense that the former have fewer units and more time periods, whereas the latter usually have fewer time periods

and more units. This structure of the data has informed the assumptions that are typically made in these settings, where panel data models tend to rely on asymptotics as $N \to \infty$ and TSCS models rely on asymptotics as $T \to \infty$. Thus, TSCS models often rely on assumptions and models first developed for time-series analysis. One important assumption is that the time-series defined by both Y_t and X_t are weakly or covariance stationary, which means that they have (1) time-constant expectation, (2) finite variance, and (3) cross-time covariances that only depend on the distance between the variables in time. This assumption implies that the time series are relatively stable over time so that inference about one part of the series can be projected out into the future. Of course, this type of assumption is more suitable outside of the fixed-time-window approach that take in this paper, but it is useful to consider stationarity in the causal inference context.

Unfortunately, stationarity has a few drawbacks when trying to identify and estimate causal effects with minimal assumptions. First, stationarity requires the covariates, including the treatment, to be stationary, which can be limiting in many settings. For instance, a violation of stationarity would include a binary treatment variable that increases from 0 to 1 at some point in the series and never reverts. Many variables of interest have this step process over time. Second, stationarity becomes more difficult to interpret when treatment effects can vary over time and between units. For instance, suppose that we have a single unit with the following model:

$$Y_t = \beta_0 + \beta_t X_t + \varepsilon_t,$$

where the ε_t are i.i.d. draws from a mean-zero distribution and X_t is assumed to be weakly stationary. Here, the outcome time-series Y_t will only be stationary if the vector $\{\beta_t, X_t, \varepsilon_t\}$ is strictly stationary, which may be a strong assumption. When treatment effects are assumed constant over time, so that $\beta_t = \beta$ for all t, the joint distribution of the treatment, X_t, and error, ε_t, only need to be covariance stationary.[5] Note that stationarity of the treatment effects would rule out a good deal of interesting variation in treatment effects over time. Finally, stationarity imposes restrictions on how the outcomes and covariates can depend on the past. If, for instance, the distribution of the outcome depends on the cumulative sum of a covariate up to t, this might violate stationarity since that cumulative sum will obviously have time trend.

Stationarity is restrictive, yet it does allow one to leverage over-time variation to learn about causal parameters. As an alternative approach, and consistent with our assumption of a fixed time window, we focus on assumptions that make use of cross-sectional variation.

[5] With hetergeneous effects over time, weak stationarity of the variables is not sufficient to imply weak stationarity of Y_t because it is a *non-linear* function of the covariates, treatment effect, and the error.

This allows us to side-step many of the thorny time-series issues and use standard results to identify and estimate causal effects of treatment histories.

3.2 Baseline randomized treatments

A powerful, if rare, research design for TSCS data are those in which the entire history of treatment, $\underline{X}$, are selected at random from the entire set of possible histories 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:

$$\{Y_{it}(\underline{x}_t) : t = 1, \ldots, T\} \perp\!\!\!\perp \underline{X}_i \quad \forall t. \tag{15}$$

This assumes that the entire treatment is independent of all potential outcomes, which can be achieved, for example, if the entire treatment history is randomly assigned at baseline.[6] Hernán, Brumback and Robins (2001) called $\underline{X}_i$ *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 (15) implies that these variables play no part in the assignment of treatment or that they are causally unrelated to the outcome. Furthermore, this assumption assumes no feedback from the outcome to future values of the treatment. We could generalize this assumption to allow the randomization to depend on baseline covariates:

$$\{Y_{it}(\underline{x}_t) : t = 1, \ldots, T\} \perp\!\!\!\perp \underline{X}_i | Z_{i0} \quad \forall t, \tag{16}$$

where $A \perp\!\!\!\perp B|C$ is defined as "A is independent of B conditional on C." Even here the independence is only conditional on variables realized prior to the start of the sequence. Under either of these baseline randomization assumptions, the data become a classical randomized experiment, albeit with a possibly large number of possible treatment values. But in this case the assumption justifies the use of the usual approaches to estimating treatment effects in randomized studies.

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

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \varepsilon_{it} \tag{17}$$

[6] Randomization actually implies a stronger statement: the treatment sequence will be independent of the joint distribution of all potential outcome sequences across $\underline{x}_t$. For simplicity, we state the ignorability assumptions as independence separately for each possible treatment history since the practical differences between these assumptions are fairly minor and not relevant to the current paper.

Here, baseline randomization of the treatment history implies the usual identifying assumption in linear ADL models: *strict exogeneity* of the errors. This is a mean independence assumption about the relationship between the errors, ε_{it}, and the treatment history, $\underline{X}_i$:

$$E[\varepsilon_{it}|\underline{X}_i] = E[\varepsilon_{it}] = 0. \tag{18}$$

Thus, this baseline randomization assumption can justify assumptions that are common in the econometric literature.

While randomized experiments o are growing in popularity in political science, it is extremely rare to encounter them in TSCS data. They may be most common in panel designs with randomized rollout of a treatment. In Gerber et al. (2011), for example, the authors randomly assigned various media markets in Texas to receive a sequence of weekly advertising campaigns. The experiment ran for three weeks, and each media market could receive advertising for none, some, or all of these weeks. But the key component of the research design that justified assumption (15) was that they "randomly assigned [markets] into an ordering that indicated the start date of the broadcast television campaign" (p. 139). The process of determining $\underline{X}_i$ occurred before the start of the experiment and so time-varying covariates could have no impact on the treatment history.

For most TSCS applications, baseline randomization will be closer to an ideal type than a realistic data generating mechanism. Treatments are not typically randomized and when they are, they typically depend on a host of time-varying covariates, including lagged outcomes. For example, baseline randomization would imply that tax policies and any other time-varying covariates at time t have no effect on trade openness in the future. Thus, we view baseline randomization and strict exogeneity as unsuitable for most observational TSCS studies.

3.3 Sequentially randomized treatments

Beginning with Robins (1986), scholars in epidemiology have expanded the potential outcomes framework to handle weaker identifying assumptions than baseline randomization. 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 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 describes the relationship between the treatment history and a set of time-varying confounders, Z_{it}, and their history, $\underline{Z}_{it}$. Unlike baseline randomization and strict exogeneity, it allows for feedback between these histories so that the covariates can

affect and be affected by the treatment. The assumption states that, conditional on the covariate and treatment histories up to time t, the treatment at time t is independent of the potential outcomes at time t:

Assumption 1 (Sequential Ignorability)**.** *For every treatment history $\underline{x}_t$, covariate history $\underline{Z}_{it}$, and times t, if $\underline{X}_{i,t-1} = \underline{x}_{t-1}$, then*

$$\{Y_{is}(\underline{x}_s) : s = t, , \ldots, T\} \perp\!\!\!\perp X_{it} | \underline{Z}_{it}, \underline{X}_{i,t-1} = \underline{x}_{t-1}. \tag{19}$$

In the context of econometric TSCS models such as (7), sequential ignorability implies the *sequential exogeneity* assumption:

$$E[\varepsilon_{it} | \underline{X}_{it}, \underline{Z}_{it}] = 0. \tag{20}$$

According to the model in (7), 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 X_{it} conditional only on past values. Thus, this allows the errors to be related to future values of the treatment.

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 trade liberalization of a country to impact future economic policy and for policies to affect future trade openness. Thus, in this dynamic case, treatments can affect the covariates and so the covariates also have potential responses: $Z_{it}(\underline{x}_{t-1})$. This dynamic feedback is what complicates the estimation of ATHEs. Because the treatment can affect these time-varying confounders, the total effect of treatment becomes the amalgam of effects we see in Figure 1. We must consider not only the direct effect of the treatment history on the outcome, but also its indirect effects through the covariates. For example, trade openness might affect the effective tax rate directly through its effect on capital mobility but also indirectly through its effect on the overall levels of trade.

An important type of time-varying covariate in TSCS data is the lagged dependent variable, or LDV. Usually, scholars worry whether or not the LDV has an effect on the current value of the dependent variable and, if so, how to model that relationship. But if we view the LDV as potentially a member of $\underline{Z}_{it}$, then sequential ignorability requires us to know if the LDV has an effect on the treatment history as well. For instance, it may be the case that while trade openness has a strong effect on effect tax rates, these same tax rates might have an effect on future trade openness. If this type of feedback exists, then a lagged dependent variable must be in the conditioning set $\underline{Z}_{it}$ and strict exogeneity must be violated. This structure is common in TSCS data and implies that sequential ignorability may be the weakest possible assumption for many applications.

3.4 Unmeasured confounding and fixed effects assumptions

In this paper, we focus on sequential ignorability, which is, essentially, a selection-on-observables assumption: the researcher is able to choose a (time-varying) conditioning set to eliminate any unmeasured confounding. A oft-cited benefit of having repeated observations is that it allows scholars to estimate causal effect in spite of unmeasured heterogeneity in the outcome. One way to interpret this argument is that the above baseline randomization or sequential ignorability assumptions fail to hold, but that they may hold within a unit. For example, each country might have its own normal level of trade openness that is determined by idiosyncratic factors, but that the year-to-year variation within a country is exogeneous. If this is true and observed covariates are not sufficient to control for the unmeasured heterogeneity, then both baseline randomization and sequential ignorability will be false.

There are modified versions of the above causal assumptions that allow for unmeasured heterogeneity. In order to define these, let $D_i = (D_i^1, \ldots, D_i^N)$ be a vector of binary indicator variables for each unit, $D_i^j = 1(i = j)$. Note that this variable is constant with respect to time and could be considered a baseline covariate. Then we can define the *within-unit* baseline randomization:

$$\{Y_{it}(\underline{x}_t) : t = 1, \ldots, T\} \perp\!\!\!\perp \underline{X}_i | D_i. \tag{21}$$

We can also define within-unit sequential ignorability:

$$\{Y_{is}(\underline{x}_s) : s = t, \ldots, T\} \perp\!\!\!\perp X_{it} | \underline{X}_{it}, \underline{Z}_{it}, D_i. \tag{22}$$

Both of these are weaker versions of their between-unit counterparts because they make no assumptions about the comparability of units. Each unit can have its own baseline level of treatment as long as period-to-period variation is random. In the traditional TSCS literature, these two assumptions would imply, respectively, the within-unit strict exogeneity assumption:

$$E[\varepsilon_{it} | \underline{X}_i, D_i] = 0, \tag{23}$$

and the within-unit sequential exogeneity assumption:

$$E[\varepsilon_{it} | \underline{X}_{it}, \underline{Z}_{it}, D_i] = 0. \tag{24}$$

These fixed effects assumptions, though, are typically not sufficient to identify the above causal parameters since those quantities can vary over time in an arbitrary way (Chernozhukov et al., 2013). Intuitively, estimation under these fixed effect assumptions will depend on within-unit variation over time, so there must be some stability in effects over time in order to make headway. Thus, estimation will depend on the quantities being time-constant,

so the IRF would be $\tau_r(t, j) = \tau_r(j)$, the CEF would be $\tau_c(t) = \tau_c$, and so on. In some parametric models we may be able to allow for some time heterogeneity, but such inference will depend heavily on the modeling choices. Furthermore, most approaches to identifying and estimating ATHEs in fixed-effects models require strict exogeneity within units and so would preclude any feedback between time-varying covariates (including the outcome) and the treatment (Imai and Kim, 2012; Sobel, 2012). 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.

4 Modeling and estimation of treatment history effects

In this section, we show how to estimate the causal quantities of interest in Section 2 under sequential ignorability using a variety of approaches. In settings with a single-shot binary treatment and only a few discrete covariates, it can be possible to proceed nonparametrically by matching, weighting, or subclassification. In the present situation, we must deal with both the high-dimensional nature of the covariates and their history, but also the history of the treatment itself. Thus, we will require some modeling to estimate ATHEs, and this raises the possibility of model dependence in our estimates. Below we highlight two approaches that are *semiparametric* in the sense that they model part of the data generating process, but leave other parts unspecified. These approaches have the advantage of being robust to different assumptions about these unspecified aspects of the data generating process.

The approaches outlined below differ in what parts of the joint distribution of the data are left unspecified and which parts are modeled. We first review the fully parametric regression models that are common in the social science. The first semiparametric approach focuses on the modeling the relationship between the treatment and the covariates, leaving the outcome-covariate relationships unspecified. The second semiparametric approach incorporates both a model for the relationship between the outcome (or the treatment effect) and the covariates and a model for the treatment-covariate relationship. Both of these semiparametric approaches can be seen as incoporating information about the treatment process in order to help estimate causal effects and protect against bias from mispecification in the outcome regression model.

4.1 Regression-Based Approaches

The first set of techniques for estimating ATHEs focuses on modeling the relationship between the outcome or the treatment effect and the covariates. Historically, regression-based methods have been the most common way scholars have approached TSCS data, with scholars imposing a functional form for the conditional expectation, $E[Y_{it}|\underline{X}_{it}, \underline{Z}_{it}]$, such as in the ADL model:

$$E[Y_{it}|\underline{X}_{it}, \underline{Z}_{it}] = \beta_0 + \alpha Y_{i,t-1} + \beta_1 X_{it} + \beta_2 X_{i,t-1} + Z'_{it}\beta, \tag{25}$$

This form for the conditional expectation of Y_{it} imposes linear relationships for both the treatment, lagged treatment, and the lagged outcome. Estimation usually proceeds by ordinary least squares, possibly with various corrections for the standard errors due to the panel nature of the data. This approach obviously relies on correctly specifying the conditional mean of the outcome, including lag lengths and functional form. What is less obvious is that even when sequential ignorability holds, the coefficients on lagged values of the treatment in this case will not have a causal interpretation. There are two reasons for this. First, the conditioning set for this regression is usually focused on X_{it}, so that in most specifications, there will be omitted variable bias for the effect of $X_{i,t-1}$ unless its confounding set is also included in the regression. Second, there will be bias due to conditioning on variables that are post-treatment relative to $X_{i,t-1}$ such as the lagged outcome, $Y_{i,t-1}$, and contemporaneous covariates, Z_{it} (Rosenbaum, 1984).

The modeling assumptions of this approach may be restrictive. First, these models will generally assume constant treatment effects both over time and across units. Under this assumption, β_1 will be equivalent to the CET. If this assumption is violated, then the least squares estimator based on this model may converge to a weighted average of causal effects that is not equivalent to any average treatment effect (Angrist and Pischke, 2008; Aronow and Samii, 2016). Furthermore, OLS may be inconsistent for even this estimand if the functional form of the time-varying covariates is incorrectly specified. Thus, the traditional approach to TSCS relies heavily on modeling assumptions for the conditional expectation of the outcome.

4.2 Marginal structural models

An alternative to regression modeling is to write a model for the marginal mean of the potential outcomes, called a marginal structural model or MSM (Robins, Hernán and Brumback,

2000).[7] Again, focusing on the effect on the final outcome, the MSM would be

$$E[Y_{it}(\underline{x}_t)] = g(\underline{x}_t; \beta), \tag{26}$$

where the function g operates similarly to a link function in a generalized linear model.[8] For instance, we might take g to be linear for a continuous outcome and depend only on an additive combination of the treatment for the current period and the first two lags,

$$g(\underline{x}_t; \beta) = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2}, \tag{27}$$

or we might take g to have a logistic form for a binary outcome,

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

In both of these cases, we have made restrictions on how the history of treatment affects the outcome. In particular, treatments more than 2 periods before the final outcome are assumed to have no impact on that outcome. There are other ways to map the treatment history to the outcome, such as the cumulative number of treated periods, $\text{sum}(\underline{x}_t) = \sum_{s=1}^{t} x_{is}$. This allows for the entire history of treatment to affect the outcome in a structured, low-dimensional way. Under any of these models, an ATHE becomes:

$$\tau(\underline{x}_t, \underline{x}'_t) = g(\underline{x}_t; \beta) - g(\underline{x}'_t; \beta). \tag{29}$$

Of course, the choice of the MSM will place restrictions on the ATHEs that we can estimate. A MSM that is a function of only the cumulative treatment, for instance, implies that $\tau(\underline{x}_t, \underline{x}'_t) = 0$ if $\underline{x}_t$ and $\underline{x}'_t$ have the same number of treated periods, even if their sequence differs.

These MSMs lack any reference to time-varying covariates, $\underline{Z}_{it}$. Thus, if one simply estimates these models with observed data using, say, ordinary least squares, there will be omitted variable bias in the estimated effects. Fortunately, the causal parameters of these models are estimable using a extension of the propensity score weighting approach (Robins, Hernán and Brumback, 2000). In this MSM approach, we adjust for time-varying covariates using the propensity score weights, not the outcome model itself because, as described above, including such covariates in that model induces post-treatment bias. The weighting removes imbalances on the time-varying covariates across values of the treatments, so that omitting these variables in the reweighted data produces no omitted variable bias.

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

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

This approach works because, similar to nonparametric matching, weighting ensures that the there is balance in the time-varying covariates across different treatment histories.

Of course, this inverse probability of treatment weighting (IPTW) approach to estimating marginal structural models depends on a number of assumptions, which may be quite strong in some applications. First, sequential ignorability must hold for an observed set of covariates, $\underline{Z}_{it}$. Second, we must assume that *positivity* holds, here defined to mean that

$$0 < \Pr[X_{it} = 1 | \underline{Z}_{it} = \underline{z}_t, \underline{X}_{i,t-1} = \underline{x}_{t-1}] < 1 \quad \forall t, \underline{z}_t, \underline{x}_{t-1}, \tag{30}$$

so that it is possible for units to receive treatment at every time period and every possible combination of covariate and treatment histories. This assumption is similar to the common support and overlap conditions in the matching literature. Third, we assume that we have a consistent model for the probability of treatment, conditional on the past:

$$\widehat{\Pr}[X_{it} = 1 | \underline{Z}_{it}, \underline{X}_{i,t-1}; \widehat{\alpha}_N] \to_p \Pr[X_{it} = 1 | \underline{Z}_{it}, \underline{X}_{i,t-1}]. \tag{31}$$

Here $\hat{\alpha}_N$ is an estimator for the coefficients of a model for the probability of X_{it} conditional on the covariate and treatment histories. This might be simply a pooled logit model, 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). To establish consistency of the estimator, we need a model that is correct in the sense that its predicted values converge to the true propensity scores.[9] 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, 2013, 2015).

We use these predicted probabilities to construct weights for each unit-period:

$$\widehat{SW}_{it} = \prod_{t=1}^{t} \frac{\widehat{\Pr}[X_{it} | \underline{X}_{i,t-1}; \widehat{\gamma}]}{\widehat{\Pr}[X_{it} | \underline{Z}_{it}, \underline{X}_{i,t-1}; \widehat{\alpha}]}. \tag{32}$$

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 that unit's observed treatment and covariate histories. When we multiply this over time, it is the probability of seeing this unit's treatment history conditional on the time-varying covariates. This feature of the IPTW—weighting by the inverse of the probability of the observed treatment—is what inspires its name. The numerators here are the marginal probability of the observed treatment history

[9] 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.

and stabilize the weights to make sure they are not too variable which can lead to poor finite sample performance (Cole and Hernán, 2008). The numerator is the estimated marginal probability of the treatment history, estimated from a model for current treatment as a function of the treatment history omitting any covariates. While this choice of numerator is not required for consistency of the estimator (it can be replace with 1, for instance), it can help to stabilized weights that are highly variable.

Under these assumptions, the expectation of Y_i conditional on $\underline{X}_i$ in the reweighted data is equal to the MSM:

$$E_{SW}[Y_{it}|\underline{X}_{it} = \underline{x}_t] = E[Y_{it}(\underline{x}_t)]. \tag{33}$$

Here $E_{SW}[\cdot]$ is the expectation in the reweighted data. This implies that we can estimate ATHEs by simply running a weighted least squares regression of the outcome on the treatment history with $\widehat{SW}_i$ as the weights. The coefficients on the components of $\underline{X}_i$ 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). With TSCS data, it is important to address the possible lack of independence between units in this estimation strategy either through cluster-robust standard errors or a block bootstrap.[10]

One benefit of this MSM and IPTW approach is that does not require a model for mean of the outcome conditional on the time-varying covariates. This is advantageous because we often have very little guidance for the correct specification of the conditional relationship between controls and the outcome. In the above regression model, for instance, the $Z'_{it}\beta_z$ must be correctly specified in functional form, which may be very difficult to satisfy in practice. Of course, with the IPTW approach, one must have the correct model specification for the relationship between the treatment and the covariates. Thus, there is no free lunch, but it may be the case that substantive knowledge provides more insight into one type of relationship than another. For instance, if the treatment is a governmental policy, there might be considerable knowledge about the decision-making process for that policy that is amenable to modeling.

4.3 Structural nested mean models

A second class of models developed in biostatistics and called structural nested mean models (SNMMs) directly models the average effect of treatment rather than the conditional

[10] In applications using a MSM, it is common for researchers to specify a model that conditions on baseline covariates, $E[Y_{it}(\underline{x})|Z_{i0}]$. Because these baseline covariates are being conditioned on in this model, it is not necessary for the weights to balance them and these papers include the baseline covariates in the numerator of the stabilized weights (32). This approach has the advantage of increasing efficiency and reducing the variability of the weights with the potential downside of incorrecto functional form assumptions for the baseline covariates in the MSM.

expectation of the outcome (Robins, 1986, 1997). More specifically, these models focus on a conditional version of the impulse response function. That is, instead of writing a model for the conditional distribution of Y_{it} given all information up to time t, the SNMM approach is to parameterize the effect of a blip of treatment:[11]

$$b_t(\underline{x}_t, \underline{z}_t, j) = E[Y_{i,t+j}(\underline{x}_t) - Y_{i,t+j}(\underline{x}_{t-1}, 0)|\underline{X}_t = \underline{x}_t, \underline{Z}_t = \underline{z}_t] \tag{34}$$

This function gives the effect of a change from 0 to x_t in the treatment, conditional on a covariate and treatment history. For example, we might parameterize the impulse response function as:

$$b_t(\underline{x}_t, \underline{z}_t, j; \gamma) = \gamma_j x_t. \tag{35}$$

Here, γ_j is the impulse effect of X_{it} on $Y_{i,t+j}$ which does not depend on the past treatment history, $\underline{x}_{t-1}$ or the time period t. We could generalize this specification and have an impulse response that depended on past values of the treatment:

$$b_t(\underline{x}_t, \underline{z}_t, j; \gamma) = \gamma_{1j} x_t + \gamma_{2j} x_t x_{t-1}, \tag{36}$$

where γ_{2j} captures the interaction. 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. If Y_{it} is not continuous, it is possible to choose an alternative blip-down function (such as one that uses a log link) that restricts the effects to the proper scale (Vansteelandt and Joffe, 2014).

The key to SNMMs is that a particular transformation of the outcome can lead to easy estimation of the these conditional impulse responses. This transformation is

$$\widetilde{Y}_{it}^{j} = Y_{it} - \sum_{s=0}^{j} b_{t-s}(\underline{X}_{i,t-s}, \underline{Z}_{i,t-s}, s), \tag{37}$$

which, under the modeling assumptions of equation (35), would be

$$Y_{it} - \sum_{s=0}^{j} \gamma_s X_{i,t-s}. \tag{38}$$

Intuitively, this transformation subtracts off the effect of the treatment from period $t - j$ to period t, creating a new outcome that is similar to the potential outcome under treatment equaling 0 during those same periods. To see how this quantity is useful in the TSCS context, suppose that the ADL model in (7) is correct and perform the first transformation when

[11] Robins (1997) refers to impulse responses functions as "blip-down functions."

$s = 0$, noting that the contemporaneous effect is the same for both models $\gamma_0 = \beta_1$:

$$
\begin{aligned}
Y_{it} - \gamma_0 X_{it} &= Y_{it} - \beta_1 X_{it} && (39)\\
&= \beta_0 + \alpha Y_{i,t-1} + \beta_2 X_{i,t-1} + \varepsilon_{it} && (40)\\
&= \beta_0 + \alpha(\beta_0 + \alpha Y_{i,t-2} + \beta_1 X_{i,t-1} + \beta_2 X_{i,t-2} + \varepsilon_{i,t-1}) + \beta_2 X_{i,t-1} + \varepsilon_{it} && (41)\\
&= (\beta_0 + \alpha\beta_0) + \alpha^2 Y_{i,t-2} + \underbrace{(\alpha\beta_1 + \beta_2)}_{\gamma_1} X_{i,t-1} + \alpha\beta_2 X_{i,t-2} + (\alpha\varepsilon_{i,t-1} + \varepsilon_{it}) && (42)
\end{aligned}
$$

From this, we can see that the coefficient on $X_{i,t-1}$ for this transformed outcome is simply the IRF evaluated at lag 1, which is exactly the quantity that the SNMM models. 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 lags of the IRF and Table 1 shows how the two sets of quantities relate for various lags of the IRF.

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 impulse response function at various lags under the ADL (1,1) in (7) and SNMM in (35).

By combining the SNMM and sequential ignorability, we can derive an estimation strategy for the parameters of the SNMM. Robins (1994) and Robins (1997) show that, under sequential ignorability, the transformed outcome is a good proxy for the counterfactual of 0 treatment from $t - j$ to t:

$$E[\widetilde{Y}_{it}^{j} | \underline{X}_{t-j} = \underline{x}_{t-j}, \underline{Z}_{t-j}] = E[Y_{i,t}(\underline{x}_{t-j-1}, 0) | \underline{X}_{t-j} = \underline{x}_{t-j}, \underline{Z}_{t-j}]. \tag{43}$$

This statement says that the mean of this transformed outcome is the same as the mean of the potential outcome with the observed treatment values replaced with zeros for j periods before t. Thus, if we can estimate the impulse responses j periods back, then we can use this to identify the impulse response $j+1$ periods back. This recursive structure of the modeling is what gives SNMM the "nested" moniker.

The estimation strategy of SNMMs leverages these results about the transformations of Y_{it} in a straightforward way. The first, and most closely aligned with the traditional regression approach, is to place a model on the relationship between $\widetilde{Y}_{it}^{j}$ (the transformed outcome at time t) and the covariate and treatment histories. Applying this estimation

strategy to each lag sequentially and with linear models has been referred to as *sequential g-estimation* in the biostatistics literature (Vansteelandt, 2009).[12] With $j = 0$, one simply regresses the untransformed outcomes at time t on the treatment at time t and the set of covariates and the treatment history. This regression will provide estimates of the blip-down parameters, γ_0, which can be used to construct the one-lag blipped-down outcome, $\widetilde{Y}^1_{i,t}$. This blipped-down outcome can be used in the same way to estimate the second set of blip-down parameters, γ_1, and so on. In the Appendix, we show that this estimation strategy with no covariates except a lagged dependent variable is nearly mechanically equivalent to a traditional ADL model with one lag. The difference is that the traditional ADL model relies on an assumption that the contemporaneous effect is constant over time, whereas the blip-down approach relaxes this assumption. This provides an useful interpretation of the ADL model in terms of counterfactual causal effects.

The sequential g-estimation approach requires a correct specification of the relationship between the (transformed) outcome and the covariate and treatment histories. It thus requires a similar model to the traditional regression TSCS approaches described above. Similarly to MSMs and IPTW, it is possible to weaken the dependence on this modeling, if it is possible to model the treatment process. 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.

4.4 Comparing the methods

What are the advantages and disadvantages of each of these approaches? If one can correctly specify the outcome model, a regression-based approach will often be more efficient than other approaches. But the MSM and SNMM approaches will be robust to misspecification of the outcome model, *provided that the treatment model can be correctly specified.* Between these, the SNMM approach can be more stable when there treatment outliers—for example, a treated unit that has very low probability of receiving treatment given the covariate and treatment histories—whereas the IPTW approach may have large and unstable weights (Imai and Ratkovic, 2015). Relatedly, SNMMs 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). One feature of the SNMM approach is that it identifies treatment effects that have

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

a specific form: they are only effects of setting treatment to 0 from a given point forward. It is possible to combine the parameters of the blip-down functions to estimate all possible treatment history effects, but this requires either no interactions with time-varying covariates or a high-dimensional model for those covariates (Vansteelandt and Joffe, 2014, pp. 726-727).

5 Empirical illustration

5.1 The effect of trade on taxation in OECD countries

Should the details of ATHEs burden the applied researcher? Or are they mere technicalities? In this section we address the peril of ignoring the subtleties of causality in TSCS data with the analysis of political economy data on the OECD nations. The above weighting 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:[13]

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

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 4 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.[14] These results are consistent with the findings of Swank and Steinmo (2002). 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).

An alternative to both of these approaches is the above weighting method. To implement 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 (32). 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 4 shows the IPTW estimates are not only significant and positive, but also far larger in magnitude than either of the other approaches. Thus, the particulars of ATHEs and IPTW are more than mere technicality—they can transform seemingly robust substantive results.

[13]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).

[14]We estimate both of these models using a GEE approach with robust standard errors.

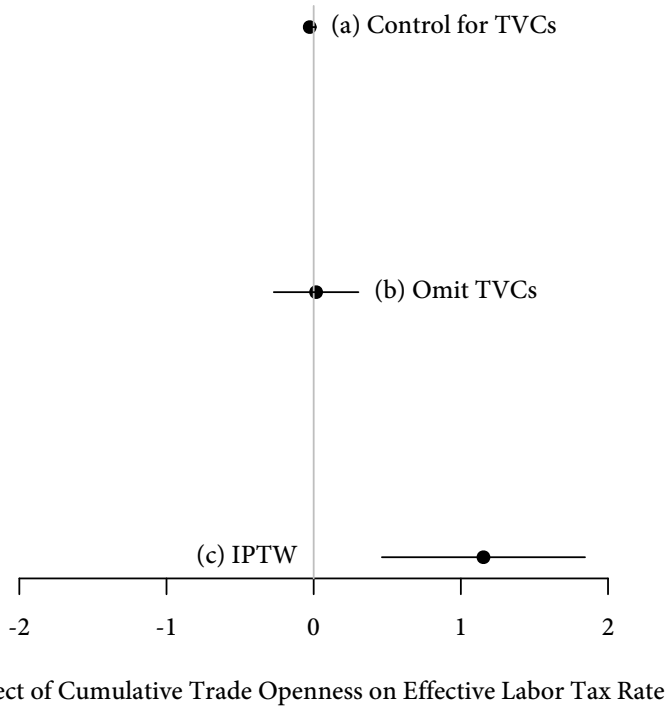

Figure 4: Estimated effect on labor tax rates of cumulative trade openness using three models. They represent the estimated effect and 95% confidence interval (a) when controlling for variables that trade openness affects, (b) when omitting those variables from the model, and (c) using the recommended IPTW approach.

5.2 Welfare spending and terrorism

Burgoon (2006) studied the effect of government and welfare spending on terrorist activity within countries and used TSCS data to show that increasing spending leads to lower levels of terrorist activity within a country. One natural question arising from such a country is over what time-frame do these effects accrue? That is, can we assess the effects of lagged government spending on future values of terrorist activity for fixed levels of spending in the interim? We can use the SNMM approach to answer these types of 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 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).[15] We follow Burgoon and use a set of regional dummies as baseline covariates. For time-varying covariates, we use 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. For the treatment, we include the log of total government spending and its one-year lag. Finally, we include a lagged dependent variable in each model. As we move backward through the SNMM to estimate the effect of government spending at each lag, we lag each of the time-varying covariates by one year. This way, there are never post-treatment variables in the model for the effect of interest at the given lag.

For the IRF, we use a simple functional form as in (35), $\gamma_j x_j$, with no interactions between treatment and past treatment. We compare our SNMM approach to an ADL approach, where use the formulas for calculating long-run effects from an ADL model, as described in Section 2.5.[16] This approach only relies on estimates from a single model and we follow standard practice by including the time-varying covariates in the model.

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 fixed effects. Though this is a very strong assumption, it is possible to weaken it though a sensitivity analysis to determine how much of a violation in this assumption would lead to substantively different inferences. This assumption, however, is either the same as or even weaker than what is needed to justify causal claims in the baseline specification of Burgoon (2006).

We use the SNMM methodology to estimate the contemporaneous effect of government spending on terrorist activity and the lagged effect up to 4 years prior. To Figure 5 shows the results of these analyses. For instance, the 1-year lag has $\widehat{\gamma}_1$ for the SNMM and $\widehat{\alpha}\widehat{\beta}_1 + \widehat{\beta}_2$ for the ADL approach. The differences between the two sets of results are notable for several reasons. First, while they both have the same insignificant result for the contemporaneous effect of spending, they differ greatly in their lagged effect. In the ADL model, the negative coefficient on lagged spending drives the impulse response into negative territory and is even statistically significant. This result depends heavily, however, on the negative effect of lagged spending, which could be driven by posttreatment bias as most of the controls are measured a year after that variable. The SNMM model, on the other hand, estimates that the one-year lag of spending to be roughly zero.

Second, the SNMM approach picks up very little effect of spending other than in the two-period lag. The ADL estimated effect, on the other hand, is strong and decays over lags.

[15] One could use a multiplicative SNMM (Robins, 1997) instead of a linear SNMM to accommodate the negative binomial model.

[16] We approximate standard errors for the ADL-derived effects via simulation.

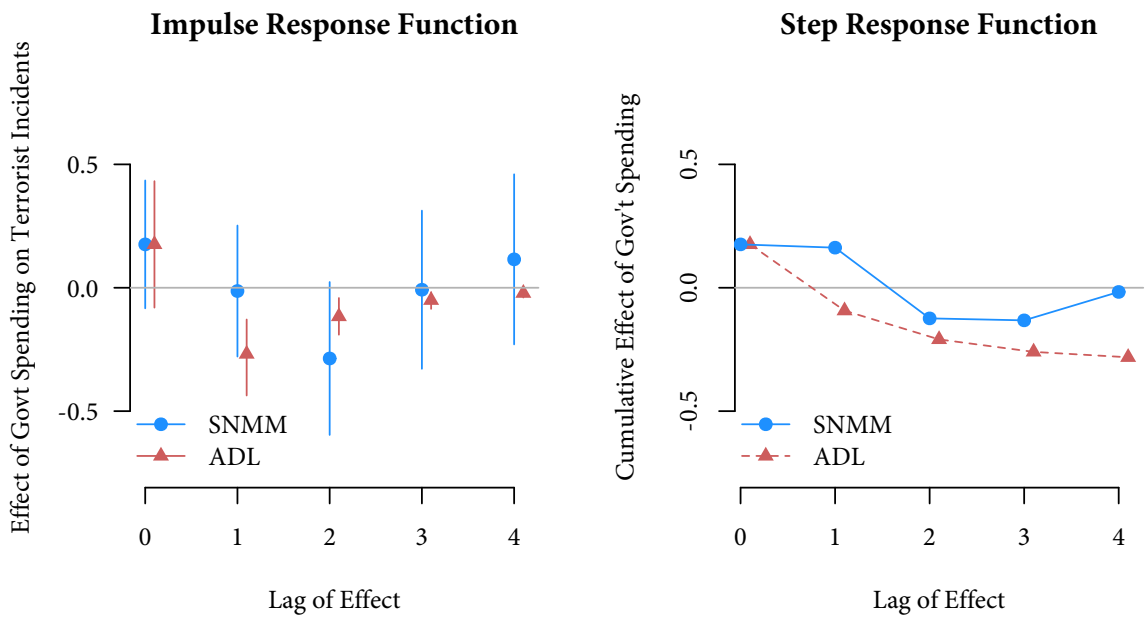

Figure 5: Left: Estimated effect of government spending on the terrorist incidents at various lags, along with 95% confidence intervals. Right: implied step response function at various lags. Data from Burgoon (2006).

One explanation for these differences might be the form of the ADL model: any effect in one period, must distributed over all of the lags due to satisfy the functional form assumptions. Thus, possible misspecification of the lag structure could result in misleading inferences. Finally, the two approaches give very different pictures of the SRF, or cumulative effect, for 4 lags. While the ADL model gives an SRF of -0.28, the SNMM approach gives a much smaller -0.017. Thus, weakening the assumptions on the ADL model appears to lead to different inferences for both the impulse and step response functions.

6 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 treatment history effects. But with an expanded scope comes complications. The usual TSCS regression methods break down for historical effects. Nevertheless, we have shown that an alternative approach from biostatistics can overcome these difficulties and recover effect estimates across a wide variety of settings.

SNMM methods have their own drawbacks, of course. Even though sequential ignorabil-

ity nonparametrically identifies any ATHE, the SNMM approach 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 this approach. 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). Under a within-unit version of sequential ignorability, one could estimate the SNMM effects within each unit separately and then average across the units to recover a consistent estimate of the average treatment effects. Bootstrapping the units and repeating this process would provide valid standard errors. The asymptotic validity of this approach would rely on $T \to \infty$, and so would likely perform best when T is large relative to N.

Bibliography

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490):493–505.

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.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. An Empiricist's Companion Princeton University Press.

Aronow, Peter M. and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1):250–267.
URL: *http://dx.doi.org/10.1111/ajps.12185*

Beck, Nathaniel and Jonathan N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6(1):1–36.
URL: *http://pan.oxfordjournals.org/content/6/1/1.abstract*

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

Blackwell, Matthew. 2013. "A Framwork for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2):504–520.
URL: *http://www.mattblackwell.org/files/papers/dynci.pdf*

Blackwell, Matthew. 2014. "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." *Political Analysis* 22(2):169–182.
URL: *http://pan.oxfordjournals.org/content/22/2/169*

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, April):176–203.

Chernozhukov, Victor, Ivn Fernndez-Val, Jinyong Hahn and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81(2):535–580.
URL: *http://dx.doi.org/10.3982/ECTA8405*

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, March):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, November):1049–1066.

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

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.
URL: *http://pubs.amstat.org/doi/abs/10.1198/016214501753168154*

Imai, Kosuke and In Song Kim. 2012. "On the Use of Linear Fixed Effects Regression Models for Causal Inference.".
URL: *http://imai.princeton.edu/research/files/FEmatch.pdf*

Imai, Kosuke and Marc Ratkovic. 2013. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* pp. n/a–n/a.
URL: *http://dx.doi.org/10.1111/rssb.12027*

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.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Behavioral Sciences*. Cambridge University Press.

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.

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.
URL: *http://biosun1.harvard.edu/ robins/new-approach.pdf*

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics* 23(8):2379–2412.
URL: *http://www.hsph.harvard.edu/james-robins/files/2013/03/correcting-1994.pdf*

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.
URL: *http://biosun1.harvard.edu/ robins/cicld-ucla.pdf*

Robins, James M., Miguel A. Hernán and Babette A. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.
URL: *http://www.jstor.org/stable/3703997*

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.
URL: *http://www.jstor.org/stable/2669978*

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)* pp. 656–666.
URL: *http://www.jstor.org/stable/10.2307/2981697*

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6(1):34–58.
URL: *http://www.jstor.org/stable/2958688*

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, June):521–529.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.2011.646917*

Swank, Duane and Sven Steinmo. 2002. "The New Political Economy of Taxation in Advanced Capitalist Democracies." *American Journal of Political Science* 46(3):pp. 642–655.
URL: *http://www.jstor.org/stable/3088405*

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." *Statist. Sci.* 29(4, 11):707–731.
URL: *http://dx.doi.org/10.1214/14-STS493*

A 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_{it}^j 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 $w_{it}^j = f_(z_{i,1}, x_{i,1}, \ldots, z_{i,t-j}, x_{i,t-j})$. Some of these covariates, $\tilde{x}_{it}^j$, are those in the impulse response function and will be used to transform the outcome for the next lag. The remaining covariates, $\tilde{z}_{it}^j$, are covariates used to satisfy sequential ignorability. These two sets of covariates partition the vector, $w_{it}^j = (\tilde{x}_{it}^j, \tilde{z}_{it}^j)$.

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:

$$E[W_{i0}'(y_i - W_{i0}\gamma_0)] = 0 \tag{47}$$

$$E[W_{i1}'(y_i - \tilde{X}_{i0}\beta_0 - W_{i1}\gamma_1)] = 0 \tag{48}$$

$$E[W_{i2}'(y_i - \tilde{X}_{i0}\beta_0 - \tilde{X}_{i1}\beta_1 - W_{i2}\gamma_2)] = 0 \tag{49}$$

$$\vdots \qquad = 0$$

$$E[W_{iJ}'(y_i - (\sum_{j=0}^{J-1} \tilde{X}_{ij}\beta_j) - W_{iJ}\gamma_J)] = 0 \tag{50}$$

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 $\widehat{\gamma}$ such that $(1/n)\sum_i g(V_i, \widehat{\gamma}) = 0$. 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 $\widehat{\gamma}_j$ and using it to transform y_i to estimate $\widehat{\gamma}_{j+1}$. The point estimate from that approach will be identical to one from estimating all

parameters jointly. The standard errors on $\widehat{\gamma}$, 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, $\widehat{\gamma}$ will be asymptotically Normal with asymptotic variance,

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

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:

$$G = E\begin{bmatrix} W_{i0}'W_{i0} & 0 & 0 & 0 & \cdots & 0 \\ W_{i1}'\tilde{W}_{i0} & W_{i1}'W_{i1} & 0 & 0 & \cdots & 0 \\ W_{i2}'\tilde{W}_{i0} & W_{i2}'\tilde{W}_{i1} & W_{i2}'W_{i2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{iJ}'\tilde{W}_{i0} & W_{iJ}'\tilde{W}_{i1} & W_{iJ}'\tilde{W}_{i2} & W_{iJ}'\tilde{W}_{i3} & \cdots & W_{iJ}'W_{iJ} \end{bmatrix}. \tag{51}$$

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:

$$\widehat{G} = N^{-1}\begin{bmatrix} W_0'W_0 & 0 & 0 & 0 & \cdots & 0 \\ W_1'\tilde{W}_0 & W_1'W_1 & 0 & 0 & \cdots & 0 \\ W_2'\tilde{W}_0 & W_2'\tilde{W}_1 & W_2'W_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ W_J'\tilde{W}_0 & W_J'\tilde{W}_1 & W_J'\tilde{W}_2 & W_J'\tilde{W}_3 & \cdots & W_J'W_J \end{bmatrix}. \tag{52}$$

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:

$$B = E\begin{bmatrix} W_{i0}'u_{i0}(\gamma)u_{i0}(\gamma)'W_{i0} & W_{i0}'u_{i0}(\gamma)u_{i1}(\gamma)'W_{i1} & \cdots & W_{i0}'u_{i0}(\gamma)u_{iJ}(\gamma)'W_{iJ} \\ W_{i1}'u_{i1}(\gamma)u_{i0}(\gamma)'W_{i0} & W_{i1}'u_{i1}(\gamma)u_{i1}(\gamma)'W_{i1} & \cdots & W_{i1}'u_{i1}(\gamma)u_{iJ}(\gamma)'W_{iJ} \\ \vdots & \vdots & \ddots & \vdots \\ W_{iJ}'u_{iJ}(\gamma)u_{i0}(\gamma)'W_{i0} & W_{iJ}'u_{iJ}(\gamma)u_{i1}(\gamma)'W_{i1} & \cdots & W_{iJ}'u_{iJ}(\gamma)u_{iJ}(\gamma)'W_{iJ} \end{bmatrix}. \tag{53}$$

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

$$\widehat{B} = N^{-1}\sum_{i=1}^{N}\begin{bmatrix} W_{i0}'\widehat{u}_{i0}\widehat{u}_{i0}'W_{i0} & W_{i0}'\widehat{u}_{i0}\widehat{u}_{i1}'W_{i1} & \cdots & W_{i0}'\widehat{u}_{i0}\widehat{u}_{iJ}'W_{iJ} \\ W_{i1}'\widehat{u}_{i1}\widehat{u}_{i0}'W_{i0} & W_{i1}'\widehat{u}_{i1}\widehat{u}_{i1}'W_{i1} & \cdots & W_{i1}'\widehat{u}_{i1}\widehat{u}_{iJ}'W_{iJ} \\ \vdots & \vdots & \ddots & \vdots \\ W_{iJ}'\widehat{u}_{iJ}\widehat{u}_{i0}'W_{i0} & W_{iJ}'\widehat{u}_{iJ}\widehat{u}_{i1}'W_{i1} & \cdots & W_{iJ}'\widehat{u}_{iJ}\widehat{u}_{iJ}'W_{iJ} \end{bmatrix}. \tag{54}$$

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

$$\widehat{\text{Var}}[\widehat{\gamma}] = (\widehat{G}'\widehat{G})^{-1}\widehat{G}'\widehat{B}\widehat{G}(\widehat{G}'\widehat{G})^{-1}. \tag{55}$$

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.

B 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 $\widehat{\beta}$ be the coefficient vector from the regression of Y_t on X so that $\widehat{\beta} = (X'X)^{-1}X'Y_t$ and has entries, $\widehat{\beta} = (\widehat{\beta}_2, \widehat{\beta}_1, \widehat{\alpha})'$. 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.

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

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

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

Note that $\widehat{\psi}_1 = \widehat{\beta}_2 + (X'_{t-1}X_{t-1})^{-1}(X'_{t-1}Y_{t-1})\widehat{\alpha}$ is close to the estimated impulse response from the ADL approach $(\widehat{\beta}_2 + \widehat{\beta}_1\widehat{\alpha})$. 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.