How to Make Causal Inferences with Time-Series Cross-Sectional Data*

Matthew Blackwell†
Adam Glynn‡

July 12, 2013

Abstract

Time-series cross-sectional (TSCS) data have played an increasingly important role in empirical political science research. The appeal of such repeated measurements is obvious: TSCS data allow for more powerful estimation of causal effect by pooling across time periods. However, this can also lead to confusion regarding which causal questions can be answered and which methods can answer them. In this paper, we demonstrate that a weighting approach to causal inference can estimate a wide array of causal quantities of interest and that other, more commonly used approaches tend to perform poorly in the TSCS setting. We demonstrate these methods with an application to the relationship between democracy and war.

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

†Department of Political Science, University of Rochester, Harkness Hall, Rochester, NY 14627. web: http://www.mattblackwell.org email: m.blackwell@rochester.edu

‡Department of Government, Harvard University, 1737 Cambridge St., Cambridge, MA 02138 email: aglynn@gov.harvard.edu

1 Introduction

Time-series cross-sectional (TSCS) data have played an increasingly important role in empirical political science research. The appeal of such repeated measurements is obvious: TSCS data allow for the identification and estimation of causal effects drawing on a larger pool of information than purely cross-sectional data. They also give researchers the power to ask a richer set of causal questions. However, this multitude of options can also lead to confusion regarding which causal questions can be answered and which methods can answer them.

In this paper, we demonstrate that a weighting approach to causal inference can estimate a wide array of causal quantities of interest and that other more commonly used approaches tend to perform poorly in the TSCS setting. We distinguish between the immediate impact of contemporaneous effects and the more general, potentially cumulative effect of treatment histories. Weighting approaches can consistently estimate either of these effects, while matching and regression—including regressions with fixed effects—can only estimate the former. We discuss the array of modeling choices that come with such a rich set of causal quantities. Nonparametric approaches are difficult with more than a few time periods, so how we model the effect of a treatment's history will have strong consequences for the estimated effects.

In terms of identification, we discuss research designs that assume sequential ignorability, which is the TSCS analogue of selection on the observables. The statistical literature has shown that, even under this assumption, regression and matching methods cannot recover the cumulative effects of treatment over time (Robins, Hernán, and Brumback, 2000; Blackwell 2013). We illustrate this point by replicating the results from Beck, Katz, and Tucker (1998) on the relationship between democracy and war in country pairs. Naively analyzing the cumulative effect of democracy over time leads to the conclusion that democracy actually *increases* the probability of war. However, when properly adjusted for dynamic confounding with inverse probability of treatment weighting, the cumulative effect of democracy again appears to be pacifying.

This paper proceeds as follows. Section 2 clarifies the causal quantities of interest available with TSCS data. In Section 3 we describe the assumptions necessary to identify these causal effects. Section 4 discusses the estimation of these quantities for a given time period, and in Section 5 we highlight the modeling choices necessary when we generalize multiple time periods. We present the replication of Beck, Katz, and Tucker (1998) in Section 6. Finally, Section 7 concludes with thoughts on future research.

2 Quantities of interest in TSCS data

With TSCS data, we have a treatment or variable of interest and an outcome measured at various points in time for the same unit, which allows researchers to ask and potentially answer a broader set of causal questions. In cross-sectional data with a binary treatment, there are a limited number of counterfactual comparisons to make. For example, if we take country dyads as our unit of analysis, and we are willing to measure the democracy at the country level as binary, then at a given point in time, a pair of countries in a dyad are either both democracies or not. As we gather data on these pairs over time, more interesting possibilities arise: how does the history of institutions in these countries affect trade or war between them? Does their democratic status *today* only affect their relations today or do their recent histories matter as well? The variation over time provides the opportunity and the challenge of answering these more complex questions.

To fix ideas, let A_{it} be the treatment or independent variable of interest for unit i in time period t. For simplicity, we focus on the case of a binary treatment so that $A_{it} = 1$ if the unit is treated in period t and $A_{it} = 0$ if the unit is untreated in period t. We collect all of the treatments for a given unit into a *treatment history*, $\underline{A}_i = (A_{i1}, \ldots, A_{iT})$, 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 $\underline{A}_{it} = (A_{i1}, \ldots, A_{it})$ to be the partial treatment up through time t.

The goal is to estimate causal effects of the treatment on an outcome, Y_{it}, that also varies over time. We take a counterfactual approach (Rubin, 1978) and define potential outcomes for each time period, $Y_{it}(\underline{a}_t)$, where $\underline{a}_t$ is a representative treatment history up through time t.[1] This potential outcome

[1] The definition of potential outcomes in this manner requires the Stable Unit Treatment Value Assumption (SUTVA).

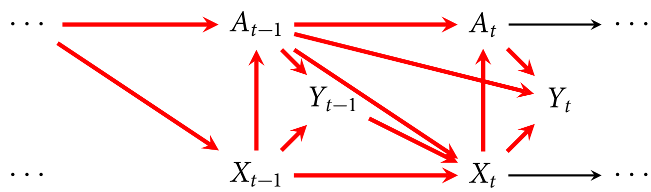

Figure 1: Treatment History Effect at Time t

represents the value that the outcome would take in period t if country-pair i had followed history $\underline{a}_t$. Obviously, for any country-pair in any time period, we only observe one of these potential outcomes. To connect the potential outcomes to the observed outcomes, we make a *consistency asssumption*. Namely, we assume that the observed outcome is the potential outcome for the observed history: $Y_{it} = Y_{it}(\underline{a}_t)$ when $\underline{A}_{it} = \underline{a}_t$.

With these potential outcomes in hand, we can define the causal quantities of interest available with TSCS data.[2] The most basic quantity is simply the average treatment history effect, or ATHE:

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

This quantity is the average difference between the world where all units had history $\underline{a}_t$ and the world where all units had history $\underline{a}'_t$. For example, we might be interested in the effect of two countries having always been democracies versus two countries never being democracies. A graphical depiction of 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 A_t, A_{t-1}, A_{t-2}, etc. that end up at Y_t. Note that many of these effects flow through the time-varying covariates, X_t. This point greatly 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 study grows, so does the number

This assumption is questionable for the democratic peace application, but we avoid discussing this complication in this paper in order to focus on the issues regarding TSCS data.

[2]For each of the quantities we present here, there are parallel estimands that condition on baseline (that is, time-invariant) covariates.

of possible comparisons. In fact, with T time periods, there are 2^T different values of the ATHE for the outcome in the T period. This large number of comparisons allows a host of causal questions: does the stability of democracy over time matter for the impact of democracy on war? Is there a cumulative impact of democratic institutions or is it only the current institutions that matter?

We can define other causal quantities beyond the general ATHE. For instance, one specific class of treatment history effects is the *blip effect* for two histories that agree up to time t:

$$\tau_b(\underline{a}_{t-1}) = E[Y_{it}(\underline{a}_t^1) - Y_{it}(\underline{a}_t^0)|\underline{A}_{i,t-1} = \underline{a}_{t-1}], \tag{2}$$

where $\underline{a}_t^1 = (\underline{a}_{t-1}, 1)$ and $\underline{a}_t^0 = (\underline{a}_{t-1}, 0)$, so that τ_b represents the effect of a treatment "blip" in the last period. For example, this might be the effect of two countries becoming democracies at time t after being autocracies for their entire history. This is obviously a special case of the more general ATHE, where we restrict the two histories to agree up to time t. Every treatment history at time $t-1$, then, has its own blip effect for time t. We can average across these blip effects to estimate the *contemporaneous effect of treatment* (CET) in period t:

$$\begin{aligned}\tau_t &= \sum_{\underline{m}\in\underline{a}_{t-1}} E[Y_{it}(\underline{m}, 1) - Y_{it}(\underline{m}, 0)|\underline{A}_{i,t-1} = \underline{m}] \Pr[\underline{A}_{i,t-1} = \underline{m}], \\ &= \sum_{\underline{m}\in\underline{a}_{t-1}} E[Y_{it}(1) - Y_{it}(0)|\underline{A}_{i,t-1} = \underline{m}] \Pr[\underline{A}_{i,t-1} = \underline{m}],\end{aligned} \tag{3}$$

where the second line holds due to the consistency assumption. This quantity reflects the effect of treatment in period t on the outcome in period t, averaging across all of the treatment histories. Thus, it would be the expected effect of switching a random country-pair from non-democracies to democracies in period t. A graphical depiction of a CET is presented in Figure 2, where the red arrows correspond to components of the effect. These arrows represent all of the effects of A_t in the graph that end up at Y_t. It is common to assume that this effect is constant over time so that $\tau_t = \tau$, but we could alternatively attempt to estimate the average of this effect over time.

It is crucial to distinguish between these various estimands because the various approaches to causal inference in TSCS data will identify some of these and not others. In general, we will see that blip effects and, thus, the CET will be easier to estimate because they require no fundamental

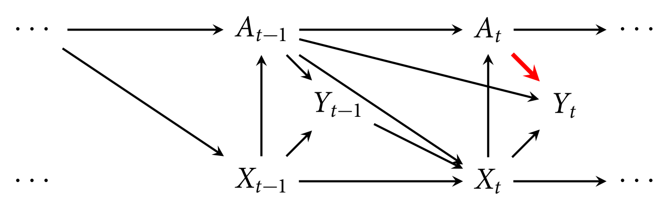

Figure 2: Contemporaneous Effect at Time t

changes to current TSCS estimation. Furthermore, some quantities may be more or less useful for testing theories in political science. In particular, it is important to distinguish whether the ATHE or the CET more directly addresses a particular theory.

3 Causal Assumptions with TSCS Data

One approach to causal inference in TSCS data relies on a "selection on the observables" assumption, similar to those commonly invoked for cross-sectional data. Beginning with Robins (1986), scholars in epidemiology have extended the usual cross-sectional causal inference framework to handle treatments that can vary over time. These approaches rely on *sequential ignorability*, which is an assumption that weakens the usual cross-sectional ignorability assumption to allow for time dependence. At its core, sequential ignorability describes the relationship between the treatment history and a set of time-varying confounders, X_{it}, and their history, $\underline{X}_{it}$. 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 action sequences $\underline{a}_t$, covariate history $\underline{X}_{it}$, and time t, if $\underline{A}_{i,t-1} = \underline{a}_{t-1}$, then $Y_{it}(\underline{a}_t) \perp\!\!\!\perp A_{it} | \underline{X}_{it}, \underline{A}_{i,t-1} = \underline{a}_{t-1}$.*

This assumption is weaker than the so-called *strict ignorability* assumption, which requires the entire treatment history to be independent of the potential outcomes only conditional on baseline (or cross-sectional) variables. Strict ignorability rules out the possibility of feedback between the time-varying covariates and the treatment. For example, this implies that a dispute between two countries at time t has no effect on their democratic institutions, bilateral trade or alliance status in the future. For these reasons, this strict ignorability assumption is typically unsuitable for use in TSCS applications in political science.

Sequential ignorability, on the other hand, allows for feedback between the treatment status and the time-varying covariates, including the outcome. For instance, sequential ignorability allows for the democracy of two countries to impact future trade between those countries and for trade to affect future democratization. Thus, in this dynamic case, treatments can affect the covariates and so the covariates also have potential responses: $X_{it}(\underline{a}_{t-1})$. 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, democracy might affect disputes directly through its institutional features but also indirectly through its effect on economic interdependence.

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 $\underline{X}_{it}$, 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 democracy has a strong effect on disputes, these same disputes might have an effect on future democracy. If this type of feedback exists, then a lagged dependent variable must be in the conditioning set $\underline{X}_{it}$ and strict ignorability must be violated. This structure is common in TSCS data and implies that sequential ignorability may be the weakest possible assumption for many applications.

4 How to Estimate Causal Effects for a Single Time Period

4.1 Contemporaneous Effects

To highlight the relevant issues with the identification and estimation of the above causal quantities, we first investigate the effects for a specific time period—namely, the effect on the outcome in the final period, $Y_i = Y_{iT}$. This choice is for exposition and we will generalize it in the next section. Under sequential ignorability, the last period of treatment, a_T, is exactly the same as a typical single-shot treatment, with the past history of treatment and the covariate history as the usual confounders. And it is clear that sequential ignorability simplifies to the more usual conditional ignorability assumption if we evaluate it at time period T. Thus, under sequential ignorability, the blip effect of the final period, $\tau_b(\underline{a}_{T-1})$, and the CET for that period, τ_T, are identified by the standard, single-shot causal inference framework. Of course there is nothing special about the final time period; we could repeat the same analysis for the effect of A_t on any Y_t, conditional on the covariate and treatment history.

The above discussion shows that the CET, τ_T, is just the average treatment effect if we regard a_T, the treatment in the final period, as the only treatment of interest. Thus, we can use standard approaches to estimate τ_T. One could use a nonparametric estimator such as a matching analysis on both the treatment history, $\underline{A}_{T-1}$, and the covariate history, $\underline{X}_{T-1}$ (Ho et al. 2007). An alternative to matching would be to run a regression of the final treatment on the final outcome, conditional on the past:

$$Y_i = \beta_0 + \beta_1 A_{iT} + \underline{A}'_{i,T-1}\beta_2 + \underline{X}'_{i,T-1}\beta_3 + \varepsilon_{iT}, \tag{4}$$

where each unit contributes only their last observation to the regression. Note that the coefficient on the final treatment, β_1, in this regression will not, in general, equal the CET, τ_T, or even the blip effect, $\tau_b(\underline{a}_{T-1})$, unless these effects are constant across units due to what we call the *regression weighting problem* or RWP. The RWP is that regression coefficients averages over a different distribution of the covariates than the average treatment effect. Rather, coefficients are averages weighted by the a unit's treatment variance conditional on the covariates, while the ATE averages over the marginal distribution of the covariates. Unless there are constant treatment effects across units, these two weighting schemes will produce distinct estimates. This is a standard result that applies to regression estimates of the average treatment effect as well (see, for example Angrist and Krueger 1999; Angrist and Pischke 2008; Aronow and Samii 2013). It is possible to use these regression estimators to estimate the CET, though, just as regression estimators can estimate the ATE in the single-shot framework (Imbens 2004). This approach estimates the CET by averaging the predicted effect estimates over the empirical distribution of the covariates ($\underline{A}_{i,T-1}$ and $\underline{X}_{i,T-1}$ in this case) instead of simply interpreting β_1 as an estimate of the CET.

4.2 Treatment History Effects and IPTW

When we move from contemporaneous effects to cumulative effects, the standard regression and matching techniques from the last section break down. This is because the conditioning on the covariate history, $\underline{X}_{i,T-1}$, in (4) induces post-treatment bias for the effect of the lagged treatment values. Intuitively, this conditioning blocks the indirect causal pathways from treatment history to outcome, such as $A_{i,t-1} \rightarrow X_{it} \rightarrow Y_{it}$. Thus, using a regression or matching model from the last section and interpreting the coefficients on lagged values of the treatment as causal will be very misleading. But if we remove the time-varying covariates to avoid the post-treatment bias, we will surely induce omitted variable bias because the time-varying covariates are important confounders. In general, methods that condition on time-varying covariates such as regression and matching will be inappropriate for estimating the effect of the treatment history on the outcome.[3] Note that this bias is present even if there are constant treatment effects across units so that the regression weighting problem is not an issue.

There are several methods for estimating ATHEs, though they are quite rare in political science. The easiest to implement of these is to write a model for the marginal mean of the potential outcomes, called a marginal structural model or MSM. Again, focusing on the effect on the final outcome, the MSM would be

$$E[Y_i(\underline{a})] = g(\underline{a}; \beta), \tag{5}$$

[3] This includes lags of the treatment, but also any function of the lags, such as the cumulative sum of treatment or the number periods since the last treated period.

where the function g operates similarly to a link function in a generalized linear model. 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,

$$g(\underline{a}; \beta) = \beta_0 + \beta_1 A_T + \beta_2 A_{T-1} + \beta_3 A_{T-2}, \tag{6}$$

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

$$g(\underline{a}; \beta) = \frac{\exp(\beta_0 + \beta_1 A_T + \beta_2 A_{T-1} + \beta_3 A_{T-2})}{1 + \exp(\beta_0 + \beta_1 A_T + \beta_2 A_{T-1} + \beta_3 A_{T-2})}. \tag{7}$$

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, $\text{sum}(\underline{A}_i) = \sum_{t=1}^{T} A_{it}$. 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:

$$\tau(\underline{a}, \underline{a}') = g(\underline{a}; \beta) - g(\underline{a}'; \beta). \tag{8}$$

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 $\tau(\underline{a}, \underline{a}') = 0$ if $\underline{a}$ and $\underline{a}'$ have the same number of treated periods, even if their sequence differs.

These MSMs lack any reference to time-varying covariates, $\underline{X}_i$. Thus, if one simply estimates these models with observed data, 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; Blackwell 2013). 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 the causal effects from the time-varying covariates to the treatments, so that omitting these variables in the reweighted data produces no omitted variable bias. 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, $\underline{X}_i$. Second, we must assume that *positivity* holds, here defined to mean that

$$0 < \Pr[A_{it} = 1 | \underline{X}_{it} = \underline{x}_t, \underline{A}_{i,t-1} = \underline{a}_{t-1}] < 1 \quad \forall t, \underline{x}_t, \underline{a}_{t-1}, \tag{9}$$

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:

$$\widehat{\Pr}[A_{it} = 1 | \underline{X}_{it}, \underline{A}_{i,t-1}; \hat{\alpha}_N] \to_p \Pr[A_{it} = 1 | \underline{X}_{it}, \underline{A}_{i,t-1}]. \tag{10}$$

Here $\hat{\alpha}_N$ is an estimator for the coefficients of a model for the probability of A_{it} conditional on the covariate and treatment histories. This might be simply a pooled logit model or a generalized additive model with a flexible functional form. In general, though, we need a model that is correct in the sense that its predicted values converge to the true propensity scores.

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

$$\widehat{SW}_i = \prod_{t=1}^{T} \frac{\widehat{\Pr}[A_{it} | \underline{A}_{i,t-1}; \hat{\gamma}]}{\widehat{\Pr}[A_{it} | \underline{A}_{i,t-1}, \underline{X}_{it}; \hat{\alpha}]}. \tag{11}$$

The denominator of each term in the product is the predicted probability of observing unit i's observed treatment status in time t (A_{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 stabilize the weights to make sure they are not too variable, which can lead to poor finite sample performance. The numerator is of the same form as the denominator, but with time-varying covariates omitted from the propensity score model. Note that we must build up these weights over time even though we are focusing on the ATHEs for the last outcome.

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

$$E_{SW}[Y_i | \underline{A}_i = \underline{a}, X_{i0}] = E[Y_i(\underline{a}) | X_{i0}]. \tag{12}$$

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 and any baseline covariates with $\widehat{SW}_i$ as the weights. The coefficients on the components of $\underline{A}_i$ from this regression will have a causal interpretation (Robins, Hernán, and Brumback, 2000). For this case of a single time-period outcome, the usual standard errors estimates, ignoring the estimation of the weights, will be conservative estimates of the true standard errors. This holds because the estimation of the weights actually increases the efficiency of the MSM estimator so that ignoring their estimation only increases the standard error estimates (Robins, 2000).

IPTW and MSMs are not the only way to estimate historical effects. Other estimation strategies rely explicitly on the g-computational formula, which uses the entire joint distribution of the data, outcomes and time-varying covariates, to estimate any causal effect (Robins, Greenland, and Hu, 1999). There are a number of ways to implement the g-computational formula, including structurally nested models and Bayesian simulation. This approach is very flexible, but it generally requires a model for the distribution of the covariates over time, which can be a large burden for empirical researchers who view these covariates as simply a tool to control for potential bias. Moreover, the dimension of these covariates can be quite large. Robins (2000) and Robins, Greenland, and Hu (1999) discuss some of the tradeoffs involved in choosing among these methods. Note that all of these approaches assume sequential ignorability.

5 The Promises and Pitfalls of Repeated Outcome Measurements

Of course, most TSCS studies use more than just the outcome for the last time period when estimating causal effects. There are a number of reasons for this, but on a basic level, the choice of the last period outcome as *the* outcome is a bit arbitrary. Nothing would stop us from choosing the outcome from any other year as the outcome variable, and this along with other considerations may lead the analyst toward a model with repeated outcomes. However, the use of a repeated outcomes creates a number of complications for the estimation of CETs and ATHEs, the most immediate of which is that these quantities are no longer uniquely defined: there is a CET and an ATHE defined for outcome variable at each point in time. In this section, we clarify exactly how to use and explore the existence of repeated outcomes to strengthen the estimation of causal effects.

5.1 Contemporaneous Effects with Repeated Measurements

The estimation of CETs with repeated outcomes has been covered extensively in the political science literature (see Beck and Katz (2011) for a review), but we review certain aspects here that are relevant to the estimation ATHEs with repeated outcomes. First, as noted above, one could estimate different CETs for the outcome in each year, although usually an assumption is made that allows some amount of pooling across years. Most simply, we might assume that the CET is the same for all years, although this can be relaxed. Second, even if we assumed that the CET were the same for all years, when estimating this time-constant CET, we would have to account for the dependence in the outcome variable across time. To some extent, this concern may be ameliorated by the control variables that are included in order to justify sequential ignorability. For example, we might assume that the lagged dependent variable must be included in the model in order to identify the causal effect and the inclusion of this variable would also alleviate some of the concerns regarding dependence over time (Beck and Katz, 1996). Further concerns about independence can often be addressed using a robust approach such as GEE (Liang and Zeger 1986; Zeger and Liang 1986).

5.2 Treatment History Effects with Repeated Measurement

The estimation of CETs with repeated outcomes is simplified by the fact that the treatment regimes associated with a CET are the same up to $t-1$. Although the CET for time t will be associated with a longer treatment regime than the CET for time $t-1$, both CETs will be concerned with a change in the treatment only in the final period, and hence it is interpretable to pool these CETs across time. In contrast, ATHEs will not be generally comparable across time, because the ATHE for time t will be associated with a longer treatment regime than the CET for time $t-1$. For example, when $t=2$

we might consider the ATHE $E[Y_{i2}(1,0) - Y_{i2}(0,1)]$, but it is not possible to construct an analogous ATHE for $t = 1$ with $\text{sum}(\underline{A}_i) = \sum_{k=1}^{t} A_{it} = 1$ for two different treatment histories. Therefore, in some sense when we decide to pool ATHEs across time, we restrict the space of possible contrasts and MSMs that we might consider.

It is helpful to consider two basic MSMs (and extensions) that allow pooling across time, and to consider what assumptions these MSMs imply about the ATHEs. For simplicity we present only linear models here, but the issues are analogous for non-linear models.

Recall the linear MSM considered above,

$$E[Y_{it}(\underline{a})] = \beta_0 + \beta_1 A_t + \beta_2 A_{t-1} + \beta_3 A_{t-2}, \tag{13}$$

This MSM assumes that treatments more than two periods prior do not affect the mean outcome in time t. Therefore, as long as the outcomes from $t = 1$ and $t = 2$ are not used in the analysis, this MSM will be comparable across repeated outcomes.

In contrast, consider the cumulative MSM,

$$E[Y_{it}(\underline{a})] = \beta_0 + \beta_1 \sum_{k=1}^{t} A_{ik}, \tag{14}$$

This MSM allows treatments from more than two periods ago to affect the mean outcome by assuming that only the sum of treatments will affect the mean outcome in time t. This MSM is also comparable for different t, although the sum will have more terms as t grows, and therefore it is typical to include a time term in this model. For example, we might include time additively in the following manner,

$$E[Y_{it}(\underline{a})] = \beta_0 + \beta_1 \sum_{k=1}^{t} A_{it} + \beta_2 t, \tag{15}$$

This allows some heterogeneity of the MSM across time periods, although it still assumes a constant ATHE across time.

Finally, we might want to allow the current treatment to be modeled differently than past treatments, so we could use hybrids of these MSMs.

$$E[Y_{it}(\underline{a})] = \beta_0 + \beta_1 A_{it} + \beta_1 \sum_{k=1}^{(t-1)} A_{ik} + \beta_2 t, \tag{16}$$

This MSM allows the treatment in the current period to have an effect separate from the cumulative effect for the other periods.

Finally, note that we can allow much more heterogeneity in time for the MSM. It is possible to include splines or other smoothers for time in order to allow non-linear heterogeneity in the MSM. It is also possible to include interactions between the treatment terms and the time terms in order to allow heterogeneity in the ATHEs across time. Both of these options will be explored in the application.

6 Application: The long arm of history in the Liberal Peace

Beck, Katz, and Tucker (1998) (BKT, hereafter) present a method for handling TSCS data with a binary outcome that allows for flexible temporal dependence in the outcome. They demonstrate this approach on the "Liberal Peace" models of Oneal and Russet (1997), showing that democracy has a strong impact on disputes between countries. BKT, however, focus on the contemporaneous effect of treatment: what is the effect of the democratic institutions of two countries now on conflict in the two countries now? In this section, we show that, if a scholar were to naively use BKT's model to estimate the historical, cumulative effect of democracy, they would reach a fundamentally different conclusion than the previous literature: that democracy has no effect on disputes. We show, however, that this conclusion is entirely due to the incorrect handling of time-varying confounders in a TSCS model with lagged treatment variables. Once we adjust for these time-varying confounder via an IPTW approach, the ameliorating effects of long-term democracy are quite clear.[4]

BKT build on the basic model of Oneal and Russet (1997) (OR) that models the presence of a dispute between two dyad members as a function of various features of that dyad. The main variables of interest are the Polity score of the less democratic member of the dyad ("Democracy") and the ratio of dyadic trade to GDP for the less trade-oriented member ("Trade"). Both papers control for several control variables, including, the lesser of the rates of economic growth ("Growth"), whether the member of the dyad are allies ("Allies"), whether or not the countries share a border ("Contiguous"), and the ratio of material capabilities as measured by the Correlates of War ("Capability Ratio").

[4] Of course, we are stipulating to all of the modeling assumptions being made in BKT.

Note that, of these controls, only Contiguous remains fixed over time; all others vary over time, albeit sometimes very slowly. Both OR and BKT employ a logistic regression pooled over time and BKT advocate for and include cubic splines of the time since last dispute ("Peace Years") to allow for temporal dependence.[5] These splines allow for the time to last dispute to have a flexible relationship with current disputes. In what follows below, we adopt the basic approach of Table 2, Column II in BKT, which includes Peace Year splines and drops ongoing years of a dispute from the data.

To highlight the issues involved, we make a few adjustments to the basic BKT model. First, we first dichotomize democracy into two categories: democracy and autocracy, and let the treatment, A_{it}, be one 1 if both countries in a dyad are democracies.[6] We call this variable Democracy Blip as it refers to one blip of democracy in that dyad's history, or A_{it} in the notation above. In Table 1, we run the above BKT specification with Democracy Blip in place of Democracy. Notice that the contemporaneous democratic peace result is represented by the parameter in the first row, and the standard result holds. Democratic dyads in a given year are less likely to have conflicts in that same year—conditional on the other variables in the model. Now suppose we want to address the question of whether a long history of a dyad being democratic makes conflict less likely. To assess this, we create a variable that measure the past history of democracy. For this application, we choose the cumulative sum of previous years with democratic blips (Cumulative Democracy$_{t-1}$): $\text{sum}(\underline{A}_{i,t-1}) = \sum_{s=1}^{t-1} A_{is}$. This represents a relative simple, but powerful dynamic hypothesis: how does conflict vary with the rich history of democracy of two nations?

Suppose however that we naively include the cumulative democracy variable in the regression model. The results for this misspecified model are presented in the second column of Table 1, and the effect of the history of democracy appears to *increase* the probability of war. This would be an odd result, implying that democracy this year has a negative effect, but past democracy actually has a positive effect. However, the problem with this analysis is the inclusion of time varying confounders

[5] We note that Beck (2011) has recently updated this advice to account for more modern approaches to smoothing. As the type of smoothing is not pertinent to our discussion, we will continue with the original recommendation from BKT.

[6] We use the Polity score of 0 as the cutoff value for this treatment. The major claims of Beck, Katz, and Tucker (1998) still hold after creating this binary treatment variable.

such as Growth in the model. Even if we assume that Growth in year t is causally prior to Democracy in year t, Growth in year t is certainly post-treatment to previous years of Democracy. Furthermore, the Peace-Year Splines actually represent a strong post-treatment confounder because they are a function of the lagged dependent variable. If past democracy affected these past disputes (and, thus, the number of peace years), then there is strong post-treatment bias in the estimates from the second column of Table 1.

In order to address the problem of variables that are simultaneously pre- and post-treatment to cumulative democracy, we remove all time varying conditioning variables from the BKT model in order to specify an MSM as in (5). In order to take account of the time varying conditioning variables, we specify IPTWs as in (11). We use a generalized additive model with a logit link for the probability of a dyad being Democratic in each time period, with the time varying conditioning variables: Growth, Allies, Capability Ratio, Trade, measures of Past Democracy, a spline for Calendar Year, and a spline for Peace Years. The results for this model are presented in Column 4 of Table 1. For the numerator of the weights, we run the same model, only including Past Democracy and the spline for Calendar Year. We then run a GEE model with a logisitic link based on (5), weighted according to (11). Standard errors are the so-called "robust" standard errors, which take into consideration the clustering by dyad.[7] This model can be seen in the third column of Table 1, where the only conditioning variables remaining are the constant and a spline for years since the dyad's first appearance in the sample. We also explored specifications that allowed for heterogeneous ATHEs across time, but none of these showed significant deviations across time.

In the MSM, countries with a more robust history of democracy are less likely to have conflicts after adjusting for the other time-varying variables, as we would expect. This model highlights the bias of the misspecified model, where the implicit inclusion of the lagged dependent variable (through the Peace Year Splines) induced post-treatment bias for the effect of Cumulative Past Democracy. To see why this bias occurs, imagine we condition on a specific number of years since the last dispute, say one. The number of Peace Years refers to the number of years either since the last dispute or since the

[7] Robins (2000), relying on Robins, Rotnitzky, and Zhao (1995), shows that this procedure is a conservative estimate of the true standard errors, since estimating the weights actually increases the efficiency of the estimator.

	DV: Dispute			DV: Democracy Blip
	Blip Model	Misspecified Cumulative Model	Weighted MSM	Weighting Model
	(1)	(2)	(3)	(4)
Democracy Blip	-0.80^{***}	-0.89^{***}	-0.53^{*}	
	(0.13)	(0.17)	(0.28)	
Cumulative Democracy$_{t-1}$		0.03^{***}	-0.04^{**}	
		(0.01)	(0.01)	
Growth	-0.91	-2.60^{**}		11.00^{***}
	(0.99)	(1.14)		(2.44)
Allies	-0.49^{***}	-0.44^{***}		0.66^{***}
	(0.09)	(0.10)		(0.14)
Contiguous	0.78^{***}	0.70^{***}		-0.28^{*}
	(0.09)	(0.10)		(0.15)
Capability Ratio	-0.28^{***}	-0.19^{***}		0.01
	(0.04)	(0.04)		(0.01)
Trade	-18.50	-22.60^{*}		53.60^{***}
	(11.40)	(11.80)		(10.40)
Democracy Blip$_{t-1}$				7.41^{***}
				(0.15)
Cumulative Democracy$_{t-2}$				0.09^{***}
				(0.01)
Constant	-1.00^{***}	-0.27^{***}	-3.28^{***}	-5.19^{***}
	(0.09)	(0.10)	(0.26)	(0.35)
Peace-Year Splines	✓	✓		✓
Calendar-Year Splines				✓
Dyad-Year Splines			✓	
IPTW Weights			✓	
N	21,641	20,826	20,022	20,022

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table 1: Estimating the cumulative historical effect of democracy on disputes between countries. Dependent variable for the first three models is whether or not there was a dispute between the dyad members in that year. First two columns are logit models and the third column is a logistic marginal structural model, fit by generalized estimating equations with inverse probability of treatment weights and cluster-robust standard errors. The weights are based on the model in the last column.

dyad entered the sample, if no conflict has yet occurred. Thus, among these dyads with one year of peace, there are two types: the dyads who just went to war a year ago and the dyads that that entered the sample a year ago. In this group, then, a high number of years with democracy indicates that they are likely many years into the sample and, thus, more likely to have just experienced a dispute. And because past disputes are correlated with future disputes, this creates a positive, spurious correlation between Cumulative Past Democracy and Disputes, conditional on Peace Years.[8] The MSM with IPTW never conditions on the number of Peace Years in the outcome model and adjusts for their bias through the weights. Thus, it avoids these types of spurious post-treatment bias issues while still eliminating omitted variable bias.

In this example, we took the continuous (or ordered categorical) variable Democracy and replaced with its binary counterpart Democracy Blip. An IPTW approach can accommodate non-binary treatments, but the weights produced with continuous treatments tend to be unstable and lead to inefficiency. Causal inference is about ensuring units across treatment statuses are comparable, either by randomization or by adjustment. The more treatment statuses there are, the more difficult this task will be. Dichotomization can introduce additional bias if the effect of the original, continuous treatment has a direct effect that does not flow through its binary counterpart. Thus, the choice of how exactly to model the treatment involves a bias-variance tradeoff. Here, for exposition, we show the simplest possible way to weight, with a binary treatment, but future research should focus on how this tradeoff might affect our inferences.

7 Conclusion

In this paper we have demonstrated how causal inference becomes complicated with time-series cross-sectional data. We have shown that weighting approaches can recover effect estimates across a wide array of estimands. We reviewed some of the different causal parameters of potential interest in TSCS data—clarifying the difference between contemporaneous effects and treatment history effects. We also reviewed different approaches to the estimation of these effects, highlighting the inability of regression based methods to identify treatment history effects, both in theory and in an important application to the literature on the Liberal Peace.

[8]There are ways to avoid this issue by changing the modeling strategy or by dropping any observations that follow a first dispute. These approaches, though, change the fundamental quantities of interest.

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 strict (not sequential) ignorability (Chernozhukov et al. 2009; Imai and Kim 2012; Sobel 2012). Future work should investigate how causal estimation could proceed under a unit-specific version of sequential ignorability, where the key assumption only holds within units. Estimators that exploit this assumption could bring together the rich set of causal quantities with the attractive within-unit variation of TSCS data.

References

Angrist, Joshua D., and Alan B Krueger. 1999. "Empirical strategies in labor economics." In *Handbook of labor economics,* 1277–1366. Elsevier. (Cited on page 9).

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics.* An Empiricist's Companion. Princeton University Press, December. (Cited on page 9).

Aronow, Peter M., and Cyrus Samii. 2013. "Does Regression Produce Representative Estimates of Causal Effects?" Presented at the Annual Meeting of the European Political Science Assocation, 2013, Barcelona, Spain. (Cited on page 9).

Beck, Nathaniel. 2011. "Time is Not A Theoretical Variable." *Political Analysis* 19 (1): 293–294. (Cited on page 16).

Beck, Nathaniel, and Jonathan N. Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6 (1): 1–36. (Cited on page 13).

———. 2011. "Modeling Dynamics in Time-Series–Cross-Section Political Economy Data." *Annual Review of Political Science* 14, no. 1 (June): 331–352. (Cited on page 13).

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42 (4): 1260–1288. (Cited on pages 2, 3, 15, 16).

Blackwell, Matthew. 2013. "A Framwork for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57 (2): 504–520. (Cited on pages 2, 10).

Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey. 2009. "Average and Quantile Effects in Nonseparable Panel Models." *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* stat.ME (April). (Cited on page 20).

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236. (Cited on page 8).

Imai, Kosuke, and In Song Kim. 2012. "On the Use of Linear Fixed Effects Regression Models for Causal Inference." July. (Cited on page 20).

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86, no. 1 (February): 4–29. (Cited on page 9).

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

Oneal, John R., and Bruce M. Russet. 1997. "The classical liberals were right: Democracy, interdependence, and conflict, 1950–1985." *International Studies Quarterly* 41 (2): 267–294. (Cited on page 15).

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. (Cited on page 6).

———. 2000. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials,* edited by M. Elizabeth Halloran and Donald Berry, 95–134. Vol. 116. The IMA Volumes in Mathematics and its Applications. New York: Springer-Verlag. (Cited on pages 12, 17).

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): pages. (Cited on page 12).

Robins, James M., Miguel A. Hernán, and Babette A. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5): 550–560. (Cited on pages 2, 10, 12).

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data." *Journal of the American Statistical Association* 90 (429): 106–121. (Cited on page 17).

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58. (Cited on page 3).

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, no. 498 (June): 521–529. (Cited on page 20).

Zeger, Scott L, and Kung-Yee Liang. 1986. “Longitudinal data analysis for discrete and continuous outcomes.” *Biometrics:*121–130. (Cited on page 13).