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

Matthew Blackwell†
Adam Glynn‡

May 2, 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 the identification and estimation of causal effects under weaker conditions than with purely cross-sectional data. 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. Furthermore, we extend the usual weighting techniques to handle situations where unit-specific unmeasured confounding exists. We demonstrate these methods with a set of simulation results and an application to the relationship between democracy and war.

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

*We are grateful to Jake Bowers and Simo Goshev 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

the identification and estimation of causal effects under weaker conditions than with 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. Furthermore, we extend the usual weighting techniques to handle situations where there is unit-specific unmeasured confounding.

In terms of identification, we discuss two scenarios. The first are 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 has no effect on war. However, when properly adjusted for dynamic confounding with inverse probability of treatment weighting, the cumulative effect of democracy again appears to be pacifying.

The second scenario involves instances where sequential ignorability may only be assumed within units. That is, there might be unmeasured variables that confound comparisons between units. Here, recent results demonstrate that "fixed effects" estimators may at least partially identify contemporaneous effects in a wide variety of circumstances, but even under extremely favorable assumptions, such estimators cannot recover treatment history effects (Chernozhukov et al. 2009; Imai and Kim 2012; Sobel 2012). To alleviate this, we propose a fixed effects version of inverse probability of treatment weighting that can recover estimates of treatment history effects under a set of assumptions similar to those commonly invoked with TSCS data.[1] Furthermore, we present the results of Monte Carlo simulations that show the finite sample performance of this estimator.

This paper proceeds as follows. Section 2 clarifies the causal quantities of interest available with TSCS data. In Section 3 we describe identification and estimation of those effects under sequential ignorability. We present the replication of Beck, Katz, and Tucker (1998) in Section 4. Section 5 discusses identification of causal effects under "fixed effects" assumptions and describes our weighting approach to this problem. To demonstrate the usefulness of this approach, Section 6 provides simulation results on its finite sample performance. 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: at a given point in time, a pair of countries 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

[1] A similar approach, in the context of g-computation, was proposed, but not studied, in Robins, Greenland, and Hu (1999).

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. This potential outcome 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 of possible comparisons. In fact, with T time periods, there are 2^T different values of the ATHE. This large number of comparisons allows a host of causal questions: does the stability of democracy

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

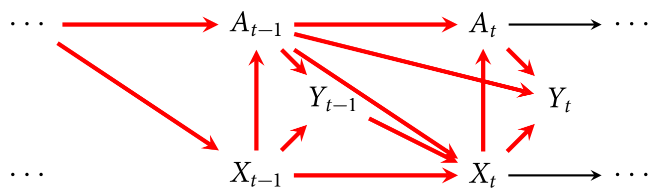

Figure 1: Treatment History Effect at Time t

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 basic ATHE. For instance, there 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. Every treatment history at time $t-1$ 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.

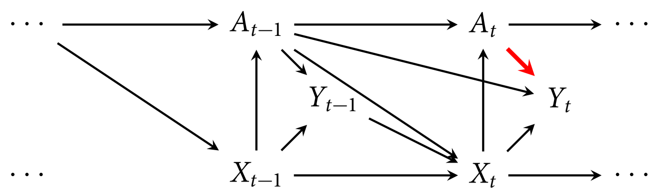

Figure 2: Contemporaneous Effect at Time t

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 identify because they require no assumptions on the relationship between the treatment history and any time-varying covariates. 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.

In general, there are two approaches to estimating the above causal quantities, one that extends cross-sectional approaches and the other unique to TSCS data. These approaches require different sets of assumptions, both untestable, that help to identify the various effects from this section. Which set of assumptions is more plausible depends on the empirical problem at hand.

3 Identification with Sequential Ignorability

The first 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 (possibly including a lagged dependent variable) 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.

Regression Models and Contemporaneous Effects

What does the above causal complexity mean for the estimation of causal quantities? One glimmer of hope comes from the contemporaneous effect, for which there can be no such indirect effects. This is because, by definition, they only involve direct effects. To see this, note that the CET only compares units with the same treatment history up to time $t-1$. Therefore, any contemporaneous effect can only be due to the change in treatment status at time t not to any affect of the treatment on the covariates. This allows us to estimate CETs with the standard cross-sectional estimators that control for the treatment and covariate history. This is because, under the assumption of sequential ignorability, the mean of the potential outcomes, conditional on the covariate and treatment histories, is equal to the mean of the observed outcomes, conditional on treatment in time t:

$$E[Y_{it}(1)|\underline{X}_{it}, \underline{A}_{i,t-1}] = E[Y_{it}|A_{it} = 1, \underline{X}_{it}, \underline{A}_{i,t-1}]. \tag{4}$$

The right hand side of (4) is a function of the observed data and the left hand side is a component of the CET, which implies that we can estimate the CET from the observed data. To estimate this conditional expectation, we might use a nonparametric regression estimator as defined in Imbens (2004), and if functional form assumptions hold, we might use parametric regression estimators. Under the additional assumption of constant treatment effects (across units), the effect of A_{it} in a regression of Y_{it} on A_{it}, $\underline{X}_{it}$, and $\underline{A}_{i,t-1}$ is the CET. If the effects are heterogeneous, we can always estimate separate effects for each treatment history and combine them according to (3). In addition, one could also estimate the above conditional expectation by matching on the treatment and covariate histories.

Most approaches to TSCS data in political science either implicitly or explicitly take the CET as their parameter of interest. Beck and Katz (1995), Beck and Katz (1996), Beck, Katz, and Tucker (1998), all take a time-constant version of the CET as the parameter of interest and investigate various ways to improve inference for that parameter. In order for these estimates to have an interpretation in terms of a causal effect, however, we must assume that the effects are constant across units. The bias of causal estimates from regression in the face of heterogeneous effects is a common problem, not limited to TSCS data (Angrist and Pischke 2008).

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 conditioning on the covariate history, $\underline{X}_{it}$, 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]

There are several methods for estimating ATHEs, though they are quite rare in political science. The most basic of these is to write a model for the marginal mean of the potential outcomes, called a marginal structural model or MSM,

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

and estimate the causal parameters of this model using a extension of the propensity score weighting approach (Robins, Hernán, and Brumback, 2000; Blackwell 2013). Under this model, an ATHE becomes:

$$\tau(\underline{a}_t, \underline{a}'_t) = g(\underline{a}_t; \beta) - g(\underline{a}'_t; \beta). \tag{6}$$

Note that time-varying confounders are missing from this model. 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.

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, $\underline{X}_{it}$. Second, we must assume that *positivity* holds, here defined to mean that

$$0 < \Pr[A_{it} = 1 | \underline{X}_{it} = \underline{x}_t, \underline{A}_{i,t-1} = \underline{a}_{t-1}] < 1 \quad \forall t, \underline{x}_t, \underline{a}_{t-1}, \tag{7}$$

[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.

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:

$$\widehat{\Pr}[A_{it} = 1 | \underline{X}_{it}, \underline{A}_{i,t-1}; \hat{\alpha}_N] \rightarrow_p \Pr[A_{it} = 1 | \underline{X}_{it}, \underline{A}_{i,t-1}]. \tag{8}$$

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:

$$\widehat{SW}_{it} = \prod_{s=1}^{t} \frac{\widehat{\Pr}[A_{is} | \underline{A}_{i,s-1}, X_{i0}; \hat{\gamma}]}{\widehat{\Pr}[A_{is} | \underline{A}_{i,s-1}, \underline{X}_{is}; \hat{\alpha}]}. \tag{9}$$

The denominator of each term in the product is the predicted probability of observing unit *i*'s observed treatment status in time s (A_{is}), 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. If baseline covariates, X_{i0}, are left in the numerator model, as they are here, they must be included in the outcome to control possible confounding.

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

$$E_{SW}[Y_{it} | \underline{A}_{it} = \underline{a}_t, X_{i0}] = E[Y_{it}(\underline{a}_t) | X_{i0}]. \tag{10}$$

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 $\widehat{SW}_{it}$ as the weights. The coefficients on the components of $\underline{A}_{it}$ from this regression will have a causal interpretation (Robins, Hernán, and Brumback, 2000). To get standard errors, one approach is to use Huber-White standard errors, which Robins (2000) shows to be conservative in this setting. These "sandwich" variance estimators are robust to time dependence as well as cluster heteroskedasticity. Alternatively, one can use the block bootstrap with units as blocks to calculate standard errors that are typically narrower than the Huber-White standard errors.

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.

4 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 the BKT 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]

To implement the BKT model with binary treatment, 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.[5] The results from this alteration of the BKT model are presented in the first column of Table 1. 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 series of variables that measure the past history of democracy. For this application, we choose the cumulative sum of years under democracy: $\text{sum}(\underline{A}_{it}) = \sum_{s=1}^{t} A_{is}$. 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 BKT model (replacing the single year democracy variable). 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 be indistinguishable from zero. This would be an odd result, implying that a blip of democracy had a larger effect than a history of democracy.[6] However, the problem with this analysis is the inclusion of time varying confounders 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. This means that there is 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). This model can be seen in the third column of Table 1, where the

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

[5]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. Thus, differences between our results and theirs are not likely due to this choice.

[6]It might also be the case that the sum of previous democratic years is not the correct functional form. Once could explore more complicated historical effects, but this is outside the scope of our illustrative example.

	Dependent variable: Dispute		
	BKT Model	Misspecified Cumulative Model	IPTW + MSM
	(1)	(2)	(3)
Democracy Blip	-0.651^{***}		
	(0.160)		
Cumulative Democracy		-0.010	-0.045^{***}
		(0.012)	(0.013)
Growth	-3.837^{***}	-4.360^{***}	
	(1.239)	(1.243)	
Allies	-0.356^{***}	-0.399^{***}	
	(0.114)	(0.113)	
Contiguous	0.996^{***}	1.047^{***}	1.374^{***}
	(0.124)	(0.122)	(0.169)
Capability Ratio	-0.230^{***}	-0.228^{***}	
	(0.052)	(0.052)	
Trade	-5.035	-15.738	
	(9.652)	(10.930)	
Constant	-4.170^{***}	-4.243^{***}	-4.451^{***}
	(0.107)	(0.111)	(0.132)
Observations	20,448	20,448	20,448
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 1: Estimating the cumulative historical effect of democracy on disputes between countries. A naive approach that simply adds a cumulative variable to a the usual TSCS model leads to the conclusion that democracy is irrelevant (Column 2). When the time-varying covariates are handled properly though IPTW, we see that history matters for disputes.

only conditioning variables remaining are the constant and an indicator for geographic contiguity. In order to take account of the time varying conditioning variables, we specify IPTWs as in (9). We use a logit model for the probability of a dyad being Democratic in each time period, with the time varying conditioning variables: Growth, Allies, Capability Ratio, and a spline for time since last dispute. We then run a regression based on (5), weighted according to (9). The results from this MSM + IPTW approach are reported in the third column of Table 1, where w now see that countries with a more robust history of democracy are less likely to have conflicts after adjusting for the other variables in the BKT model.

5 Identification with Fixed Effects Sequential Ignorability

The second broad approach to causal inference with TSCS data, linear models with unit fixed effects, is already a common tool for political scientists. In fact, one of the main proposed benefits of having TSCS data is it allows for the removal of unobserved heterogeneity through fixed effects. But given the richness of the causal questions possible in TSCS data, the scope of fixed effects estimators is actually quite limited.

The typical linear fixed effects models proposes a causal model for the dependent variable that includes additive and separable effects for each unit:

$$Y_{it} = \alpha_i + \tau A_{it} + X_{it}'\beta + \varepsilon_{it} \tag{11}$$

The α_i term accounts for any time-constant shift in the outcome that is unique i. This model is usually paired with strict exogeneity, which is a strong assumption about the errors:

$$E[\varepsilon_{it} | \alpha_i, \underline{A}, \underline{X}] = 0. \tag{12}$$

This assumption states that the errors of the linear fixed effect model are uncorrelated with past and future values of the treatment and the covariates. With this model and this assumption in hand, we can consistently estimate the coefficient on the treatment, τ, in spite of not observing the unit fixed effects. This is powerful because, in general, we do not need to measure and include these time-invariant covariates in order to estimate the causal parameters of (11) using a linear model. This has led many scholars to favor fixed effects models when estimating causal effects.

How does the coefficient estimated from (11) compare to the causal effects defined above? First note that this model implicitly assumes the presence of a comptemporaneous effect of A_t on Y_t, but no effect of the treatment history beyond this. Furthermore, since this contemporaneous effect is assumed to be constant, there is no need to weight it across the distribution of the treatment histories as in (3). Thus, in this model, τ is the CET.

While the usual fixed effects estimator allows for the estimation of the CET, it fails to identify the effects of treatment history without much stronger assumptions. Thus, even with constant treatment effects (over units and time), a fixed effects model cannot identify the "cumulative" effects of treatment. One might think that adding lagged values of A_{it} to (11) might alleviate this problem, but Sobel (2012) shows that this too fails to identify the ATHE. The nature of this failure is the same as with regression and matching under sequential ignorability: conditioning on time-varying confounders is necessary to remove omitted variable bias, but also induces post-treatment bias for the ATHE. That is, part of the effect of a treatment history might flow through its impact on X_{it} and fixed effects estimators remove these "indirect paths" by conditioning on these time-varying covariates. As Sobel (2012) points out, the ATHE is only identified in a fixed effect model if past treatment history has no effect on the time-varying covariates. But this is often a tenuous assumption because it implies that there might be effects of the treatment on the outcome, but not on other time-varying processes.

A Fixed Effects IPTW Estimator

Given the above, it appears as though fixed-effects estimators are, in general, of limited use with sequentially ignorable data. Yet, under a broader definition of "fixed-effects" and an assumption about the memory of the causal processes, we can consistently estimate (marginal) causal effects in TSCS data even in the face of lagged dependent variables and unmeasured heterogeneity.

In general, scholars often find sequential ignorability rather implausible, but are more comfortable with ignorability within units. We can capture this generalization by allowing ignorability to hold only conditional on an unmeasured, time-constant variable, U_i. We call this assumption *fixed effects sequential ignorability*:

Assumption 2 (Fixed Effect 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* $Y_{it}(\underline{a}_t) \perp\!\!\!\perp A_{it} | \underline{X}_{it}, \underline{A}_{i,t-1} = \underline{a}_{t-1}, U_i$.

Note the treatment assignment is sequentially ignorable only once we condition on the unmeasured confounder, U_i. Thus, this assumption does not guarantee the usual sequential ignorability assumption holds. Recently, Sobel (2012) and Chernozhukov et al. (2009) have considered the estimation of causal effects under a roughly similar assumption.

If the probabilities of treatment assignment were known, then the fact that these probabilities differed across units would not prevent us from the estimation of a marginal structural model by inverse probability of treatment weighting. However, if these probabilities are not known, then it is not possible to estimate the IPTW weights, because we would only have one unit to estimate each set of weights. That is, the ignorability assumption, as stated, requires the probability of treatment to depend on the unmeasured confounder and the entire covariate and treatment history. Obviously, these histories grow with time, so that the conditioning set for time t and time $t+1$ will differ, rendering us unable to use one to help estimate a model for the other. Under the usual sequential ignorability assumption, we could draw on other units to help us estimate these weights. Here, though, the weights are unit-specific.

To avoid these issues, we introduce an assumption that strengthens fixed effects sequential ignorability in a manner consistent with the current use of fixed effects models. Namely, we restrict the dependence of the treatment history and the outcome, so that the conditioning set does not grow with time. We define k-order sequential ignorability, which requires the treatment assignment to depend on only the last k time periods:

Assumption 3 (k-Order Sequential Ignorability). *For every action sequences $\underline{a}_t$, covariate history $\underline{X}_{i,t-k}$, and time $t > k$, if $\underline{A}_{i,t-1:t-k} = \underline{a}_{t-1:t-k}$, then $Y_{it}(\underline{a}_t) \perp\!\!\!\perp A_{it} | \underline{X}_{i,t:t-k}, \underline{A}_{i,t-1:t-k} = \underline{a}_{t-1:t-k}, U_i$.*

This assumption states that treatment assignment in time period t is unrelated to treatment or time-varying covariates from more than k periods ago, conditional on the history of these variables between $t-k$ and t.[7] Many implementations of fixed effects have this "Markovian" property inherent in their model specification. What this assumption allows us to do, however, is estimate unit-specific probability of treatment. One way to see this: break up each unit into T/k blocks of size k. Define these NT/k blocks as the units and then the k-order sequential ignorability assumption becomes the

[7] We use $\underline{A}_{i,t:s}$ to be the partial history of treatment between times t and s: $(A_{is}, A_{i,s+1}, \ldots, A_{it})$.

standard sequential ignorability assumption, conditional only on observables (since now the original unit to which the block belongs is observed).

More specifically, under this assumption, we can weight units to remove the confounding due to both the time-varying covariates and the unmeasured confounding, where we define the weights as:

$$\widehat{SW}_{it} = \frac{\prod_{s=t-k}^{t} \widehat{\Pr}[A_{is} | \underline{A}_{i,s-1:s-k}; \hat{\gamma}]}{\prod_{s=t-k}^{t} \widehat{\Pr}[A_{is} | \underline{A}_{i,s-1:s-k}, \underline{X}_{i,s:s-k}, U_i; \hat{\alpha}_i]}. \tag{13}$$

The denominator of these weights is simply the probability of receiving the treatment that unit i did receive at time t, conditional on the last k periods. The numerator is the probability of the same event, marginalizing across the time-varying covariates and the unit effects. This function has no effect on the consistency of the estimator, rather it simply stabilizes the weights and increases efficieny (Robins, Hernán, and Brumback, 2000). Because we have multiple observations per unit, we can estimate the denominator either by including a unit dummy variable or by separately estimating a model for each unit. Under either approach, U_i becomes a measured covariate and we can apply the standard proofs from Robins (2000) as long as the models in (13) are consistent for the true probabilities of treatment.[8] Obviously, a pooled estimator will work perfectly fine for the numerator.

Robins (2000) proves that the data reweighted by $\widehat{SW}_{it}$ will be free of confounding due to $\underline{X}_{it}$ and U_i and yet have the same distribution of the potential outcomes, $Y_{it}(\underline{a}_t)$, as the original data. Thus, in the reweighted data we can estimate the marginal mean of the potential outcome, $E[Y_{it}(\underline{a}_t)]$, with the conditional mean, $E_{SW}[Y_{it} | \underline{A}_t]$. This approach, which we call IPTW Fixed Effects (IPTW-FE), avoids both the need to condition on the time-varying covariates *and* the need to use a "within" estimator for the outcome. The procedure for this estimation strategy is as follows:

1. For each unit, run a unit-specific logit or probit model to estimate the probability of treatment in time t, conditional on covariates and treatment histories. Use fitted values from these models for the denominator of the weights.

2. Run a pooled logit or probit model to estimate the components numerator of the weights. Use fitted values from this model estimate the numerator of the weights.

[8] This consistency obtains as both $T \to \infty$ and $N \to \infty$.

3. Use the estimated weights, $\widehat{SW}_{it}$, in a weighted regression or weighted generalized linear model to estimate the marginal structural model, $E[Y_{it}(\underline{a}_t)]$.

This procedure will perform better as we increase both the number of units and the number of time periods. Intuitively, the more time-periods we have, the more precisely we will be able to estimate the propensity scores for each unit. And as we increase the number of units, the precision of our causal estimates will increase. For standard errors, one can simply use the conservative Huber-White standard errors.

6 Simulations: Fixed Effects and Weights

In this section, we provide evidence for the finite-sample properties of our proposed estimator. This is especially important in this case because the asymptotic behavior of the estimator relies on both the sample size and the number of time periods. Because the weights require a separate time-series model for each unit, it is important to know how many over-time observations are needed before the estimator gives reasonable performance. Further, we can compare this estimator to other approaches to causal estimation: a pooled model, a traditional fixed-effects model, and an IPTW model with known weights. In general, each of the simulations below tell a similar story with three standout features. First, while fixed effects outcome models do a good job estimating the CET, they are very poor at estimating any other effects of treatment histories even as the number of time periods grows. Second, even with a small number of time periods, IPTW and marginal structural models with known weights can (somewhat noisily) recover both the CET and any ATHE. Third, the IPTW-FE approach does a decent job estimating the CET and improves upon fixed effects outcome models for the ATHE. Furthermore, this improvement grows in the number of time periods.

Continuous Outcomes

Our first set of simulation results comes from a model with continuous Y_{it}, a situation where fixed effects outcome models are commonly used. In this set of simulations, we tried to create a simple, but powerful dynamic setting. Figure 3 show a DAG of the assumed structure in the simulations. The basic idea is that the time-varying covariate, X_t affects the treatment A_t, which affects the outcome Y_t, which in turn affects the time-varying covariate X_{t+1}. In addition, each of these variables affects the next three in the causal ordering, so that Y_{t-1} affects X_t, A_t, and Y_t. Finally, there is an unmeasured confounder U that affects all variables. It's important to note that this setup is Markovian, conditional on the unmeasured confounder, so that it satisfies first-order sequential ignorability. This structure might be less restrictive than it sounds, since it allows for a lagged dependent variable effect and for feedback between the outcome and the independent variables. Furthermore, while this structure is Markovian conditional on the unit fixed-effect, there can be quite complicated dynamic structures in the marginal distribution of the data.[9] One way to think about the IPTW estimators is that the reweighted data has the same DAG as Figure 3 with the arrows into A_t and A_{t-1} removed, but all other arrows remaining.

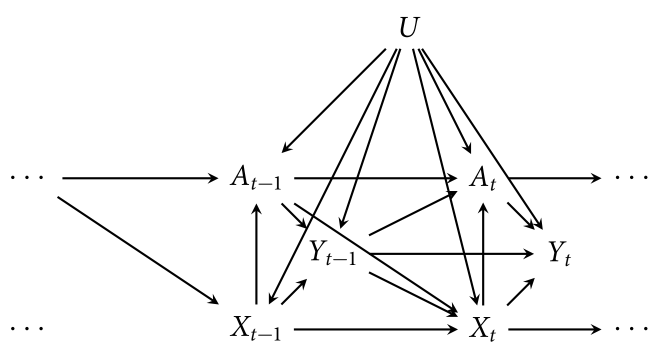

Figure 3: Simulation Structure

	Formula
Unit effect	$U_i \sim \mathcal{N}(0, 0.5^2)$
Covariate	$X_{it} = 0.45 * U_i + 0.1 * X_{i,t-1} + 0.7 * A_{i,t-1} + 0.2 * Y_{i,t-1} + \mathcal{N}(0,1)$
Treatment	$A_{it} = 1$ w.p. $\Phi(1 * U_i + 0.1 * A_{i,t-1} + 0.25 * X_{i,t-1} + 0.25 * Y_{i,t-1})$
Outcome	$Y_{it} = 1 - 0.6 * U_i + 0.2 * Y_{i,t-1} + 0.5 * X_{it} + 0.5 * A_{it} + \mathcal{N}(0, \sigma_y^2)$

Table 2: Simulation parameters with a continuous outcome.

[9] Various restrictions of this model give qualitatively similar results, with IPTW-FE possessing slightly better finite sample performance.

Table 2 present the specific data generating process and parameters for this simulation. It is important to note that the simulations here have constant effects in both time and units, which is favorable to the outcome fixed effects models. While the contemporaneous treatment value, A_t, has a direct effect on the outcome, the lagged value of treatment, $A_{i,t-1}$ only has indirect effects. But it does have a number of different ways it can affect the outcome: through the lagged dependent variable ($A_{i,t-1} \rightarrow Y_{i,t-1} \rightarrow Y_{it}$), through the lagged dependent variable's effect on the covariates ($A_{i,t-1} \rightarrow Y_{i,t-1} \rightarrow X_{it} \rightarrow Y_{it}$), and through the covariate itself ($A_{i,t-1} \rightarrow X_{it} \rightarrow Y_{it}$). These indirect paths are effectively cut off by fixed effects models that condition on the lagged dependent variable and the time-varying covariate. In these simulations, we fix the number of units to be 500, the number of simulations to be 1,000, and allow each of the models to be correctly specified.[10]

Figure 4 presents the results of this simulation. The columns show each of three estimands: the CET, the last period effect, and the overall ATHE. There are several items to note in these results. First, as expected, the pooled estimators exhibit strong bias for all of the parameters. Second, the fixed effects outcome model does the best job at estimating the contemporaneous effect, with roughly the same mean as IPTW with known propensity scores, but far more efficient. This advantage breaks down when we look at the effect of the lagged treatment, A_{t-1}. Here, we see that both of the IPTW estimators perform reasonably well, with the estimated weights improving efficiency. This efficiency improvement from estimating weights is a common result in propensity score theory (Hirano, Imbens, and Ridder, 2003). Our novel estimation approach vastly improves estimation over the alternative when considering estimands other than the CET, but does poorly when estimating the CET with just a handful of time periods. As the time periods increase, we can see that the IPTW-FE performance improves considerably.

A restricted model

A potential concern here is that the complexity of the simulations is causing problems for outcome fixed effects models. For instance, it is well-known that fixed effects models have difficulty with lagged

[10] By correctly specified, we mean that in the estimation models we include all variables with non-zero coefficients in the data generating process.

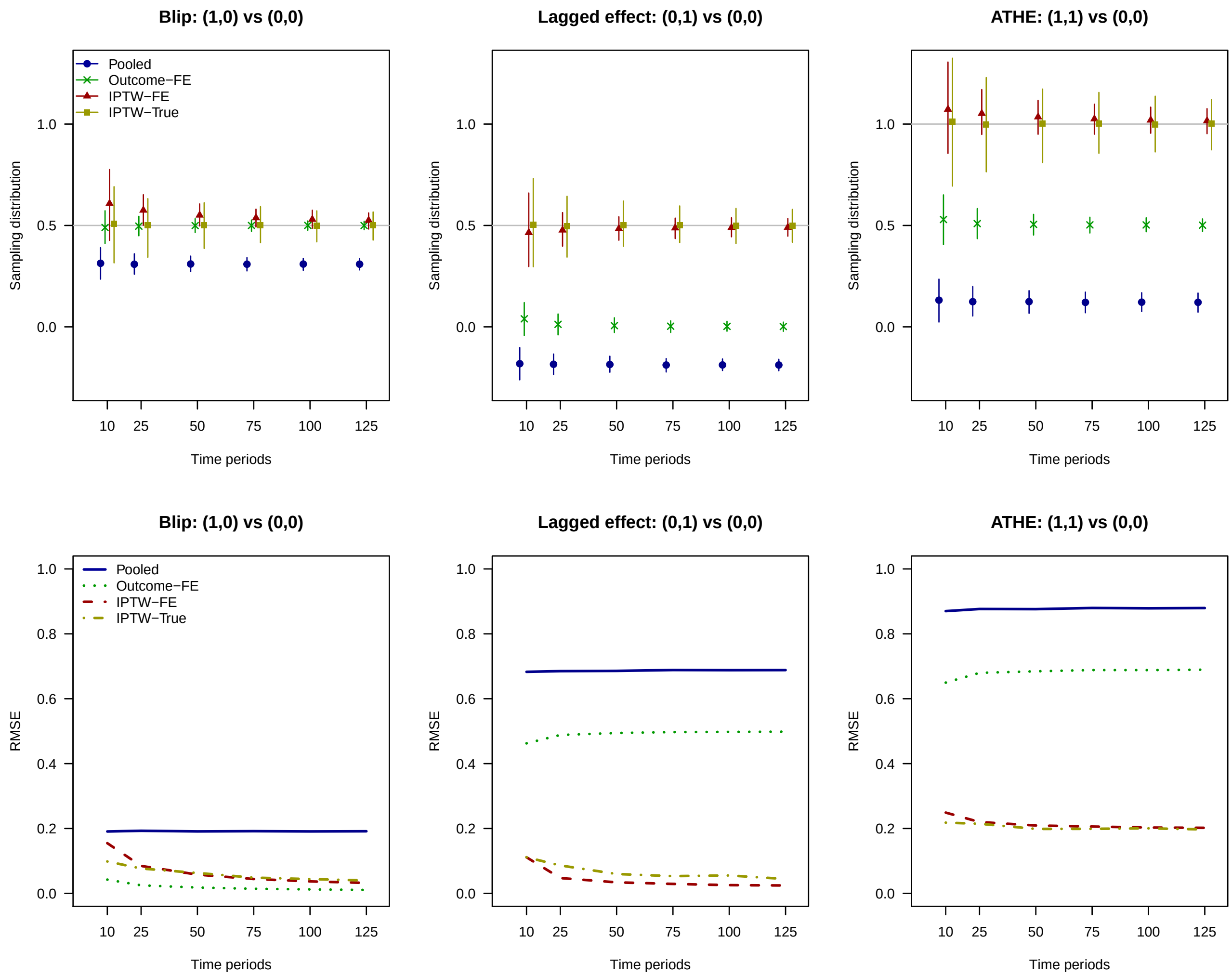

Figure 4: Results for a Monte Carlo simulation with continuous outcomes. The top panels are the sampling distributions of the various estimators. The bottom panels presents a comparison of the root mean-squared error of the various estimators.

dependent variables. Or, perhaps the number of causal pathways from A_{t-1} to Y_t stacks the deck against fixed effects. To show that the problems with fixed effects are unrelated to these points, we ran another simulation with the restricted structure of Figure 5: one where the random effect only impacts the treatment and the outcome, and the only connection between periods is a link between

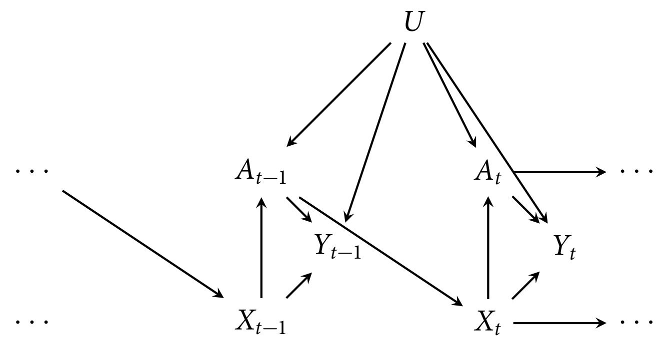

Figure 5: Simulation with Restricted Structure

A_{t-1} and X_t.[11] Thus, here we have removed the lagged dependent variable from the data generating process and only have one indirect effect of the lagged treatment. And yet the results of this simulation in Figure 6 show that outcome fixed effects models cannot recover ATHEs even here. Note also that in this setting, the IPTW-FE reduces bias more quickly in T than in the more general setting. This indicates that there are a host of factors that could influence the finite sample performance of the weighting estimator, including how one variable affects another over time.

Binary outcomes

The IPTW-FE model makes fixed effect estimation possible in situations where it was difficult or impossible beforehand. To demonstrate this, we investigate the performance of IPTW-FE with binary outcomes, a situation that should be very challenging for estimating treatment history effects. In many cases, researchers either have to impose strong assumptions on the distribution of the unobserved effect or forgo the estimation of the usual causal estimands. To show how the weighting approach works in such a setting, we take a slightly restricted structure from the last setting and change the continuous variables to be binary instead. All other parameters (Table 3) are the same. One complication here are the effects themselves. Although we generate the data with probit coefficients, these coefficients do not correspond to actual causal effects such as the CET or the ATHE. To calculate these marginal effects, we need to apply the g-computational formula (Robins, 1986; Robins,

[11]The parameters of the simulations are the same with zeros replacing the coefficients for the dropped arrows in Figure 5.

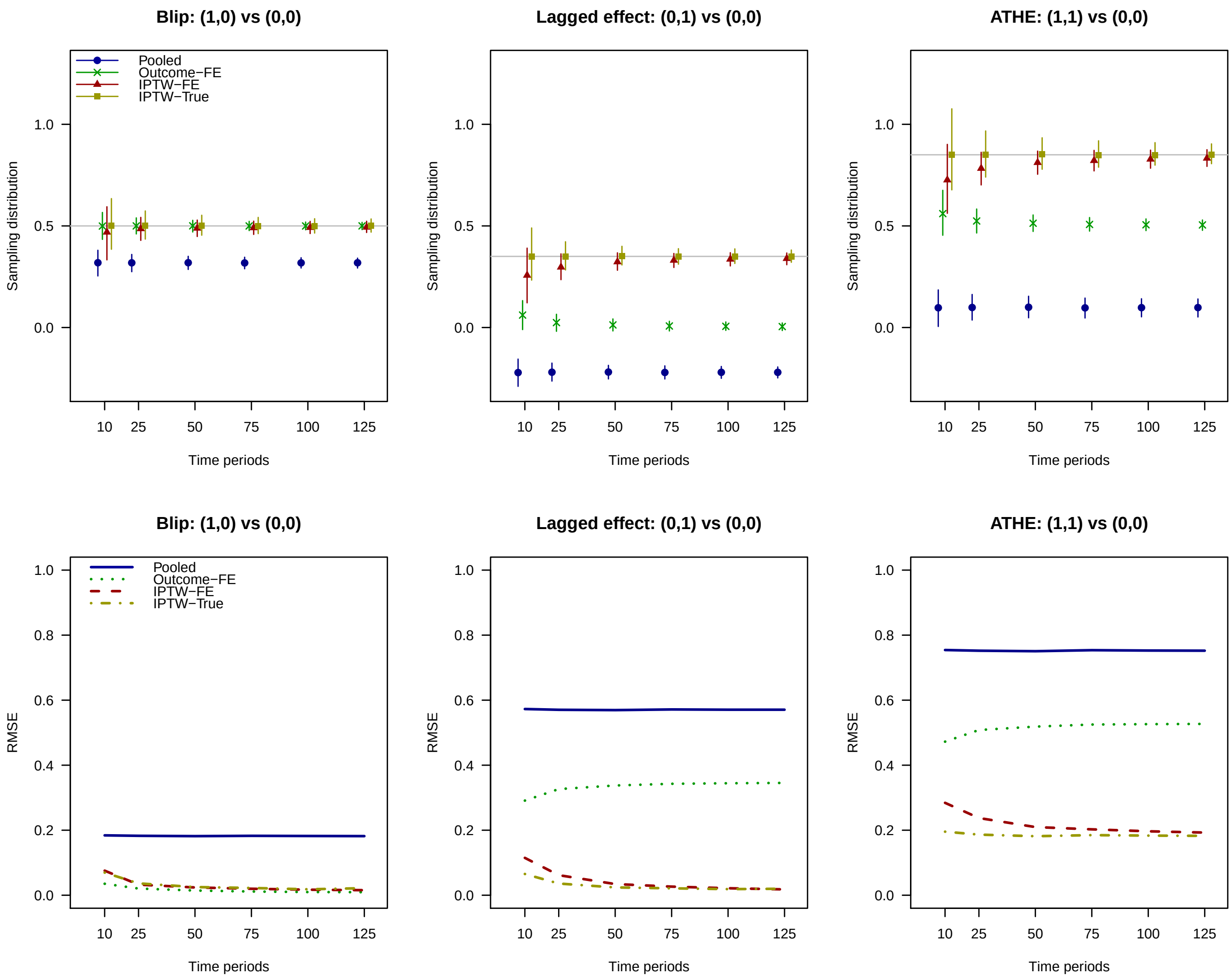

Figure 6: Results for a Monte Carlo simulation with continuous outcomes and causal structure from Figure 5. The top panels are the sampling distributions of the various estimators. The bottom panels presents a comparison of the root mean-squared error of the various estimators.

Greenland, and Hu, 1999) and integrate over the distribution of the time-varying covariate, which is relatively straightforward in this case because it is binary. More difficult, however, is the integration of the unobserved effect. To calculate the "true" value of the causal effects, we leverage the simulations to numerically integrate out the U_i. For the estimators, we rely on the probit functional form to calculate the predicted probabilities that go into the estimated causal effects.[12]

	Formula
Unit effect	$U_i \sim \mathcal{N}(0, 0.5^2)$
Covariate	$X_{it} = 1$ w.p. $\Phi(0.45 * U_i + 0.7 * A_{i,t-1})$
Treatment	$A_{it} = 1$ w.p. $\Phi(1 * U_i + 0.1 * A_{i,t-1} + 0.25 * X_{i,t-1})$
Outcome	$Y_{it} = 1$ w.p. $\Phi(1 - 0.6 * U_i + 0.5 * X_{it} + 0.5 * A_{it})$

Table 3: Simulation parameters with all binary variables.

Figure 7 presents the results of this binary simulation. It is clear that the general pattern from the first two simulations carries over to this binary case. The IPTW model with the true weights noisily estimates the various effects, but these estimates become more precise as the number of time periods increases. The estimates from the IPTW-FE approach have small variances, but seem to also converge more slowly (in T) to the true weights. Even so, with a moderate number of time periods, the bias is small relative to the bias of the pooled model.[13] Thus, while IPTW-FE performs slightly worse with binary outcomes, it is still able to recover decent causal estimates in a situation where outcome fixed effect models are generally difficult or impossible.

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 and assumptions. 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. Finally, we investigated the performance of a fixed effects IPTW approach to the estimation of treatment his-

[12] For instance, with the IPTW-FE approach, if we have a probit coefficients of $\hat{\beta}_0$ for the constant, $\hat{\beta}_1$ on A_t, and $\hat{\beta}_2$ on A_{t-1}, then the estimated effect of $(A_t, A_{t-1}) = (1, 1)$ versus $(0, 0)$ would be $\Phi(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) - \Phi(\hat{\beta}_0)$.

[13] We do not compare these estimates to an outcome fixed effects model because those models either have computational problems (as in the case of a probit model with unit dummy variables) or cannot be used to estimate marginal effects (as in the case of the conditional logit estimator).

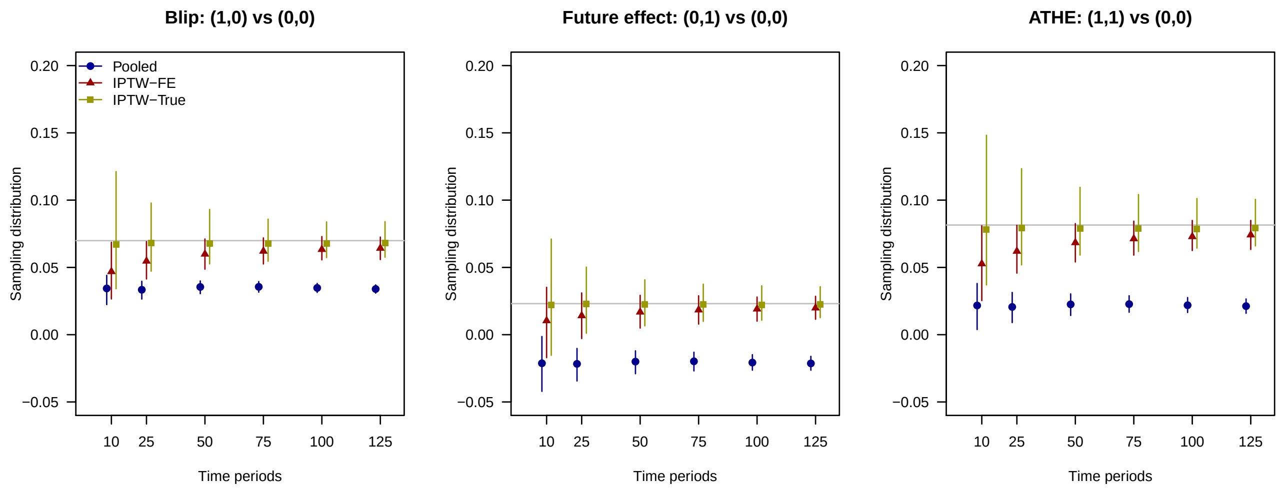

Figure 7: Results for a Monte Carlo simulation with binary outcomes, treatments and covariates. Sampling distributions are the vertical lines with their means as points.

tory effects, using Monte Carlo experiments. This approach was shown to have relatively good performance in the case where the true parametric model is known, the true parameters are constant, and the time dependence is first order. Future work should address the sensitivity of this approach when these conditions are relaxed.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics.* An Empiricist's Companion. Princeton University Press, December. (Cit. on p. 8).

Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to do (and not to do) with Time-Series Cross-Section Data." *American Political Science Review* 89 (3): 634–647. (Cit. on p. 8).

———. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models." *Political Analysis* 6 (1): 1–36. (Cit. on p. 8).

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. (Cit. on pp. 2, 3, 8, 11, 12).

Blackwell, Matthew. 2013. "A Framwork for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57 (2): 504–520. (Cit. on pp. 2, 9).

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). (Cit. on pp. 2, 16).

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): (cit. on p. 20).

Imai, Kosuke, and In Song Kim. 2012. "On the Use of Linear Fixed Effects Regression Models for Causal Inference." July. (Cit. on p. 2).

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. (Cit. on p. 8).

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. (Cit. on p. 11).

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. (Cit. on pp. 6, 22).

———. 2000. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials,* ed. M. Elizabeth Halloran and Donald Berry, 95–134. Vol. 116. The IMA Volumes in Mathematics and its Applications. New York: Springer-Verlag. (Cit. on pp. 11, 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): (cit. on pp. 3, 11, 22).

Robins, James M., Miguel A Hernán, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5): 550–560. (Cit. on pp. 2, 9, 11, 17).

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58. (Cit. on p. 4).

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. (Cit. on pp. 2, 15, 16).