The Effect of Political Advertising after *Citizen's United*: Adjusting for Unmeasured Confounding in Marginal Structural Models*

Matthew Blackwell† Soichiro Yamauchi‡

February 5, 2024

Abstract

Corporations, unions, and other interest groups have become key sponsors of television advertising in United States elections after the Supreme Court's decision in *Citizen's United v. FEC* that eliminated restrictions on such spending. This paper estimates the partisan effects of ads sponsored by these groups to obtain a more complete picture of voter behavior and electoral politics. Advertising strategies vary over the course of the campaign, and so marginal structural models are a natural tool to estimate these effects. Unfortunately, this approach requires an assumption of no unobserved confounders between the treatment and outcome, which may not be plausible with observational electoral data. To address this, we propose a novel inverse probability of treatment weighting estimator with propensity-score fixed effects to adjust for time-constant unmeasured confounding in marginal structural models of fixed-length treatment histories. We show that these estimators are consistent and asymptotically normal when the number of units and time periods grow at a similar rate. Unlike traditional fixed effect models, this approach works even when the outcome is only measured at a single point in time as in our setting, though the method does rely on some degree of treatment switching within units. Against conventional wisdom, we find that interest group ads are only effective when run by groups supporting Democratic candidates and that these effects are most prominent after Donald Trump became a presidential candidate in 2016.

*Thanks to Adam Glynn for extensive discussions and feedback. We also thank Dmitry Arkhangelsky, Gary King, and Jacob Montgomery for generous comments. Any errors remain our own. Comments welcome.

†Associate Professor, Department of Government, Harvard University. Email: mblackwell@gov.harvard.edu. Web: www.mattblackwell.org

‡PhD candidate, Department of Government, Harvard University. Email: syamauchi@g.harvard.edu. Web: https://soichiroy.github.io

1 Introduction

Television advertisements have been a cornerstone of United States politics since they first aired in the 1950s (Benoit, 2013). In the 2019–2020 election cycle, there were 2.35 million ad airings in the presidential race, 2.33 million in Senate races, and 1.36 million in U.S. House races, each number setting a record (Ridout et al., 2021). These ads are one of the main ways in which candidates, political parties, and outside interest groups attempt to influence voter behavior, electoral outcomes, and, ultimately, public policy. A large literature in political science and related fields has attempted to estimate the effect of these ads on various outcomes and for various political offices (Jacobson, 1975; Goldstein and Ridout, 2004; Huber and Arceneaux, 2007; Ridout and Franz, 2011; Blackwell, 2013; Hill et al., 2013; Sides et al., 2022). The findings from these studies vary but generally point to ad airings having persuasive effects on voter behavior with larger effects at lower levels of political office.

In this paper, we focus on the effects of advertisements sponsored by independent interest groups. The U.S. Supreme Court decision in *Citizen's United v. FEC* (2010) removed campaign finance restrictions on independent expenditures by outside interest groups, including corporations and labor unions. Since that time, the share of television ad airings sponsored by outside groups has grown from roughly 10% pre-*Citizen's United* to over 25% in the period after the decision (Ridout et al., 2021). This growth has worried citizens and political observers since these groups were no longer required to disclose their donors, a phenomenon known as "dark money," potentially providing a large benefit to corporations, wealthy individuals, and, ultimately, the Republican party. Indeed, early research into the effects of the ruling found that *Citizen's United* increased Republican vote share in state legislative races (Klumpp et al., 2016), though these studies tended to focus on aggregate effects without focusing on advertising directly. Our goal is to measure how independent group television ads have affected U.S. Senate and Gubernatorial races in the post-*Citizen's* era.

A major challenge to assessing the effects of political advertising is the dynamic nature of its deployment. Candidates and groups change the amount and content of advertising in response to how other groups advertise, which then affects the decisions of opponents. The feedback cycle of political advertising implies the potential for time-varying confounding that can bias our estimates of the effectiveness of advertising. Unfortunately, most studies of advertising ignore these issues and simply rely on aggregate campaign-level measures of advertising. One exception, Blackwell (2013), applied the combination of marginal structural models (MSM) and inverse probability of treatment weighting (IPTW) (Robins et al., 2000) to estimate the time-varying effects of negative advertising, showing considerable differences with estimates that ignore or poorly handle time-varying confounding. More generally, the use of marginal structural models for time-varying treatments in the social sciences has grown over the last few decades (Sampson et al., 2006; Sharkey and Elwert, 2011; Wodtke et al., 2011; Bacak and Kennedy, 2015; Ladam et al., 2018; Obikane et al., 2018; Creamer and Simmons, 2019; Baćak and Karim, 2019; Kurer, 2020).

One limitation of the IPTW approach to marginal structural models is that it usually relies on an assumption of sequential ignorability, which states that there are no unmeasured confounders between the treatment at time t and the outcome conditional on the treatment and covariate history up to that point. In social science studies, this assumption could be suspect when units select into treatment based on data not available to the researcher. In our setting, we might worry that groups will be more likely to advertise in certain media markets where voters are known by the campaigns to be favorable to the supported candidate. To overcome these issues, this article extends the IPTW approach to estimating the effects of time-varying treatments to allow for time-constant unmeasured confounding. To do so, we propose a straightforward modification to IPTW: to include unit-specific fixed effects in the propensity score model used to construct the inverse-probability weights. While this approach will lead to an incidental parameters problem for the propensity score model (Neyman and Scott, 1948),

we show that if this model is correctly specified and the number of time periods grows at the same rate as the number of units, the IPTW with fixed effects estimator (IPTW-FE) will lead to a consistent and asymptotically normal estimator for the parameters of the marginal structural model. This is true even when we only have a single measurement of the outcome after the final instance of treatment, as is the case in our setting. This approach relies on a within-unit version of sequential ignorability, which allows the type of feedback between the treatment and outcome usually ruled out by linear outcome fixed effects estimators (Sobel, 2012; Imai and Kim, 2019). The essential logic of the IPTW-FE is quite simple. If the propensity score model is stable over time and we have a number of time periods, we can allow for each unit to have a unique offset to the propensity score model that should incorporate any time-constant variables, measured or unmeasured.

To prove our main results, we rely on a robust literature on nonlinear panel models that has established the asymptotic distribution of our propensity score estimator when the number of time periods grows at a similar rate to the number of units (Hahn and Newey, 2004; Arellano and Hahn, 2007; Fernández-Val, 2009; Hahn and Kuersteiner, 2011; Fernández-Val and Weidner, 2016, 2018). Many of these approaches have developed bias correction techniques since these estimators are often asymptotically biased. Our approach avoids this issue with these estimators for two reasons. First, we follow the MSM literature and focus on estimating the parameters of the MSM at the slower $\sqrt{N}$ rate rather than the $\sqrt{NT}$ rate so that the asymptotic bias described in this literature converges to 0. Second, we focus on the effect of a finite number of lags of treatment, which limits how much the bias from noisy fixed effect estimation can affect the estimates of the MSM parameters.

Applying these methods to data on U.S. Senate and Gubernatorial elections from 2010 to 2020, we find that each additional week of ads from independent groups supporting Democratic candidates increases the Democratic share of the two-party vote, increases Democratic turnout, and decreases Republican turnout. We find no such effects for ads from independent

groups supporting Republican candidates, in spite of the conventional wisdom that spending from interest groups would generally favor Republicans. We additionally find that the effectiveness of pro-Democratic independent group ads is driven mostly by the post-2016 era after Donald Trump became a presidential candidate. Finally, we show that our method has null effects on a number of placebo tests, increasing our confidence that these results are not driven by unmeasured confounding.

Our methodological approach is also related to recent work on causal inference in fixed effects settings. Arkhangelsky and Imbens (2018) is most closely related to our approach here. They investigate how to use inverse probability weighting with fixed effects when a set of sufficient statistics for the treatment process is available, though in a fixed-T setting with no dynamic feedback between the treatment and the outcome and no time-varying covariates. Other work has explained how this dynamic feedback stymies estimation of both contemporaneous effects and the effects of treatment histories with fixed effects assumptions (Sobel, 2012; Imai and Kim, 2019). In contrast, our approach allows for feedback between the treatment and the outcome, so long as sequential ignorability holds conditional on the unit-specific effect. Finally, a large literature has grown recently to explain how and when difference-in-difference methods may be used to estimate the effects of time-varying treatments on outcomes when a panel of treatments and outcomes are observed together (Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021). In our application (and many others in the MSM literature), we only have a single endpoint measure of the outcome, so there are no "pre-treatment" or baseline outcomes to leverage for removing unmeasured confounding.

The paper proceeds as follows. Section 2 introduces the data and notation for our setting. In Section 3, we review marginal structural models and inverse probability of treatment weighting as they are currently deployed in applied research. We then introduce our fixed-effect approach in Section 4, describing both the assumptions that justify its use and its large-sample properties under these assumptions. In Section 5, we present simulation evidence of

the finite-sample performance of this estimator, which shows that it works well, especially when the amount of unmeasured heterogeneity is limited. Finally, we present our results in Section 6 and conclude with some ideas for future research in Section 7.

2 Data and Notation

Our data consists of Senate and Gubernatorial general election campaigns in the United States from 2010 until 2020. These are state-wide races, but we analyze the data on the level of the media market, the lowest level at which we can obtain advertising data. Media markets consist of clusters of counties that where a single group of broadcast television channels can reach. Our advertising data comes from the Wesleyan Media Project and contains political ads on all broadcast television stations in all media markets in the United States (Fowler et al., 2019). Each ad is coded for its sponsor and the nature of its content, which allows us to determine if an ad is an attack ad or not, a fact we use in constructing some of our covariates. For our outcome data on electoral returns, we used data from CQ's Voting and Elections Collection combined with data on the citizen voting-age population (CVAP) from the US Census. Finally, we obtain polling data from the website RealClearPolitics. We map these county-level outcomes to media markets using the mapping provided by Sides et al. (2022).

Our primary treatment of interest is the presence or absence of independent group (IG) ads. We define D_{it} to be a binary indicator if an IG ran ads in media market i in week t of the campaign. Independent groups are any interest or advocacy group other than the candidate or political party and include so-called "dark money" groups in addition to political action committees. Let $\overline{D}_{it} = \{D_{i1}, \ldots, D_{it}\}$ be the treatment history up to time t and $\underline{D}_{it} = \{D_{it}, \ldots, D_{iT}\}$ be the history from t to T. Let $\overline{D}_i \equiv \overline{D}_{iT}$, where these take values in $\mathcal{D}_T \in \{0, 1\}^T$. We investigate several outcomes, including the share of the two-party vote for the Democratic candidate and the share of the eligible vote won by the Democrat and Republican. The latter two outcomes

use the CVAP as a denominator, which allows us to explore the possibility that advertising mobilizes each party differently. We denote these outcomes as Y_i and define the potential outcomes $Y_i(\overline{d})$, where $\overline{d} \in \mathcal{D}_T$, which is the outcome that unit i would have if they had followed treatment history $\overline{d}$.

We also have a number of time-varying confounders, including various measures of past advertising by other groups and other candidates and polling averages on support for the Democratic candidate and percent undecided or backing third-party candidates. We denote the measure of these covariates in week t as X_{it}, and we lag these measures carefully to ensure they are causally prior to D_{it}. We define $\overline{X}_{it}$, $\underline{X}_{it}$, and $\overline{X}_i$ similarly to the treatment history.

3 A Review of Marginal Structural Models

The combination of marginal structural models and inverse probability of treatment weighting was developed by Robins (1998a) and has since become an important method across a number of scientific domains. Robins et al. (2000) provides a general introduction to the method. A robust methodological literature has built up around the method, focusing on stabilizing the construction of the weights (Cole and Hernán, 2008; Xiao et al., 2013; Imai and Ratkovic, 2015; Kallus and Santacatterina, 2019), using machine learning methods to make estimation more flexible (Muñoz and van der Laan, 2011; Gruber et al., 2015), or developing doubly robust versions of the approach (Bang and Robins, 2005; Rotnitzky et al., 2012). Our contribution to this literature is to show how these methods may be applied when a researcher suspects there may be time-constant unmeasured confounding.

The MSM methodology is based on a sequential ignorability assumption that treatment at time t is unrelated to the potential outcomes conditional on (some function of) the history of treatment and the time-varying covariates. In particular, there is some vector of time-varying

covariates, such that,

$$Y_i(\overline{d}) \perp\!\!\!\perp D_{it} \mid \overline{X}_{it}, \overline{D}_{i,t-1}, \quad \forall\, \overline{d} \in \{0,1\}^T.$$

This assumption is a time-varying version of a selection-on-observables assumption applied repeatedly to treatment in each period. One drawback of this approach in the social sciences is that units may have differing baseline probabilities of treatment based on traits that are difficult to measure. In the context of advertising, groups may target ads at media markets that have more persuadable voters by some metric unknown to the researcher. This limitation of sequential ignorability is one motivation for developing the fixed-effects approach we introduce below.

A marginal structural model is a model for the marginal mean of the potential outcomes as a function of the treatment history

$$\mathbb{E}[Y_i(\overline{d})] = g(\overline{d}; \gamma_0), \tag{3.1}$$

parameterized as a function of γ. Throughout, we use a zero subscript (γ_0, for example) to indicate the true values of parameters. The dimensionality of $\overline{d}$ grows quickly in T, so even when T is moderate, $g(\cdot)$ will usually impose some parametric restrictions on the response surface. Even if these modeling restrictions are correct, the observed conditional expectation function $\mathbb{E}[Y_i \mid \overline{D}_i = \overline{d}] \neq g(\overline{d}; \gamma_0)$ due to confounding by X_{it}. On the other hand, including the covariates in the conditional expectation will lead to post-treatment bias so that $\mathbb{E}[Y_i \mid \overline{D}_i = \overline{d}, \overline{X}_i] \neq g(\overline{d}; \gamma_0)$. Robins (1999) showed how an inverse probability of treatment weighting scheme could avoid these two biases. In particular, he showed that a weighted conditional expectation can recover the parameters of the MSM when the weights are proportional to the inverse of the conditional probability of the unit's treatment history given their covariate history. Let $\pi_t(\overline{d}_{t-1}, \overline{x}_t) = \mathbb{P}(D_{it} = 1 \mid \overline{D}_{i,t-1} = \overline{d}_{t-1}, \overline{X}_{it} = \overline{x}_t)$ and let $\pi_{it} = \pi_t(\overline{D}_{i,t-1}, \overline{X}_{it})$. Then,

the IPTW weights for our MSM become

$$W_i = \prod_{t=1}^{T} \pi_{it}^{-D_{it}} (1 - \pi_{it})^{-(1-D_{it})} \tag{3.2}$$

With these weights, Robins (1999) showed that $\mathbb{E}[\mathbf{1}\{\overline{D}_i = \overline{d}\} W_i Y_i] = g(\overline{d}; \gamma_0)$.

In observational studies, the propensity scores used to construct the weights are not usually known to the analyst and so must be estimated. The standard approach to this in the MSM literature is to specify a parametric model for treatment and estimate its parameters via maximum likelihood. Define a parametrization of the propensity score $\pi_t(\overline{x}_t, \overline{d}_t; \beta)$, where we define the true value of this parameter as $\pi_t(\overline{x}_t, \overline{d}_t; \beta_0) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it} = \overline{x}_t, \overline{D}_{it} = \overline{d}_t)$. We then define the estimated propensity scores as $\widehat{\pi}_{it} = \pi_t(\overline{X}_{it}, \overline{D}_{it}; \widehat{\beta})$, where $\widehat{\beta}$ is the MLE. These estimated propensity scores can then be used to generate estimated weights, $\widehat{W}_i = \prod_{t=1}^{T} \widehat{\pi}_{it}^{-D_{it}} (1 - \widehat{\pi}_{it})^{-(1-D_{it})}$. With these estimated weights, an IPTW estimator for the MSM can be constructed by solving the empirical version of the following estimating equation for γ,

$$\mathbb{E}\left\{\widehat{W}_i h(\overline{D}_i)(Y_i - g(\overline{D}_i; \gamma))\right\} = 0,$$

where $h(\cdot)$ is a researcher-specified $\dim(\gamma) \times 1$ vector of fixed functions of $\overline{d}$. This approach finds the value of γ that makes the MSM residuals approximately uncorrelated with $h(\overline{D}_i)$ in the weighted data, and it simplifies to standard estimation techniques in many cases. For example, when $g(\cdot)$ and $h(\cdot)$ are the identity functions, then this approach reduces to weighted least squares. Robins (1998b) established this procedure as producing a consistent and asymptotically normal estimator for the parameters of the MSM.

The weights in equation (3.2) can often be unstable when the true or estimated propensity scores are close to one or zero, which can lead to highly variable estimates. A common practice, in this case, is to include a stabilizing numerator that is the marginal probability of the

treatment history, $\overline{\pi}_{it} = HP(D_{it} = 1 \mid \overline{D}_{i,t-1})$. In this case, the stabilized weights become

$$\widetilde{W}_i = \prod_{t=1}^{T} \left(\frac{\overline{\pi}_{it}}{\pi_{it}} \right)^{D_{it}} \left(\frac{1 - \overline{\pi}_{it}}{1 - \pi_{it}} \right)^{1 - D_{it}}.$$

Another common practice is to trim the weights to additionally guard against unstable causal parameter estimates (Cole and Hernán, 2008), though other propensity score estimation techniques also help with this problem (Imai and Ratkovic, 2015).

4 Fixed-effect Propensity Score Estimators

4.1 Setting and Assumptions

We now focus on estimating propensity scores with fixed effects for MSMs when time-constant unmeasured confounding exists. As with the traditional MSM case, we assume that $(Y_i, \overline{D}_i, \overline{X}_i)$ are independent across observations. In order to adjust for unit-specific heterogeneity, we do require restrictions beyond the typical MSM case. First and foremost, we focus on marginal structural models for a treatment history of a fixed length rather than the entire treatment history, which we call *truncated* MSMs. In particular, truncated MSMs focus on modeling only the last k periods of treatment, $\mathbb{E}[Y_i(\underline{d}_{T-k})] = g(\underline{d}_{T-k}; \gamma)$, where $\underline{d}_{T-k} = (d_{T-k}, \ldots, d_T)$, k is fixed, and the parameter vector γ is of length J. Truncation is a restriction on what quantities of interest can be consistently estimated in this setting, not a substantive assumption about the effect of the treatment before the truncation point. By the usual consistency assumption, we can define these "shorter" potential outcomes as $Y_i(\underline{d}_{T-k}) \equiv Y_i(\overline{D}_{i,T-k-1}, \underline{d}_{T-k})$, so that treatment history before k lags acts more like a baseline confounder. In particular, our use of a truncated MSM does not invoke a "no carryover" assumption as in Imai and Kim (2019). Compared to typical MSM practice, the main limitation of this restriction is to rule out functional forms where the cumulative sum of the entire treatment history is included as part of the MSM. Intuitively, this

restriction implies that analysts cannot simultaneously use long treatment histories to estimate long-term effects and adjust for unmeasured confounding.

We now describe the key identification assumption of the IPTW-FE approach, which combines the concept of a unit-specific randomized experiments with the standard MSM framework in Section 3. Let $\underline{X}_{i,t+1}(\overline{d}) = (X_{i,t+1}(\overline{d}_t), X_{i,t}(\overline{d}_{t+2}), \ldots, X_{i,T}(\overline{d}_{T-1}))$ represent the potential outcomes of the future covariates under a particular treatment history, where we truncate the full treatment history $\overline{d} = (d_1, \ldots, d_T)$ for each time period, $\overline{d}_k = (d_1, \ldots, d_k)$, since future treatments values cannot affect past covariates.

Assumption 1 (Unit-specific Sequential Ignorability)**.** *Let α_i be an unmeasured, time-constant random variable. For all i, t and $\overline{d}$,*

$$\{Y_i(\overline{d}), \underline{X}_{i,t+1}(\overline{d})\} \perp\!\!\!\perp D_{it} \mid \overline{X}_{it}, \overline{D}_{i,t-1} = \overline{d}_{t-1}, \alpha_i.$$

Assumption 1 states that conditional on the unit-specific effect, the treatment history, and (a function of) the covariate history, treatment is independent of future potential outcomes for both the outcome and the covariate process. In essence, treatment is randomized with respect to future covariates and the outcome, conditional on the past and time-constant features of the unit. This assumption allows for both time-varying confounding by measured covariates and time-constant confounding by measured and unmeasured covariates. We do assume that the time-constant unmeasured confounding can be captured by the unidimensional, α_i, which might represent a combination of several unit-specific factors.

Assumption 1 involves potential outcomes of the entire treatment history, $Y_i(\overline{d})$, but above, we defined our main marginal structural models in terms of truncated treatment histories, $\mathbb{E}[Y_i(\underline{d}_{T-k})]$. Thus, the requirements of sequential ignorability go beyond the treatments of interest in the marginal structural model and apply to the potential outcomes for the entire treatment history. This allows for the fixed-effect propensity score estimators to be consistent

even without a no-carryover assumption that would assume that treatment before $T-k$ has no effect on the outcome.

Under Assumption 1, we can nonparametrically identify the mean of the potential outcomes under a given history with unit-specific propensity scores. Let $\pi_{it}(\overline{x}_t, \overline{d}_{t-1}, \alpha_i) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it} = \overline{x}_t, \overline{D}_{i,t-1} = \overline{d}_{t-1}, \alpha_i)$ and let $\pi_{it} = \pi_{it}(\overline{X}_{it}, \overline{D}_{i,t-1}, \alpha_i)$. Then, we can use the usual techniques to arrive at the nonparametric identification of

$$\mathbb{E}[Y_i(\underline{d}_{T-k})] = \mathbb{E}\left[\frac{\mathbf{1}(\underline{D}_{i,T-k} = \underline{d}_{T-k})Y_i}{\prod_{t=T-k}^{T} \pi_{it}^{d_t}(1-\pi_{it})^{1-d_t}}\right],$$

where d_t denotes the corresponding entry in $\underline{d}_{T-k}$. Thus, under Assumption 1 (and a positivity assumption), treatment history effects are nonparametrically identified since we can write them as functions of quantities that are in principle observable as $N, T \to \infty$.

As is common with nonparametric identification, however, the sampling details across units and time will play an important role in actually obtaining valid estimates of these causal effects. We can see this even in static causal inference settings. If units in that setting are not independent across units, for example, standard IPW approaches might not be estimable at standard rates without further assumptions. While we assume i.i.d. data across units, this assumption would be unrealistic for the time dimensions. We now lay out the sampling assumptions for our setting.

Assumption 2 (Sampling Assumptions)**.** *Let $\nu > 0$, $\mu > 4(8+\nu)/\nu$, and $\mathcal{B}_0(\epsilon)$ is an ϵ-neighborhood of (β_0, α_{i0}) for all i, t, N, T.*

(i) (Asymptotics) Let $N, T \to \infty$ such that $N/T \to \rho$ where $0 < \rho < \infty$.

(ii) (Across/Within-Unit Dependence) For all N and T, $\{(Y_i(\bullet), \overline{D}_i, \overline{X}_i, \alpha_i) : i = 1, \ldots, N\}$ are i.i.d. across i, where $Y_i(\bullet) = \{Y_i(\overline{d}); \overline{d} \in \{0,1\}^T\}$. Letting $Z_{it} = (D_{it}, X_{it})$ for $t = 1, \ldots, T$ and $Z_{i,T+1} = (Y_i(\overline{d}))$, then for each i, $\{Z_{it} : t = 1, \ldots, T+1\}$ is α-mixing conditional on α_i

with mixing coefficients satisfying $\sup_i a_i(m) = O(m^{-\mu})$ *as* $m \to \infty$ *where*

$$a_i(m) \equiv \sup_t \sup_{A \in \mathcal{A}_{it}, B \in \mathcal{B}_{i,t+m}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

and $\mathcal{A}_{it}$ *is the sigma field generated by* $(Z_{it}, Z_{i,t-1}, \ldots)$ *and* $\mathcal{B}_{i,t}$ *is the sigma field generated by* $(Z_{it}, Z_{i,t+1}, \ldots)$.

Assumption 2(i) establishes the large-N, large-T asymptotic framework, which has been widely used for nonlinear panel models in econometrics (Hahn and Newey, 2004; Arellano and Hahn, 2007; Fernández-Val, 2009; Hahn and Kuersteiner, 2011; Fernández-Val and Weidner, 2016, 2018). The strong mixing process in Assumption 2(ii) allows us to rely on the laws of large numbers and the central limit theorem in the time dimension. It essentially states that dependence over time is sufficiently weak that as the distance between two periods increases, information in the two periods becomes approximated uncorrelated. That is, data over time within a unit may be dependent, but there is new information as time goes on. This assumption is substantially weaker than independence over time or even stationarity. In particular, it allows for time trends, which are a common feature of propensity score models in MSMs. The i.i.d. nature of the distribution of the data and the fixed effects across units is common to IPTW approaches and allows us to take averages over the unit-specific heterogeneity and has been used before for average partial effects in nonlinear panel models (Fernández-Val and Weidner, 2016). It is possible to replace this assumption with stationarity of X_{it} over time, but this would rule out lagged treatment in the propensity score model along with time trends.

To determine the asymptotic properties of our approach, we assume researchers will specify a correct parametric model for the propensity score (up to the unmeasured heterogeneity) as $\pi_{it}(\overline{x}_t, \overline{d}_{t-1}; \beta, \alpha_i) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it} = \overline{x}_t, \overline{D}_{i,t-1} = \overline{d}_{t-1}; \beta, \alpha_i)$, where β is a $k \times 1$ parameter vector, α_i is the time-constant unmeasured confounder, and $\pi_{it}(\beta, \alpha_i) = \pi_{it}(\overline{X}_{it}, \overline{D}_{i,t-1}; \beta, \alpha_i)$.

We write the log-likelihood of this model as

$$\ell_{it}(\beta, \alpha) = D_{it} \log \pi_{it}(\beta, \alpha) + (1 - D_{it}) \log \{1 - \pi_{it}(\beta, \alpha)\}$$

Let $\alpha_0 = (\alpha_{10}, \ldots, \alpha_{N0})$ and β_0 be the values of the parameters that generate the treatment process. In particular, we assume that these values are the solution to the following population conditional maximum likelihood condition

$$(\beta_0, \alpha_0) = \underset{(\beta,\alpha)\in\mathbb{R}^{d_\beta+N}}{\arg\max} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}[\ell_{it}(\beta, \alpha) \mid \alpha_i], \tag{4.1}$$

where the expectation is with respect to the distribution of the data conditional on the unobserved effect (see, for example, Fernández-Val and Weidner, 2016, equation 2.1). Our approach requires a correctly specified parametric model for the covariates in the propensity score (which is common in the MSM literature) but is semiparametric in that we make no assumptions about the relationship between the unmeasured heterogeneity and the covariates. We assume a fixed-length parameter vector, but it may be possible to allow this vector to grow with N and T and thus allow for more flexible estimation strategies. As this is beyond the scope of the current paper, we leave it to future research.

With this propensity score model in hand, we can construct weights that can adjust for both observed time-varying confounding and unobserved time-constant confounding. In particular, we use the following weights

$$W_i(\beta, \alpha_i) = \prod_{j=T-K}^{T} \left(\frac{1}{\pi_{ij}(\beta, \alpha_i)}\right)^{D_{ij}} \left(\frac{1}{1 - \pi_{ij}(\beta, \alpha_i)}\right)^{1-D_{ij}},$$

where we only take the product over the last k time periods because our quantities of interest focus on those periods. As with the standard MSM case, we can replace the numerator with the marginal probability of the treatment history, $\overline{\pi}_{it}$, which can stabilize the variance of the

estimator without affecting identification.

The IPTW approach to estimating this MSM is to rely on the estimating equation

$$0 = \frac{1}{N}\sum_{i=1}^{N} U_i(\gamma, \beta, \alpha_i) = \frac{1}{N}\sum_{i=1}^{N} \left\{ W_i(\beta, \alpha_i) h(\underline{D}_{i,T-k})(Y_i - g(\underline{D}_{i,T-k}; \gamma)) \right\},$$

where $h(\cdot)$ is a function with J-length output chosen by the researcher as in the standard MSM case. For example, if Y_i is continuous and g is linear and additive, it is common to use $h(\underline{D}_{i,T-k}) = \underline{D}'_{i,T-k}$. Under the fixed-effects sequential ignorability assumption and the MSM, we have $\mathbb{E}[U_i(\gamma_0, \beta_0, \alpha_{i0})] = 0$, which is a semiparametric identification result because the restriction identifies the causal parameters, γ_0, solely in terms of sample quantities (up to the propensity score parameters). This result follows the standard g-computation algorithm with the unit-specific heterogeneity, α_i, included in the place of a baseline covariate (Robins, 1999, 2000). We make the following regularity conditions on the marginal structural model and outcome.

4.2 Proposed Method

We propose a two-step approach to estimating the parameters of the marginal structural model using inverse probability of treatment weighting. These two steps are:

1. Obtain estimates of the parameters of the propensity score model, $(\widehat{\beta}, \widehat{\alpha}_i)$, using conditional maximum likelihood treating the unit-specific effects α_i as fixed parameters to be estimated. Construct estimated weight $W_i(\underline{D}_{i,T-k}; \widehat{\beta}, \widehat{\alpha}_i)$.

2. Pass the estimated weights to a weighted estimating equation $N^{-1}\sum_{i=1}^{N} U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i) = 0$ to obtain estimates of the MSM parameters, γ.

The first step in this procedure can be implemented with a sample conditional maximum

likelihood estimator. Letting $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N)$, we have

$$(\widehat{\beta}, \widehat{\alpha}) = \underset{(\beta,\alpha)\in\mathbb{R}^{d_\beta+N}}{\arg\max} \sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{it}(\beta, \alpha_i) \tag{4.2}$$

Under these assumptions, we use the following maximum likelihood estimators:

$$\widehat{\beta} = \arg\max_{\beta} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{it}(\beta, \widehat{\alpha}_i(\beta)), \qquad \widehat{\alpha}_i(\beta) = \arg\max_{\alpha} \frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\beta, \alpha).$$

These maximum likelihood estimates are subject to the usual incidental parameters problem that results in bias that shrinks as $T \to \infty$. Even when N and T grow at the same rate, Hahn and Newey (2004) showed that these types of MLE estimators are not $\sqrt{NT}$-consistent, and a large literature has developed proposing several bias correction techniques (Arellano and Hahn, 2007; Fernández-Val and Weidner, 2018). We sidestep these issues in our results because we target the slower convergence rate of $\sqrt{N}$ because we only have a single outcome per unit, which is common in the MSM literature.

To obtain estimates of the MSM parameters, $\widehat{\gamma}$, we use the sample version of the MSM moment condition, $N^{-1}\sum_{i=1}^{N} U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i) = 0$. This estimator depends on the link function for the marginal structural model and a function $h(\cdot)$. One particularly straightforward estimator in this class is weighted least squares for the identity link with continuous outcomes. Often, $h(\cdot)$ can be chosen to enhance the efficiency of the estimator (Robins, 1999), but we do not explore that here. We now show in Theorem 1 that under regularity conditions and the above assumptions, this estimator is consistent and asymptotically normal. The proof and precise statements of the regularity conditions are in Appendix A. Let $G = \mathbb{E}\{\partial U_i(\gamma, \beta, \alpha)/\partial\gamma\}_{\gamma=\gamma_0}$, and $U_i = U_i(\gamma_0, \beta_0, \alpha_{i0})$.

Theorem 1. *Under Assumptions 1, 2, and suitable regularity conditions,* $\widehat{\gamma} \xrightarrow{p} \gamma_0$ *and*

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} N(0, V_{\gamma_0}), \tag{4.3}$$

where $V_{\gamma_0} = G^{-1}\mathbb{E}[U_i U_i^\top]G^{-1}$.

We can build a consistent variance estimator in the usual way with $\widehat{V}_\gamma = \widehat{G}^{-1}\widehat{\Omega}\widehat{G}^{-1}$, where

$$\widehat{G} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \widehat{U}_i}{\partial \gamma}, \qquad \widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} \widehat{U}_i \widehat{U}_i^\top, \qquad \widehat{U}_i = U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i).$$

This is a standard sandwich estimator for estimators based on estimating equations.

Theorem 1 establishes that the IPTW-FE for MSMs is asymptotically normal and that we can asymptotically ignore the estimation of the weights. In the standard IPTW case, the estimation of the weights does impact the distribution of the MSM estimates. Here, however, the estimation of the weights doesn't affect the first-order asymptotic distribution because we are using NT observations to estimate the propensity score parameters but only using a fraction of the observations, Nk, to create the weights, where k is fixed as $T \to \infty$. Thus, the $\widehat{\beta}$ converges much faster than $\widehat{\gamma}$ and so we can ignore its estimation error. Of course, this is an approximation that might be less accurate when T is small, so a bootstrap of units might yield more accurate variance estimates in that case.

In typical nonlinear panel models, plugging in noisy estimates of the fixed-effect parameters leads to a bias that converges to 0 slowly enough to create asymptotic bias. In our setting, however, the strong mixing property of the treatment process ensures that this bias fades over time, and so allows us to ignore the estimation of the fixed-effect parameters as well. In the literature on nonlinear panel models, there is a similar result for estimating partial effects or differences in the conditional expectation, as opposed to parameters of the nonlinear model. For example, Fernández-Val and Weidner (2018) showed how these average partial effects can converge at a

slower rate with parameter estimation not having a first-order effect on the asymptotic distribution (see also Fernández-Val and Weidner, 2016). The current approach is similar since we are only interested in the parameters of the weighting model insofar as they provide consistent estimates of the IPTW weights.

This result establishes that it is possible to adjust for unmeasured baseline confounding in MSMs when the time dimension is long and provides sufficiently new information within units. The quality of this adjustment will depend on both how long the panels are and how severe the unmeasured heterogeneity is. A second-order expansion of the estimator shows that second-order bias (which can be ignored in our asymptotic analysis) is inversely related to the propensity scores. Thus, strong unit-specific heterogeneity will push propensity scores close to zero or one and create more finite-sample bias. Longer panels help with this finite-sample bias since these second-order terms will be of order $O_P(1/\sqrt{T})$. A fruitful avenue for future research would be to use analytic or computational approaches like the jackknife to adjust for these second-order terms as in Hahn and Newey (2004).

What about doubly robust estimation? In traditional MSM settings, it is possible to develop doubly robust estimators that depend both on the correct modeling of the propensity scores and a series of outcome regression models (Bang and Robins, 2005). In our setting, however, this would require an outcome regression model that had unobserved heterogeneity, and without multiple observations of the outcome over time, it is not possible to estimate such a model without overly strong assumptions.

4.3 Trimming Weights

One drawback of the IPTW-FE approach is that the fixed-effect parameters of the propensity score model are not identified when units are either always treated or always control. Even when we maintain the population-level positivity assumption, this in-sample positivity viola-

tion means that some units will have undefined weights. We propose three ways to address this issue. First, one could simply omit the no-treatment-variance units and estimate the parameters of the MSM for the units that have at least one treated or control period. This is the simplest procedure but could induce confounding bias, especially if the α_i has a nonlinear relationship with the outcome. Second, we could use an ad hoc rule for imputing propensity scores of the no-treatment-variance units. For example, we could set these units to have $\widehat{\pi}_{it} = 0.01$ if $D_{it} = 0$ for all t and $\widehat{\pi}_{it} = 0.99$ if $D_{it} = 1$ for all t. Depending on the lag length k in the MSM and the exact trimming, this may lead to extreme weights, which themselves could require trimming. Alternatively, one could place bounds on the range of the unit-specific effects in the MLE estimation to $\alpha_i \in [a_0, a_1]$ and set the estimates of those effects as $\widehat{\alpha}_i = a_0$ or $\widehat{\alpha}_i = a_1$ if $D_{it} = 0$ or $D_{it} = 1$ for all t, respectively. The amount of trimming of the weights in this approach amounts to a bias-variance trade-off similar to weight trimming in standard IPTW estimators for MSMs (Cole and Hernán, 2008).

Finally, one alternative approach to handling positivity violations would be to focus on a different quantity of interest. Kennedy (2019) proposed estimating the effect of incremental propensity score interventions, which are interventions that shift the propensity score rather than set treatment histories to specific values. The identification and estimation of these effects do not depend on positivity, and under the assumption of a correctly specified propensity score model, a simple inverse probability weighting estimator is available (Kennedy, 2019, p. 650).

5 Simulation Evidence

In this section, we conduct simulation studies to evaluate the finite sample performance of the proposed approach.

5.1 Setup

We simulate a balanced panel of n units with T time points where the number of units varies $n \in \{200, 500, 1000, 3000\}$. We fix the ratio of the number of units to the number of time periods $n/T = \rho \in \{5, 10, 50\}$. This setup mimics the key asymptotic approach of our theoretical results, and the larger value of ρ implies the small number of time points, $T = n/\rho$. The treatment sequence is generated as a function of individual unobserved effect α_i, the past treatment $D_{i,t-1}$ and the time-varying covariates, X_{it}.

$$D_{it} \sim \text{Bernoulli}(\text{expit}(\alpha_i + \varphi D_{i,t-1} + \beta^\top X_{it}))$$

where $\text{expit}(x) = 1/(1+\exp(-x))$ is the inverse logistic function. The individual heterogeneity is drawn from a uniform distribution with support on $[-a, a]$ for $a \in \{1, 2\}$. The value of a is chosen such that the variance of individual heterogeneity explains $1/3$ ($a = 1$) or $2/3$ ($a = 2$) of the variance of the linear predictor. The time-varying covariates X_{it} are generated exogenous to the treatment, drawn from the multivariate normal distribution, $X_{it} \sim \mathcal{N}(-1/2\mathbf{1}, \Sigma)$ where $\Sigma_{jj} = 1$ and $\Sigma_{jj'} = 0.2$ for $j \neq j'$. Finally, we set $\varphi = 0.3$ and $\beta = (-0.5, -0.5)$ when the number of covariates is two or $\beta = (-0.5, -0.5, 1.0, -0.5)$ when the number of covariates is four.

The outcome is generated by the linear model with individual unobserved variable α_i, the final treatment D_{iT}, the cumulative lagged treatments $\sum_{t=T-1}^{T-3} D_{it}$ and the average of the time-varying covariates, $\overline{X}_i = \sum_{t=1}^{T} X_{it}/T$, all of which are generated in the previous step.

$$Y_i = \alpha_i + \tau_F D_{iT} + \tau_C \sum_{t=T-1}^{T-3} D_{it} + \gamma^\top \overline{X}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

where we set $\tau_F = 1$, $\tau_C = 0.3$, and $\gamma = (1.0, 0.5)$ or $\gamma = (1.0, 0.5, 1.0, 1.0)$ depending on the number of covariates used in each simulation.

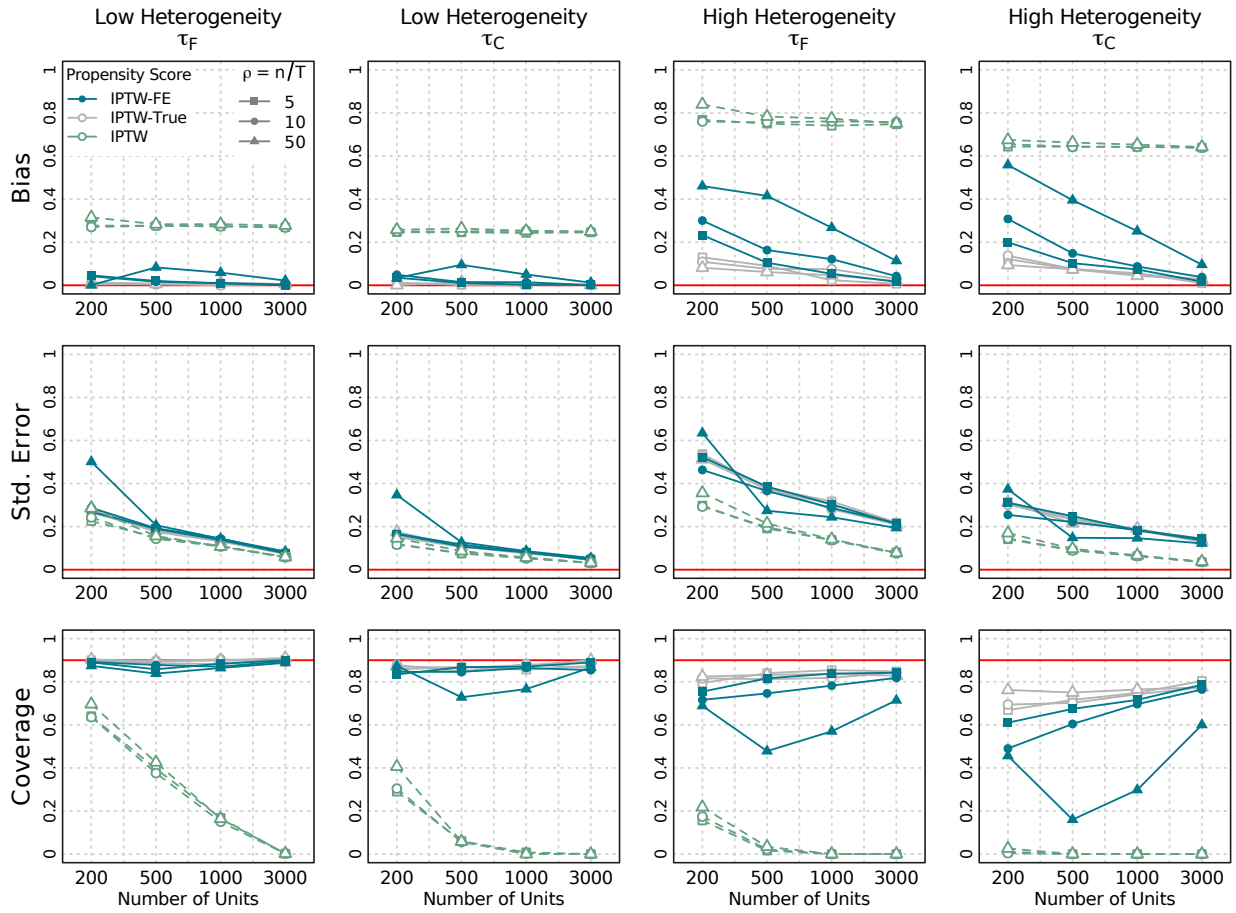

Figure 1: Bias, standard error (Std. Error) and coverage probability of 90% confidence intervals (Coverage) for the estimation of the final period effect τ_F and the cumulative effect τ_C under the "low" heterogeneity ($a = 1$) – first two columns – and the "high" heterogeneity ($a = 2$) – last two columns – scenario. Solid lines in blue show the proposed estimator (`IPTW-FE`), solid lines in grey show the estimator based on the true propensity score (`IPTW-True`), and dashed lines in green show the estimator based on the estimated propensity score without fixed effects (`IPTW`). Shapes correspond to the n to T ratio ρ such that squares represent $\rho = 5$ (the largest number of time periods), circles represent $\rho = 10$, and triangles represent $\rho = 50$ (the smallest number of time periods)

5.2 Results

We compared the performance of the proposed method in terms of estimating two causal quantities: the final period effect τ_F and the cumulative lagged effect τ_C. We estimate two quantities

together in the framework of weighted least squares,

$$(\widehat{\tau}_F, \widehat{\tau}_C) = \underset{\tau_F, \tau_C}{\arg\min} \sum_{i=1}^{n} \widehat{W}_i \left\{ Y_i - \alpha - \tau_F D_{iT} - \tau_C \sum_{t=T-1}^{T-3} D_{it} \right\}^2$$

where $\widehat{W}_i$ is constructed as described in the previous section. We focus on this correctly specified MSM to isolate the effects of unmeasured heterogeneity on estimator performance. The variance of $\widehat{\tau}_F$ and $\widehat{\tau}_C$ is estimated using the standard sandwich formula with the HC2 option, which is an adjustment to improve finite-sample properties of the variance estimator (MacKinnon and White, 1985).

In addition to the fixed effect approach, we consider two other strategies to obtain the weights $\widehat{W}_i$ as benchmarks to the proposed method. First, we use the true propensity score to construct the weights. Second, the estimated propensity score without the fixed effect is used to construct weights. We expect that the weights with known propensity scores is least biased and the weights without the fixed effect is most biased.

Figure 1 shows the results for the two-covariate case. Bias (first row), standard errors (second row) and coverages (third row) are computed based on 500 Monte Carlo simulations. Additional simulation results are presented in Supplemental Materials C. The first two columns correspond to the "low" heterogeneity case where the support of the fixed effect is $[-1, 1]$, whereas the last two columns correspond to the "high" heterogeneity scenario where the support of α_i is set to $[-2, 2]$. Solid lines in blue show the proposed estimator (`IPTW-FE`), solid lines in grey show the estimator based on the true propensity score (`IPTW-True`), and dashed lines in green show the estimator based on the estimated propensity score without fixed effects (`IPTW`). Shapes correspond to the n to T ratio ρ such that squares represent $\rho = 5$ (the largest number of time periods), circles represent $\rho = 10$, and triangles represent $\rho = 50$ (the smallest number of time periods).

We can see that under the low heterogeneity setting, where the unobserved individual het-

erogeneity explains roughly 1/3 of the variance of the treatment assignment, the bias of the proposed estimator (`IPTW-FE`) is indistinguishable from the estimator that is based on the true propensity score (`IPTW-True`) and the confidence interval estimates maintain the nominal coverage across different values of n and ρ. Under this scenario, even a case of $n = 200$ and $T = 4$, the proposed method performs well.

When the variance of the individual heterogeneity is high ($a = 2$) such that it explains roughly 2/3 of the variance of the treatment assignments, the proposed estimator shows relatively larger bias compared with `IPTW-True`, while bias of the estimator without fixed effects (`IPTW`) is substantially larger. Under this setting, the coverage results are mainly driven by the bias, thus the figure shows that as n increases, the coverage results also improves thanks to the reduction in bias. We can also see that in general the estimator without fixed effects shows smaller standard errors than `IPTW-FE`. This implies that the proposed method (`IPTW-FE`) trade-off the efficiency with lower bias. Finally, we highlight that small Monte Carlo bias is observed even for `IPTW-True` under this scenario. This is possibly due to the high variability of the weights, which are produce of inverse probabilities over four time periods with stabilization.

Overall, these results point to two key tensions in controlling for time-constant unmeasured heterogeneity through fixed effects in the propensity score models. First, high degrees of unmeasured heterogeneity in the propensity scores may lead to near violations of the positivity assumption that could lead to the kind of instability we see when $a = 2$. Second, larger magnitudes of heterogeneity may require more time periods to achieve good finite sample performance compared to when the heterogeneity is relatively small.

6 Results

6.1 Specification and balance

We now apply these techniques to estimate the effectiveness of independent group advertising in U.S. Senate and Gubernatorial elections from 2010 until 2020. We build on Blackwell (2013), who investigated the effects of negative advertising using an MSM approach without fixed effects for elections over the period from 2000 to 2008. Our primary results focus on three outcomes: the Democratic percentage of the two-party vote, percent of the voting-eligible population casting Democratic votes (which we call "Democratic Turnout"), and percentage of the voting-eligible population casting Republican votes (which we call "Republican Turnout").

To calculate the propensity scores, we organize the data into a market-week panel, where an example of a market-race would be the 2010 California Gubernatorial election in the Santa Barbara media market (as distinct from the media markets of San Diego, Fresno, and so on). We focus on the time period between the primary election for the race and the general election so that we have campaign lengths ranging from 8 to 40 weeks with a median of 20 weeks. After dropping market-races that have no variation in the treatment, we have $N = 467$ market-races for Democratic IG ads and $N = 623$ market-races for Republican IG ads. Most of the dropped races are very uncompetitive races or media markets with smaller audiences. In Supplemental Materials D, we investigate an alternative approach to handling no-treatment-variation market-races that uses extreme values of the unit fixed effects to obtain weights. We find results from this approach are very similar to our own below.

Table 1 shows the distribution of our aggregated treatment variable across different election years. Even with only a handful of time periods and a high level of aggregation, we can see that empirical positivity violations in the final weeks of the race are fairly common, which is the main motivation for using a marginal structural model. Note that even if a market-race had zero weeks of ads in the final five weeks of the campaign, they may have had IG ads earlier in

Table 1: Number of weeks with Democratic IG ads in the last 5 weeks of the campaign

# of ads weeks	2010	2012	2014	2016	2018	2020	All
0	12	14	25	6	5	6	68
1	10	9	13	5	16	3	56
2	19	8	17	8	14	15	81
3	14	4	6	7	4	8	43
4	13	15	7	4	6	6	51
5	11	34	28	21	50	24	168

the campaign, allowing us to estimate propensity scores for these units.

Our marginal structural model is

$$\mathbb{E}[Y_i(\overline{d}) \mid R_i] = \gamma_{R_i} + \gamma_1 \left(\sum_{k=0}^{4} d_{T-k} \right),$$

where the time index here is weeks of the campaign and R_i is the electoral race associated with market-race i. This MSM allows for race-specific intercepts, which helps to purge any remaining race-specific confounding from our estimates. The main quantity of interest, γ_1, can be interpreted as the effect of an additional week of IG advertising in the last five weeks on the outcomes, conditional on the state-wide race. For the outcome MSM, we restrict our attention to races with multiple markets to accommodate the race-specific intercepts.

We apply several different estimation approaches to this MSM: the proposed IPTW-FE approach, a standard IPTW approach without fixed effects, and a naive approach that ignores time-varying covariates altogether. For the weighting model, we included various time-varying covariates: average Democratic share of the two-party preferences in polls in the previous week (and the square of this term), the average percentage reporting undecided or voting for third-party candidates in the previous week, measures of Republican negativity over the last six weeks, the cumulative number of ads shown by the Democrat and Republican (and their squared terms). For the fixed effects approach, we additionally include a market-race fixed effect term in the specification. For the IPTW approach, we only include fixed effects at the race

level.

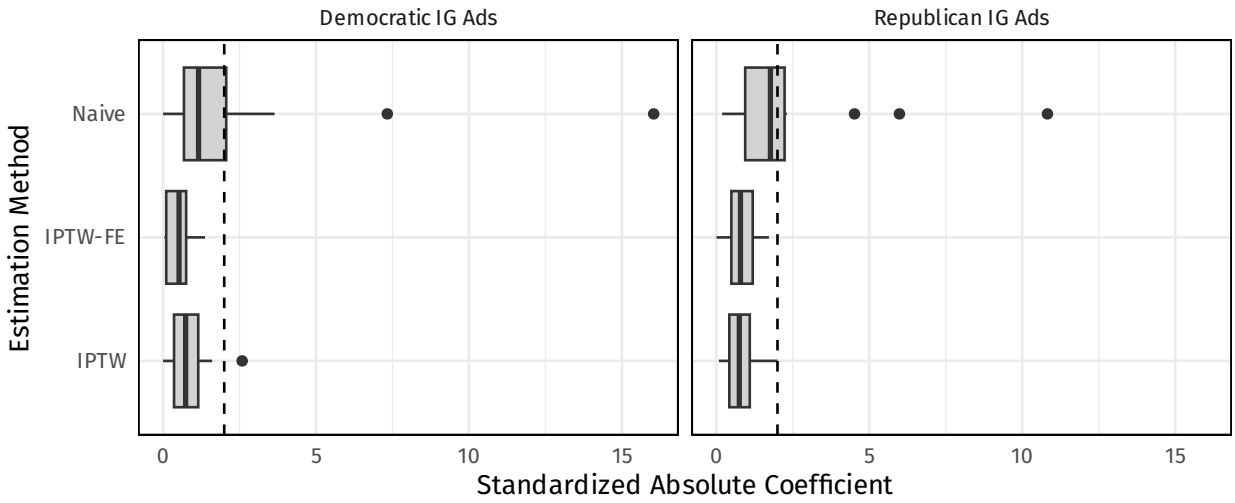

Figure 2: Balance of baseline covariates under different weighting approaches.

Assessing balance with the IPTW-FE approach is difficult because we care about the balance with respect to both measured and unmeasured confounders. Of course, we cannot assess balance with respect to unmeasured confounders. We can, however, investigate how well IPTW-FE balances the measured time-varying covariates. To do so, we regress each of these covariates on the treatment indicator, the lagged cumulative sum of treatment, and a race-specific intercept (all variables included in the MSM and the numerator of our weighting models) in the five-week period of our MSM. Figure 2 shows the distribution of standardized partial correlations of the treatment indicator and the various covariates under different weighting schemes (no weighting, IPTW, and IPTW-FE). Both IPTW and IPTW-FE vastly reduce the conditional imbalance on these covariates relative to the naive approach. In the unweighted approach, there are a few extremely unbalanced time-varying confounders.

6.2 Main results

Figure 3 shows the results of these methods for each of the outcomes. Substantively, the methods generally agree that there is a positive effect of Democratic IG ads on Democratic electoral

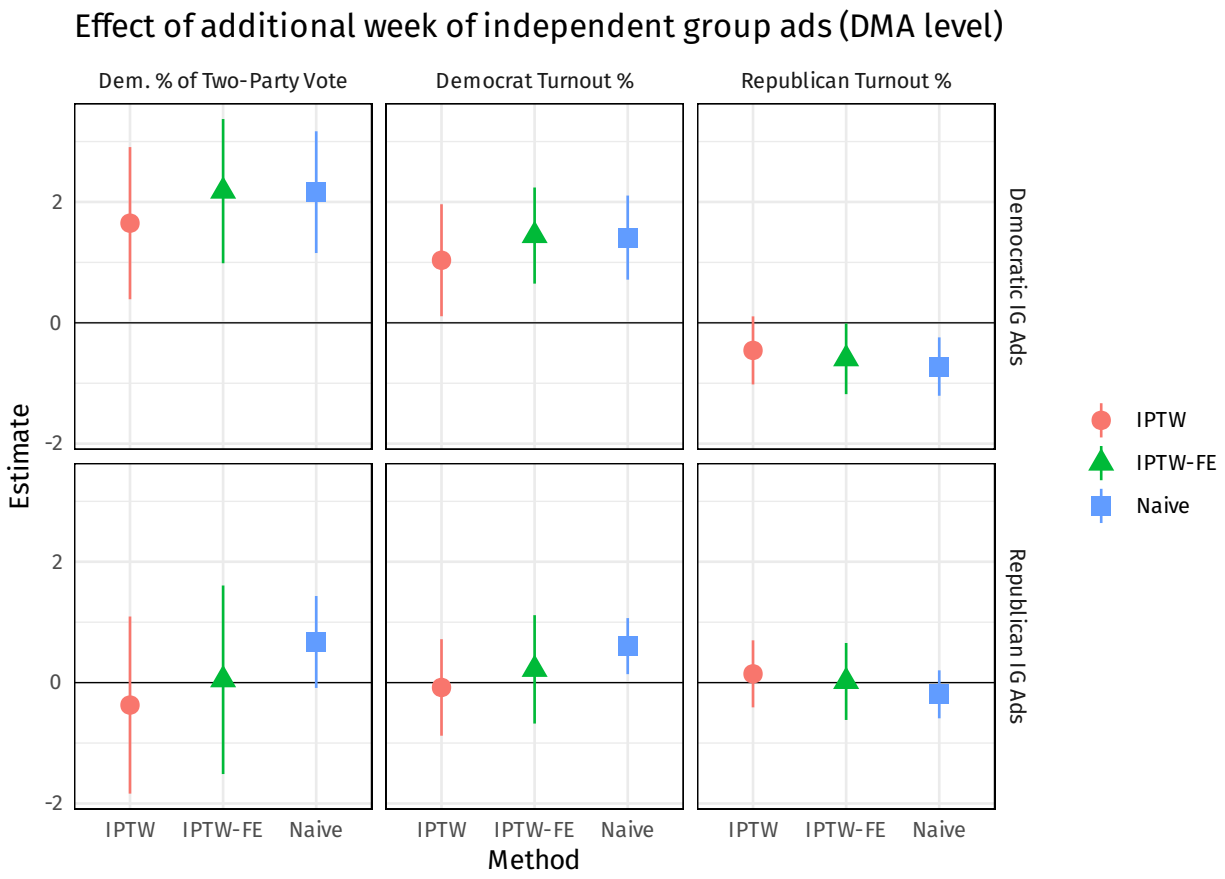

Figure 3: Estimated effects of the number of weeks of independent group advertising in the last five weeks of the campaign with different methods.

performance. This effect is driven by a positive effect on Democratic turnout and a weaker negative effect on Republican turnout. Thus, it appears that Democratic independent group ads mobilize Democratic voters and perhaps demobilize Republican voters. Republican IG ads, on the other hand, have no estimated effect on any of these measures, indicating that these ads are not very effective. The different methods here generally agree on the direction and significance of the effects, though IPTW-FE estimates a larger effect for Democratic groups than the basic IPTW approach.

The effectiveness of Democratic independent group ads runs counter to the conventional wisdom about what party would benefit the most from the *Citizen's United* decision. To understand what drives this effect, we estimated differential effects by election era. Specifically,

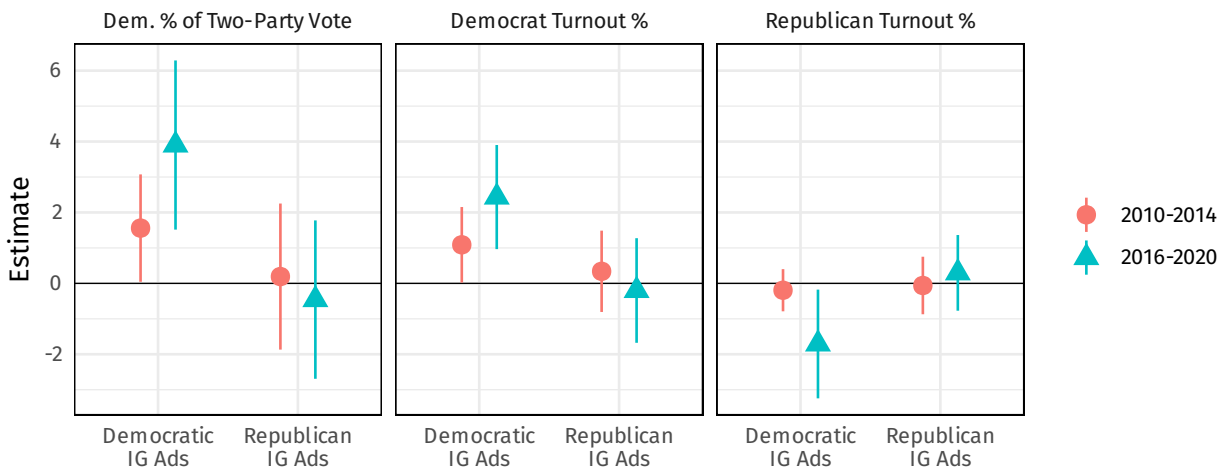

Figure 4: Treatment effect heterogeneity before and after Donald Trump enters the 2016 Presidential Race.

we included an interaction between our cumulative treatment measure and an indicator for whether the election was before or after Donald Trump became a candidate for president in 2015. Figure 4 shows that the effectiveness of Democratic group ads is driven in large part by the post-Trump era. These ads are more effective at increasing Democratic votes and reducing Republican votes, and all of these effect differences are statistically significant at the $\alpha = 0.1$ level. In particular, the demobilizing effect of Democratic group ads on Republican voters is a feature of the Trump era. These patterns are consistent with how Trump alienated large segments of Republicans and perhaps made them more vulnerable to ads that encouraged them to stay home or vote for Democrats.

The flexible structure of marginal structural models allows us to investigate which weeks of the campaign are driving the effects on these outcomes. To do so, we can break up the cumulative sum of treated weeks into the number of treated weeks within three weeks of election day and the number of treated weeks 4-5 weeks before election day. Under our assumptions, this is another valid way to parameterize the MSM, allowing us to summarize the causal response surface in a different way. Figure 5 shows the estimated effects of independent group ads at

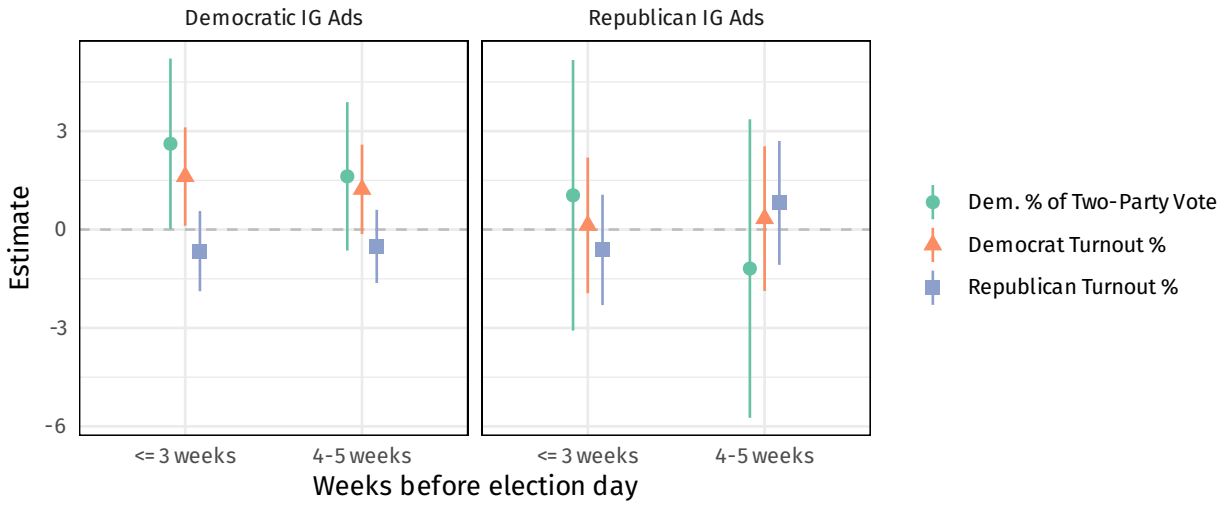

Figure 5: IPTW-FE estimated effects of IG ads by week of the campaign for various outcomes.

various weeks before election day, as estimated by the IPTW-FE with the baseline covariates. The only major difference is that the positive effect of Democratic IG ads appears stronger in the last few weeks of the campaign compared to earlier weeks. This increased effectiveness of more recent ads is consistent with previous experimental studies of television ads (Gerber et al., 2011).

6.3 Robustness checks

Given that advertising is not randomized across markets, we may worry about residual unmeasured confounding that our approach may miss. To investigate if we can detect any potential biases in our estimation strategy, we use the same designs as above on placebo outcomes. First, we obtain the outcomes for the same media market for the most recent previous election for the same office and use those as outcomes. If our IPTW-FE approach was unable to adjust for unmeasured confounding at the market level, then these estimates would detect bias since future independent group ads cannot affect past electoral outcomes. We also investigate the effects of our estimates on the baseline polling for the Democratic candidate before the five-week period in our MSMs.

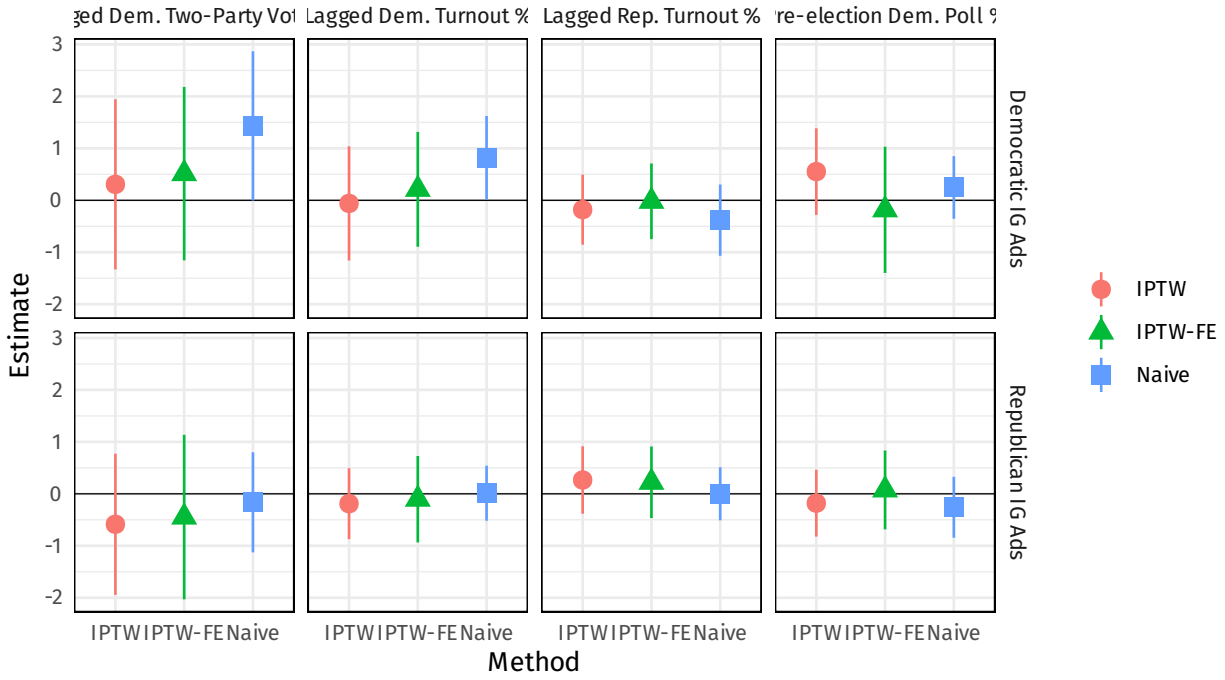

Figure 6: Falsification test results. These are estimated effects on outcomes from the previous election in that market for that office and pre-election polling results.

Figure 6 shows these results. Both of the IPTW approaches result in estimates very close to zero for all outcomes, which is consistent with our identifying assumptions. Interestingly, the naive approach does show some residual confounding for some of the effects of Democratic group ads. Taken together, these results give us some confidence that further unmeasured confounding is not a major source of bias in our estimates.

7 Conclusion

In this paper, we estimated the effects of independent group advertising on electoral outcomes in U.S. state-wide elections. To do so, we developed a method to control for time-constant unmeasured confounding in marginal structural models by using a fixed effects approach to estimate the propensity score of the time-varying treatment. We derived the large-sample properties of this estimator under an asymptotic setup where the number of time periods and

the number of units grow together. Simulations showed that the proposed method outperforms a naive approach that omits fixed effects and performs well overall, especially when the magnitude of the heterogeneity is moderate. An obvious place for future research would be to apply these methods to data where we have repeated measurements of the outcomes as well as the treatment. In those situations, it may be possible to develop doubly-robust estimators under fixed effects assumptions.

References

Arellano, M. and J. Hahn (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. Newey, and T. Persson (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Volume 3 of *Econometric Society Monographs*, Chapter 12, pp. 381–409. Cambridge University Press.

Arkhangelsky, D. and G. Imbens (2018). The role of the propensity score in fixed effect models. Available at https://arxiv.org/pdf/1807.02099.pdf.

Baćak, V. and M. E. Karim (2019, March). The effect of serious offending on health: A marginal structural model. *Society and Mental Health 9*(1), 18–32.

Bacak, V. and E. H. Kennedy (2015, February). Marginal structural models: An application to incarceration and marriage during young adulthood. *Journal of Marriage and Family 77*(1), 112–125.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–972.

Benoit, W. L. (2013, October). *Televised Political Advertisements*. Oxford University Press.

Blackwell, M. (2013). A framwork for dynamic causal inference in political science. *American Journal of Political Science 57*(2), 504–520.

Callaway, B. and P. H. Sant'Anna (2021, December). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Cole, S. R. and M. A. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiolology 168*(6), 656–664.

Cox, D. D. and T. Y. Kim (1995). Moment bounds for mixing random variables useful in non-parametric function estimation. *Stochastic Processes and their Applications 56*(1), 151–158.

Creamer, C. D. and B. A. Simmons (2019). Do self-reporting regimes matter? evidence from the convention against torture. *International Studies Quarterly 63*(4), 1051–1064.

Fan, J. and Q. Yao (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics 150*(1), 71–85.

Fernández-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics 192*(1), 291–312.

Fernández-Val, I. and M. Weidner (2018). Fixed effects estimation of large-T panel data models. *Annual Review of Economics 10*(1), 109–138.

Fowler, E. F., M. M. Franz, T. N. Ridout, L. M. Baum, and C. Bogucki (2019). Political advertising in 2020. The Wesleyan Media Project, Department of Government at Wesleyan University.

Gerber, A. S., J. G. Gimpel, D. P. Green, and D. R. Shaw (2011). How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *The American Political Science Review 105*(1), 135–150.

Goldstein, K. and T. N. Ridout (2004, May). Measure the effects of televised political advertising in the united states. *Annual Review of Political Science 7*(1), 205–226.

Goodman-Bacon, A. (2021, December). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Gruber, S., R. W. Logan, I. Jarrín, S. Monge, and M. A. Hernán (2015). Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in Medicine 34*(1), 106–117.

Hahn, J. and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory 27*(6), 1152–1191.

Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica 72*(4), 1295–1319.

Hill, S. J., J. Lo, L. Vavreck, and J. Zaller (2013, October). How quickly we forget: The duration of persuasion effects from mass communication. *Political Communication 30*(4), 521–547.

Huber, G. A. and K. Arceneaux (2007). Identifying the persuasive effects of presidential advertising. *American Journal of Political Science 51*(4), 957–977.

Imai, K. and I. S. Kim (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science 63*(2), 467–490.

Imai, K. and M. Ratkovic (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association 110*(511), 1013–1023.

Jacobson, G. C. (1975, August). The impact of broadcast campaigning on electoral outcomes. *The Journal of Politics 37*(3), 769–793.

Kallus, N. and M. Santacatterina (2019). Optimal balancing of time-dependent confounders for marginal structural models.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association 114*(526), 645–656.

Klumpp, T., H. M. Mialon, and M. A. Williams (2016, February). The business of american democracy: *Citizens United* , independent spending, and elections. *The Journal of Law and Economics 59*(1), 1–43.

Kurer, T. (2020). The declining middle: Occupational change, social status, and the populist right. *Comparative Political Studies 53*(10-11), 1798–1835.

Ladam, C., J. J. Harden, and J. H. Windett (2018). Prominent role models: High-profile female politicians and the emergence of women as candidates for public office. *American Journal of Political Science 62*(2), 369–381.

MacKinnon, J. and H. White (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*(3), 305–325.

Muñoz, I. D. and M. J. van der Laan (2011). Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics 7*(1), 1–20.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, Chapter 36, pp. 2111–2245. Elsevier.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*(1), 1–32.

Obikane, E., T. Shinozaki, D. Takagi, and N. Kawakami (2018, July). Impact of childhood abuse on suicide-related behavior: Analysis using marginal structural models. *Journal of Affective Disorders 234*, 224–230.

Ridout, T. M. and M. M. Franz (2011). *The Persuasive Power of Campaign Advertising*. Temple University Press.

Ridout, T. N., E. F. Fowler, and M. M. Franz (2021, April). Spending fast and furious: Political advertising in 2020. *The Forum 18*(4), 465–492.

Robins, J. M. (1998a). Correction for non-compliance in equivalence trials. *Statistics in Medicine 17*(3), 269–302.

Robins, J. M. (1998b). Marginal structural models. In *1997 Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, pp. 1–10. American Statistical Association.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese 121*(1/2), 151–179.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Volume 116 of *The IMA Volumes in Mathematics and its Applications*, pp. 95–134. New York: Springer-Verlag.

Robins, J. M., M. A. Hernán, and B. A. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Rotnitzky, A., Q. Lei, M. Sued, and J. M. Robins (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika 99*(2), 439–456.

Sampson, R. J., J. H. Laub, and C. Wimer (2006, August). Does marriage reduce crime? a counterfactual approach to within-individual causal effects. *Criminology 44*(3), 465–508.

Sharkey, P. and F. Elwert (2011, May). The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology 116*(6), 1934–81.

Sides, J., L. Vavreck, and C. Warshaw (2022, May). The effect of television advertising in united states elections. *American Political Science Review 116*(2), 702–718.

Sobel, M. E. (2012). Does marriage boost men's wages?: Identification of treatment effects in fixed effects regression models for panel data. *Journal of the American Statistical Association 107*(498), 521–529.

Sun, L. and S. Abraham (2021, December). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Wodtke, G. T., D. J. Harding, and F. Elwert (2011, October). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review 76*(5), 713–736.

Xiao, Y., E. E. Moodie, and M. Abrahamowicz (2013). Comparison of approaches to weight truncation for marginal structural cox models. *Epidemiologic Methods 2*(1), 1–20.

A Proofs

We let $V_{it}(\beta,\alpha) = \partial\ell_{it}(\beta,\alpha)/\partial\alpha$ and $S_{it}(\beta,\alpha) = \partial\ell_{it}(\beta,\alpha)/\partial\beta$ be the score functions for the propensity score model with additional subscripts indicating higher-order partial derivatives, $V_{it\alpha}(\beta,\alpha) = \partial V_{it}(\beta,\alpha)/\partial\alpha$. As above, without arguments, these functions are evaluated at their true values, $V_{it} = V_{it}(\beta_0,\alpha_{i0})$.

Here, we state the regularity conditions on the treatment process.

Assumption 3 (Treatment Regularity Conditions)**.** *Let $\nu,\epsilon > 0$ and $\mathcal{B}_0(\epsilon)$ is an ϵ-neighborhood of (β_0,α_{i0}) for all i,t,N,T.*

(i) We assume that for all i,t,N,T, we have $\pi_{it}(\beta_0,\alpha_{i0}) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it}, \overline{D}_{i,t-1}, \alpha_i)$. The realization of the parameters and unobserved effects that generate the observed data are denoted β_0 and α_0.

(ii) We assume that $(\beta,\alpha) \mapsto \ell_{it}(\beta,\alpha)$ is four-times continuously differentiable over $\mathcal{B}_0(\epsilon)$ almost surely. The partial derivatives of $\ell_{it}(\beta,\alpha)$ with respect to the elements of (β,α) are bounded in absolute value uniformly over $(\beta,\alpha) \in \mathcal{B}_0(\epsilon)$ by a function $M(Z_{it}) > 0$ almost surely and $\max_{i,t} \mathbb{E}[M(Z_{it})^{8+\nu}]$ is almost surely uniformly bounded over N,T.

(iii) For all i,t we have $\pi_{it}(\beta,\alpha)$ bounded away from 0 and 1 uniformly over $(\beta,\alpha) \in \mathcal{B}_0(\epsilon)$.

(iv) (Concavity) For all N, T $(\beta,\alpha) \mapsto \ell_{it}(\beta,\alpha)$ is strictly concave over $\mathbb{R}^{dim(\beta)+1}$ almost surely. Furthermore, there exists $b_{\min}$ and $b_{\max}$ such that for all $(\beta,\alpha) \in \mathcal{B}_0(\epsilon)$, $0 < b_{\min} \leq -\mathbb{E}[\partial^2\ell_{it}(\beta,\alpha)/\partial\alpha_i\alpha_j \mid \alpha_i] \leq b_{\max}$ almost surely uniformly over i, t, N, and T.

Assumption 3 mostly derives from Fernández-Val and Weidner (2016), who used them to establish the asymptotic properties of nonlinear panel models with unit- and time-specific effects, though we focus only on unit effects. Assumption 3(i) establishes the parametric component of the model, which will help us derive the asymptotic behavior of our estimation strategy

under the sampling assumption. As noted above, this assumption is not required for nonparametric identification, but it does reflect the common correctly specified parametric propensity score model often invoked in applications of IPW estimators. If this model is misspecified, we can view the results propensity scores as projections onto the assumed parametric family and the resulting bias will depend on the size of the error of that projection and how it correlates with the outcome. Assumption 3(ii) requires the log-likelihood of the propensity score model and its derivatives to be sufficiently smooth to allow for the higher-order asymptotic expansions we use. With a binary response, this assumption could be replaced by a moment condition on the distribution of the covariates. We invoke a locally uniform version of positivity in Assumption 3(iii). Note that Assumption 3(iii) implicitly restricts α_i, since if α_i were completely unrestricted, then we may have $\pi_{it}^{-1} \to \infty$. Furthermore, bounded propensity scores also rules out staggered adoption designs where a unit can only switch into treatment once and cannot revert. Finally, Assumption 3(iv) ensures that the MLE is identified and should be satisfied in the usual parametric models used for binary data when the covariates, X_{it} vary in the time and unit dimensions.

We now turn to the regularity conditions for the outcome model

Assumption 4 (Outcome Regularity Conditions)**.** *Let $\nu > 0$ and $\mathcal{B}_0(\epsilon)$ is an ϵ-neighborhood of $(\gamma_0, \beta_0, \alpha_{i0})$ for all i, N.*

(i) (Bounded outcome moments) $\mathbb{E}[|Y_i(d)|^{4+\nu}]$ and $\mathbb{E}[|Y_i(d)|^{4+\nu} \mid \alpha_i, \overline{D}_i, \overline{X}_i]$ are bounded by finite constants, uniformly over i.

(ii) (MSM regularity) The parameters $\phi = (\gamma, \beta, \alpha)$ are in the interior of Φ, which is a compact, convex subset of $\mathbb{R}^{J+R+1}$ with $J = \dim(\gamma)$ and $R = \dim(\beta)$. The map $\gamma \mapsto U_i(\gamma, \beta, \alpha)$ is continuously differentiable over $(\gamma, \beta, \alpha) \in \mathcal{B}_0(\epsilon)$ with $\mathbb{E}[\sup_{\gamma \in \mathcal{B}_0(\epsilon)} \|\partial_\gamma U_i(\gamma, \beta, \alpha)\|] < \infty$.

Assumption 4(i) ensures the potential outcomes have sufficiently bounded (conditional)

moments. Assumptions 4(ii) is a set of standard regularity conditions for the marginal structural model.

Proof of Theorem 1. Suppose now we are interested in an MSM $g(\underline{d}_k;\gamma) = \mathbb{E}[Y_{iT}(\underline{d}_k)]$, where $\underline{d}_k = (d_{T-k},\ldots,d_T)$ and k is fixed. The parameter vector γ is of length J. Define the probability of a particular treatment history as a function of the propensity score parameters as

$$W_i(\underline{d}_k;\beta,\alpha_i) = \prod_{j=0}^{k} \pi_{i,T-j}(\beta,\alpha_i)^{d_{T-j}}(1-\pi_{i,T-j}(\beta,\alpha_i))^{1-d_{T-j}}.$$

Generally, we can define an MSM as the solution to the following:

$$0 = \mathbb{E}\left\{\frac{h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik};\gamma))}{W_i(\underline{D}_{ik};\beta,\alpha_i)}\right\}$$

where $h(\cdot)$ is a function with J-length output, chosen by the researcher. For example, if Y_i is continuous and g is linear and additive, it is common to use $h(\underline{D}_{ik}) = \underline{D}'_{ik}$. If we knew the propensity scores, we could derive an estimator of γ based on the sample moment condition:

$$0 = \frac{1}{N}\sum_{i=1}^{N} \frac{h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik};\widehat{\gamma}))}{W_i(\underline{D}_{ik};\beta,\alpha_i)} = \frac{1}{N}\sum_{i=1}^{N} U_i(\widehat{\gamma},\beta,\alpha_i)$$

Because the propensity score is never known in observational studies, we define our estimator based on the estimated propensity scores:

$$0 = \frac{1}{N}\sum_{i=1}^{N} \frac{h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik};\widehat{\gamma}))}{W_i(\underline{D}_{ik};\widehat{\beta},\widehat{\alpha}_i)} = \frac{1}{N}\sum_{i=1}^{N} U_i(\widehat{\gamma},\widehat{\beta},\widehat{\alpha}_i)$$

Consistency Note that within $(\beta,\alpha_i) \in \mathcal{B}_0(\epsilon)$, $|\partial_\beta \ell_{it}(\beta,\alpha_i)| < M(Z_{it})$ implies that $|\partial_{\beta_k}\pi_{it}(\beta,\alpha_i)| < CM(Z_{it})$ for some constant C since,

$$|\partial_{\beta_k}\ell_{it}(\beta,\alpha_i)| = \left|\left(\frac{D_{it}-\pi_{it}(\beta,\alpha_i)}{\pi_{it}(\beta,\alpha_i)(1-\pi_{it}(\beta,\alpha_i))}\right)\partial_{\beta_k}\pi_{it}(\beta,\alpha_i)\right|,$$

and the propensity scores are uniformly bounded from below over $\mathcal{B}_0(\epsilon)$. The same applies to $|\partial_\alpha \pi_{it}(\beta, \alpha)|$. With these results, we can also bound the partial derivatives of the weights:

$$|\partial_{\beta_k} W_i(\beta, \alpha_i)| = \left| \sum_{t=T-k}^{T} \partial_{\beta_k} \pi_{it}(\beta, \alpha_i) \sum_{s \neq t} \pi_{is}(\beta, \alpha_i) \right| \leq k(k-1) |\partial_{\beta_k} \pi_{it}(\beta, \alpha_i)| \leq Ck(k-1) M(Z_{it}),$$

where k is fixed as $N, T \to \infty$. Again, a similar expression holds for $|\partial_\alpha W_i(\beta, \alpha_i)|$. Thus, by the mean value theorem, we have

$$\begin{aligned} |W_i(\widehat{\beta}, \widehat{\alpha}_i) - W_i(\beta_0, \alpha_{i0})| &\leq \|\partial_\beta W_i(\beta, \alpha_i)\| \, \|\widehat{\beta} - \beta_0\| + |\partial_\alpha W_i(\beta, \alpha_i)| \, |\widehat{\alpha}_i - \alpha_{i0}| \\ &\leq C_\beta M(Z_{it}) \|\widehat{\beta} - \beta_0\| + C_\alpha M(Z_{it}) |\widehat{\alpha}_i - \alpha_{i0}|, \end{aligned}$$

for some constants C_β and C_α.

Using this result we can uniformly bound the convergence of the estimating equation in terms of the parameters of the weighting model.

$$\begin{aligned} &\sup_{\gamma \in \Gamma} |N^{-1} \sum_{i=1}^{N} U_i(\gamma, \widehat{\beta}, \widehat{\alpha}_i) - U_i(\gamma, \beta_0, \alpha_{i0})| \\ &\quad \leq N^{-1} \sum_{i=1}^{N} \sup_{\gamma \in \Gamma} |U_i(\gamma, \widehat{\beta}, \widehat{\alpha}_i) - U_i(\gamma, \beta_0, \alpha_{i0})| \\ &\quad \leq N^{-1} \sum_{i=1}^{N} \sup_{\gamma \in \Gamma} |h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik}, \gamma))(W_i(\widehat{\beta}, \widehat{\alpha}_i)^{-1} - W_i(\beta_0, \alpha_{i0}))| \\ &\quad \leq N^{-1} \sum_{i=1}^{N} \frac{|W_i(\widehat{\beta}, \widehat{\alpha}_i) - W_i(\beta_0, \alpha_{i0})|}{W_i(\widehat{\beta}, \widehat{\alpha}_i) W_i(\beta_0, \alpha_{i0})} \sup_{\gamma \in \Gamma} |h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik}, \gamma))| \\ &\quad < N^{-1} \sum_{i=1}^{N} |W_i(\widehat{\beta}, \widehat{\alpha}_i) - W_i(\beta_0, \alpha_{i0})| \sup_{\gamma \in \Gamma} |h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik}, \gamma))| \\ &\quad \leq C_\beta \|\widehat{\beta} - \beta_0\| N^{-1} \sum_{i=1}^{N} |M(Z_{it})| \sup_{\gamma \in \Gamma} |h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik}, \gamma))| \\ &\qquad + C_\alpha \max_i |\widehat{\alpha}_i - \alpha_{i0}| N^{-1} \sum_{i=1}^{N} |M(Z_{it})| \sup_{\gamma \in \Gamma} |h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik}, \gamma))| \end{aligned}$$

The fourth inequality holds because under Assumption 3(iii), we have that $W_i(\underline{d}_k; \beta, \alpha) \in (\epsilon, 1-\epsilon)$ where $\epsilon > 0$ near (β_0, α_{i0}). By the bounded moments of $M(Z_{it})$ and Y_i, we have that $N^{-1}\sum_{i=1}^N |M(Z_{it})| \sup_{\gamma\in\Gamma} |h(\underline{D}_{ik})(Y_i - g(\underline{D}_{ik}, \gamma))| = O_p(1)$. Combined with the consistency of $\widehat{\beta}$ and $\widehat{\alpha}_i$ from Lemma 1, we have $\sup_{\gamma\in\Gamma} |N^{-1}\sum_{i=1}^N U_i(\gamma, \widehat{\beta}, \widehat{\alpha}_i) - U_i(\gamma, \beta_0, \alpha_{i0})| = o_p(1)$. Thus, we have

$$
\begin{aligned}
|N^{-1}\sum_{i=1}^N U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i)| &\leq |N^{-1}\sum_{i=1}^N U_i(\widehat{\gamma}, \beta_0, \alpha_{i0})| + |N^{-1}\sum_{i=1}^N U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i) - U_i(\widehat{\gamma}, \beta_0, \alpha_{i0})| \\
&\leq |N^{-1}\sum_{i=1}^N U_i(\widehat{\gamma}, \beta_0, \alpha_{i0})| + \sup_{\gamma\in\Gamma} |N^{-1}\sum_{i=1}^N U_i(\gamma, \widehat{\beta}, \widehat{\alpha}_i) - U_i(\gamma, \beta_0, \alpha_{i0})| \\
&= |N^{-1}\sum_{i=1}^N U_i(\widehat{\gamma}, \beta_0, \alpha_{i0})| + o_p(1).
\end{aligned}
$$

This establishes that $\widehat{\gamma} \xrightarrow{p} \gamma_0$, by standard results of estimating equations.

Asymptotic expansion Let $G_i(\gamma, \beta, \alpha) = \partial U_i(\gamma, \beta, \alpha)/\partial\gamma$, then we have the following expansion:

$$
0 = \frac{1}{\sqrt{N}}\sum_{i=1}^N U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i) = \frac{1}{\sqrt{N}}\sum_{i=1}^N U_i(\gamma_0, \widehat{\beta}, \widehat{\alpha}_i) + \sqrt{N}(\widehat{\gamma} - \gamma_0)\left(\frac{1}{N}\sum_{i=1}^N G_i(\overline{\gamma}, \widehat{\beta}, \widehat{\alpha}_i)\right)
$$

where $\overline{\gamma}$ is a value between $\widehat{\gamma}$ and γ_0. This implies the following influence-function representation for the estimator:

$$
\sqrt{N}(\widehat{\gamma} - \gamma_0) = \left(\frac{1}{N}\sum_{i=1}^N G_i(\overline{\gamma}, \widehat{\beta}, \widehat{\alpha}_i)\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^N U_i(\gamma_0, \widehat{\beta}, \widehat{\alpha}_i) \tag{A.1}
$$

$$
= G^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^N U_i(\gamma_0, \widehat{\beta}, \widehat{\alpha}_i) + o_p(1), \tag{A.2}
$$

where $G = \mathbb{E}[G_i]$ and noting that function without arguments are evaluated at the true values of the parameters, $G_i = G_i(\gamma_0, \beta_0, \alpha_{i0})$. The second equality here follows from Lemma 2.4

of Newey and McFadden (1994) after noting that $\overline{\gamma}$ is between $\widehat{\gamma}$ and γ_0, that $\widehat{\gamma}$, $\widehat{\beta}$, and $\widehat{\alpha}_i$ are all consistent and from Assumption 4.

We can expand the rth element of U_i, U_{ir} in Equation (A.2) as

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_{ir}(\gamma_0,\widehat{\beta},\widehat{\alpha}_i) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_{ir} &+ \underbrace{\sqrt{N}(\widehat{\beta}-\beta_0)^\top\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}U_{ir}\right)}_{\text{(I)}} + \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widehat{\alpha}_i-\alpha_{i0})\frac{\partial}{\partial\alpha}U_{ir}}_{\text{(II)}} \\
&+ \underbrace{\sqrt{N}(\widehat{\beta}-\beta_0)^\top\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\beta\partial\beta}U_{ir}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\right)(\widehat{\beta}-\beta_0)}_{\text{(III)}} \\
&+ \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widehat{\alpha}_i-\alpha_i)^2\frac{\partial^2}{\partial\alpha\partial\alpha}U_{ir}(\gamma_0,\overline{\beta},\overline{\alpha}_i)}_{\text{(IV)}} \\
&+ \underbrace{\sqrt{N}(\widehat{\beta}-\beta)^\top\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\beta\partial\alpha}U_{ir}(\gamma_0,\overline{\beta},\overline{\alpha}_i)(\widehat{\alpha}_i-\alpha_{i0})\right)}_{\text{(IV)}}
\end{aligned}$$

where $\overline{\alpha}_i$ is the value between $\widehat{\alpha}_i$ and α_{i0} ($\overline{\beta}$ is defined similarly).

First order terms We first show that Term (I) and (II) in the above expression are both $o_p(1)$.

We denote derivatives of U_{ir} with subscripts so that,

$$U_{ir,\alpha}(\gamma,\beta,\alpha) = \frac{\partial}{\partial\alpha}U_{ir}(\gamma,\beta,\alpha), \quad \text{and} \quad U_{ir,\beta}(\gamma,\beta,\alpha) = \frac{\partial}{\partial\beta}U_{ir}(\gamma,\beta,\alpha)$$

and simply write $U_{ir,\alpha}$ when evaluated at the true values of the parameters. Letting $h(\underline{D}_{ik})_{[r]}$ be the rth entry of that vector, the expression of $U_{ir,\alpha}$ is given by

$$U_{ir,\alpha} = \frac{h(\underline{D}_{ik})_{[r]}(Y_i - g(\underline{D}_{ik};\gamma_0))}{W_i^2(\underline{D}_{ik};\beta,\alpha_i)}$$

$$
\begin{aligned}
&\times \sum_{t=T-k}^{T} \left\{ (2D_{it}-1) \frac{\partial}{\partial \alpha_i} \pi_{it}(\beta_0, \alpha_{i0}) \prod_{\substack{s=T-k \\ s \neq t}}^{T} \pi_{is}(\beta_0, \alpha_{i0})^{D_{is}} \left[1 - \pi_{is}(\beta_0, \alpha_{i0}) \right]^{1-D_{is}} \right\} \\
&= U_{ir} \sum_{t=T-k}^{T} \frac{(2D_{it}-1)\pi_{it} V_{it}}{\pi_{it}^{D_{it}} (1-\pi_{it})^{1-D_{it}}} \\
&= U_{ir} \sum_{t=T-k}^{T} \left[(2D_{it}-1) V_{it} \left\{ \frac{\pi_{it}}{1-\pi_{it}} \right\}^{1-D_{it}} \right] \\
&\equiv U_{ir} \overline{V}_i
\end{aligned}
$$

Here, we have used

$$
\frac{\partial}{\partial \alpha_i} \pi_{it}(\beta_0, \alpha_{i0}) = \pi_{it} \frac{\partial}{\partial \alpha_i} \log \ell_{it} = \pi_{it} V_{it}.
$$

Similarly, the expression of $U_{ir,\beta}$ can be derived as

$$
\begin{aligned}
U_{ir,\beta} &= U_{ir} \sum_{t=T-k}^{T} \left[(2D_{it}-1) S_{it} \left\{ \frac{\pi_{it}}{1-\pi_{it}} \right\}^{1-D_{it}} \right] \\
&\equiv U_{ir} \overline{S}_i
\end{aligned}
$$

where S_{it} is the score function $\partial \log \ell_{it} / \partial \beta$.

To control (I), we use the following results:

$$
|(\mathrm{I})| \leq \sqrt{N} \|\widehat{\beta} - \beta_0\| \left\| \frac{1}{N} \sum_{i=1}^{N} U_{ir} \overline{S}_i \right\|
$$

Because $\sqrt{NT}\|\widehat{\beta} - \beta_0\| = O_p(1)$, we have $\sqrt{N}\|\widehat{\beta} - \beta_0\| = O_p(1/\sqrt{T}) = o_p(1)$. Let $\overline{S}_{iq}$ be the qth entry in the $\overline{S}_i$ vector. Note that $\overline{S}_{iq}$ has bounded fourth moments by Lemma 5 since it a

function of the qth score vector. Thus, for the second term bounding (I), we have:

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} U_{ir}\overline{S}_{iq}\right)^2\right] \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(U_{ir}\overline{S}_{iq})^2] \leq \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}[U_{ir}^4]\right)^{1/2}\left(\mathbb{E}[\overline{S}_{iq}^4]\right)^{1/2} = O(1),$$

where the first inequality holds because for any i.i.d. set of random variables $X_1, \ldots, X_n$, we have $\mathbb{E}[(n^{-1}\sum_i X_i)^2] \leq n^{-1}\sum_i \mathbb{E}[X_i^2]$. The second inequality holds by Cauchy-Schwarz, and the last equality holds by Assumption 3(ii) and 4(i). Because the same holds for all entries in $\overline{S}_i$, we have $\|N^{-1}\sum_{i=1}^N U_{ir}\overline{S}_i\| = O_p(1)$ by the Markov inequality and so (I) is $o_p(1)O_p(1) = o_p(1)$.

By Lemma 1, Term (II) can be written as

$$\begin{aligned}(\text{II}) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_{ir,\alpha}(\widehat{\alpha}_i - \alpha_{i0}) = \frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} U_{ir,\alpha}\psi_{it} + o_p(1) \\ &= \underbrace{\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[U_{ir,\alpha}\psi_{it} \mid \alpha_i]}_{\text{(II.a)}} + \underbrace{\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(U_{ir,\alpha}\psi_{it} - \mathbb{E}[U_{ir,\alpha}\psi_{it} \mid \alpha_i]\right)}_{\text{(II.b)}} + o_p(1),\end{aligned}$$

where the $o_p(1)$ term in the first line is due to $N^{-1}\sum_{i=1}^N U_{ir,\alpha}$ being $O_p(1)$ and the remained of the $\widehat{\alpha}_i$ expansion from Lemma 1 being $\max_i |R_i| = o_p(T^{-1/2})$.

Note that $\psi_{it} = \mathbb{E}_T\{\mathbb{E}_\alpha[V_{it\alpha}]\}^{-1}V_{it}$. In an abuse of notation, we define $\overline{V}_i(\underline{d}_k, D_{it})$ to be $\overline{V}_i$ with all covariates and the outcome replaced with their potential outcomes setting $\underline{d}_k$ and we leave D_{it} as an argument to emphasize that this function depends on D_{it}. Furthermore, let $U_{ir,\alpha}(\underline{d}_k, D_{it}) = \overline{V}_i(\underline{d}_k, D_{it})h(d_k)_{[r]}(Y_i(\underline{d}_k, D_{it}) - g(\underline{d}_k; \gamma_0))$. Then, applying the g-computational formula, we have for all $t < T - k$,

$$\mathbb{E}[U_{ir,\alpha}\psi_{it} \mid \alpha_i] = \sum_{\underline{d}_k}\mathbb{E}_T\{\mathbb{E}_\alpha[V_{it\alpha}]\}^{-1}\mathbb{E}\left[V_{it}U_{ir,\alpha}(\underline{d}_k, D_{it}) \mid \alpha_i\right]$$

Note that because V_{it} is a score, we can use iterated expectations to show that $\mathbb{E}[V_{it} \mid \alpha_i] = 0$. Thus, the inner expectation in the above expression is the covariance between V_{it} and

$U_{ir,\alpha}(\underline{d}_k, D_{it})$ conditional on α_i. Using Lemma 2, we have

$$\begin{aligned}
\left|\mathbb{E}\left[V_{it}U_{ir,\alpha}(\underline{d}_k, D_{it}) \mid \alpha_i\right]\right| &= \left|\mathrm{Cov}(V_{it}, U_{ir,\alpha}(\underline{d}_k, D_{it}) \mid \alpha_i)\right| \\
&\leq 8a(T-k-t)^{[1-1/(8+\nu)-1/(2+\nu)]} \\
&\quad \times \left[\mathbb{E}[|V_{it}|^{8+\nu}]\right]^{1/(8+\nu)} \left[\mathbb{E}[|U_{ir,\alpha}(\underline{d}_k, D_{it})|^{2+\nu}]\right]^{1/(2+\nu)} \\
&\leq C(T-k-t)^{-\mu[1-1/(8+\nu)-1/(2+\nu)]} \\
&\leq C(T-k-t)^{-4}
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\left|\sum_{t=1}^{T}\mathbb{E}[U_{ir,\alpha}\psi_{it} \mid \alpha_i]\right| &\leq \sum_{t=1}^{T}\sum_{\underline{d}_k}\left|\mathbb{E}_T\{\mathbb{E}_\alpha[V_{it\alpha}]\}^{-1}\mathbb{E}\left[V_{it}U_{ir,\alpha}(\underline{d}_k, D_{it}) \mid \alpha_i\right]\right| \\
&\leq 2^k \sum_{t=1}^{T}\mathbb{E}\left\{\left|\mathbb{E}\left[V_{it}U_{ir,\alpha}(\underline{d}_k, D_{it}) \mid \alpha_i\right]\right|\right\} \\
&\leq C2^k \sum_{t=1}^{T}(T-k-t)^{-4} \\
&\leq C2^k \sum_{m=1}^{\infty} m^{-4} = \frac{C2^k\pi^4}{90}
\end{aligned}$$

Thus, we can establish that (II.a) is $O_p(1/\sqrt{T})$. By Lemma 3 and the bounded moment conditions for the outcome and the partial derivatives, we have

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(U_{ir,\alpha}\psi_{it} - \mathbb{E}[U_{ir,\alpha}\psi_{it} \mid \alpha_i]\right) = O_p(1/\sqrt{NT}).$$

This implies that (II.b) is $\sqrt{N}O_p(1/\sqrt{NT}) = O_p(T^{-1/2}) = o_p(1)$. Then, it follows that

$$\text{(I)} + \text{(II)} = o_p(1) + o_p(1) = o_p(1).$$

Second order terms We will show that Term (III), (IV) and (V) are all $o_p(1)$.

Define the following second derivatives of the MSM estimating equation:

$$U_{ir\alpha\alpha}(\gamma,\beta,\alpha) = \frac{\partial}{\partial\alpha}U_{ir\alpha}(\gamma,\beta,\alpha), \quad U_{ir\beta\beta}(\gamma,\beta,\alpha) = \frac{\partial}{\partial\beta}U_{ir\beta}(\gamma,\beta,\alpha),$$

$$U_{ir\beta\alpha}(\gamma,\beta,\alpha) = \frac{\partial}{\partial\alpha}U_{ir\beta}(\gamma,\beta,\alpha).$$

Lemma 1 implies estimation error in the fixed effects are uniformly bounded so (IV) is bounded as

$$|(\text{IV})| \leq \sqrt{N}\max_i |\widehat{\alpha}_i - \alpha_i|^2 \left|\frac{1}{N}\sum_{i=1}^{N} U_{ir\alpha\alpha}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\right|. \tag{A.3}$$

Note that Lemma 1 also implies that $\max_i |\widehat{\alpha}_i - \alpha|^2 = O_p(T^{-3/4})$.

Next we bound the second term in (A.3). First, we derive an expression for $U_{ir\alpha\alpha}$ using the derivation of $U_{ir\alpha}$ above:

$$\begin{aligned} U_{ir\alpha\alpha}(\gamma_0,\overline{\beta},\overline{\alpha}_i) &= U_{ir\alpha}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\overline{V}_i(\overline{\beta},\overline{\alpha}) + U_{ir}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\frac{\partial}{\partial\alpha}\overline{V}_i(\overline{\beta},\overline{\alpha}) \\ &= U_{ir}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\overline{V}_{i\alpha}(\overline{\beta},\overline{\alpha}), \end{aligned}$$

where we define $\overline{V}_{i\alpha}(\beta,\alpha) = \overline{V}_i(\beta,\alpha)^2 + \partial\overline{V}_i(\beta,\alpha)/\partial\alpha$.

With this, we bound the second moment of the second term in a neighborhood around the truth:

$$\begin{aligned} \mathbb{E}\left[\sup_{(\alpha,\beta)\in B_0(\epsilon)}\left(\frac{1}{N}\sum_{i=1}^{N} U_{ir\alpha\alpha}(\gamma_0,\beta,\alpha_i)\right)^2\right] &= \mathbb{E}\left[\sup_{(\alpha,\beta)\in B_0(\epsilon)}\left(\frac{1}{N}\sum_{i=1}^{N} U_{ir}(\gamma_0,\beta,\alpha_i)\overline{V}_{i\alpha}(\beta,\alpha)\right)^2\right] \\ &\leq \mathbb{E}\left[\sup_{(\alpha,\beta)\in B_0(\epsilon)}\left(\frac{1}{N}\sum_{i=1}^{N} |U_{ir}(\gamma_0,\beta,\alpha_i)||\overline{V}_{i\alpha}(\beta,\alpha)|\right)^2\right] \\ &\leq \mathbb{E}\left[\sup_{(\alpha,\beta)\in B_0(\epsilon)}\left(\frac{1}{N}\sum_{i=1}^{N} |Y_i - g(\underline{D}_{ik};\widehat{\gamma})||\overline{V}_{i\alpha}(\beta,\alpha)|\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}|Y_i - g(\underline{D}_{ik};\widehat{\gamma})|M_i\right)^2\right] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(|Y_i - g(\underline{D}_{ik};\widehat{\gamma})|M_i\right)^2\right] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}[(Y_i - g(\underline{D}_{ik};\widehat{\gamma}))^4]\right)^{1/2}\left(\mathbb{E}[M_i^4]\right)^{1/2} = O(1)
\end{aligned}
$$

The second inequality here is due to bounded propensity scores, the third due to Lemma 5, the fourth due to i.i.d. data, the fifth is Cauchy-Swarchz, and the final equality is due to bounded outcome moments and Lemma 5. This implies that

$$
\sup_{(\alpha,\beta)\in B_0(\epsilon)}\left|\frac{1}{N}\sum_{i=1}^{N}U_{ir\alpha\alpha}(\gamma_0,\beta,\alpha_i)\right| = O_p(1).
$$

Note that $(\overline{\beta},\overline{\alpha}_i) \in \mathcal{B}(\epsilon)$ with probability approaching 1 due to these values between the consistent estimators and the true values of the parameters. This implies that

$$
\left|\frac{1}{N}\sum_{i=1}^{N}U_{ir\alpha\alpha}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\right| = O_p(1).
$$

Combining this with the above, we have that (IV) is $\sqrt{N}O_p(T^{-3/4})O_p(1) = o_p(1)$.

For (V), we follow a similar strategy. First note that we have:

$$
|(\mathrm{V})| < \sqrt{N}\|\widehat{\beta} - \beta_0\|\max_i|\widehat{\alpha}_i - \alpha_i|\left\|\frac{1}{N}\sum_{i=1}^{N}U_{ir\beta\alpha}(\gamma_0,\overline{\beta},\overline{\alpha}_i)\right\|
$$

As above, $\sqrt{N}\|\widehat{\beta} - \beta_0\|\max_i|\widehat{\alpha}_i - \alpha_i| = \sqrt{N}O_p(1/\sqrt{NT})O_p(T^{-3/8}) = O_p(T^{-7/8})$. Let $U_{irq\alpha} = \partial U_{ir\alpha}/\partial\beta_q$ be the qth entry of $U_{ir\beta\alpha}$. By a similar argument to $U_{ir\alpha\alpha}$, we have

$$
U_{irq\alpha}(\beta,\alpha) = U_{ir}(\beta,\alpha)\left(\overline{V}_i(\beta,\alpha)\overline{S}_{iq}(\beta,\alpha) + \partial\overline{V}_i(\beta,\alpha)/\partial\beta_q\right) \equiv U_{ir}(\beta,\alpha)\overline{V}_{iq}(\beta,\alpha).
$$

By Lemma 5 and the argument for $U_{ir\alpha\alpha}$, we have

$$\left| \frac{1}{N} \sum_{i=1}^{N} U_{irq\alpha}(\gamma_0, \overline{\beta}, \overline{\alpha}_i) \right| = O_p(1),$$

for all q which in turn implies,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} U_{ir\beta\alpha}(\gamma_0, \overline{\beta}, \overline{\alpha}_i) \right\| = O_p(1)$$

Thus, we have that (V) is $O_p(T^{-7/8})O_p(1) = o_p(1)$. The proof for (III) being $o_p(1)$ follows similarly.

Combining all results Combining all results, we have that

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) = G^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U_i + o_p(1)$$

□

B Supporting Lemmas

B.1 Lemmas

The following lemma is a restatement of results in Fernández-Val and Weidner (2016) with an additional result on the uniform rate of converge for the propensity score model.

Lemma 1. *Under Assumption 2, the following hold:*

(i) $\|\widehat{\beta} - \beta_0\| = O_p(1/\sqrt{NT})$

(ii) Letting $\psi_{it} = \mathbb{E}_T\{\mathbb{E}_\alpha[V_{it\alpha}]\}^{-1}V_{it}$*, we have:*

$$\widehat{\alpha}_i = \alpha_{i0} + \frac{1}{T}\sum_{t=1}^{T}\psi_{it} + R_i,$$

where $R_i = O_p(1/T)$*.*

(iii) $\max_i |\widehat{\alpha}_i - \alpha_{i0}| = \max_i |T^{-1}\sum_{t=1}^{T}\psi_{it}| = O_p(T^{-3/8})$ *and* $\max_i |R_i| = o_p(T^{-1/2})$*.*

The following can be found as Proposition 2.5 of Fan and Yao (2005).

Lemma 2. *Let* $\{\xi_t\}$ *be an* α*-mixing process with mixing coefficient* $a(m)$*. Let* $\mathbb{E}|\xi_t|^p < \infty$ *and* $\mathbb{E}|\xi_{t+m}|^q < \infty$ *for some* $p, q \geq 1$ *and* $1/p + 1/q < 1$*. Then,*

$$|Cov(\xi_t, \xi_{t+m})| \leq 8a(m)^{1/r}\left[\mathbb{E}|\xi_t|^p\right]^{1/p}\left[\mathbb{E}|\xi_{t+m}|^q\right]^{1/q},$$

where $r = (1 - 1/p - 1/q)^{-1}$*.*

The following lemma comes from Theorem 1 of Cox and Kim (1995).

Lemma 3. *Let* $\{\xi_t\}$ *be an* α*-mixing process with mixing coefficient* $a(m)$ *and* $\mathbb{E}[\xi_t] = 0$*. Let* $r \geq 1$ *be an integer, and let* $\delta > 2r$*,* $\mu > r/(1 - 2r/\delta)$*,* $c > 0$*, and* $C > 0$*. Assume that* $\sup_t \mathbb{E}[|\xi_t|^\delta] \leq C$ *and that* $a(m) \leq cm^{-\mu}$ *for all* $m \in \{1, 2, 3, \ldots\}$*. Then there exists a constant* $B > 0$ *depending on*

r, δ, μ, c and C, but not depending on T or any other distributional characteristics of ξ_t such that for any $T > 0$,

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_t\right)^{2r}\right] \leq B.$$

Lemma 4. *Let $\ell_i^*(\beta, V)$ be the Legendre transformation of the objective function $\ell_i(\beta, \alpha) = T^{-1}\sum_{t=1}^{T}\ell_{it}(\beta, \alpha)$ such that $\ell_i^*(\beta, V) = \max_{\alpha\in\mathcal{B}_\alpha(\epsilon_\alpha)}[\ell_i(\beta, \alpha) - \alpha V]$ and $A_i(\beta, V) = \arg\max_{\alpha\in\mathcal{B}_\alpha(\epsilon_\alpha)}[\ell_i(\beta, \alpha) - \alpha V]$ where $\beta \in \mathcal{B}_\beta(\epsilon_\beta)$ and V denotes the dual parameter to α. Suppose Assumption 2 holds. Then, for some $\nu > 0$*

(i)

$$\sup_{(\beta,\alpha_i)\in\mathcal{B}_0(\epsilon)} |\partial_{VVVV}\ell_i^*(\beta, \alpha_i)| = O_p(1)$$

$$\sup_{(\beta,\alpha_i)\in\mathcal{B}_0(\epsilon)} \|\partial_{V\beta\beta'}\ell_i^*(\beta, \alpha_i)\| = O_p(1)$$

$$\sup_{(\beta,\alpha_i)\in\mathcal{B}_0(\epsilon)} \|\partial_{VV\beta}\ell_i^*(\beta, \alpha_i)\| = O_p(1)$$

(ii)

$$\sup_{(\beta,\alpha_i)\in\mathcal{B}_0(\epsilon)} \max_i |\partial_{VVVV}\ell_i^*(\beta, \alpha_i)| = O_p(T^{1/(8+\nu)})$$

$$\sup_{(\beta,\alpha_i)\in\mathcal{B}_0(\epsilon)} \max_i \|\partial_{V\beta\beta'}\ell_i^*(\beta, \alpha_i)\| = O_p(T^{2/(8+\nu)})$$

$$\sup_{(\beta,\alpha_i)\in\mathcal{B}_0(\epsilon)} \max_i \|\partial_{VV\beta}\ell_i^*(\beta, \alpha_i)\| = O_p(T^{2/(8+\nu)})$$

Lemma 5. *Let* $V_{it}(\beta,\alpha) = \partial \ell_{it}(\beta,\alpha)/\alpha$ *and* $S_{itq}(\beta,\alpha) = \partial \ell_{it}(\beta,\alpha)/\partial \beta_q$. *Define the following:*

$$\begin{aligned}
\overline{V}_i(\beta,\alpha) &= \sum_{t=T-k}^{T} \left[(2D_{it}-1) V_{it}(\beta,\alpha) \left\{ \frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)} \right\}^{1-D_{it}} \right] \\
\overline{S}_{iq}(\beta,\alpha) &= \sum_{t=T-k}^{T} \left[(2D_{it}-1) S_{itq}(\beta,\alpha) \left\{ \frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)} \right\}^{1-D_{it}} \right] \\
\overline{V}_{i\alpha}(\beta,\alpha) &= \overline{V}_i(\beta,\alpha)^2 + \partial \overline{V}_i(\beta,\alpha)/\partial \alpha, \\
\overline{V}_{iq}(\beta,\alpha) &= \overline{V}_i(\beta,\alpha)\overline{S}_{iq}(\beta,\alpha) + \partial \overline{V}_i(\beta,\alpha)/\partial \beta_q, \\
\overline{S}_{iq\alpha}(\beta,\alpha) &= \overline{V}_i(\beta,\alpha)\overline{S}_{iq}(\beta,\alpha) + \partial \overline{S}_i(\beta,\alpha)/\partial \alpha, \\
\overline{S}_{iqm}(\beta,\alpha) &= \overline{S}_{iq}(\beta,\alpha)\overline{S}_{im}(\beta,\alpha) + \partial \overline{S}_{iq}(\beta,\alpha)/\partial \beta_m
\end{aligned}$$

Suppose Assumption 2 holds. Then each of these is uniformly bounded in absolute value for $(\beta,\alpha) \in \mathcal{B}_0(\epsilon)$ *by a function* $\widetilde{M}_i$ *such that* $\max_{i,t} \mathbb{E}[\widetilde{M}_i^4]$ *is almost surely uniformly bounded over* N.

B.2 Proof of Lemmas

Proof of Lemma 1. We take the convention here that any function with the t subscript omitted is the over-time average of that quantity. For example, $V_i(\beta,\alpha) = \mathbb{E}_T[V_{it}(\beta,\alpha)] = T^{-1}\sum_{t=1}^{T} V_i(\beta,\alpha)$.

Part (i) follows from the results of Fernández-Val and Weidner (2016) without period effects.

To derive an asymptotic expansion of $\widehat{\alpha}_i$ in part (ii), we largely follow the Legendre transformation approach of Fernández-Val and Weidner (2016). Our discussion follows theirs closely, though in a more specialized setting. We define

$$\ell_i^*(\beta,V) = \max_{\alpha \in \mathcal{B}_\alpha(\epsilon_\alpha)} [\ell_i(\beta,\alpha) - \alpha V], \qquad A_i(\beta,V) = \arg\max_{\alpha \in \mathcal{B}_\alpha(\epsilon_\alpha)} [\ell_i(\beta,\alpha) - \alpha V]$$

where $\beta \in \mathcal{B}_\beta(\epsilon_\beta)$. The function $\ell_i^*(\beta,V)$ is the Legendre transformation of the objective function $\ell_i(\beta,\alpha) = T^{-1}\sum_{t=1}^{T} \ell_{it}(\beta,\alpha)$. We use V to denote the dual parameter to α and $\ell_i^*(\beta,V)$ as the dual function to $\ell_i(\beta,\alpha)$. The relationship between α and V is one-to-one since the optimal $\alpha =$

$A_i(\beta, V)$ satisfies the first-order condition $V_i(\beta, \alpha) = V$, where $V_i(\beta, \alpha) = T^{-1} \sum_{t=1}^{T} V_{it}(\beta, \alpha)$. We can write $\ell_i^*(\beta, V) = \ell_i(\beta, A_i(\beta, V)) - A_i(\beta, V)V$ when $A_i(\beta, V)$ solves the FOC, $V_i(\beta, A_i(\beta, V_i)) = V_i$. Taking the derivative of the last identity on both sides of the equality gives:

$$[\partial_V A_i(\beta, V)][V_{i\alpha}(\beta, A_i(\beta, V))] = 1$$
$$V_{i\beta}(\beta, A_i(\beta, V)) + [\partial_\beta A_i(\beta, V)]V_{i\alpha}(\beta, A_i(\beta, V)) = 0$$

so we have

$$\partial_V A_i(\beta, V) = \frac{1}{V_{i\alpha}(\beta, A_i(\beta, V))}$$
$$\partial_\beta A_i(\beta, V) = -\frac{V_{i\beta}(\beta, A_i(\beta, V))}{V_{i\alpha}(\beta, A_i(\beta, V))}$$

When $V = 0$, then the optimization in ℓ_i^* is just over ℓ_i, so we have $A_i(\beta, 0) = \widehat{\alpha}_i(\beta)$ and $\ell_i^*(\beta, 0) = \ell_i(\beta, \widehat{\alpha}_i(\beta))$. The latter is the profile likelihood for β. Note that

$$\begin{aligned}\partial_V \ell_i^*(\beta, A_i(\beta, V)) &= [V_i(\beta, A_i(\beta, V))][\partial_V A_i(\beta, V)] - A_i(\beta, V) - [\partial_V A_i(\beta, V)]V \\ &= -A_i(\beta, V)\end{aligned}$$

Thus, we have $\widehat{\alpha}_i(\beta) = -\partial_V \ell_i^*(\beta, 0)$ and $\partial_V \ell_i^*(\beta_0, 0) = -\alpha_{i0}$.

We now expand $\partial_V \ell_i^*(\beta, 0)$ around (β_0, V_i) for $\beta \in \mathcal{B}_{\beta_0}(\epsilon_\beta)$, which gives:

$$\widehat{\alpha}_i(\beta) = -\partial_V \ell_i^*(\beta, 0) = -\partial_V \ell_i^* - (\partial_{V\beta'} \ell_i^*)(\beta - \beta_0) + (\partial_{VV} \ell_i^*)V_i - \frac{1}{2}(\partial_{VVV} \ell_i^*)V_i^2 + R(\beta),$$

where

$$\begin{aligned}R_i(\beta) = \frac{1}{2}(\beta - \beta_0)^\top (\partial_{V\beta\beta'} \ell_i^*(\overline{\beta}, V_i))(\beta - \beta_0) + (\partial_{VV\beta'} \ell_i^*(\beta_0, \widetilde{V}_i))(\beta - \beta_0)V_i \\ + \frac{1}{6}(\partial_{VVVV} \ell_i^*(\beta_0, \ddot{V}_i))V_i^3\end{aligned}$$

where $\overline{\beta}$ is between β and β_0 and $\widetilde{V}_i$ and $\ddot{V}_i$ are between V_i and 0.

Using the above identities, it is possible to derive the following:

$$\partial_{V}\ell_i^* = -1/V_{i\alpha}, \qquad \partial_{VVV}\ell_i^* = V_{i\alpha\alpha}/V_{i\alpha}^3, \qquad \partial_{\beta}\ell_i^* = S_i$$

$$\partial_{V\beta}\ell_i^* = V_{i\beta}/V_{i\alpha}, \qquad \partial_{\beta\beta'}\ell_i^* = S_{i\beta} + V_{i\beta}V_{i\beta}^{\top}/V_{i\alpha}$$

$$\partial_{V\beta\beta'}\ell_i^* = \frac{S_{i\beta\alpha}}{V_{i\alpha}} + \frac{V_{i\beta\alpha}V_{i\beta}^{\top}}{V_{i\alpha}^2} + \frac{V_{i\beta}V_{i\beta\alpha}^{\top}}{V_{i\alpha}^2} - \frac{V_{i\beta}V_{i\beta}^{\top}}{V_{i\alpha}^3}$$

$$\partial_{VV\beta}\ell_i^* = \frac{V_{i\beta\alpha}}{V_{i\alpha}^2} - \frac{V_{i\alpha\alpha}V_{i\beta}}{V_{i\alpha}^3}$$

$$\partial_{VVVV}\ell_i^* = \frac{V_{i\alpha\alpha\alpha}}{V_{i\alpha}^4} - 3\frac{V_{i\alpha\alpha}}{V_{i\alpha}^5}$$

Combining these with the above expansion gives us:

$$\widehat{\alpha}_i(\beta) - \alpha_{i0} = -\frac{V_i}{V_{i\alpha}} - \frac{V_{i\beta}^{\top}(\beta - \beta_0)}{V_{i\alpha}} + \frac{V_{i\alpha\alpha}V_i^2}{2V_{i\alpha}^3} + R(\beta)$$

Let $\mathbb{E}_\alpha$ is the expectation conditional on the fixed effects. We have $\mathbb{E}_\alpha[V_i] = 0$ and so by Lemma 3, we have $V_i = O_p(T^{-1/2})$, $V_i^2 = O_p(T^{-1})$, and $V_i^3 = O_p(T^{-3/2})$. For the partial derivative terms, we define $\overline{V}_{it\alpha} = \mathbb{E}_\alpha[V_{it\alpha}]$ and $\overline{V}_{i\alpha} = \mathbb{E}_T[\overline{V}_{it\alpha}]$. We also define the mean deviations as $\widetilde{V}_{i\alpha} = V_{i\alpha} - \overline{V}_{i\alpha}$. We define similar quantities for $V_{i\alpha\alpha}$ and $V_{i\beta}$. As above, we have $\widetilde{V}_{i\alpha} = O_p(T^{-1/2})$. We can derive the following using standard asymptotic results:

$$V_{i\alpha}^{-1} = \overline{V}_{i\alpha}^{-1} + \overline{V}_{i\alpha}^{-2}\widetilde{V}_{i\alpha} + O_p(T^{-1})$$

$$V_{i\alpha}^{-2} = \overline{V}_{i\alpha}^{-2} + \overline{V}_{i\alpha}^{-3}\widetilde{V}_{i\alpha} + O_p(T^{-1})$$

$$V_{i\alpha}^{-3} = \overline{V}_{i\alpha}^{-3} + \overline{V}_{i\alpha}^{-4}\widetilde{V}_{i\alpha} + O_p(T^{-1})$$

Let $\psi_{i1} = V_i/\overline{V}_{i\alpha}$. We have $V_i/V_{i\alpha} = \psi_{i1} + \psi_{i1}(\widetilde{V}_{i\alpha}/\overline{V}_{i\alpha}) + O_p(T^{-3/2})$. For the next two terms,

we have

$$\frac{V_{i\beta}^{\top}(\beta-\beta_0)}{V_{i\alpha}} = \frac{V_{i\beta}^{\top}(\beta-\beta_0)}{\overline{V}_{i\alpha}} + O_P(T^{-1/2}\|\beta-\beta_0\|)$$
$$\frac{V_{i\alpha\alpha}V_i^2}{2V_{i\alpha}^3} = \frac{V_{i\alpha\alpha}V_i^2}{2\overline{V}_{i\alpha}^3} + O_P(T^{-3/2}) = \frac{V_{i\alpha\alpha}\psi_{i1}^2}{2\overline{V}_{i\alpha}} + O_P(T^{-3/2})$$

Again based on the bounded derivative and moment conditions and Lemma 3, we have $\widetilde{V}_{i\alpha\alpha} = O_P(T^{-1/2})$ and $\|\widetilde{V}_{i\beta}\| = O_p(T^{-1/2})$. Combined with the above, we have:

$$\widehat{\alpha}_i(\beta) - \alpha_{i0} = -\psi_{i1} - \psi_{i2} + R_i(\beta) + O_p\left(T^{-1/2}\|\beta-\beta_0\| + T^{-3/2}\right),$$

where

$$\psi_{i2} = \frac{1}{\overline{V}_{i\alpha}}\left(\widetilde{V}_{i\alpha}\psi_{i1} + \overline{V}_{i\beta}^{\top}(\beta-\beta_0) + \overline{V}_{i\alpha\alpha}\psi_{i1}^2\right),$$

and $\psi_{i2} = O_p(T^{-1})$.

We now plug in $\widehat{\beta}$ to get $\widehat{\alpha}_i = \widehat{\alpha}_i(\widehat{\beta})$. Note that $\|\widehat{\beta}-\beta_0\| = O_p((NT)^{-1/2}) = O_p(T^{-1})$. Now by part *(i)* of Lemma 4 we have

$$\begin{aligned}|R_i(\widehat{\beta})| &\le \|\widehat{\beta}-\beta_0\|^2\|\partial_{V\beta\beta'}\ell_i^*(\overline{\beta},V_i)\| + |V_i|\ \|\widehat{\beta}-\beta_0\|\ \|\partial_{VV\beta'}\ell_i^*(\beta_0,\widetilde{V}_i)\| \\ &\quad + \left|V_i^3\right|\left|\partial_{VVVV}\ell_i^*(\beta_0,\ddot{V}_i)\right| \\ &= O_P(T^{-2})O_p(1) + O_p(T^{-1/2})O_p(T^{-1})O_p(1) + O_p(T^{-3/2})O_p(1) = O_P(T^{-3/2})\end{aligned}$$

Combining this with the above, we have

$$\widehat{\alpha}_i - \alpha_i = -\psi_{i1} - \psi_{i2} + O_p(T^{-3/2})$$

For part (iii), we now derive a maximal inequality over units. We have

$$\max_i|\widehat{\alpha}_i - \alpha_{i0}| \le \max_i|\psi_{i1}| + \max_i|\psi_{i2}| + \max_i|R_i(\beta)|$$

By Lemma 3, we have:

$$\begin{aligned}\mathbb{E}_\alpha\left[\left(\max_i |V_i|\right)^8\right] &= T^{-4}\mathbb{E}_\alpha\left[\max_i\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right)^8\right] \\ &\leq T^{-4}\sum_i \mathbb{E}_\alpha\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right)^8\right] \\ &\leq T^{-4}NB = O(T^{-3})\end{aligned}$$

Thus, we have $\max_i |V_i| = O_p(T^{-3/8})$, $\max_i |V_i^2| = O_p(T^{-3/16})$, and $\max_i |V_i^3| = O_p(T^{-9/8})$. Furthermore, recall that $\widehat{\alpha}_i = \widehat{\alpha}_i(\widehat{\beta})$ and that $\|\widehat{\beta} - \beta_0\|$ is $O_p((NT)^{-1/2}) = O_p(T^{-1})$. Finally, recall that $\inf_i |V_{it\alpha}| > 0$, which implies that $\inf_i |\overline{V}_{i\alpha}| > 0$. These facts combined with part *(ii)* of Lemma 4 imply for some constants C_1

$$\begin{aligned}\max_i|\psi_{i1}| &= \max_i\left|\overline{V}_{i\alpha}^{-1}V_i\right| < C\max_i|V_i| = O_p(T^{-3/8}) \\ \max_i|\psi_{i2}| &= \max_i\left|\overline{V}_i^{-1}\left(\overline{V}_{i\beta}(\widehat{\beta}-\beta_0) + \overline{V}_{i\alpha\alpha}\psi_{i1}^2\right)\right| \\ &< C\left(\|\widehat{\beta}-\beta_0\|\max_i\|\overline{V}_{i\beta}\| + \max_i\left|\overline{V}_{i\alpha\alpha}\right|\max_i|\psi_{i1}^2|\right) \\ &= O_p(T^{-1}+T^{-3/4}) = o_p(T^{-1/2}), \\ \max_i|R_i(\widehat{\beta})| &= o_p(T^{-1}),\end{aligned}$$

Here, we used the fact that $\max_i |\overline{V}_{i\alpha\alpha}| < E_\alpha[M(Z_{it})]$, which is uniformly bounded over i and t. These three combined implies $\max_i|\widehat{\alpha}_i - \alpha_{i0}| = O_p(T^{-3/8})$. □

Proof of Lemma 4. Part *(i)*. Let ξ_{it} be one of $V_{it\beta_k}$, $V_{it\alpha\alpha}$, $V_{it\alpha\alpha\alpha}$, $V_{i\beta_k\alpha}$, or $V_{it\beta_k\beta_j\alpha}$ and note that

$E[|\xi_{it}|^{8+\nu}] < E[M(Z_{it})^{8+\nu}] < \infty$ by Assumption 3(ii). We have:

$$\begin{aligned}\mathbb{E}_\alpha\left[\left(\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\frac{1}{T}\sum_{t=1}^{T}|\xi_{it}(\beta,\alpha)|\right)^{(8+\nu)}\right] &\leq \mathbb{E}_\alpha\left[\left(\frac{1}{T}\sum_{t=1}^{T}M(Z_{it})\right)^{(8+\nu)}\right] \\ &\leq \mathbb{E}_\alpha\left[\frac{1}{T}\sum_{t=1}^{T}M(Z_{it})^{(8+\nu)}\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_\alpha[M(Z_{it})^{(8+\nu)}] = O_p(1)\end{aligned}$$

Thus, $\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\frac{1}{T}\sum_{t=1}^{T}|\xi_{it}(\beta,\alpha)| = O_p(1)$. From the expression for $\partial_{VV\beta}$ given above, we have

$$\begin{aligned}\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\left|\partial_{VV\beta_k}\ell_i^*(\beta,\alpha)\right| &\leq \sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\left(\left|\frac{V_{i\beta_k\alpha}(\beta,\alpha)}{V_{i\alpha}(\beta,\alpha)^2}\right| + \left|\frac{V_{i\alpha\alpha}(\beta,\alpha)V_{i\beta_k}(\beta,\alpha)}{V_{i\alpha}^3(\beta,\alpha)}\right|\right) \\ &< \sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\left(\left|V_{i\beta_k\alpha}(\beta,\alpha)\right| + \left|V_{i\alpha\alpha}(\beta,\alpha)V_{i\beta_k}(\beta,\alpha)\right|\right) \\ &= O_p(1)\end{aligned}$$

The second inequality follows from $\inf_i |V_{it\alpha}(\beta,\alpha)| > 0$. The other statements in part *(i)* follow analogously.

For the maximal results in part *(ii)*, we follow a similar strategy:

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\max_i\frac{1}{T}\sum_{t=1}^{T}|\xi_{it}(\beta,\alpha)|\right)^{(8+\nu)}\right] &= \mathbb{E}\left[\max_i\left(\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\frac{1}{T}\sum_{t=1}^{T}|\xi_{it}(\beta,\alpha)|\right)^{(8+\nu)}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{N}\left(\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\frac{1}{T}\sum_{t=1}^{T}|\xi_{it}(\beta,\alpha)|\right)^{(8+\nu)}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}M(Z_{it})\right)^{(8+\nu)}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}M(Z_{it})^{(8+\nu)}\right]\end{aligned}$$

$$= \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[M(Z_{it})^{(8+\nu)}] = O(N)$$

Thus, $\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)} \max_i \frac{1}{T}\sum_{t=1}^{T} |\xi_{it}(\beta,\alpha)| = O_p(N^{1/(8+\nu)}) = O_p(T^{1/(8+\nu)})$. From above we have,

$$\begin{aligned}
&\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)} \max_i \left|\partial_{VV\beta_k} \ell_i^*(\beta,\alpha)\right| \\
&\quad \leq \sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)} \max_i \left(\left| \frac{V_{i\beta_k\alpha}(\beta,\alpha)}{V_{i\alpha}(\beta,\alpha)^2} \right| + \left| \frac{V_{i\alpha\alpha}(\beta,\alpha) V_{i\beta_k}(\beta,\alpha)}{V_{i\alpha}^3(\beta,\alpha)} \right| \right) \\
&\quad < \sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)} \max_i \left(\left|V_{i\beta_k\alpha}(\beta,\alpha)\right| + \left|V_{i\alpha\alpha}(\beta,\alpha) V_{i\beta_k}(\beta,\alpha)\right| \right) \\
&\quad = O_p\left(T^{1/(8+\nu)} + T^{2/(8+\nu)}\right) = O_p(T^{2/(8+\nu)})
\end{aligned}$$

The other results follow analogously. □

Proof of Lemma 5. We prove this for $\overline{V}_i(\beta,\alpha)$ and $\overline{V}_{i\alpha}(\beta,\alpha)$, the rest follow from very similar arguments. Let C_π be a uniform bound on $(\pi_{it}(\beta,\alpha)/(1-\pi_{it}(\beta,\alpha)))$ in $(\beta,\alpha) \in \mathcal{B}_0(\epsilon)$, which exists by the virtue of bounded propensity scores (Assumption 3(iii)). Furthermore, let $M_i = \max_t M(Z_{it})$ be a uniform bound for all $V_{it}(\beta,\alpha)$, which exists due to Assumption 3(ii). Then, within $\mathcal{B}_0(\epsilon)$, for $\overline{V}_i$ we have:

$$\begin{aligned}
\left|\overline{V}_i(\beta,\alpha)\right| &= \left| \sum_{t=T-k}^{T} \left[(2D_{it}-1) V_{it}(\beta,\alpha) \left\{ \frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)} \right\}^{1-D_{it}} \right] \right| \\
&\leq \sum_{t=T-k}^{T} |V_{it}(\beta,\alpha)| \left\{ \frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)} \right\}^{1-D_{it}} \\
&< C_\pi \sum_{t=T-k}^{T} |V_{it}(\beta,\alpha)| < C_\pi k M_i
\end{aligned}$$

Thus, $\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)} \left|\overline{V}_i(\beta,\alpha)\right| < \widetilde{M}_i = C_\pi k M_i$. Since $\mathbb{E}[\widetilde{M}_i^4] = C_\pi^4 k^4 \mathbb{E}[M_i^4]$, we obtain the result for the first quantity.

For $\overline{V}_{i\alpha}$ there are two terms. For the first term, we have

$$\begin{aligned}
\left|\overline{V}_i^2(\beta,\alpha)\right| &= \left|\sum_{t=T-k}^{T}\sum_{s=T-k}^{T}(2D_{it}-1)(2D_{is}-1)V_{it}(\beta,\alpha)V_{is}(\beta,\alpha)\right. \\
&\qquad \left.\times\left(\frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)}\right)^{1-D_{it}}\left(\frac{\pi_{is}(\beta,\alpha)}{1-\pi_{is}(\beta,\alpha)}\right)^{1-D_{is}}\right| \\
&\leq \sum_{t=T-k}^{T}\sum_{s=T-k}^{T}|V_{it}(\beta,\alpha)V_{is}(\beta,\alpha)|\left(\frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)}\right)^{1-D_{it}}\left(\frac{\pi_{is}(\beta,\alpha)}{1-\pi_{is}(\beta,\alpha)}\right)^{1-D_{is}} \\
&< C_\pi^2\sum_{t=T-k}^{T}\sum_{s=T-k}^{T}|V_{it}(\beta,\alpha)V_{is}(\beta,\alpha)| \\
&< C_\pi^2 k^2 M_i^2
\end{aligned}$$

Now, for the second term, we have:

$$\begin{aligned}
\frac{\partial\overline{V}_i(\beta,\alpha)}{\partial\alpha} &= \sum_{t=T-k}^{T}(2D_{it}-1)V_{it\alpha}(\beta,\alpha)\left(\frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)}\right)^{1-D_{it}} \\
&\qquad + \sum_{t=T-k}^{T}\frac{(1-D_{it})}{1-\pi_{it}(\beta,\alpha)}V_{it}(\beta,\alpha)^2\left(\frac{\pi_{it}(\beta,\alpha)}{1-\pi_{it}(\beta,\alpha)}\right)^{1-D_{it}}
\end{aligned}$$

Let $\overline{C}_\pi > (1-\pi_{it}(\beta,\alpha))^{-1}$, which exists by Assumption 3(iii). By the above argument, we have:

$$\left|\frac{\partial\overline{V}_i(\beta,\alpha)}{\partial\alpha}\right| < C_\pi k M_i + \overline{C}_\pi C_\pi k M_i^2$$

Then, we have

$$\sup_{(\beta,\alpha)\in\mathcal{B}_0(\epsilon)}\left|\overline{V}_{i\alpha}(\beta,\alpha)\right| < \widetilde{M}_i = (C_\pi^2 k^2 + \overline{C}_\pi C_\pi k)M_i^2 + C_\pi k M_i$$

and

$$\mathbb{E}[\widetilde{M}_i^4] = C_1\mathbb{E}[M_i^8] + C_2\mathbb{E}[M_i^7] + C_3\mathbb{E}[M_i^6] + C_4\mathbb{E}[M_i^5] + C_5\mathbb{E}[M_i^4],$$

where C_1 through C_5 are constants that depend on C_π, $\overline{C}_\pi$, and k. By Assumption 4(i), each of the expectations on the right-hand side is uniformly bounded in N, which implies that $\mathbb{E}[\widetilde{M}_i^4]$ is also uniformly bounded in N. □

C Additional Simulation Results

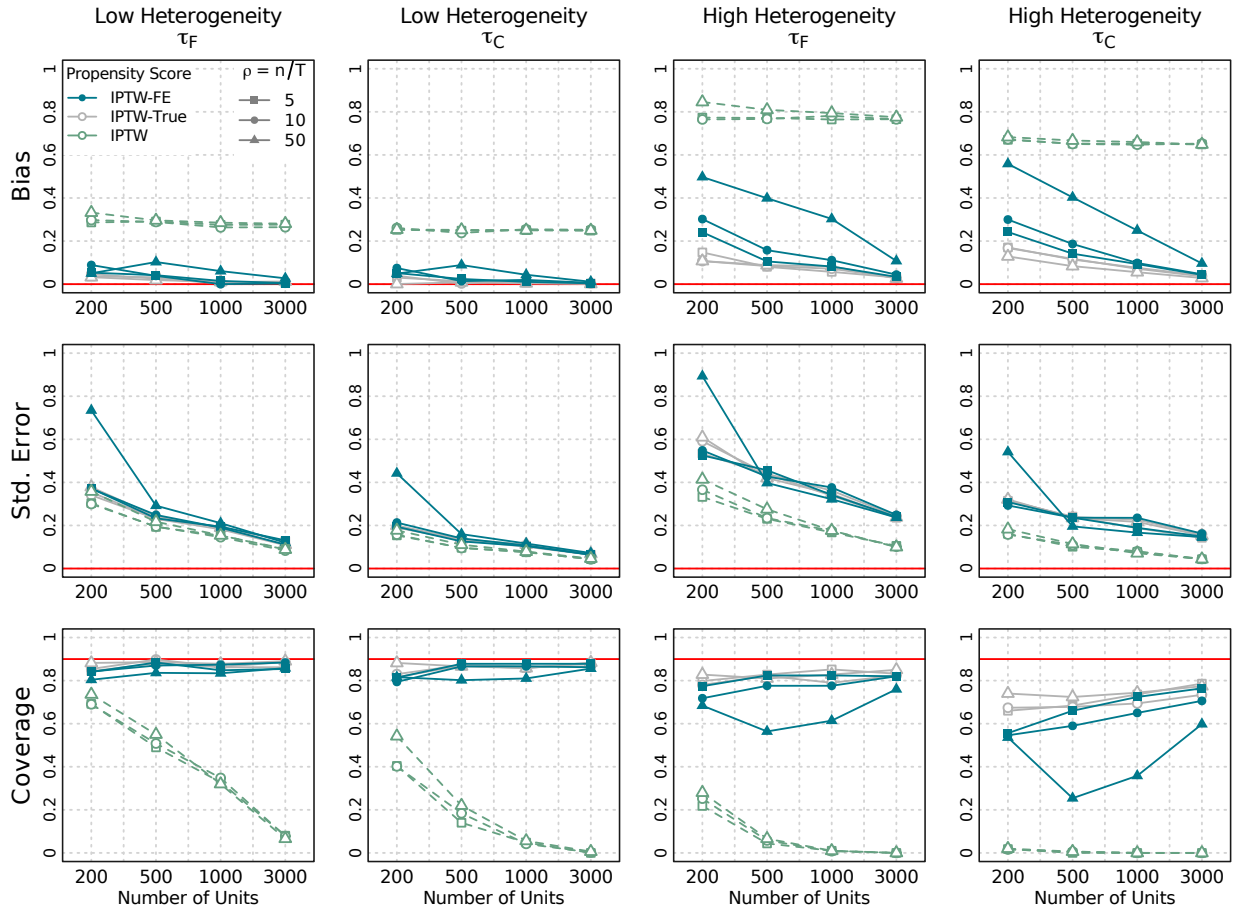

Figure 7: Bias, standard error (Std. Error) and coverage probability of 90% confidence intervals (Coverage) for the estimation of the final period effect τ_F and the cumulative effect τ_C under the "low" heterogeneity ($a = 1$) – first two columns – and the "high" heterogeneity ($a = 2$) – last two columns – scenario. The number of time-varying covariates is four. Solid lines in blue show the proposed estimator (`IPTW-FE`), solid lines in grey show the estimator based on the true propensity score (`IPTW-True`), and dashed lines in green show the estimator based on the estimated propensity score without fixed effects (`IPTW`). Shapes correspond to the n to T ratio ρ such that squares represent $\rho = 5$ (the largest number of time periods), circles represent $\rho = 10$, and triangles represent $\rho = 50$ (the smallest number of time periods)

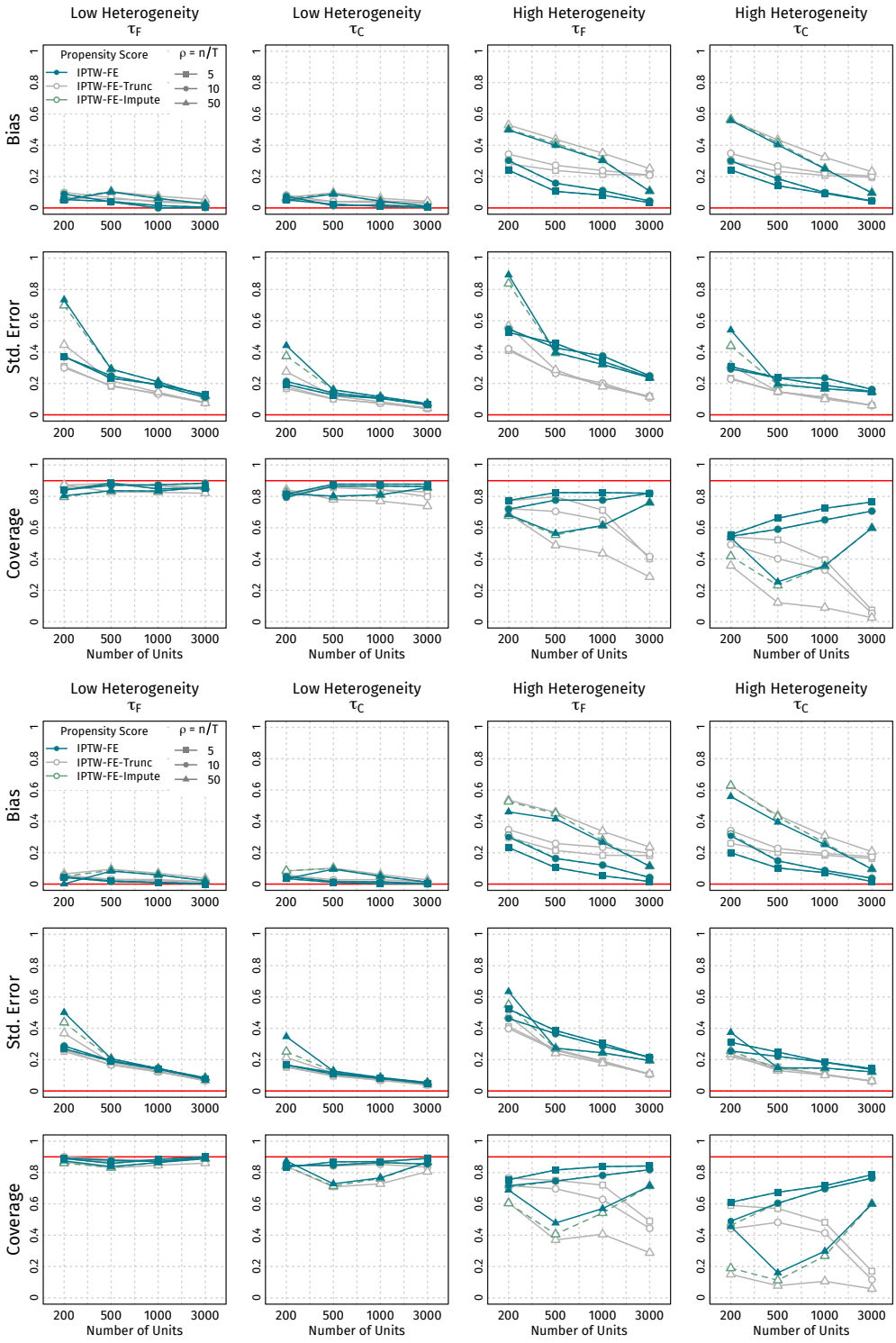

Figure 8: Simulation results for imputing the non-identified fixed effect estimates.

D Additional Empirical Results

Here we present the main results when including all market-races that have no variation in their treatment over time. To do so, we take the maximum and minimum values of $\widehat{\alpha}_i$ from our propensity score model and use them as the α_i for the units that have are always treated or never treated, respectively. Figure 9 shows these results, which are very similar to the original findings, at least for our approach. We can see that the regular IPTW approach is no longer significant for the Democratic IG ads.

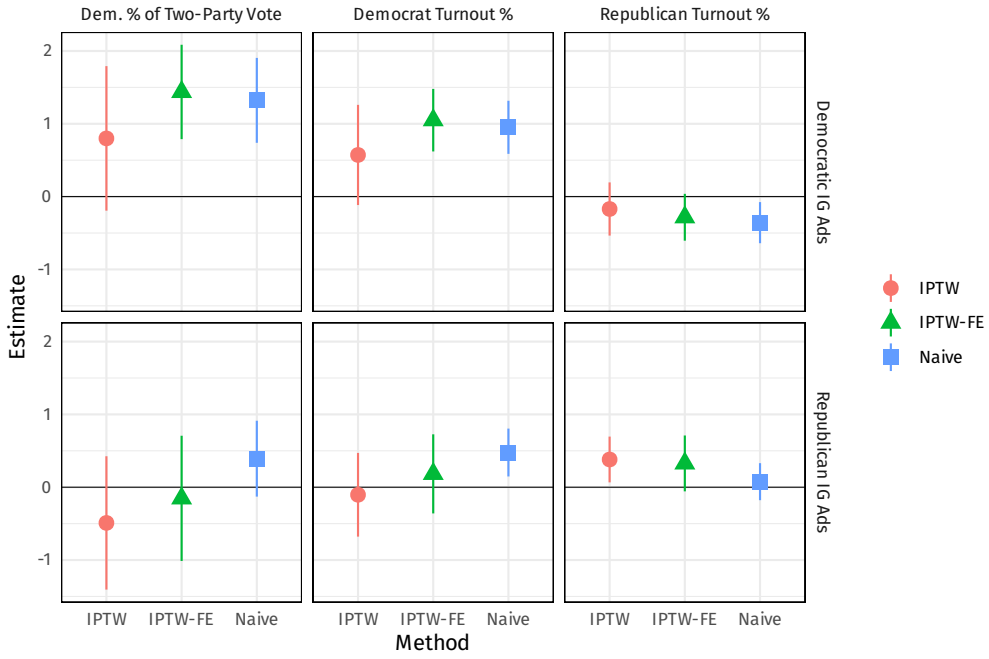

Figure 9: Estimated effects from the MSM where the IPTW weights are imputed for no-variation in treatment market-races.