Sequential Heterogeneous Treatment Effect Estimation for Dynamic Time-Series Cross-Sectional Data*

Adam Kaplan†

April 2022

Abstract

Increasingly more granular measurements and modern machine learning methods allow political scientists to flexibly estimate treatment effect heterogeneities in cross-sectional settings. These methods, however, are not appropriate for an important source of heterogeneity, time. In this paper, I develop a sequential and model-agnostic heterogeneous treatment effect estimation procedure under the sequential ignorability assumption in dynamic time-series cross-sectional settings. My method allows researchers to estimate complex heterogeneities, including for lagged outcomes, for both contemporary and long-term treatment effects by iteratively re-estimating the heterogeneous effect using blipped-down outcomes. I illustrate this procedure through simulations and an application.

*This is very much a work in progress, so please do not distribute without permission and take simulation/application results with a grain of salt as there are most definitely implementation issues present at this point in time.

†PhD student in Political Science. Email: akapl@mit.edu.

1 Introduction

Treatment effects in Political Science rarely affect all the recipients the same way, making heterogeneity often a quantity of interest. Heterogeneous treatment effects, contemporaneous and long-term, are particularly interesting in time-series cross-sectional (TSCS) settings, as now, past treatments, past outcomes, and (past) covariates can and usually do affect the impact of any particular treatment regime. Although researchers are working with increasingly more granular measurements, such as monthly drone strikes (Rigterink, 2021) or global sub-national analyses of refugees (Zhou and Shaver, 2021), the method of choice for estimating heterogeneous treatment effects is still a linear regression (often with 2-way fixed effects) with interaction effects.

Unfortunately, this comes with two important limitations. First, for both contemporaneous and long-term effects, linearity is a strong assumption. Simply replacing the linear model with a modern machine learning method (e.g. Hahn, Murray, and Carvalho (2017), Athey, J. Tibshirani, and Wager (2018), Künzel et al. (2019), and Nie and Wager (2020)) is also not appropriate, as most rely on the independent and identically distributed (i.i.d.) assumption. My proposed procedure, however, through its sequential blipping-down of the outcome allows for a model-agnostic and asymptotically unbiased estimation in dynamic TSCS settings.

The second limitation, concerns long-term effect estimation. In applied works, e.g. Rigterink (2021), it is common to estimate long-term treatment effects by simply including the past treatment as a covariate in the model, in addition to future treatments and covariates, interpreting its coefficient as the long-term effect of a treatment. This approach yields biased estimates, due to post-treatment bias, of long-term effects. Dropping all the future variables is not a solution either, as that will often induce an omitted variable bias in dynamic TSCS settings (Blackwell and Glynn, 2018).

Additionally, an important caveat to long-term effects is that, whereas in the contemporaneous case, the coefficient on the treatment measures all possible paths to the outcome, when you control for intermediate variables, future treatments, covariates, etc., only the direct effect is estimated. That is the effect that flows through all variables other than future treatments. Although my method cannot help with the problem of estimating the long-term total effects, it provides an asymptotically unbiased estimate of the long-term direct effect while allowing researchers to express the regression in the same way they do now.

I rely on the sequential ignorability assumption (J. Robins, 1986) to identify the quantities of interest and introduce a sequential estimation procedure inspired by sequential g-estimation (Vansteelandt and Joffe, 2014; Vansteelandt and Sjolander, 2016). My method first estimates preliminary heterogeneous treatment effects, uses them to construct a blipped down outcome, and finally uses those to update the preliminary estimates. Although the procedure is similar to sequential g-estimation and both rely on sequential ignorability for identification, I do not assume a linear functional form, but weaken it by assuming the functional form of the treatment effect is correctly specified. For most treatment effect estimation approaches, such as meta-learners (Künzel et al., 2019), this assumptions is equivalent to assuming the correct outcome model, which can be made more credible by choosing a flexible enough model. In practice, the researcher can further choose whether they want to estimate a time-invariant treatment effect function $\tau\left(\bar{\mathbf{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}\right)$ or a time-variant one

$\tau_{st}\left(\bar{\mathbf{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}\right)$. Here, time-invariant simply means, the heterogeneous treatment effect function is the same across time, but it can still rely on time covariates s, t.

This estimation procedure yields asymptotically unbiased estimates for both contemporaneous and long-term effects, assuming the preliminary estimator is asymptotically unbiased. In practice, for an appropriate and flexible model, a researcher can estimate treatment effects as functions of not only covariates, past treatments, but also past outcomes, describing potentially very complicated treatment effect functions.

To illustrate the utility of my proposed method, I run Monte Carlo simulations and apply it to a recent piece by Rigterink (2021). In the application, I extend their analysis of the effect of drone strikes killing leaders on future terrorist activity, by 1) estimating heterogeneous long-term effects as a function of past aggressiveness, proxied by the lagged outcome, and 2) correctly estimate the average long-term effects of drone strikes on terrorist attacks. Albeit statistically insignificant, I find a negative effect of drone strikes as past aggressiveness increases, and increasing average long-term effects about half the size smaller than originally estimated by Rigterink (2021).

2 Motivation

To motivate my procedure I introduce a recent piece by Rigterink (2021), which I use throughout for examples and extend in section 5. Rigterink (2021) examines the impact of drone strikes killing terrorist group leaders on future terrorist attacks against civilians using a novel data set. She finds a rather striking result that drone strikes killing leaders actually increase the number of attacks against civilians by double digit percentages. She uses a two-way fixed effect regression with lagged treatments and a parallel trends assumption to estimate the contemporaneous and long-term effects of drone strikes on subsequent attacks against civilians.

To estimate the effect in question, Rigterink (2021) relies on a novel data set consisting of 13 terrorist organizations i with monthly t observations for 12 years (2004-2015). In the paper, she decomposes the treatment into two binary variables, encoding whether a drone strike targeted a leader and if it also killed them. Though this allows her to estimate the effect of a kill and miss separately, for the purposes of this paper, I only consider the leader being killed as the treatment. The outcome of interest Y_{it} is the logged number of terror attacks[1].

"To credibly identify the causal impact of targeted leader killing, the probability of a hit must not be driven by prior trends in the outcome variable. This would, for example, be the case if counterterrorist organizations would accept a higher or lower probability of a hit for terrorist groups that commit an increasing number of terrorist attacks (Rigterink, 2021, p. 40)." With this form of a parallel trends assumption she estimates the short-term and long-term effect jointly using the 2-way fixed effect model (Rigterink, 2021, eq. 1):

$$Y_{it} = \sum_{k=-6}^{6} \beta_{i,t-k}\text{hit}_{i,t-k} + \sum_{k=-6}^{6} \delta_{i,t-k}\text{targeted}_{i,t-k} + \sum_{k=-6}^{6} \gamma_{i,t-k}X_{i,t-k} + \mu_i + \theta_t + \varepsilon_{it}, \qquad (1)$$

[1]To be precise it is $\log(\text{count}_{it} + 1)$, where count_{it} is the number of terror attacks against civilians.

where X_{it} are additional covariates. She finds that two-, three-, and six-months ago drone strikes lead to approximately 39% to 53% increases in terror attacks against civilians. I replicate the results in table 1.

	Log # of Terrorist Attacks	
	(1)	(2)
Leader Killed	0.298	0.233
	(0.188)	(0.186)
Leader Killed (1 months ago)	0.209	0.120
	(0.188)	(0.187)
Leader Killed (2 months ago)	0.390**	0.374**
	(0.188)	(0.186)
Leader Killed (3 months ago)	0.533***	0.537***
	(0.188)	(0.187)
Leader Killed (4 months ago)	0.119	0.130
	(0.188)	(0.186)
Leader Killed (5 months ago)	0.095	0.036
	(0.185)	(0.184)
Leader Killed (6 months ago)	0.422**	0.300
	(0.192)	(0.184)
Leads?	Yes	No
N	1,577	1,655
R^2	0.886	0.883
Adjusted R^2	0.871	0.870
Residual Std. Error	0.495 (df = 1394)	0.500 (df = 1484)
F Statistic	59.197*** (df = 183; 1394)	65.807*** (df = 171; 1484)

*p < .1; **p < .05; ***p < .01
Newey-West standard errors not reported.

Table 1: Truncated regression table for the (Rigterink, 2021) application. The first model is the replication and the second is the same, but without leads of the outcome included. I omitted replicating the Newey-West standard errors, so they differ from the original results (table 5, model 1 in the original paper), but the significance levels are still retained.

3 Proposed Methodology

In this section I introduce the quantities of interest, the contemporaneous total average effect and the long-term direct average effect, and present the proposed estimation method. It is an adapted sequential g-estimation estimation procedure (Vansteelandt and Sjolander, 2016). It consists of 1) iteratively estimating the quantity of interest, 2) calculating the blipped down outcome H_{ist} for all $i, s \geq t$, and 3) re-estimating the quantity of interest by replacing the outcome Y_{is} with its blipped down version H_{ist} for all $s \geq t$.

3.1 Notation

Throughout this paper I follow the Rubin-Neyman potential outcome framework for causal inference (Neyman, 1923; Rubin, 1974). Let $Y_{it}(\mathbf{h})$ denote the potential outcome at time $t \in \{1, \ldots, T\}$ for unit $i \in \{1, \ldots, N\}$ under the treatment regime $\mathbf{h}$. For exposition purposes, I am assuming a balanced panel, though this is not required for the validity of the method. I also follow the convention to represent a vector with a bold letter, with the notable exception of a history of a variable (introduced below). Denote the observed outcome for unit i at time t as Y_{it}. In addition, let D_{it} and $\boldsymbol{X}_{it}$ be the actual treatment and covariates for unit i at time t. For exposition, I assume binary treatment $D_{it} \in \{0, 1\}$, though this procedure can be generalized for categorical and continuous treatments.

Furthermore, I denote the history of a variable V_{it} with a bar, e.g. $\bar{V}_{it} = \{V_{i1}, \ldots, V_{it}\}$, which is a random variable itself, following J. Robins (1986). I use this notation to express the potential outcomes in an alternate, but equivalent form as $Y_{is}(\bar{D}_{it}, \mathbf{0})$, which is the potential outcome for unit i at time $s \geq t$ for the treatment regime $\mathbf{h}$ where the first t treatments are given by $\bar{D}_{it}$ and in the remaining $s - t$ periods, no treatment is received.[2]

3.2 Quantities of Interest

The three quantities of interest are all conditional average treatments, but differ on whether they concern short-term, long-term for specific treatment and outcome times, or average short-/long-term treatment effects. The first quantity of interest is the contemporaneous conditional total average treatment effect,

$$\tau_{tt}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) = \mathbb{E}\left[Y_{it}(\bar{D}_{it}, \mathbf{0}) - Y_{it}(\bar{D}_{it-1}, \mathbf{0}) \mid \bar{\boldsymbol{X}}_{it} = \bar{\boldsymbol{x}}, \bar{Y}_{it-1} = \bar{y}, \bar{D}_{it-1} = \bar{d}\right], \tag{2}$$

where I abuse notation slightly, as $\bar{\boldsymbol{x}}, \bar{y}, \bar{d}$ do not have to be full histories, but can be any subset of them, even only containing the most recent occurrences $\boldsymbol{X}_{it}, Y_{it-1}, D_{it-1}$. Note that this quantity is simply the contemporaneous conditional average treatment and I refer to it as the contemporaneous treatment effect (CTE) from hereon. In my drone strike example, the CTE is the immediate (1 month) effect of a drone strike on the likelihood of a terrorist attack today.

The second quantity of interest is the long-term conditional direct average treatment effect (LDTE) for $s > t$,

$$\tau_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) = \mathbb{E}\left[Y_{is}(\bar{D}_{it}, \mathbf{0}) - Y_{is}(\bar{D}_{it-1}, \mathbf{0}) \mid \bar{\boldsymbol{X}}_{it} = \bar{\boldsymbol{x}}, \bar{Y}_{it-1} = \bar{y}, \bar{D}_{it-1} = \bar{d}\right], \tag{3}$$

where the same caveat about the full histories holds. Although very similar, this quantity is different from the CTE in two important ways. First, the outcome time s is now strictly larger than t, meaning I am looking at long-term effects. Second, and more importantly, the effect is now only a direct effect, referred to as the controlled direct effect in the mediation literature (Pearl, 2001; James M Robins, 2003), as opposed to a total effect. This can be seen by the fact that after period t, the potential outcome sets all treatments to 0. In the drone

[2] In the categorical and continuous treatment case, it is necessary to encode the treatment in such a way that "0" denotes no treatment. See Vansteelandt and Sjolander (2016, section 4.2) for an example of how to encode a continuous treatment

strike example, the LDTE is the effect of a drone strike 6 months ago on today's likelihood of a terrorist attack, without any other strikes in the last 6 months.

This does not mean that in reality no treatment can happen after time t, rather, it ensures that all of the indirect effects, i.e. effects that flow from D_{it} through future treatments, are not included in the estimate, ruling out any "accumulation effects." I present this visually in figure 1. The dashed arrow shows the familiar CTE, and the sum of the red arrows the LDTE for a one-period lag.

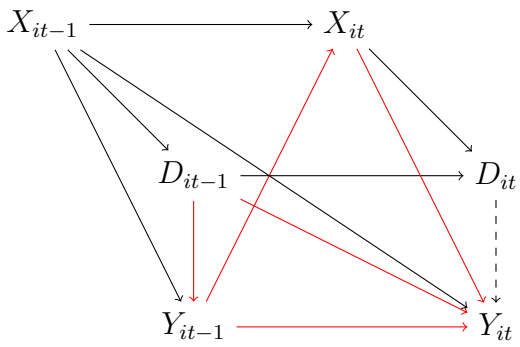

Figure 1: Causal directed acyclic graph (DAG) (Pearl, 2009) that satisfies sequential ignorability (assumption 3) and showcases the two different quantities of interest. Here the dashed arrow is the CTE and the sum of the red arrows is the LDTE for a one-period lag.

This distinction highlights a "time-series bias conundrum" my procedure can help solve. In most applied work with TSCS data interested in long-term effects, researchers estimate a statistical model, often a linear regression, with the treatment of interest D_{it} included as a lagged variable, alongside future treatments $D_{it'}$ for $t' > t$. Unfortunately, the coefficient of D_{it} cannot be interpreted as the LDTE. This estimation approach yields a biased estimate of the LDTE, due to post-treatment bias of including covariates affected by the past treatment D_{it}. In the case of non-dynamic TSCS data, i.e. no lagged outcome, this can be alleviated by omitting all post-treatment variables in the estimation. In the setting this paper considers, dynamic TSCS, however, omitting all post-treatment variables in the estimation leads to omitted variable bias in the contemporaneous effect, which affects the LDTE through the lagged outcome (Blackwell and Glynn, 2018). I call it the "bias conundrum", because no matter what the researcher tries to do in the standard regression framework, the LDTE estimate will be biased.

My estimation procedure outlined below, allows one to estimate the LDTE without bias (asymptotically), by iteratively "blipping-down" the current outcome, essentially removing the contemporaneous treatment effect sequentially until the desired lag.

Finally, the third quantity of interest is a slight modification of the previous two quantities. As opposed to having to specify an outcome and treatment time of interest, the average CTE or LDTE only requires specifying the effect length (or lag) L one is interested in. Formally for any $s > L \geq 0$,

$$\tau_L(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) = \mathbb{E}\left[\tau_{t+L,t}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) \mid \bar{\boldsymbol{x}}, \bar{y}, \bar{d}\right], \tag{4}$$

averages the CTE (if $L = 0$) or LDTE (if $L > 0$) over treatment time t, holding the covariates at the prespecified values $(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$. Intuitively, the average CTE/LDTE is the *average* effect

of a treatment at any point in point in time. This quantity of interest makes sense for treatments that can potentially happen at any point in time, such as drone strikes, but is not appropriate for analyzing effects of a one-time policy change, for which I would use the CTE/LDTE. For the drone strike example with $L = 6$, this is the average effect of a drone strike at any point in time on the likelihood of a terrorist attack in 6 months, without any further drone strikes in the meantime.

3.3 Identification

I rely on the well-established sequential ignorability or randomization (J. Robins, 1986; Vansteelandt and Joffe, 2014), consistency, and positivity assumptions to non-parameterically identify both quantities of interest. The consistency assumption rules out any interference across units, positivity ensures that at any point in time, each unit can potentially receive treatment, and sequential ignorability requires that treatment in each period is assigned randomly conditioned on the past treatments, outcomes, and covariates. I only state the assumptions and the identification result in this section and leave the proof for the appendix.

Assumption 1 (Consistency)**.** *Let* $\mathbf{h}_i$ *be any treatment history for unit* i *and* $\mathbf{h}$ *the treatment history for all units. Since outcomes, covariates, and treatments are affected by past treatments, I need to make a consistency assumption akin to SUTVA in the cross-sectional case. That is,* $\bar{Y}_{it}(\mathbf{h}) = \bar{Y}_{it}(\mathbf{h}_i)$, $\bar{\boldsymbol{X}}_{it}(\mathbf{h}) = \bar{\boldsymbol{X}}_{it}(\mathbf{h}_i)$, $\bar{D}_{it}(\mathbf{h}) = \bar{D}_{it}(\mathbf{h}_i)$ *for all* i, t*. In other words, the only treatment history that affects potential outcomes, covariates, and treatments is one's own. The second part of this assumptions links potential values to observed ones. For any observed treatment history up until time* t *it holds that* $\bar{Y}_{it}(\bar{D}_{it}) = \bar{Y}_{it}$, $\bar{\mathbf{X}}_{it}(\bar{D}_{it}) = \bar{\mathbf{X}}_{it}$ *for all* i, t*.*

This assumption also justifies the equivalent potential outcome notation, $Y_{is}(\bar{D}_{it}, \mathbf{0})$, I introduced earlier.

Assumption 2 (Positivity)**.** *In order to properly define an expectation, I need to ensure that the treatment density is non-zero conditional on past covariates, outcomes, and treatments. That is,* $f(\bar{D}_{it} \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}) \neq 0$ *for all* i, t*.*

Note that in the discrete treatment case, this is equivalent to the treatment probability being non-zero for all possible treatment values.

Assumption 3 (Sequential ignorability)**.** *Treatment assignment must be randomized at any point in time* t *conditioned on the history of covariates, outcomes, and treatments. Specifically,* $\{Y_{is}(\mathbf{h}_i) | s \geq t\} \perp\!\!\!\perp D_{it} \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}$ *for all* i, t*, and* $\mathbf{h}_i$*.*

Sequential randomization is a strong assumption and needs to be carefully justified. To put this into a more familiar context, in a linear model, this assumption implies the sequential exogeneity assumption (Blackwell and Glynn, 2018). Under sequential ignorability, common confounders such as past outcomes, treatments, covariates, and time fixed-effects are all allowed. Note that these assumption non-parameterically identify the causal estimands in this paper, but estimation model selection can impose stricter assumptions, such as strict exogeneity for two-way fixed effect linear models. I discuss this for some common models in the model choice discussion in section 3.5.

Proposition 1. *Under assumptions 1, 2, and 3, the three quantities of interest introduced in section 3.2 are non-parameterically identified.*

See appendix A for the proof.

3.4 Estimation

The estimation procedure is the key contribution of this paper. It is model-agnostic (both in terms of treatment effect and/or outcome model) allowing for highly complex models, its iterative nature is inspired by sequential g-estimation (Vansteelandt and Sjolander, 2016), and is asymptotically unbiased assuming the treatment effect model is correctly specified. In this section I introduce the necessary parametric assumption for estimation and the actual procedure itself.

Before stating the assumption, I define the blipped-down outcome H_{ist}. Formally, adapted from Vansteelandt and Sjolander (2016),

$$H_{ist} = Y_{is} - \sum_{u=t+1}^{s} \hat{\tau}_{su}(\bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}). \tag{5}$$

The key intuition behind this quantity is that under expectation it equals the potential outcome under control[3],

$$\mathbb{E}\left[H_{ist} \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}\right] = \mathbb{E}\left[Y_{is}(\bar{D}_{it-1}, \mathbf{0}) \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}\right]. \tag{6}$$

Hence the name, blipped-down, as it blips-down the outcome by one treatment. Importantly, by definition, $H_{iss} = Y_{is}$. With this, I can present the final assumption.

Assumption 4 (Treatment effect model)**.** *The treatment effect model as identified by proposition 1,*

$$\begin{aligned}\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) = \mathbb{E}\left[H_{ist} \mid \bar{\boldsymbol{X}}_{it} = \bar{\boldsymbol{x}}, \bar{Y}_{it-1} = \bar{y}, \bar{D}_{it-1} = \bar{d}, D_{it} = 1\right] \\ - \mathbb{E}\left[H_{ist} \mid \bar{\boldsymbol{X}}_{it} = \bar{\boldsymbol{x}}, \bar{Y}_{it-1} = \bar{y}, \bar{D}_{it-1} = \bar{d}, D_{it} = 0\right].\end{aligned} \tag{7}$$

is correctly specified for all $t \leq s$.

Note that $\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ is an estimand, since the two expected values on the left-hand-side are population expectations, and thus need to be estimated. I denote the estimate of $\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ as $\hat{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$. Therefore, the estimand $\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ is defined as a difference in means of the blipped down outcome H_{ist}, which are themselves defined by a *preliminary* estimate $\hat{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ of $\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$.

This seemingly cyclic definition is valid upon closer inspection of the blipped down outcome. For any fixed s, t, $\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ is defined by preliminary estimates $\hat{\tau}_{su}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ for $s \geq u > t$. That is, at each point in time I only use *future estimates* $\hat{\tau}_{su}(\cdot)$ to define the current *estimand* $\tilde{\tau}_{st}(\cdot)$.

Furthermore, assumption 4 is more general by design to encapsulate different approaches of modeling $\tilde{\tau}_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$, though the most common ones, specifying an outcome model and

[3]See for example Vansteelandt and Joffe (2014) for proof of this claim.

using the difference in means, or any other meta-learner (Künzel et al., 2019), boil down to correctly specifying the outcome model. For example, one can assume that the outcome model $\mathbb{E}\left[Y_{is} \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}\right]$ is correctly specified and estimate $\hat{\tau}_{st}(\cdot)$ as,

$$\begin{aligned}&\hat{\mathbb{E}}\left[Y_{is} \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}, D_{it}=1\right] \\ &\quad - \hat{\mathbb{E}}\left[Y_{is} \mid \bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it-1}, D_{it}=0\right].\end{aligned} \tag{8}$$

That being said, it is not advisable to estimate $\hat{\tau}_{st}$ directly by regressing Y_{is} on $\bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it}$ as it will likely be plagued by omitted variable bias. Rather, the sequential estimation approach proposed here relies on iteratively re-estimating the contemporaneous treatment effect with blipped-down outcomes.

The sequential estimation procedure can be divided into two main categories, depending on the modeling choice of $\tilde{\tau}_{st}(\cdot)$. Although both CTE and the LDTE are indexed by s, t, thus requiring a fixed outcome time and treatment time respectively, I can choose to model this in different ways. The most direct way is using a time-varying treatment effect function. That is, estimating $\tilde{\tau}_{st}(\cdot)$ with $\hat{\tau}_{st}(\cdot)$, a separate function for each s, t pair. A simpler approach is to assume a time-invariant, or common function across all s, t pairs and estimate $\tilde{\tau}_{st}(\cdot)$ with $\hat{\tau}(\cdot)$. Although the function itself is time-invariant, I can include the outcome s and treatment time t as covariates to still allow the treatment effect to vary across time. I discuss the advantages and disadvantages of each modeling choice in section 3.5.

3.4.1 Time-varying treatment effect $\hat{\tau}_{st}(\cdot)$

1. Regress Y_{it} on $\bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it}$ pooled for all i, t to get an (asymptotically) unbiased estimate $\hat{\tau}_{tt}$ of $\tilde{\tau}_{tt}$ for all t.

2. Calculate $H_{ist} = Y_{is} - \sum_{u=t+1}^{s} \hat{\tau}_{su}(\bar{\boldsymbol{X}}_{iu}, \bar{Y}_{iu-1}, \bar{D}_{iu-1})$ for all possible s and $t \leq s$.

3. Regress H_{ist} on $\bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it}$ pooled for all i, t, s to get an (asymptotically) unbiased estimate $\hat{\tau}_{st}$ of $\tilde{\tau}_{st}$.

4. Repeat steps 2 and 3 as often as needed (always using the newest estimate for $\hat{\tau}_{st}$) until I have an estimate for all the desired t, s time points.

Although this procedure is designed to calculate the CTE or LDTE, it can be easily extended to the average CTE/LDTE. Rather than repeating steps 2 and 3 as often as needed, repeat them until you have an estimate $\hat{\tau}_{st}(\cdot)$ for all s, t such that $s - t = L$. Then, the average CTE/LDTE can be estimated as the average of the different CTEs, LDTEs as follows,

$$\hat{\tau}_L(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) = \frac{1}{T-L} \sum_{t=1}^{T-L} \hat{\tau}_{t+L,t}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}). \tag{9}$$

3.4.2 Time-invariant treatment effect $\hat{\tau}(\cdot)$

1. Regress Y_{it} on $\bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it}$ pooled for all i, t to get an (asymptotically) unbiased estimate $\hat{\tau}$ of $\tilde{\tau}_{tt}$.

2. Calculate $H_{ist} = Y_{is} - \sum_{u=t+1}^{s} \hat{\tau}(\bar{\boldsymbol{X}}_{iu}, \bar{Y}_{iu-1}, \bar{D}_{iu-1})$ for all s and $t \leq s$.

3. Regress H_{ist} on $\bar{\boldsymbol{X}}_{it}, \bar{Y}_{it-1}, \bar{D}_{it}$ pooled for all i, t, s to get an (asymptotically) unbiased estimate $\hat{\tau}$ of $\tilde{\tau}_{st}$ for all $s, t \leq s$.

Steps 2 and 3 are necessary here, as the regression in step 1 yields an inefficient estimate for $\tilde{\tau}_{st}(\cdot)$ since it only uses the immediate effect (Vansteelandt and Sjolander, 2016).

The time-invariant approach also allows for two ways of estimating the average CTE/LDTE. Most directly, one can include the lag L as a covariate, essentially estimating $\tilde{\tau}_L(\cdot)$ as $\hat{\tau}_L(\cdot) = \hat{\tau}(\cdot, L)$. Indirectly, I can use the same sample average approach as with the time-variant procedure. That is, estimate $\tilde{\tau}_{st}(\cdot)$ as $\hat{\tau}_{st}(\cdot) = \hat{\tau}(\cdot, s, t)$ for all s, t pairs, which I get after step 3 already, and then estimate the average CTE/LDTE as,

$$\hat{\tau}_L(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}) = \frac{1}{T-L} \sum_{t=1}^{T-L} \hat{\tau}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d}, t+L, t). \tag{10}$$

3.5 Model Selection

Because the estimation procedure proposed here is fully model-agnostic, it is important to highlight the costs and benefits of different modeling choices. In this section I go over some of the decisions, such as including a lagged outcome and/or fixed effects in the estimation model, as well as when to use the time-variant or time-invariant procedure. Although I provide some preliminary overview of the costs and benefits, formalizing and expanding this is an important area of future research.

3.5.1 Lagged Outcomes and Fixed Effects

The sequential estimation procedure can theoretically handle lagged outcomes and fixed effects, as they both comply with the sequential ignorability assumption, but different model choices can impose even stronger assumptions.

I tackle the case of lagged outcomes *without* fixed effects first. Using a simple OLS regression with lagged outcomes, assuming there is no leftover autocorrelation, yields consistent (and asymptotically unbiased), but biased (in small samples) estimates under sequential exogeneity, which is implied by sequential ignorability, and some additional stationarity assumptions (Keele and Kelly, 2006). More flexible non-linear models, such as additive models (Hastie and R. J. Tibshirani, 1990) or random forests[4] (Breiman, 2001), are often used in practice and perform well with time-series data, but their performance depends on including the appropriate transformations of lagged outcomes and time variables. If the suspected time-series is not too complicated, any of these methods should work, but one can also rely on more time-series specific ones such as Arellano and Bond (1991) or Hausman and Pinkovskiy (2017) for linear or Capitaine, Genuer, and Thiébaut (2021) for non-linear models.

A common model in practice are two-way fixed effects without lagged outcomes. In this case, the estimation method, two-way fixed effects, actually imposes stronger assumptions

[4]Note that technically random forests are inappropriate due to their non-blocked bootstrapping, but seem to perform well in practice. Time-series appropriate forests are currently being developed with some recent work by Goehry (2020).

than implied by the necessary identification ones (sequential ignorability). Recent work by Imai and Kim (2021) and others has also shown that under the parallel trends in multi-period settings, the two-way fixed effect model also imposes a linear functional form assumption to recover causal estimates. Including a lagged outcome in addition to unit (and optionally time) fixed effects leads to the Nickell (1981) bias and is thus not recommended. For non-linear models, including fixed effects often leads to the incidental parameter problem, regardless of the inclusion of a lagged dependent variable.

Mixed effect models can help alleviate issues of time-invariant and unit-invariant effects by replacing strict exogeneity with potentially weaker parametric assumptions about the random effects. Linear, generalized additive models, and random forests (Capitaine, Genuer, and Thiébaut, 2021) all have mixed effect extensions. Alternatively, generalized estimating equations (GEE) and their additive model extension (Yee and Wild, 1996) also yield asymptotically unbiased estimates.

3.5.2 Time-variant vs. Time-invariant

A second important modeling decision is whether to use the time-variant or time-invariant treatment effect model. The decision boils down to efficiency and desired flexibility. In all models, linear or non-linear, one can express time-invariant models as time-variant ones by including all possible interactions with the outcome s and treatment time t (and potentially other covariates) in the model.

Linear models benefit, from the added flexibility time-variancy gives them, more than non-linear ones. In linear models, including all possible interactions between s and t in the time-invariant model can be computationally inefficient as there are $O(T^2)$ interactions. Simply omitting the interactions on the other hand reduces the expressiveness of the model compared to the time-variant one.

For non-linear models, especially highly flexible ones like random forests, time-variant estimation can actually reduce statistical efficiency, as at each step in the procedure, the forest has only a subset of the data to learn from. In a time-invariant model, the random forest already implicitly interacts s, t with all other variables, and uses the whole data set to learn these interactions.

Thus, the choice between time-variant and time-invariant comes down to two considerations. First, whether one believes the true treatment effect does actually sufficiently vary as a function of s, t, or L. Second, is the estimation model already flexible enough that the common treatment function assumption made by the time-invariant procedure does not matter.

4 Simulation Study

In this section I display the performance of the sequential estimation procedure on three simulated data sets. All data sets are AR(1) processes and only differ in their heterogeneity. The first is linear and varies with a covariate, the second is non-linear and varies with a covariate, and third is non-linear and varies as a function of the past outcome. I compare the time-variant, time-invariant, time-invariant with lag, with three baseline models.

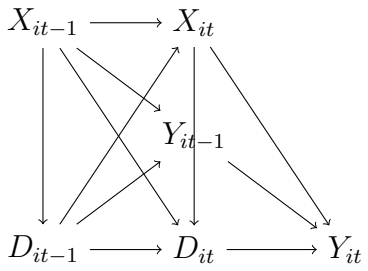

Figure 2: Slice of the causal DAG (Pearl, 2009) of the simulated data generating processes outlined in this section.

4.1 Simulation Setup

All three data generating processes (DGPs) share a common data generating process. There are $N = 1000$ units recorded for a total of $T = 10$ time periods. For each observation, I observe a binary treatment $D_{it} \sim \text{Bern}(\sigma(X_{it} - 0.2D_{it-1}))$ and a covariate $X_{it} = (0.3D_{it-1} - 0.3 * (1 - D_{it-1})) \cdot X_{it-1} + \mathcal{N}0, 1)$ where $D_{i0} \sim \text{Bern}(0.5)$ and $X_{i0} \sim \mathcal{N}5, 16)$ for all i, t. Here, $\sigma(\cdot)$ is the standard logistic function. The outcome, $Y_{it} = Y_{it-1} + \tau_{st}(\bar{\boldsymbol{X}}_{it})$, is then calculated as follows,

$$Y_{it} = Y_{it-1} + 3D_{it-1} - 2X_{it} \cdot D_{it} + \varepsilon_{it}, \tag{DGP 1}$$

$$Y_{it} = Y_{it-1} + 3D_{it-1} - 2\sin(X_{it}) \cdot D_{it} + \varepsilon_{it}, \tag{DGP 2}$$

$$Y_{it} = Y_{it-1} + 3X_{it} \cdot D_{it-1} - 2\sqrt{|Y_{it-1}|} \cdot D_{it} + \varepsilon_{it}, \tag{DGP 3}$$

where $\varepsilon_{it} = 0.6\varepsilon_{it-1} + \mathcal{N}0, 1)$ with $\varepsilon_{i0} \sim \mathcal{N}0, 1)$ for all i, t. With this, the three models imply three different heterogeneous treatment effects,

$$\tau_{st}^{(1)}(X_{it}) = -2X_{it} + 3\mathbb{1}\left(t < s\right), \tag{DGP 1}$$

$$\tau_{st}^{(2)}(X_{it}) = -2\sin(X_{it}) + 3\mathbb{1}\left(t < s\right), \tag{DGP 2}$$

$$\tau_{st}^{(3)}(X_{it}, Y_{it-1}) = -2\sqrt{|Y_{it-1}|} + 3\mathbb{E}\left[X_{it+1} \mid X_{it}\right]\mathbb{1}\left(t < s\right). \tag{DGP 3}$$

Thus, the first two DGPs are equivalent with the exception that the heterogeneity is non-linear through the $\sin(\cdot)$ transformation. Comparing DPG 1 and 2 showcases how a flexible model, GAM in this case, in conjunction with the sequential estimation can estimate heterogeneities with non-linearities. The third DGP's goal is to showcase a highly complex dynamic heterogeneity.

I run six different models on each of the data generating processes. A time-invariant, time-invariant with lag specification, a time-variant, and three baseline models for both a linear and additive model. The first three model specifications showcase different ways of estimating the same quantity of interest with the same procedure. The baseline models are there to compare my method to standard valid and invalid approaches of estimating the heterogeneous treatment effect. They consist of a regression where I estimate the LDTE by including lagged treatment indicators as well as future treatments (post-treatment bias baseline), or omitting all future treatments and covariates (omitted-variable bias baseline), and simply interacting the treatment with time dummies (interaction baseline).

For the first DGP, I estimate $Y_{it} = X_{it} \cdot D_{it} \cdot t + X_{it-1} + D_{it_1}$ as the outcome model for all sequential and the interaction baseline specifications, using $\cdot$ as the interaction operator $*$ in `R`, i.e. including the individual terms as well as the interaction. Note that I do not have to include the lagged outcome in this model, since I know that the true treatment effect function does not vary with it. Including it would only increase the efficiency of the model. I use the S-Learner to estimate the treatment effect from the outcome model. For the time-invariant cases, the blipdown regression includes an additional interaction with the outcome time, $H_{ist} = X_{it} \cdot D_{it} \cdot t \cdot s + X_{it-1} + D_{it_1}$ or with the lag, $H_{ist} = X_{it} \cdot D_{it} \cdot L + X_{it-1} + D_{it_1}$.

Since the first and second DGP are equivalent except in the non-linearity in X_{it}, I estimate the same models as in the first case, replacing the linear model with a GAM and the three-way interaction with a tensor-product thin-plate regression spline smooth f, e.g. $Y_{it} = f(X_{it}, t) \cdot D_{it}$, and similarly for the other specifications.[5]

For the last DGP, I estimate $Y_{it} = f(Y_{it-1}, t) \cdot D_{it}$ for the additive model with some additional control variables.[6] Note that the thin-plate regression spline smooth f will approximate the highly non-linear $-2\sqrt{|Y_{it-1}|}$ function, but will struggle around 0, where the treatment effect has a discontinuity.

4.2 Simulation Results

From the treatment effect specification in the previous section, you notice that there are essentially two different treatment effects for each DGP. A contemporaneous one, where the term with the indicator function drops, and a long-term one, where the indicator function evaluates to 1. Thus, I present the results for only two outcome, treatment time pairs. I consider the effect at the last observed treatment at time $s = 10$ for the contemporaneous treatment time of $t = 10$ and the long-term effect at $t = 8$. I present a visual result and some Monte Carlo simulations in this section, and leave all other results for the appendix B. For all cases and model specifications, the confidence intervals are based on 500 block bootstrap iterations, where I block on the unit i.

Figure 3 shows the heterogeneous treatment effect (both CTE and LDTE) for the third DGP as a function of the lagged outcome. I can see that all models perform well for the CTE, as is expected, since neither the post-treatment or omitted-variable bias kick in, in this case. Though, as foreshadowed, the GAM struggles to reach the peak around 0, which is a limitation of generalized additive models, and using an even more flexible model such as a random forest (Breiman, 2001) can potentially perform better. For the LDTE, all models barring the post-treatment bias baseline, perform similarly well. Notably, the omitted-variable bias baseline model actually outperforms, with respect to point estimates, the other specifications, but is far too confident with narrow confidence intervals. As expected, the post-treatment baseline model is heavily biased and misses the treatment effect function entirely. I present similar figures for the other two DGPs in the appendix.

Although a visual analysis of figure 3 can give some insights, I also conduct a series of Monte Carlo simulations to assess the statistical properties of my proposed procedure. In tables 2 and 3 I show bias and coverage rates of the six models for each DGP based on 200

[5] I also include linear terms for X_{it}, X_{it-1}, D_{it}, and D_{it-1} in all specifications.

[6] I again include linear terms for X_{it}, X_{it-1}, D_{it}, D_{it-1}, and Y_{it-1} in all specifications.

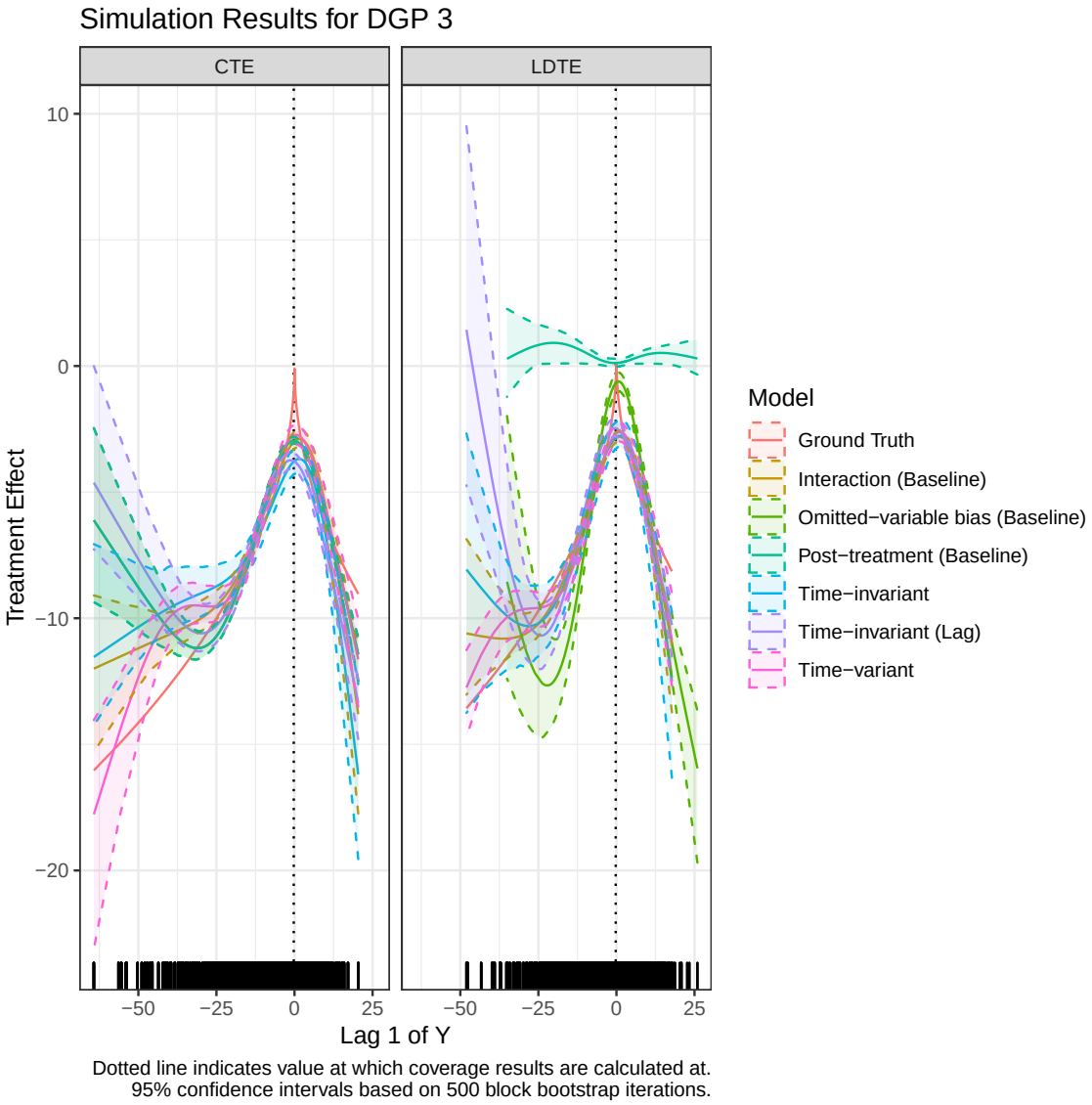

Figure 3: Simulation results for the third DGP, based on the first Monte Carlo simulation. The ground truth represents the population treatment effect function shown in the previous section. All other models are as described previously. Confidence intervals shown are based on 500 block bootstrap iterations.

	DGP 1		DGP 2		DGP 3	
	Bias	Coverage	Bias	Coverage	Bias	Coverage
Time-invariant	1.10	0.39	3.55	0.00	-1.87	0
Time-invariant (Lag)	1.49	0.00	-0.27	0.89	-1.99	0
Time-variant	0.10	0.92	0.19	0.92	-1.39	0
Interaction (Baseline)	-0.52	0.78	3.81	0.00	-1.28	0
Omitted-variable bias (Baseline)	0.31	0.74	-0.02	0.94	-1.36	0
Post-treatment (Baseline)	0.31	0.74	-0.02	0.94	-1.36	0

Table 2: Bias and coverage results of the CTE for 200 Monte Carlo simulations for the three different DGPs. Coverage is calculated based on 500 block bootstrap iterations per Monte Carlo simulation. Both bias and coverage are evaluated at the global mean of the heterogeneity inducing variable. For DGP 1, I run a linear model and evaluated at the mean of X_{it}. For DGP 2 and 3, I run a GAM and evaluate DGP 2 at the mean of X_{it}, but DGP 3 at the mean of Y_{it-1}.

Monte Carlo simulations each. Both the bias and coverage is evaluated at a single point, the global mean of X_{it} and Y_{it-1}, depending on the DGP. I also calculated an "average coverage" for each model, which is the average of proportions of true treatment effects covered in the full range of X_{it} or Y_{it-1} in each simulation, but leave the results for the appendix.

Overall, the results are disappointing. All models, except the time-variant specification are heavily biased and undercover the true effect in all specifications. Considering the CTE first, the two time-invariant specification bounce between good and bad performance among the first two DGPs. Though disappointing, this is actually a good lesson. The time-invariant specification made a simplifying assumption that one can model the treatment effect by one function and include time or the lag as a covariate. Though that is certainly possible, it requires an even more flexible model than a GAM, or at least a better additive model specification than what I have run. The time-variant model performs really well in the first two DGPs, but horribly in the third (as all others do as well). This is not necessarily a weakness of the method, but rather an artifact of the third DGP and GAM modeling choice. Referring back to figure 3, note that the coverage and bias are evaluated at a point very close to 0, a discontinuity point, unlikely to be accurately modeled by a GAM. This unfortunate issue will be addressed in future versions of this paper. For the CTE, the baselines do not perform too badly, as is expected since neither the post-treatment or omitted-variable bias kick in.

Turning to the LDTE, the results are similar, though now all the specifications seem to be performing worse on average. The time-variant model is the best for the linear case, but struggles in the non-linear one, where the time-invariant specification outperforms it. I am not sure why the time-variant model performed so poorly in the second DGP, but from visual inspection of figure 3 and the average coverage rate of 0.36 (see table B.2), I believe the confidence interval might have just slightly missed the true effect likely due to model misspecification as the bias increased too. Looking at the bias of the time-invariant procedures, they highlight another important modeling decision. Interacting the treatment

	DGP 1		DGP 2		DGP 3	
	Bias	Coverage	Bias	Coverage	Bias	Coverage
Time-invariant	-1.08	0.15	0.13	0.94	-1.08	0.02
Time-invariant (Lag)	-0.78	0.16	-0.61	0.28	-1.03	0.00
Time-variant	0.55	0.89	-0.97	0.32	-1.14	0.00
Interaction (Baseline)	-3.17	0.00	-1.21	0.02	-1.35	0.00
Omitted-variable bias (Baseline)	0.64	0.20	0.54	0.16	-1.96	0.22
Post-treatment (Baseline)	1.15	0.24	1.07	0.26	2.08	0.00

Table 3: Bias and coverage results of the LDTE for 200 Monte Carlo simulations for the three different DGPs. Coverage is calculated based on 500 block bootstrap iterations per Monte Carlo simulation. Both bias and coverage are evaluated at the global mean of the heterogeneity inducing variable. For DGP 1, I run a linear model and evaluated at the mean of X_{it}. For DGP 2 and 3, I run a GAM and evaluate DGP 2 at the mean of X_{it}, but DGP 3 at the mean of Y_{it-1}.

effect with a time covariate in the time-invariant case is a double-edged sword. On the one hand, it can help decrease the bias for the LDTE estimation in cases where the LDTE is very similar across different treatment periods. On the other hand, the CTE (or a highly variable LDTE) estimation is going to suffer because the treatment effect will not pool information from other periods. Finally, all the baselines struggled in estimating the LDTE, as the omitted-variable or post-treatment bias kicked in, similar to what I have shown in figure 3.

5 Empirical Application

In this section I motivate the use of sequential ignorability for Rigterink (2021) and extend her analysis to heterogeneous treatment effects. Using a time-invariant model I find that increasing past aggressiveness, proxied by a lagged outcome, has a negative effect on future attacks, and recover positive, albeit much smaller, long-term effects. Unfortunately all my results are statistically not significant.

Rather than relying on the parallel trends assumption, I want to motivate the use of sequential ignorability in this application. Although a very strong assumption in general, sequential ignorability is particularly well-suited for this use-case. I must convince you that a drone strike killing a leader is conditionally independent of the potential outcomes, i.e. the logged number of terrorist attacks against civilians under a killing and miss. I argue that the most important control variables for identification are the past aggressiveness, proxied by the lagged outcome, whether a leader has been killed by a drone strike in the past (lagged treatment), the number of drone strikes levied against a particular group, and whether the group was actually targeted by a drone strike at any given month t. The inclusion of the past outcome in the control set warrants further justification and is motivated by the first hypothesis I aim to test.

Group	Avg. aggressiveness	Avg. target prob.
1	2.12	0.08
2	0.06	0.01
3	0.26	0.04
4	0.05	0.00
5	3.19	0.01
6	0.42	0.00
8	0.02	0.00
10	0.04	0.00
11	0.01	0.00
12	0.12	0.00
13	0.01	0.00

Table 4: Average aggressiveness, proxied by the outcome, and average drone strike targeting (not killed) probability across all observed time periods.

5.1 Dynamic Heterogeneous Effects

I posit that more aggressive groups in the past are likely to be more aggressive in the future, particularly if their leader is killed by a drone strike. In other words, I am hypothesizing that the treatment effect of a fatal drone strike is heterogeneous, and in particular, increasing, as the lagged outcome, the number of attacks against civilians, increases. Not only is this heterogeneity substantively interesting, if my hypothesis is supported, it actually motivates the inclusion of the lagged outcome in the control set for identification reasons. I state this hypothesis formally below.

Hypothesis 1. *The contemporaneous (1-month) treatment effect of a drone strike killing a terrorist organization leader at the end of the observation period T, increases as the past aggressiveness, Y_{it-6}, increases.*

For initial, descriptive, support of my first hypothesis I calculated the average number of logged terror attacks against civilians across time for each group and their likelihood to be targeted by a drone strike (not necessarily killed) and present the results in table 4. This purely descriptive evidence seems to suggest a positive correlation between likelihood of targeting, and thus killing a leader, and aggressiveness.

To test hypothesis 1 formally, I estimate a time-invariant generalized additive model. I aim to be as close to the original study as possible, so I keep a mostly linear functional form, except for the heterogeneous effect, which I model with a tensor-product thin-plate regression spline f. Additionally, I omit the group-level fixed effects to avoid a Nickell (1981) bias. That is, I fit the following model:

$$Y_{it} = f(Y_{it-1}, t) \cdot D_{it} + \beta_0 + \phi Y_{it-1} + \tau_0 D_{it} + \beta_1 D_{it-1} + \beta_2 \mathbf{X}_{it} + \beta_3 \mathbf{X}_{it} + \delta_t. \tag{11}$$

Here the heterogeneous LDTE among the treated is modeled as $\tau_{T,T-6}(Y_{i,T-6}) = f(Y_{i,T-6}, T-6) + \tau_0$. I present the estimated results in figure 4.

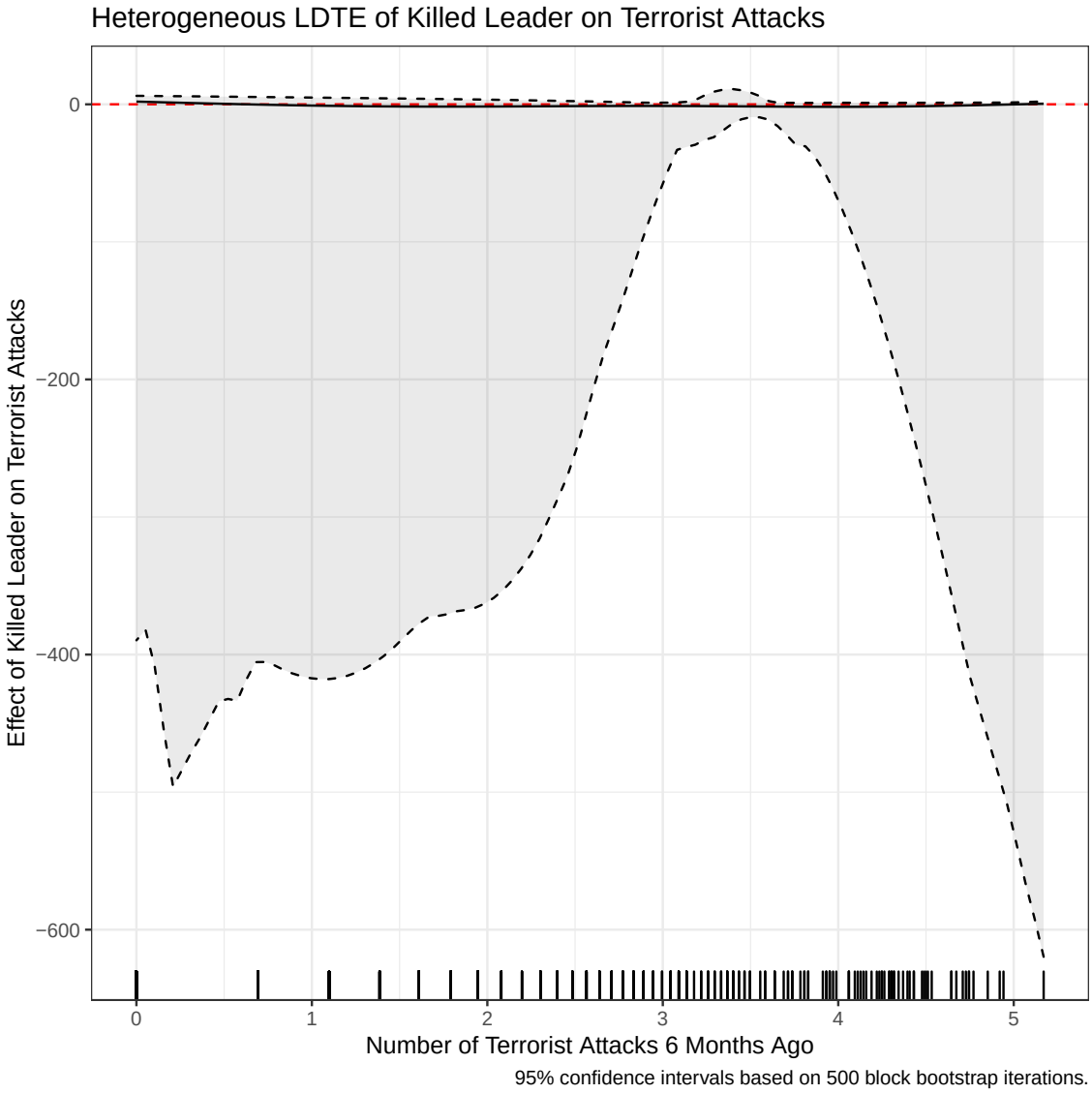

Figure 4: Heterogeneous LDTE among the treated estimates based on the model presented in equation 11. The 95% confidence intervals are based on 500 block bootstrap iterations, blocked on the group. At no level of past aggressiveness is the effect statistically significant.

Figure 4 suggests that hypothesis 1 is not supported. Although hard to see from the figure, the effect is actually negative, albeit statistically insignificant, almost throughout the entirety of the range of past aggressiveness. This result actually supports the argument that killing the leader of a terrorist organization, no matter the aggressiveness, might actually hamper the organization's ability to perform terrorist acts. To further investigate the effect of drone strikes of terrorist activity, I investigate the long-term average effects among the treated in the next section.

5.2 Long-term Effects

In my second hypothesis, I want to explicitly test whether the observation that the long-term effects of a drone strike killing a terrorist leader are in fact positive and increasing as suggested by Rigterink (2021) and seen in my replication in table 1. In this section, I estimate a time-invariant lag model to test this formally, and find that they do seem to be positive and increasing over time, but cannot attain statistical significance.

The effect of killing a leader with a drone strike on terrorist attacks is unlikely to be strong in the short-term (1-month), as carrying out an attack takes time and planning, especially after a leader is killed. With that in mind, it makes sense to look at the long-term effects, specifically the average LDTE among the treated, rather than contemporaneous ones. I express this hypothesis formally as,

Hypothesis 2. *The average LDTE among the treated of a drone strike killing a terrorist organization leader, increases as the time since the strike increases.*

Unfortunately, Rigterink (2021) two-way fixed effect model with lagged treatments is not appropriate for estimating these long-term effects. By including both past and future treatments, the estimates are plagued by post-treatment bias, as highlighted in the simulation section. My estimation procedure, however, allows one to estimate these effects asymptotically without bias.

Long-term effects are heterogeneous effects. In the previous section, I applied my method to estimate heterogeneous effects as a function of past outcomes, but another common heterogeneity in TSCS data is time itself. Assessing long-term effects is equivalent to estimating heterogeneous average LDTE, where the heterogeneity is in terms of time, or specifically lag L. Thus, to estimate the average LDTE, I model it using a time-invariant lag model. That is,

$$Y_{it} = f(t) \cdot D_{it} + \beta_0 + \phi Y_{it-1} + \tau_0 D_{it} + \beta_1 D_{it-1} + \beta_2 \mathbf{X}_{it} + \beta_3 \mathbf{X}_{it} + \delta_t. \tag{12}$$

Here the average LDTE among the treated is modeled as $\tau_L(L) = f(L) + \tau_0$. I present the estimated results in figure 5.

Figure 5 supports hypothesis 2, albeit at statistically insignificant levels. The average LDTE increases as the time since the drone strike increases, but it does so far slower than Rigterink (2021) results suggested. Not only that, but the size of the effects is drastically smaller (by about 50%) than most of the effect sizes estimated by Rigterink (2021), suggesting a possible upward post-treatment bias in the original analysis.

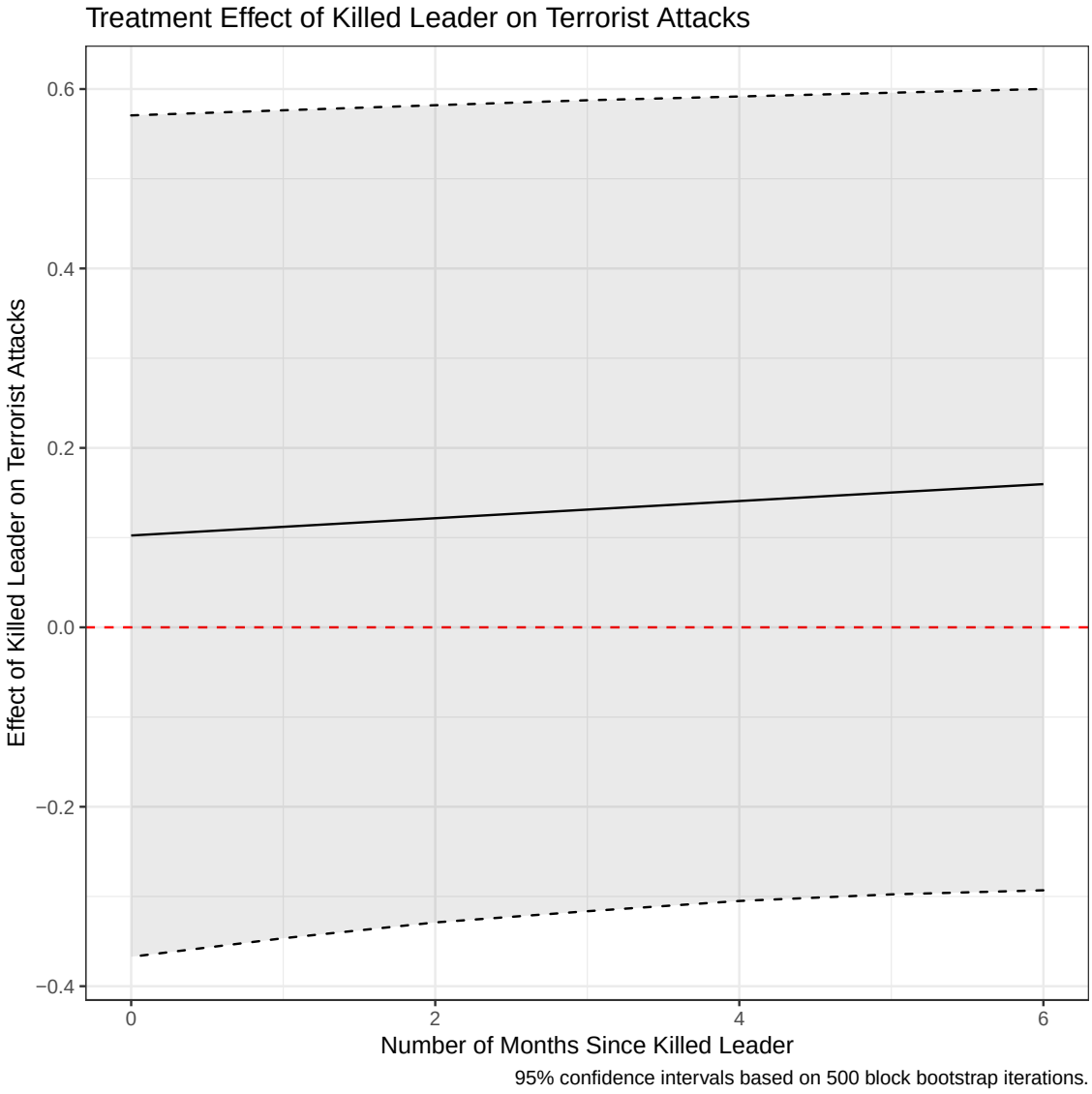

Figure 5: Average LDTE among the treated estimates for $L = 0, \ldots, 6$ based on the model presented in equation 12. The 95% confidence intervals are based on 500 block bootstrap iterations, blocked on the group. At no lag is the effect statistically significant.

6 Conclusion

I motivated this paper as a solution to two problems of estimating heterogeneous treatment effects with dynamic TSCS data. First, my model-agnostic sequential estimation procedure relaxes the usual linear functional form assumption used in applied work. Second, my method allows researchers to specify regression models the same way they are doing now, but estimate the long-term treatment effects without bias (asymptotically and under certain conditions), where standard approaches, such as including lagged and post-treatment treatment variables in the regression and reporting the coefficients, are plagued by it (Blackwell and Glynn, 2018). To make using this estimation procedure even easier, I am developing a `R` package that implements it for any model.[7] In my simulations I show that the procedure performs well for highly complex treatment effects, and motivate its use with the estimation of long-term effects of drone strikes killing terrorist organization leaders on future terrorist attacks against civilians based on Rigterink (2021).

Though this approach has all the advantages mentioned above, it still requires more work. The sequential ignorability assumption is very strong and there is so far no sensitivity analysis for my approach for the case it does not hold. I plan to tackle this problem in the future based on the sensitivity analysis for g-estimation introduced in Vansteelandt and Joffe (2014). My approach currently also only supports continuous outcomes. Since this approach is inspired by sequential g-estimation, and thus g-estimation (J. Robins, 1986; James M. Robins, 1994; Vansteelandt and Sjolander, 2016), I hope to generalize it to binary and count data in the future. Finally, at its current stage, the bias and coverage rates are highly dependent on a correct model specification, in particular for the time-invariant modeling assumption. Further work needs to be done in understanding how different modeling decisions affect the performance and validity of this estimation procedure.

[7] In its current state, the desired model needs to be able to accept a formula as its input, but can be easily extended to non-formula inputs by a translation function.

References

Arellano, Manuel and Stephen Bond (Apr. 1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". In: *The Review of Economic Studies* 58.2, p. 277. ISSN: 00346527. DOI: 10.2307/2297968. URL: https://academic.oup.com/restud/article-lookup/doi/10.2307/2297968 (visited on 04/12/2022).

Athey, Susan, Julie Tibshirani, and Stefan Wager (Apr. 5, 2018). "Generalized Random Forests". In: *arXiv:1610.01271 [econ, stat]*. arXiv: 1610.01271. URL: http://arxiv.org/abs/1610.01271 (visited on 03/30/2020).

Blackwell, Matthew and Adam N. Glynn (2018). "How to make causal inferences with time-series cross-sectional data under selection on observables". In: *American Political Science Review* 112.4. Publisher: Cambridge University Press, pp. 1067–1082. DOI: 10/gfdb4h.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324. URL: http://link.springer.com/10.1023/A:1010933404324 (visited on 04/15/2022).

Capitaine, Louis, Robin Genuer, and Rodolphe Thiébaut (Jan. 2021). "Random forests for high-dimensional longitudinal data". In: *Statistical Methods in Medical Research* 30.1, pp. 166–184. ISSN: 0962-2802, 1477-0334. DOI: 10.1177/0962280220946080. URL: http://journals.sagepub.com/doi/10.1177/0962280220946080 (visited on 04/12/2022).

Goehry, Benjamin (2020). "Random forests for time-dependent processes". In: *ESAIM: Probability and Statistics* 24, pp. 801–826. ISSN: 1262-3318. DOI: 10/gjnd8s. URL: https://www.esaim-ps.org/10.1051/ps/2020015 (visited on 04/05/2021).

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho (2017). "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3048177. URL: https://www.ssrn.com/abstract=3048177 (visited on 10/06/2020).

Hastie, Trevor J and Robert J Tibshirani (1990). *Generalized additive models*. Vol. 43. CRC press.

Hausman, Jerry A and Maxim Pinkovskiy (2017). "Estimating dynamic panel models: backing out the Nickell Bias". In: Publisher: FRB of NY Staff Report.

Imai, Kosuke and In Song Kim (July 2021). "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data". In: *Political Analysis* 29.3, pp. 405–415. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2020.33. URL: https://www.cambridge.org/core/journals/political-analysis/article/on-the-use-of-twoway-fixed-effects-regression-models-for-causal-inference-with-panel-data/F10006D0210407C5F9C7CAC1EEE3EF0D (visited on 06/03/2021).

Keele, Luke and Nathan J. Kelly (2006). "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables". In: *Political Analysis* 14.2, pp. 186–205. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mpj006. URL: https://www.cambridge.org/core/product/identifier/S1047198700001364/type/journal_article (visited on 04/13/2022).

Künzel, Sören R. et al. (Mar. 5, 2019). "Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning". In: *Proceedings of the National Academy of Sci-*

ences 116.10, pp. 4156–4165. ISSN: 0027-8424, 1091-6490. DOI: 10/gfwcnc. arXiv: 1706.03461. URL: http://arxiv.org/abs/1706.03461 (visited on 11/20/2020).

Neyman, Jerzy S (1923). "On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)". In: *Annals of Agricultural Sciences* 10, pp. 1–51.

Nickell, Stephen (1981). "Biases in dynamic models with fixed effects". In: *Econometrica: Journal of the econometric society*. Publisher: JSTOR, pp. 1417–1426. DOI: 10/cv5bt6.

Nie, Xinkun and Stefan Wager (Aug. 6, 2020). "Quasi-Oracle Estimation of Heterogeneous Treatment Effects". In: *arXiv:1712.04912 [econ, math, stat]*. arXiv: 1712.04912. URL: http://arxiv.org/abs/1712.04912 (visited on 10/06/2020).

Pearl, Judea (2001). "Direct and indirect effects". In: *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. UAI'01. Number of pages: 10 Place: Seattle, Washington. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 411–420. ISBN: 1-55860-800-1.

— (2009). *Causality: models, reasoning, and inference*. 2. ed. Cambridge: Cambridge Univ. Press. 464 pp. ISBN: 978-0-521-89560-6.

Rigterink, Anouk S. (Feb. 2021). "The Wane of Command: Evidence on Drone Strikes and Control within Terrorist Organizations". In: *American Political Science Review* 115.1. Publisher: Cambridge University Press, pp. 31–50. ISSN: 0003-0554, 1537-5943. DOI: 10.1017/S0003055420000908. URL: https://www.cambridge.org/core/journals/american-political-science-review/article/wane-of-command-evidence-on-drone-strikes-and-control-within-terrorist-organizations/E07D497271DCD4B8A9BFE37E8AF7 (visited on 04/01/2022).

Robins, James (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9, pp. 1393–1512. ISSN: 02700255. DOI: 10/d8gcgz. URL: https://linkinghub.elsevier.com/retrieve/pii/0270025586900886 (visited on 12/02/2021).

Robins, James M (2003). *Semantics of causal DAG models and the identification of direct and indirect effects*. Oxford Statistical Science Series. Publisher: OXFORD UNIV PRESS. Oxford University Press.

— (Jan. 1994). "Correcting for non-compliance in randomized trials using structural nested mean models". In: *Communications in Statistics - Theory and Methods* 23.8, pp. 2379–2412. ISSN: 0361-0926, 1532-415X. DOI: 10/dvs7pd. URL: http://www.tandfonline.com/doi/abs/10.1080/03610929408831393 (visited on 01/18/2022).

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5. Place: US Publisher: American Psychological Association, pp. 688–701. ISSN: 1939-2176. DOI: 10.1037/h0037350.

Vansteelandt, Stijn and Marshall Joffe (Nov. 1, 2014). "Structural Nested Models and G-estimation: The Partially Realized Promise". In: *Statistical Science* 29.4. ISSN: 0883-4237. DOI: 10.1214/14-STS493. URL: https://projecteuclid.org/journals/statistical-science/volume-29/issue-4/Structural-Nested-Models-and-G-estimation--The-Partially-Realized/10.1214/14-STS493.full (visited on 12/02/2021).

Vansteelandt, Stijn and Arvid Sjolander (Jan. 1, 2016). "Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding". In: *Epidemiologic Methods* 5.1. ISSN: 2194-9263, 2161-962X. DOI: 10/gn3dtz. URL: https://www.degruyter.com/document/doi/10.1515/em-2015-0005/html (visited on 01/10/2022).

Yee, T. W. and C. J. Wild (1996). "Vector generalized additive models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.3. tex.eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02095.x, pp. 481–493. DOI: 10.1111/j.2517-6161.1996.tb02095.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02095.x.

Zhou, Yang-Yang and Andrew Shaver (Nov. 2021). "Reexamining the Effect of Refugees on Civil Conflict: A Global Subnational Analysis". In: *American Political Science Review* 115.4. Publisher: Cambridge University Press, pp. 1175–1196. ISSN: 0003-0554, 1537-5943. DOI: 10.1017/S0003055421000502. URL: https://www.cambridge.org/core/journals/american-political-science-review/article/reexamining-the-effect-of-refugees-on-civil-conflict-a-global-subnational-analysis/9FB2BBEA2E2DC15560F367677C3D284E#article (visited on 04/01/2022).

Appendix

A Proof of Proposition 1

Proof. This proof uses induction on t of $\tau_{st}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ for any $s \geq t$. First, I need to show that the base case $\tau_{ss}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ is identified.

$$
\begin{aligned}
\tau_{ss}(\bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}) &= \mathbb{E}\left[Y_{is}(\bar{D}_{is}, \mathbf{0}) - Y_{is}(\bar{D}_{is-1}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] && \text{(Definition)} \\
&= \mathbb{E}\left[Y_{is}(\bar{D}_{is}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{is-1}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] && \text{(Linearity)} \\
&= \mathbb{E}\left[Y_{is} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{is-1}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] && \text{(Consistency (1))} \\
&= \mathbb{E}\left[Y_{is} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{is-1}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is-1}, D_{is} = 0\right] \\
& && \text{(Sequential ignorability (3))} \\
&= \mathbb{E}\left[Y_{is} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{is}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is-1}, D_{is} = 0\right] && \text{(Definition)} \\
&= \mathbb{E}\left[Y_{is} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] - \mathbb{E}\left[Y_{is} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is-1}, D_{is} = 0\right] && \text{(Consistency (1))} \\
&= \mathbb{E}\left[H_{iss} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is}\right] - \mathbb{E}\left[H_{iss} \;\middle|\; \bar{\boldsymbol{X}}_{is}, \bar{Y}_{is-1}, \bar{D}_{is-1}, D_{is} = 0\right].
\end{aligned}
$$

With $\tau_{ss}(\bar{\boldsymbol{x}}, \bar{y}, \bar{d})$ identified, I can assume that $\tau_{st}(\cdot)$ for all $t \leq s$ is identified. Note that an immediate implication of this assumption is that

$$
H_{ist} = Y_{is} - \sum_{r=t+1}^{s} \tau_{sr}(\bar{\boldsymbol{X}}_{ir}, \bar{Y}_{ir-1}, \bar{D}_{ir})
$$

is also identified. I now need to show that $\tau_{st-1}(\cdot)$ is identified.

$$
\begin{aligned}
\tau_{st-1}(\bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}) &= \mathbb{E}\left[Y_{is}(\bar{D}_{it-1}, \mathbf{0}) - Y_{is}(\bar{D}_{it-2}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] \\
& && \text{(Definition)} \\
&= \mathbb{E}\left[Y_{is}(\bar{D}_{it-1}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{it-2}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] \\
& && \text{(Linearity)} \\
&= \mathbb{E}\left[H_{ist-1} \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{it-2}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] \\
&= \mathbb{E}\left[H_{ist-1} \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{it-2}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-2}, D_{it-1} = 0\right] \\
& && \text{(Sequential ignorability (3))} \\
&= \mathbb{E}\left[H_{ist-1} \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] - \mathbb{E}\left[Y_{is}(\bar{D}_{it-1}, \mathbf{0}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-2}, D_{it-1} = 0\right] \\
& && \text{(Definition)} \\
&= \mathbb{E}\left[H_{ist-1} \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] - \mathbb{E}\left[H_{ist-1} \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-2}, D_{it-1} = 0\right] \\
&= \mathbb{E}\left[H_{ist} - \tau_{st}(\bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-1}\right] \\
&\quad - \mathbb{E}\left[H_{ist} - \tau_{st}(\bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, (\bar{D}_{it-2}, 0)) \;\middle|\; \bar{\boldsymbol{X}}_{it-1}, \bar{Y}_{it-2}, \bar{D}_{it-2}, D_{it-1} = 0\right] && \text{(Definition)}
\end{aligned}
$$

Which is identified using the inductive hypothesis that $\tau_{st}(\cdot)$ and H_{ist} are both identified. □

B Remaining Simulation Results

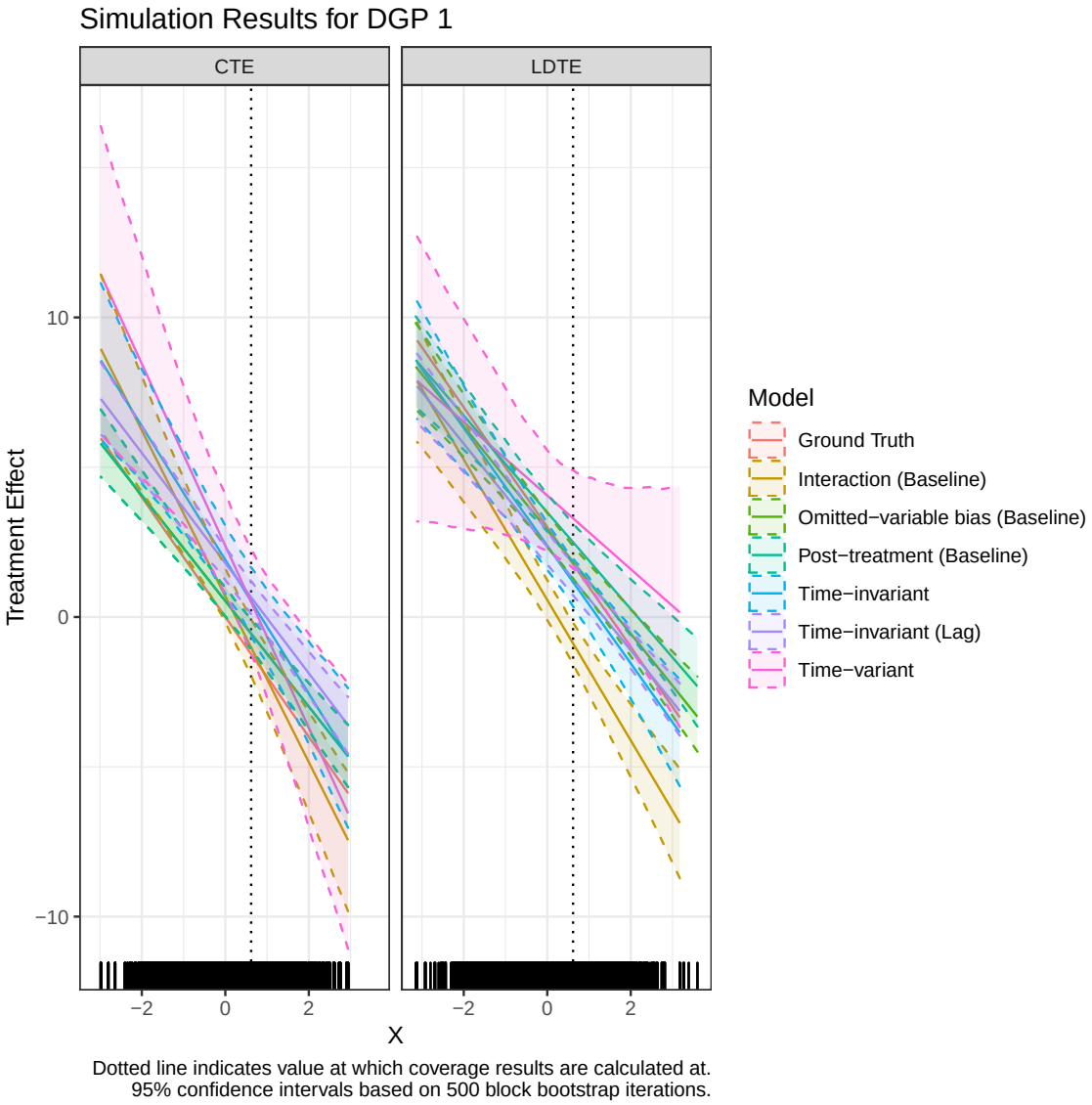

Figure B.1: Simulation results for the first DGP, based on the first Monte Carlo simulation. The ground truth represents the population treatment effect function shown in the previous section. All other models are as described previously. Confidence intervals shown are based on 500 block bootstrap iterations.

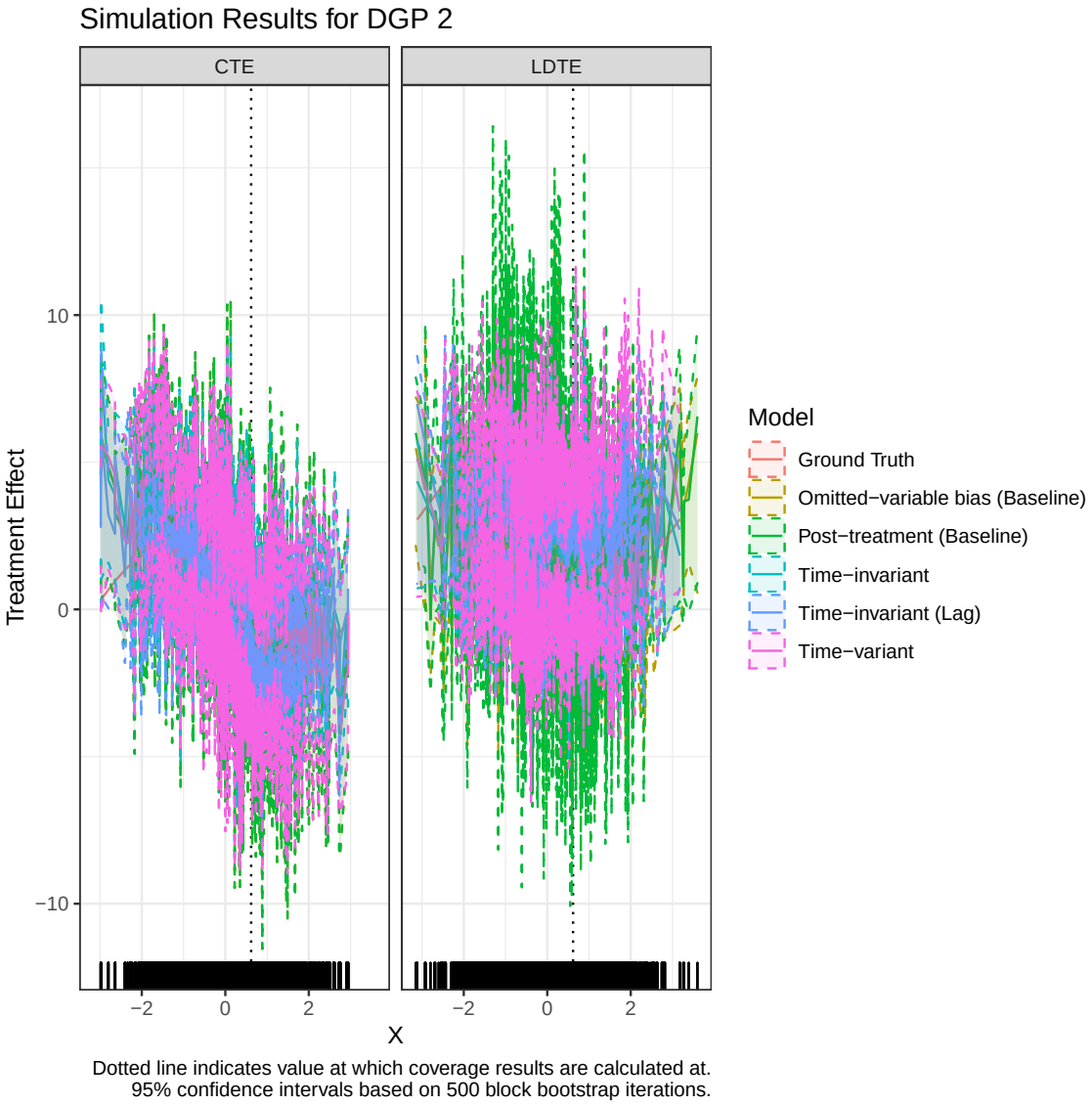

Figure B.2: Simulation results for the second DGP, based on the first Monte Carlo simulation. The ground truth represents the population treatment effect function shown in the previous section. All other models are as described previously. Confidence intervals shown are based on 500 block bootstrap iterations.

	DGP 1	DGP 2	DGP 3
	Coverage	Coverage	Coverage
Time-invariant	0.27	0.16	0.12
Time-invariant (Lag)	0.14	0.28	0.08
Time-variant	0.50	0.58	0.09
Interaction (Baseline)	0.25	0.10	0.08
Omitted-variable bias (Baseline)	0.16	0.19	0.04
Post-treatment (Baseline)	0.16	0.19	0.04

Table B.1: Global coverage results of the CTE for 200 Monte Carlo simulations for the three different DGPs. Coverage is calculated based on 500 block bootstrap iterations per Monte Carlo simulation. Coverage is an average of coverages for each Monte Carlo simulation. For DGP 1, I run a linear model and for DGP 2 and 3, I run a GAM.

	DGP 1	DGP 2	DGP 3
	Coverage	Coverage	Coverage
Time-invariant	0.22	0.42	0.17
Time-invariant (Lag)	0.16	0.23	0.11
Time-variant	0.48	0.36	0.12
Interaction (Baseline)	0.13	0.25	0.07
Omitted-variable bias (Baseline)	0.17	0.21	0.10
Post-treatment (Baseline)	0.20	0.22	0.00

Table B.2: Global coverage results of the LDTE for 200 Monte Carlo simulations for the three different DGPs. Coverage is calculated based on 500 block bootstrap iterations per Monte Carlo simulation. Coverage is an average of coverages for each Monte Carlo simulation. For DGP 1, I run a linear model and for DGP 2 and 3, I run a GAM.