Robust inference for matching under rolling enrollment

Amanda K. Glazer and Samuel D. Pimentel*

May 3, 2022

Abstract

Matching in observational studies faces complications when units enroll in treatment on a rolling basis. While each treated unit has a specific time of entry into the study, control units each have many possible comparison, or "pseudo-treatment," times. The recent GroupMatch framework (Pimentel et al., 2020) solves this problem by searching over all possible pseudo-treatment times for each control and selecting those permitting the closest matches based on covariate histories. However, valid methods of inference have been described only for special cases of the general GroupMatch design, and these rely on strong assumptions. We provide three important innovations to address these problems. First, we introduce a new design, GroupMatch with instance replacement, that allows additional flexibility in control selection and proves more amenable to analysis. Second, we propose a block bootstrap approach for inference in GroupMatch with instance replacement and demonstrate that it accounts properly for complex correlations across matched sets. Third, we develop a permutation-based falsification test to detect possible violations of the important timepoint agnosticism assumption underpinning GroupMatch, which requires homogeneity of potential outcome means across time. Via simulation and a case study of the impact of short-term injuries on batting performance in major league baseball, we demonstrate the effectiveness of our methods for data analysis in practice.

1 Introduction

Quantifying the impact of injury on player performance in professional sports is important for both managers and players themselves. Increasingly, players are valued and compensated in a manner driven by quantitative metrics of past performance, but injuries have potential to

*Amanda K. Glazer is a doctoral candidate and Samuel D. Pimentel is an Assistant Professor in the Statistics Department at University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720. Correspondence should be addressed to amandaglazer@berkeley.edu. Glazer acknowledges support from the National Science Foundation grant (DMS RTG #1745640). The major league baseball player information used in the data analysis was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at www.retrosheet.org. We thank the author and maintainer of the GitHub repository https://github.com/robotallie/baseball-injuries for making the injury data for the data analysis easily available. We also thank Eli Ben-Michael, Peng Ding, Avi Feller, Lauren Forrow, Shirshendu Ganguly, and Jiaqi Li for helpful conversations and feedback.

disrupt the continuity between past and future performance (Begly et al., 2018; Conte et al., 2016; Frangiamore et al., 2018; Wasserman et al., 2015). How should a player's expected future value to a team be adjusted in the event of injury? One way to quantify impact in this setting is as the difference between the value of a performance metric the player would have achieved in the absence of injury and the value of the same metric achieved after a given injury. Unfortunately, even post hoc, at most one of these two counterfactual outcomes is observed. This phenomenon, often known as the fundamental problem of causal inference (Holland and Rubin, 1980), is also present in impact evaluation settings across social science and medicine

One approach to impact evaluation is matching, in which individuals experiencing a treatment or condition of interest (in our case, injuries) are paired to otherwise similar control individuals who did not experience the treatment. Assuming that paired individuals are sufficiently similar on observed attributes and that no important unobserved attributes confound the comparison, the difference in outcomes approximates the impact of treatment for individuals in the pair (Stuart, 2010). When controls are plentiful, each treated unit may be matched to multiple controls, forming matched sets instead of matched pairs. In Section 7 we use this strategy to evaluate the impact of minor injury on batting performance in Major League Baseball (MLB), comparing on-base percentage between players who recently spent a short period of time on their team's injured list (IL) and otherwise similar players who did not go on the IL.

In contrast to common matched studies, the treatment in our setting is not given to all individuals at the same time. Instead, each player is observed repeatedly in games throughout the MLB season, and each treated player experiences injury at a different point in time. The longitudinal structure of the data and rolling nature of entry into treatment create complications for pairing injured players to uninjured controls. Specifically, each injured player has a date of entry onto the IL, and a similarly well-defined follow-up date at which performance is assessed based on a certain period elapsing after return from the IL. However, dates of treatment (and hence follow-up) are not defined in the data for control individuals, and the matching process involves not only selecting which control individuals will be paired to injured players but at which point in time the control unit will be measured. Two primary strategies exist for aligning control individuals in time, or equivalently selecting "pseudo-treatment" dates at which counterfactual outcomes for controls will be considered. The most common strategy is to compare players at the same point in time; for instance, if a treated player experiences injury on August 1st, then he is compared to control units only as they appear on August 1st. This approach is conceptually straightforward, although it requires specialized software as described in Witman et al. (2019); however, it implicitly prioritizes calendar time itself as the most important dimension of similarity between units, and in settings where time is not an especially important confounder it has the effect of arbitrarily limiting the pool of potential control comparisons.

The second approach is to allow comparisons between a treated unit at one calendar time and a control unit at a potentially different calendar time. For instance, a player's attributes and to-date performance in the MLB season at his time of injury on August 1 may more closely resemble the attributes and to-date performance of a an uninjured player on July 1 than it does any other player in his August 1 state, and under the second approach the first player at August 1 could be matched to the second at July 1. The resulting flexibility

has the potential to greatly improve similarity between matched units on measured variables besides time. The recent GroupMatch algorithm (Pimentel et al., 2020) constructs matches optimally across time in this manner. Our investigation focuses on this second method of matching.

In introducing the GroupMatch framework, Pimentel et al. (2020) grappled with several challenges. Whenever multiple controls are paired to a single treated unit, the presence of multiple copies of the same control individual necessitates a constraint to ensure that a treated unit is not simply paired to multiple slightly different copies of the same control. Two possible constraints were suggested but methods for inference were presented under only under one of them, in which multiple copies of a control individual are forbidden from appearing in the matched design. Furthermore, the guarantees given for causal effect estimation using GroupMatch designs rely heavily on a strong assumption that time itself is not a confounder.

In what follows, we address these challenges and provide additional tools to enhance GroupMatch. First, we introduce a new type of constraint on repeated use of control information within a GroupMatch design. This constraint has computational, analytical, and statistical advantages over existing constraints in many common settings. Secondly, we introduce a new block-bootstrap-based method for inference that applies to any GroupMatch design, motivated by related work on inference for cross-sectional matching designs by Otsu and Rai (2017). Finally, we introduce a falsification test to partially check the assumption of time agnosticism underpinning GroupMatch's validity, empowering investigators to extract evidence from the data about this key assumption prior to matching. We prove the validity of our bootstrap method under the most relevant set of constraints on reuse of controls, and we demonstrate the effectiveness of both the placebo test and the bootstrap inference approach through simulations and an analysis of MLB injury data. In particular, the bootstrap method shows similar performance to linear-regression-based approaches to inference often applied in similar settings, while making much weaker assumptions.

The paper is organized as follows. Section 2 presents the basic statistical framework and reviews the GroupMatch framework, inference approaches for matching designs, and other related literature. In Section 3 we introduce a new constraint for use of controls in GroupMatch designs, leading to a new design called GroupMatch with instance replacement. Section 4 presents a block bootstrap inference approach for GroupMatch. We apply our block bootstrap inference to a simulation study in Section 5. In Section 6 we present a falsification test for the assumption that time is not a confounder. In Secton 7 we revist our baseball example and evaluate whether short-term injury impacts short term MLB performance. We conclude with a discussion in Section 8.

2 Preliminaries

2.1 Matching in longitudinal settings and GroupMatch

Matching methods attempt to estimate average causal effects by grouping each treated unit with one or more otherwise similar controls and using paired individuals to approximate the missing potential outcomes. A number of other authors have considered matching in datasets

containing repeated measures for the same individuals over time. Some focus on the case in which only a single time of treatment is present, and the primary challenge is deciding how to construct matching distances from pre-treatment repeated measures and assess outcomes using post-treatment repeated measures. For example, in Haviland et al. (2008), the authors choose as a treatment the act of joining a gang at age 14, the age requirement ensuring that there is a single potential time of treatment for all individuals. The situation is more complex when individuals opt in to treatment at different times as in Li et al. (2001); Lu (2005); Witman et al. (2019); and Imai et al. (2020). These authors address the problem by matching each treated unit to the version of the control unit present in the data at the time of treatment. For example, in Imai et al. (2020)'s reanalysis of data from Acemoglu et al. (2019) on the impact of democratization on economic growth, countries undergoing democratizing political reforms are matched to similar control countries not undergoing such reforms in the same year. Although this method is logical whenever strong time trends are present, in other cases it may overemphasize similarity on time at the expense of other variables. For example, Bohl et al. (2010) study the impact of serious falls on subsequent healthcare expenditures for elderly adults using patient data from a large health system. While patients who fall could be matched to patients who appear similar based on recent health history on the calendar date of the fall, the degree of similarity in health histories is probably much more important than the similarity of the exact date at which each patient is measured.

GroupMatch is a new framework for matching in longitudinal settings with rolling entry into treatment that relaxes the assumption that treated units must be matched to control units at the same time (Pimentel et al., 2020). The relaxation of this assumption yields higher quality matches on other variables of interest. We focus on GroupMatch designs in what follows.

In brief, GroupMatch designs are solutions to a discrete optimization problem that constructs matched sets, each with the same number of control units, with maximum overall similarity on pre-treatment attributes between a treated unit at the time of treatment and controls, choosing freely among different possible pseudo-treatment times for controls. While GroupMatch does not require units in the same matched set to be aligned at identical timepoints, it does impose constraints on how often control information can be reused within the match. Pimentel et al. (2020) outline two possible specific forms for this constraint. In GroupMatch without replacement (a setup referred to by the original authors as Design A), each control individual may contribute at most one version of itself to any matched set. In GroupMatch with trajectory replacement (Design B in the original paper), multiple copies-in-time of a control individual may appear in the match, but no individual copy may be used twice and no two copies of the same individual may appear in the same matched set. For further discussion of these constraints, see Section 3.

2.2 Sampling framework

We observe n subjects. For each subject i in the study, we observe repeated measures $(Y_{i,t}, Z_{i,t}, \mathbf{X}_{i,t})$ for timepoints $t = 1, \ldots, T$, where $Y_{i,t}$ is an outcome of interest, $Z_{i,t}$ is equal to the number of timepoints since subject i entered treatment (inclusive of t) or equal to zero if i has not yet been treated, and $\mathbf{X}_{i,t}$ is a vector of covariates. We denote the collection

of repeated measures for each subject i as the trajectory O_i.For convenience, we also define T_i as the time t for which $Z_{i,t} = 1$ (or ∞ otherwise) and D_i as an indicator for $T_i < \infty$. We specify a burn-in period of length $L-1$ during which no individuals are treated (or allow treatment at $t = 1$ by setting $L = 1$). Let $Y_{i,t}(z)$ (with $z \leq \max\{t - L + 1, 0\}$) be the potential outcome for unit i at time t if it had been enrolled in treatment for z timepoints.We will focus on assessing outcomes at a fixed follow-up period after treatment; for simplicity of exposition, we focus on the case in which the length of this follow-up period is zero, so that outcomes are observed immediately following treatment at the same timepoint. Although in principle there are up to $t-L+2$ potential outcomes $Y_{i,t}(z)$ for each i and $t \geq L$ by choosing different z-values, we will restrict attention to the two potential outcomes $Y_{i,t}(1)$ and $Y_{i,t}(0)$. These potential outcomes allow us to define the finite sample average effect of the treatment on the treated (ATT), denoted by Δ:

$$\begin{aligned}\Delta &= \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\{t = T_i\}[Y_{i,t}(1) - Y_{i,t}(0)] \\ &= \frac{1}{N_1}\sum_{i=1}^{N} D_i\left[Y_{i,t=T_i}(1) - Y_{i,t=T_i}(0)\right]\end{aligned}$$

Finally, we consider the data-generating process. We assume that trajectories O_i are sampled independently from some infinite population, although we do not assume independence of observations within the same trajectory. Defining expectation $E(\cdot)$ with respect to sampling from this population, we may now also define the population version of the ATT as $\Delta_{pop} = E(\Delta)$. For future convenience, we also introduce a concise notation for conditional expectation (again, over the sampling distribution) of potential outcomes given no treatment through time t and the covariates observed in the previous L timepoints:

$$\mu_z^t(\mathbf{X}) = E[Y_{i,t}(z)|\{X_{i,t'}\}_{t'=t-L+1}^{t'=t} = \mathbf{X}, T_i > t]$$

Throughout the paper we abuse notation slightly by writing $\mu_0(\mathbf{X}_{i,t})$ to indicate conditional expectation given the L lagged values of $\mathbf{X}_i$ directly preceding time t.

2.3 Identification assumptions

Pimentel et al. (2020) studied the following difference-in-means estimator in GroupMatch designs where each treated unit is matched to C control observations. $M_{it,jt'}$ is an indicator for whether subject i at time t has been matched to subject j at time t':

$$\hat{\Delta} = \frac{1}{N_1}\sum_{i=1}^{n} D_i[Y_{i,t=T_i} - \frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}Y_{j,t'}]$$

Pimentel et al. (2020) show that this estimator is unbiased for the population ATT under the following conditions:

1. Exact matching:

   Matched units share identical values for covariates in the L timepoints preceding treatment.

2. L-ignorability:

   Conditional on the covariate history over the previous L timepoints and any treatment enrollment in or prior to baseline, an individual's potential outcome at a given time is independent of the individual's treatment status. Formally,

$$D_i \perp\!\!\!\perp Y_{i,t}(0) | Z_{i,t}, \{X_{i,s}\}_{s=t-L+1}^{t}, \forall i.$$

3. Timepoint agnosticism: mean potential outcomes under control do not differ for any instances with identical covariate histories at different timepoints. Formally, for any set of L covariate values $\mathbf{X}$,

$$\mu_0^t(\mathbf{X}) = \mu_0^{t'}(\mathbf{X}) = \mu_0(\mathbf{X}) \text{ for any } 1 \leq t, t' \leq T.$$

   For simplicity of notation we will drop the t superscript when discussing the conditional expectation $\mu_0(\mathbf{X})$ for the sequel, with the exception of Section 6 where we temporarily consider failures of this assumption.

4. Covariate L-exogeneity: future covariates do not encode information about the potential outcome at time t given covariates and treatment status over the previous L timepoints. Formally,

$$(X_{i,1}, ..., X_{i,T}) \perp\!\!\!\perp Y_{i,t}(0) | (Z_{i,t}, \{X_{i,s}\}_{s=t-L+1}^{t}), \forall i.$$

5. Overlap condition: given that a unit is not yet treated at time $t-1 \geq L$, the probability of entering treatment at the next time point is neither 0 nor 1 for any choice of covariates over the L timepoints at and preceding t.

$$0 < P(T_i = t \mid T_i > t-1, X_i^t, \ldots, X_i^{t-L+1}) < 1 \qquad \forall t > L$$

   While this assumption is not stated explicitly in Pimentel et al. (2020), we note that the authors rely implicitly on an overlap assumption of this type in the proof of their main result.

Assumption 1 is no longer needed for asymptotic identification of the population ATT if we modify the estimator by adding in a bias correction term. As in Otsu and Rai (2017) and Abadie and Imbens (2011), we first estimate the conditional mean function $\mu_0(\mathbf{X})$ of the potential outcomes and use this outcome regression to adjust each matched pair for residual differences in covariates not addressed by matching. As outlined in Abadie and Imbens (2011), bias correction leads to asymptotic consistency under regularity conditions on the potential outcome mean estimator $\widehat{\mu}(\cdot)$ (for further discussion of regularity assumptions on $\widehat{\mu}_0(\cdot)$ see the proof of Theorem 1 in the Appendix). Many authors have also documented benefits from adjusting matched designs using outcome models (Rubin, 1979; Antonelli et al., 2018). The specific form of our bias-corrected estimator is as follows:

$$\hat{\Delta}_{adj} = \frac{1}{N_1} \sum_{i=1}^{n} D_i[(Y_{i,t=T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - \frac{1}{C} \sum_{j=1}^{N} \sum_{t'=1}^{T} M_{iT_i,jt'}(Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'}))]$$

Large datasets with continuous variables ensure that exact matching is rarely possible in practice, and in light of this we focus primarily on estimator $\widehat{\Delta}_{adj}$ in what follows.

3 GroupMatch with instance replacement

Before discussing our method for inference in general GroupMatch designs, we introduce a new type of GroupMatch design. Pimentel et al. (2020) described two different types of designs produced by GroupMatch denoted Problems A and B, designs we refer to as GroupMatch without replacement and GroupMatch with trajectory replacement respectively.

1. **GroupMatch without replacement**: each control unit can be matched to at most one treated unit. This means that if a treated unit is matched to an instance of a control unit, no other treated unit can match to (any instance) of that control unit.

2. **GroupMatch with trajectory replacement**: each control *instance* can be matched to at most one treated unit. Each treated unit can match to no more than one instance from the same control trajectory. However different treated units can match to different instances of the same control trajectory, so a single control trajectory can contribute multiple distinct instances to the design.

As our chosen names for these designs suggest, their relative costs and benefits are similar to the relative costs and benefits of matching without and with replacement in cross-sectional settings. As discussed by Hansen (2004), matching without replacement (in which each control may appear in at most one matched set), leads to slightly less similar matches compared to matching with replacement (in which controls can reappear in many matched sets) since in cases where two treated units both share the same nearest control only can use it. On the other hand, matching without replacement frequently leads to estimators with lower variance than those from matching with replacement, since in matching with replacement an individual control unit may appear in many matched sets and the resulting large weight on a single observation makes the estimator more sensitive to random fluctuations in its response. Thus one aspect of choosing between these designs is a choice about how to strike a bias-variance tradeoff. The other important aspect distinguishing these designs is that randomization inference, which is based on permuting treatment assignments in each matched set independently of others, generally requires matching without replacement.

These same dynamics play out with slightly more complexity in comparing GroupMatch without replacement and GroupMatch with trajectory replacement. In particular, GroupMatch without replacement ensures that responses in distinct matched sets are statistically independent (under a model in which trajectories are sampled independently), allowing for randomization inference, and ensures that the total weight on observations from any one control trajectory can sum only to $1/C$, ensuring that the estimator's variance cannot be too highly inflated by a single trajectory with large weight. On the other hand, GroupMatch with trajectory replacement leads to higher-quality matches and reduced bias in matched pairs.

We suggest a third GroupMatch design which leans even further towards expanding the potential control pool and reducing bias.

3. **GroupMatch with instance replacement**: Each treated unit can match to no more than one instance from the same control unit, but control instances can be matched to more than one treated unit.

GroupMatch with instance replacement is identical to GroupMatch without trajectory replacement except that it also allows repetition of individual instances within the matched design as well as non-identical instances from the same trajectory. As such, it is guaranteed to produce higher-quality matches than GroupMatch without trajectory replacement, but may lead to higher-variance estimators since individual instances may receive weights larger than $1/C$. Figure 1 illustrates the these three GroupMatch methods with a toy example that matches injured baseball players to non-injured players based on on-base percentage (OBP).

In practice we view GroupMatch with instance replacement as a more attractive approach than GroupMatch with trajectory replacement almost without exception. One reason is that while the true variance of estimators from GroupMatch with instance replacement may often exceed that of estimators from GroupMatch with trajectory replacement by a small amount, our recommended approach for *estimating* the variance and conducting inference are not able to capture this difference. As we describe in Section 3, in the absence of a specific parametric model for correlations within a trajectory, inference proceeds in a conservative manner by assuming arbitrarily high correlations within a trajectory (much like the clustered standard error adjustment in linear regression). Since the variance advantage for GroupMatch with trajectory replacement arises only when correlations between instances within a trajectory are lower than one, the estimation strategy is not able to take advantage of them. This disconnect means that GroupMatch with trajectory replacement will not generally lead to narrower empirical confidence intervals even though it is known to be less variable in reality, much how the variance gains associated with paired randomized trials relative to less-finely-stratified randomized trials may fail to translate into improved variance estimates (Imbens, 2011).

A second important advantage of GroupMatch with instance replacement is its computational and analytical tractability relative to the other GroupMatch designs. We note that one easy way to implement GroupMatch with instance replacement as a network flow optimization problem is to remove a set of constraints in Pimentel et al. (2020)'s Network B (specifically the upper capacity on the directed edges connected to the sink node), and in Sections 5 and 7 we use this implementation for its convenient leveraging of the existing `groupmatch` package in R. However, much more computationally efficient algorithms are also possible. Crucially, the removal of the constraint forbidding instance replacement means that matches can be calculated for each treated instance without any reference to the choices made for other treated units; the C best matches for a given treated unit are simply the C nearest neighbor instances such that no two such control instances within the matched set come from the same trajectory. In principle, this allows for complete parallelization of the matching routine. On the analytical side, this aspect of the design makes it possible to characterize the matching algorithm as a generalized form of nearest neighbor matching, a strategy we adopt in the proof of Theorem 1 to leverage proof techniques used by Abadie and Imbens (2006) for cross-sectional nearest neighbor matching. In light of these considerations, we focus primarily on GroupMatch with instance replacement in what follows, although the methods derived appear to perform well empirically for other GroupMatch designs too.

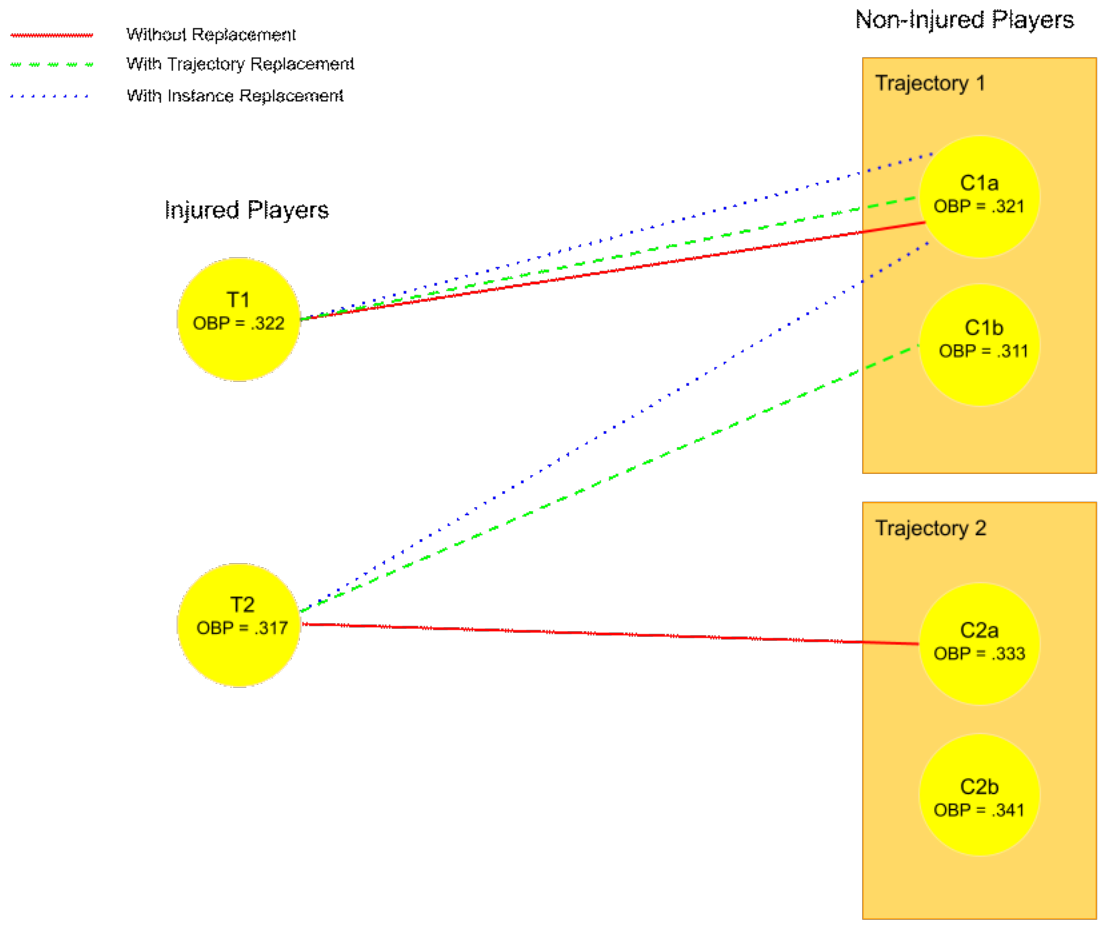

Figure 1: Toy example illustrating the three GroupMatch matching methods. Two injured baseball players (T1 and T2) are matched 1-1 to non-injured baseball players (C1a/b and C2a/b) based on player on-base percentage (OBP). Each non-injured player has two pseudo-injury times. Each instances in trajectory 1 is more similar to both treated units than either instance in trajectory 2, and instance C1a is more similar to both treated units than instance C1b. Under GroupMatch without replacement, T2 must match to an instance in Trajectory 2 because at most one instance from Trajectory 1 can participate in the match. under GroupMatch with trajectory replacement, T2 can match to C1b but not to C1a, since multiple control instances can be chosen from the same trajectory as long as they are distinct. Under GroupMatch with instance replacement, both T1 and T2 are able to match to C1a. However, if each treated instance were matched to two control instances instead of 1, Groupmatch with instance replacement would still forbid either T1 or T2 to match to a second instance in Trajectory 1 in addition to C1a.

4 Block Bootstrap Inference

4.1 Inference methods for matched designs

Broadly speaking, there are two schools of thought in conducting inference for matched designs. One approach, spearheaded by Abadie and Imbens (2006, 2008, 2011, 2012), relies on viewing the raw treated and control data as samples from an infinite population and on demonstrating that estimators based on matched designs (which in this framework are considered to be random variables, as functions of random data) are asymptotically normal. Inferences are based on the asymptotic distributions of matched estimators.

A second approach, described in detail in Rosenbaum (2002a,b) and Fogarty (2020), adopts the perspective of randomization inference in controlled experiments; one conditions on the structure of the match and the potential outcomes and considers the null distribution of a test statistic over all possible values of the treatment vector by permuting values of treatment within matched sets. This framework offers strong finite sample guarantees without assumptions on outcome variables for testing sharp null hypotheses, and asymptotic guarantees for testing weak null hypotheses. In this case the asymptotics are over a sequence of successively larger finite populations (Li and Ding, 2017). Well-studied methods of sensitivity analysis are also available. One way to understand the link between the first and second approaches to inference is to view the latter as a conditional version of the first; indeed Rosenbaum (2002b, §3) derives distributions of treatment vectors used to construct randomization tests by assuming a sampling model as in the first approach and conditioning on the matched design produced from the random data.

As described in Pimentel et al. (2020), while standard methods of inference may be applied to GroupMatch without replacement, in which control individuals contribute at most one unit to any part of the match, none have been adequately developed for GroupMatch with trajectory replacement, in which distinct matched sets may contain different versions of the same control individual. For randomization inference, the barrier appears to be quite fundamental, because permutations of treatment within one matched set can no longer be considered independently for different matched sets. In GroupMatch with trajectory replacement, a treated unit receives treatment at one time and appears in the match only once; if treatment is permuted among members of a matched set so that a former control now attains treatment status, what is to be done about other versions of this control unit that are present in distinct matched sets? We note that similar issues arise when contemplating randomization inference for cross-sectional matching designs with replacement, and we are aware of no solutions for randomization inference even in this relatively less complex case.

The problems with applying sampling-based inference to GroupMatch designs with trajectory replacement are quite distinct. Here the primary issue relates to the unknown correlation structure for repeated measures from a single control individual. The literature on matching with replacement provides estimators for pairs that are fully independent (Abadie and Imbens, 2012) and for cases in which a single observation appears identically in multiple pairs (Abadie and Imbens, 2006), but not for the intermediate case of GroupMatch with trajectory replacement where distinct but correlated observations appear in distinct matched sets.

In what follows we develop a sampling-based inference method appropriate for Group-

Match with trajectory replacement by generalizing a recent proposal of Otsu and Rai (2017) for valid sampling-based inference of cross-sectional matched studies using the bootstrap. Although the bootstrap often works well for matched designs without replacement (Austin and Small, 2014), naïve applications of the bootstrap in matched designs with replacement have been shown to produce incorrect inferences as a consequence of the failure of certain regularity conditions (Abadie and Imbens, 2008). Intuitively, if matching is performed after bootstrapping the original data, multiple copies of a treated unit will necessarily all match to the same control unit, creating a clumping effect not present in the original data. However, Otsu and Rai (2017) arrived at an asymptotically valid block bootstrap inference method for matching by bootstrapping weighted and bias-corrected functions of the original observations *after* matching rather than repeatedly matching from scratch in new bootstrap samples.

While Otsu and Rai (2017) focus their analysis on cross-sectional studies, it has been conjectured elsewhere that a similar bootstrap approach, applied to entire trajectories of repeated measures in a form of the block bootstrap, provides valid inference for certain matched longitudinal designs (Imai et al., 2020). In this section, we formalize this idea and demonstrate its applicability specifically to GroupMatch designs.

4.2 Block Bootstrap

In order to conduct inference under GroupMatch with trajectory or instance replacement we propose a weighted block bootstrap approach. We rearrange the GroupMatch ATT estimator from Section 2 as follows, letting $K_M(i,t)$ be the number of times the instance at trajectory i and time t is used as a match.

$$\begin{aligned}
\hat{\Delta}_{adj} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i[(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - \frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}(Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'}))] \\
&= \frac{1}{N_1}\sum_{i=1}^{N} D_i[(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t}))] \\
&= \frac{1}{N_1}\sum_{i=1}^{N}\hat{\Delta}_i
\end{aligned}$$

Because different instances of the same control unit are correlated, we resample information at the *trajectory* level rather than the instance level. Specifically we resample the $\widehat{\Delta}_i$; note that since these quantities are functions of the $K_M(i,t)$ weights in the original match, we do not repeat the matching process within bootstrap samples. In particular, we proceed as follows:

1. Fit an outcome regression $\widehat{\mu}_0(\cdot)$ for outcomes based on covariates in the previous L timepoints using only control trajectories.

2. Match treated instances to control instances using a GroupMatch design with instance replacements. Calculate matching weights $K_M(i,t)$ equal to the number of times the instance at time t in trajectory i appears in the matched design.

3. Calculate the bias-corrected ATT estimator $\widehat{\Delta}_{adj}$.

4. Repeat B times:

   (a) Randomly sample N elements $\widehat{\Delta}_i^*$ with replacement from $\{\widehat{\Delta}_1, \ldots, \widehat{\Delta}_N\}$.

   (b) Calculate the bootstrap bias-corrected ATT estimator $\widehat{\Delta}_{adj}^*$ for this sample of trajectories as follows:

$$\widehat{\Delta}_{adj}^* = \frac{1}{N_1} \sum_{i=1}^{N} \widehat{\Delta}_i^*$$

5. Construct a (1 - α) confidence interval based on the $\alpha/2$ and $1 - \alpha/2$ percentile of the $\widehat{\Delta}_{adj}^*$- values calculated from the bootstrap samples.

This method is essentially a block-bootstrap procedure, very similar to the method proposed in Imai et al. (2020). Our main result below shows the asymptotic validity of this approach.

First we outline several assumptions needed to prove this result, in addition to Assumptions 2-5 in Section 2.3. We summarize these assumptions verbally here, deferring formal mathematical statements to the appendix. First, we require the covariates X_i to be continous with compact and convex support and a density both bounded and bounded away from zero. Secondly, we require that the conditional mean functions are smooth in $\mathbf{X}$, with bounded fourth moments. In addition, we require that conditional variances of the treated potential outcomes and conditional variances of nontrivial linear combinations of control potential outcomes from the same trajectory of are smooth and bounded away from zero. We also require that conditional fourth moments of potential outcomes under treatment and linear combinations of potential outcomes under control are uniformly bounded in the support of the covariates. Finally, we make additional assumptions related to the conditional outcome mean estimator $\widehat{\mu}_0(\cdot)$ specifically that the kLth derivative of the true conditional mean functions $\mu_1^t(\cdot)$ and $\mu_0(\cdot)$ exist and have finite suprema, and that the $\widehat{\mu}_(\cdot)$ converges to $\mu_0(\cdot)$ at a sufficiently fast rate. To state the theorem, we also define

$$\sqrt{N_1} U^* = \frac{1}{N_1} \sum_{i=1}^{N} \left(\widehat{\Delta}_i^* - \widehat{\Delta}_{adj} \right)$$

Theorem 1. *Under assumptions Mt, W, and R presented in the Appendix,*

$$sup_r |Pr\{\sqrt{N_1} U^* \leq r | (\mathbf{Y}, \mathbf{D}, \mathbf{X})\} - Pr\{\sqrt{N_1}(\hat{\Delta}_{adj} - \Delta) \leq r\}| \xrightarrow{p} 0$$

as $N \to \infty$ with fixed control:treated ratio, C.

Remark 1. While we focus on the nonparametric bootstrap, the result holds for a wide variety of other bootstrap approaches including the wild bootstrap and the Bayesian bootstrap. For required conditions on the bootstrap algorithm see the proof in the Appendix.

We note that Assumptions M and R are modeled closely on those of Abadie and Imbens (2006) and later Otsu and Rai (2017), and that the proof technique we adopt is very similar

to the one used for the main result in Otsu and Rai (2017). Briefly, U^* is decomposed into three terms which correspond to deviations of the potential outcome variables around their conditional means, approximation errors for $\widehat{\mu}_0(\mathbf{X})$ terms as estimates of $\mu_0(\mathbf{X})$ terms, and deviations of conditional average treatment effects $\mu_1^t(\mathbf{X}) - \mu_0(\mathbf{X})$ around the population ATT Δ. Regularity conditions from Assumption M ensure that the conditional average treatment effects converge quickly to the population ATT, and Assumption R, combined with Assumption-M-reliant bounds on the largest nearest-neighbor discrepancies in $\mathbf{X}$ vectors due originally to Abadie and Imbens (2006) and adapted to our GroupMatch with instance replacement design, show that the deviation between $\widehat{\mu}_0(\cdot)$ and $\mu_0(\cdot)$ disappears at a fast rate. Finally, a central limit theorem applies to the deviations of the potential outcomes, producing the desired results. For full details, see the Appendix.

4.3 Difference-in-Differences Estimator

Our block bootstrap inference approach is similar to that of Imai et al. (2020), however the ATT estimator they consider is a difference-in-differences estimator. We can easily extend our setup and results to apply to the difference-in-differences estimator.

$$
\begin{aligned}
\hat{\Delta}_{DiD} = \frac{1}{N_1}\sum_{i=1}^{N} D_i[((Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (Y_{i,T_i-1} - \hat{\mu}_0(\mathbf{X}_{i,T_i-1})))- \\
\frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}((Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'})) - (Y_{j,t'-1} - \hat{\mu}_0(\mathbf{X}_{i,t'-1})))] \\
= \frac{1}{N_1}\sum_{i=1}^{N} D_i[((Y_{i,t=T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (Y_{i,t=T_i-1} - \hat{\mu}_0(\mathbf{X}_{i,T_i-1})))- \\
(1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}((Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t})) - (Y_{i,t-1} - \hat{\mu}_0(\mathbf{X}_{i,t-1})))] \\
= \frac{1}{N_1}\sum_{i=1}^{N}\hat{\Delta}_i^{DiD}
\end{aligned}
$$

Note that this estimator requires L lags to be measured at time $T_i - 1$, which requires a burn-in period of length L rather than length $L - 1$. Imai et al. (2020) assume exact matching, which eliminates the need for a bias correction term, $\hat{\mu}_0(\mathbf{X}_{i,t})$, and simplifies the proof of Theorem 1.

As described in the previous section, valid inference is possible if we resample the $\hat{\Delta}_i^{DiD}$. Our proof of Theorem 1 presented in the appendix requires mild modification to work for this difference-in-differences estimator. In particular, the variance estimators include additional covariance terms. For more details, see Appendix A.

5 Simulations

We now explore the performance of weighted block bootstrap inference via simulation. In particular, we investigate coverage and length of confidence intervals compared to those obtained by conducting standard parametric inference for the classical linear model (OLS), with and without cluster-robust error adjustment for controls from the same trajectory. We choose to compare to OLS with and without cluster-robust error adjustment because, to the best of our understanding, this is what is used in practice.

5.1 Data Generation

We generate eight covariates, four of them uniform across time for each individual i, (i.e., they take on the same value at every timepoint):

$$\begin{aligned} X_{1,i}, X_{3,i}, X_{4,i} &\sim N(0,1) \\ X_{2,i} &\sim N(0,1) \text{ for control units} \\ X_{2,i} &\sim N(0.25,1) \text{ for treated units} \end{aligned}$$

Additionally, for treated units:

$$\begin{aligned} X_5, X_7, X_8 &\sim N(0,1) \\ X_6 &\sim N(0.5,1) \end{aligned}$$

Four of the covariates are time-varying for control units. In particular, for each control unit, three instances are generated from a random walk process to correlate their values across time. Covariate $j = 5, 6, 7, 8$ values for an instance t in a trajectory i are generated in the following way:

$$\begin{aligned} X_{j,i1} &\sim N(0,1) \\ X_{j,i2} &= X_{j,i1} + \epsilon_{i1} \\ X_{j,i3} &= X_{j,i2} + \epsilon_{i2} \\ \epsilon_{i1}, \epsilon_{i2} &\sim N(0, 0.5^2) \end{aligned}$$

Our outcome is defined as follows fixing $a_L = log(1.25)$, $a_M = log(2)$, $a_H = log(4)$ and $a_{VH} = log(10)$ and drawing the ϵ_{it} terms independently from a standard normal distribution.

$$Y_{it} = a_L \sum_{j=1}^{4} X_{j,it} + a_{VH} X_{5,it} + a_M (X_{6,it} + X_{8,it}) + a_H (X_{7,it}) + \Delta D_i + \epsilon_{it}$$

The outcome for a unit is thus correlated across time as it is generated from some time-varying covariates. Each simulation consists of 400 treated and 600 control individuals. We consider 1:2 matching. The true treatment effect, Δ, is set at 0.25.

We consider three different ways of generating the continuous outcome variable. First, we generate the outcome based on a linear model of all the covariates with independent error terms (as described above). Second, we add correlation to the error terms, so that the error terms for a trajectory, ϵ_{it} for a given trajectory i, are generated from a normal distribution with mean 0 and covariance matrix with off diagonal values of 0.8. Finally, in addition to the correlated error terms, we square the $X_{2,it}$ term of the model, so it is no longer linear.

We compare the weighted bias-corrected bootstrap approach outlined in Section 4 to the confidence intervals obtained from weighted least squares (WLS) regression and WLS with clustered standard errors. We choose to compare to WLS because this is commonly recommended in matching literature (Ho et al., 2007; Stuart et al., 2011). However, Abadie and Spiess (2021) pointed out that standard errors from regression may be incorrect due to dependencies among outcomes of matched units, and identified matching with replacement as a setting in which these dependencies are particularly difficult to correct for. Our simulation results suggest that these difficulties carry over into the case of repeated measures. Additionally, in running our simulations we noticed that standard functions in R used to compute WLS with matching weights such as lm and Zelig (which calls lm), compute biased standard error estimates in most settings. See Appendix B for details.

5.2 Results

Tables 1 and 2 show the coverage and average 95% confidence interval (CI) length, respectively, of WLS, WLS cluster, and bootstrap bias-corrected methods for each of our three simulation settings under 10,000 simulations. As misspecification increases the bootstrap method is substantially more robust (although under substantial misspecification the bias-corrected method also fails to achieve nominal coverage). In settings where strong scientific knowledge about the exact form of the outcome model is absent, the bootstrap approach appears more reliable than its chief competitors.

Coverage	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	93.2%	94.8%	94.8%
Linear DGP, Correlated Errors	89.4%	91.5%	94.5%
Nonlinear DGP, Correlated Errors	83.4%	86.0%	89.8%

Table 1: Coverage of the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups.

Average CI Length	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	0.25	0.27	0.27
Linear DGP, Correlated Errors	0.25	0.27	0.30
Nonlinear DGP, Correlated Errors	0.26	0.28	0.31

Table 2: Average 95% confidence interval length for the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups.

6 Testing for Timepoint Agnosticism

The key advantage of GroupMatch (Pimentel et al., 2020) relative to other matching techniques designed for rolling enrollment settings (e.g., Witman et al. (2019), Imai et al. (2020), and Lu (2005)) is its ability to consider and optimize over matches between units at different timepoints, which leads to higher quality matches on lagged covariates. This advantage comes with a price in additional assumptions, notably the assumption of timepoint agnosticism. Timepoint agnosticism means that mean potential outcomes under control for any two individual timepoints in the data should be identical; in particular, this rules out time trends of any kind in the outcome model that cannot be explained by covariates in the prior L timepoints.

While in many applications scientific intuition about the data generating process suggests this assumption may be reasonable, it is essential that we consider any information contained in the observed data about whether it holds in a particular case. Accordingly, we present a falsification test for time agnosticism. Falsification tests are tests "for treatment effects in places where the analyst knows they should not exist," (Keele, 2015) and are useful in a variety of settings in observational studies (Rosenbaum, 1999). In particular, our test is designed to detect violations of time agnosticism, or "treatment effects of time" when they should be absent; rejections indicate setting in which GroupMatch is not advisable and other rolling enrollment matching techniques that do not rely on timepoint agnosticism are likely more suitable. While failure to reject may not constitute proof positive of time agnosticism's validity, it rules out gross violations, thereby limiting the potential for bias.

To test the timepoint agnosticism assumption we propose *control-control time matching*: matching control units at different timepoints and testing if the average difference in outcomes between the two timepoint groups, conditional on relevant covariates, is significantly different from zero using a permutation test. Specifically, restricting attention to trajectories i from the control group, we select two timepoints t_0 and t_1 and match each instance at one timepoint to one at the other timepoint using the GroupMatch optimization routine, based on similarity of covariate histories over the previous L timepoints. We note that since this match compares instances at two fixed time points, it is not strictly necessary for GroupMatch to be applied, and any optimal method of matching without replacement should suffice. One practical issue arises: GroupMatch and related matching routines expect one group to be designated "treated," all members of which are generally retained in the match, and the other "control," some members of which will be included, but both matching groups are controls in this case. We label whichever of the two groups has fewer instances as treated; without loss of generality, we will assume there are fewer instances at time t_1 and use these instances as the reference group to be retained.

We now define a test statistic for the falsification test, by close analogy to our ATT estimator in section 2.3. Let N_c be the overall number of control units in total and let N_{t_1} be the number of control instances at time t_1. Let $\hat{\mu}_0^{t_0}$ be a bias correction model fit on our new control group (i.e., control instances at time t_0). In addition, let $D_i' = 1$ if unit i is present at time t_1. We define the test statistic as follows:

$$\hat{\Delta}_{cc} = \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} D_i'((Y_{i,t=t_1} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_1})) - \sum_{j=1}^{N_c} M_{it_1,jt_0}(Y_{i,t=t_0} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_0})))$$

To conduct inference, we use a permutation test to test the following null hypothesis, where $E_0^{t_1}\{\cdot\}$ indicates expectation over the distribution of the covariates in control instances at time t_1.

$$E_0^{t_1}\left\{\mu_0^{t_0}(\mathbf{X})\right\} = E_0^{t_1}\left\{\mu_0^{t_1}(\mathbf{X})\right\}$$

In words, this null hypothesis says that, accounting for differences in the covariate distribution at times 0 and 1, the difference in the average outcomes of control instances at the two timepoints is zero. The test amounts to considering the tail probability of the distribution of the following test statistic $\widehat{\Delta}_{perm}$ under many draws of the random vector R, where $R = (R_1, ..., R_{N_c})$ and R_i are independent Rademacher random variables:

$$\widehat{\Delta}_{perm} = \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} R_i D_i'((Y_{i,t=t_1} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_1})) - \sum_{j=1}^{N_c} M_{it_1,jt_0}(Y_{i,t=t_0} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_0})))$$

This permutation test is identical in implementation to the standard test of a sharp null hypothesis that outcomes are unchanged by group assignment for matched designs without replacement, discussed in detail in Rosenbaum (2002b, §2-§3). In steps:

1. Randomly partition the set of control trajectories into two groups. Label control instances from the first group of trajectories at timepoint t_1 the new "treated" units, and control instances from the second grouf of trajectories at timepoint t_0 the new "control" units.

2. Fit a bias correction model on the new control units.

3. Match the new treated units to the new control units and calculate the test statistic.

4. Repeat B times:

   (a) For each matched pair randomly and independently switch which unit is labelled the treated and which is labelled the control unit with probability 0.5 (with probability 0.5 do not switch the treated/control labels). This amounts to multiplying the treatment effect by -1 for that pair with probability 0.5.

   (b) Calculate the test statistic for this randomization.

5. Calculate the proportion of permutations that result in an absolute value of the test statistic greater than or equal to the absolute value of the observed value calculated in 1. This is the P-value.

6. If the P-value is smaller than a predefined significance level α, reject the null hypothesis of no difference between groups, indicating the presence of systematic variation of outcomes with time given covariates.

We do not allow the same unit to appear in both the new control and treated group, because this would lead to dependence across matches. We employ 1-1 matching because we are testing a weak null hypothesis using a permutation test. As Wu and Ding (2020) demonstrate, issues can arise when using a randomization test to test a weak null hypothesis when treatment and control sample sizes are not equal. However, 1-1 matching allows us to avoid this issue by balancing the treatment and control sample sizes. We recommend the use of caliper matching to ensure high quality matches, especially in the case where all control units are present at both timepoints.

Note that it is important to permute treatment after matching (indeed, conditional on the matched pairs chosen) in order to preserve the covariate distribution of the treated and control units. If we permute treatment and then perform matching we risk changing the covariate distribution of the treated units under permutation. This could become an issue especially if some covariates are correlated with time, but the outcome is not. Then permuting before matching could cause an effect to appear as a result of destroying the original treated covariate distribution.

We choose to use a permutation test here rather than the bootstrap because the data split and 1:1 matching ratio ensure matches are independent under the original sampling model, making for a tractable permutation distribution. If desired, the bootstrap approach of Section 4 could be applied instead, and we expect that results would be similar given the fundamental similarity between bootstrap and randomization inference where both are viable (Romano, 1989).

A key consideration for this test is which timepoints to choose as t_0 and t_1. The choice of timepoint comparison depends largely on what a plausible time trend would be for the problem at hand. For example, if you suspect a linear time trend, it makes sense to look at the first and last timepoints. If the trend is linear, this test should have high power to detect a problem in moderate to large samples.

If one is uncertain about the specific shape of the time trend that is most likely to occur and want to test for all possible trends, we recommend testing each sequential pair of timepoints (i.e., timepoints 1 and 2, 2 and 3, 3 and 4, and so on) and combining the tests via a nonparametric combination of tests (Pesarin and Salmaso, 2010).

6.1 Simulations

We illustrate this method via simulation. We generate a dataset with 4 covariates and 1000 control units each with 2 instances occuring at time t_0 and t_1. Two of the covariates vary with time, and two are uniform across time:

$$\begin{aligned} X_{1,i}, X_{2,i} &\sim N(0,1) \\ X_{3,i,t_0}, X_{4,i,t_0} &\sim N(0,1) \\ X_{3,i,t_1} &= X_{3,i,t_0} + \epsilon_i \\ X_{4,i,t_1} &= X_{4,i,t_0} + \epsilon_i \\ \epsilon_i &\sim N(0, 0.5^2) \end{aligned}$$

The outcome variable is a linear combination of the four covariates, a time trend controlled by parameter γ, and an error term ($\epsilon_{it} \sim N(0,1)$):

$$Y_{it} = log(4)(X_{1,it} + X_{4,it}) + log(10)(X_{3,it} + X_{4,it}) + 1\{t = t_1\}\gamma + \epsilon_{it}$$

We generate data for the setting $\gamma = 0$, which does not include a time varying component, 1000 times. On each dataset, we perform the test for timepoint agnosticism outlined in this section, with 1-1 matching and bias correction. Our simulated data only contains two timepoints and the sample size is balanced so we choose the first timepoint as t_0, our new control group, and the second timepoint as t_1, our new treated group. In 0.049 of the simulations the P-value is less than 0.05, which shows that type I error is controlled. Now, we add in a time trend where there is an additional term, γ, added to the second timepoint. For $\gamma = 0.1$ for the second timepoint (resulting in a time trend of 0.1), our simulations result in a P-value less than 0.05 in 0.327 of the 1000 simulations. For $\gamma = 0.25$ for the second timepoint (resulting in a time trend of 0.25), our simulations result in a P-value less than 0.05 in 0.981 of the 1000 simulations. Table 3 summarizes these results.

Figure 2 shows two simulated datasets with $\gamma = 0.1$. The test for timepoint agnosticism detects the trend in one of the two datasets. Overall, the simulations show that the test is not a panacea for issues with time agnosticism, failing to detect small violations more often than not. However, it still adds substantial value to the analysis pipeline, detecting moderate violations of time agnosticism not especially obvious to the eye in visualization plots with a very high rate of success.

Time Trend, γ	0	0.1	0.25
Proportion P-values < 0.05	0.049	0.327	0.981

Table 3: Summary of timepoint agnosticism simulation results. Proportion of 1000 simulations where the P-value from the test is less than 0.05, for time trends, $\gamma = 0, 0.1, 0.25$.

7 Application: Baseball Injuries

A large body of literature evaluates major league baseball players' performance (Baumer, 2008), and a different body of literature analyzes injury trends and impact of injuries in athletics (Conte et al., 2016). The intersection of these two research areas is relatively small. In particular, there have been relatively few studies evaluating the impact of injury on position players' hitting performance. Studies that have evaluated the impact of injury on batters have generally been focused on specific injury types, and have not found strong evidence that injury is associated with a decline in performance (Begly et al., 2018; Frangiamore et al., 2018; Wasserman et al., 2015).

We study the impact of short-term injury on hitting performance in observational data from major league baseball during 2013-2017, by using GroupMatch to match baseball players injured at different times to similar players at other points in the season not receiving injuries. We evaluate whether players see a decline in offensive performance immediately after their return from injury. In contrast to other studies, we pool across injury types to see if there is a more general effect of short term injury on hitter performance.

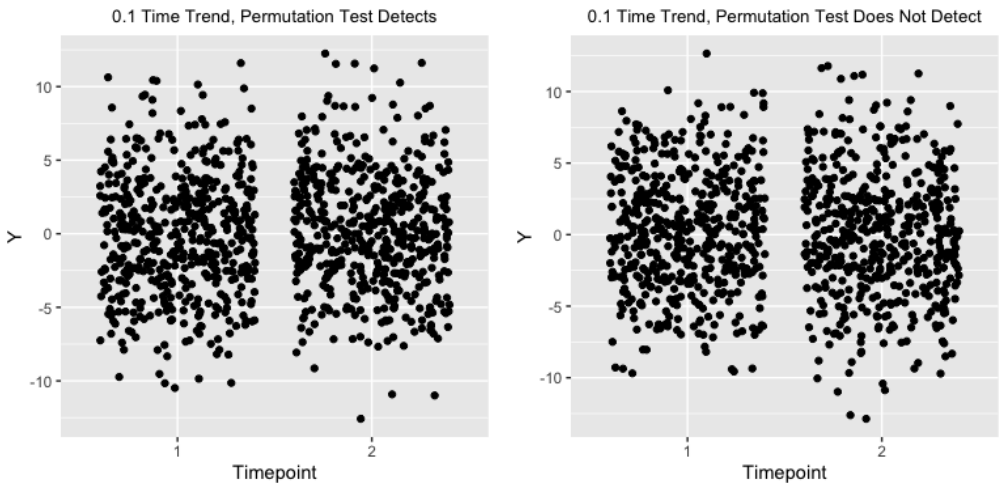

Figure 2: Simulated datasets with time trend, $\gamma = 0.1$. The figures show outcome data for a dataset where the timepoint agnosticism test detects the time trend ($P = 0.04$) and does detect the time trend ($P = 0.37$) respectively.

7.1 Data and Methodology

We use publicly available MLB player data from Retrosheet.org and injury data scraped from ProSportsTransactions.com for the years 2013-2017. Our dataset is composed of player height, weight and age, quantities that remain constant over a single season of play, as well as on-base percentage (OBP) and plate appearances (PAs) at different points in the season, and dates of short-term injuries, in which the player's team designated him for a 7-10 day stay on the team's official injured list, for each year. OBP is a common measure of batter performance and is approximately equal to the number of times a player makes it on base divided by their number of plate appearances.[1] Each plate appearance is a batter's complete turn batting.

For each non-injured player, we generate three pseudo-injury dates evenly spaced over their number of PAs. In each season, we match injured players to four non-injured players. Matches were formed using GroupMatch with instance replacement and matching on age, weight, height, number of times previously injured, recent performance as measured by OBP over the previous 100 PAs, and performance over the entire previous year as measured by end-of-year OBP after James-Stein shrinkage.[2] We choose to shrink the OBP using James-Stein instead of using raw OBP to reduce the variability for players that had a relatively small number of PAs the previous season (Efron and Morris, 1975).

Table 4 shows the balance for each of the covariates prior to matching and Table 5 shows the balance after matching. For each covariate, matching shrinks the standardized difference between the treated and control means.

We compare the results for bias-corrected block bootstrap inference, OLS and OLS with clustered standard errors.

[1] OBP = (Hits + Walks + Hit By Pitch) / (At Bats + Walks + Hit by Pitch + Sacrifice Flies)

[2] See https://chris-said.io/2017/05/03/empirical-bayes-for-multiple-sample-sizes/ for discussion on how to apply James-Stein to data with multiple sample sizes.

	Treated Mean	Control Mean	Standardized Difference
Height	73.7	73.1	0.26
Weight	213	209	0.24
2016 OBP (JS Shrunk)	.324	.328	-0.09
Lag OBP	.336	.341	-0.07
Birth Year	1988	1988	-0.08
Number Previous Injuries	2.73	1.91	0.30

Table 4: Balance table for MLB injury analysis after matching each injured player to four non-injured players.

	Treated Mean	Control Mean	Standardized Difference
Height	73.7	73.4	0.14
Weight	213	212	0.07
2016 OBP (JS Shrunk)	.324	.323	0.02
Lag OBP	.336	.338	-0.02
Birth Year	1988	1988	-0.06
Number Previous Injuries	2.73	2.16	0.21

Table 5: Balance table for MLB injury analysis after matching each injured player to four non-injured players.

7.2 Results

The ATT estimates are positive (0.010), but the 95% confidence intervals cover zero for all methods, indicating that there is not strong evidence that short term injury impacts batter performance. We present the results for 2017 in Figure 3. Results from 2013 - 2016 were substantively the same. Pooling the matched data across years and the applying the block bootstrap method also results in the same substantive conclusions. The data pass the timepoint agnosticism test, comparing the first and last pseudo-injury dates.

8 Discussion

The introduction of matching with instance replacement, a method for block bootstrap inference, and a test for timepoint agnosticism provide substantial new capability for the existing GroupMatch framework. We now discuss a number of limitations and opportunities for improvement.

Our proof of the block bootstrap approach assumes the use of GroupMatch with instance replacement. The large-sample properties of matched-pair discrepancies are substantially easier to analyze mathematically in this setting than GroupMatch with trajectory replacement or GroupMatch without replacement, designs in which different treated units may compete for the same control units, and the technical argument must be altered to account for this complexity. However, Abadie and Imbens (2012) successfully characterized similar large-sample properties in cross-sectional settings for matching without replacement. While beyond the scope of our work here, we believe it is likely that this approach could provide an

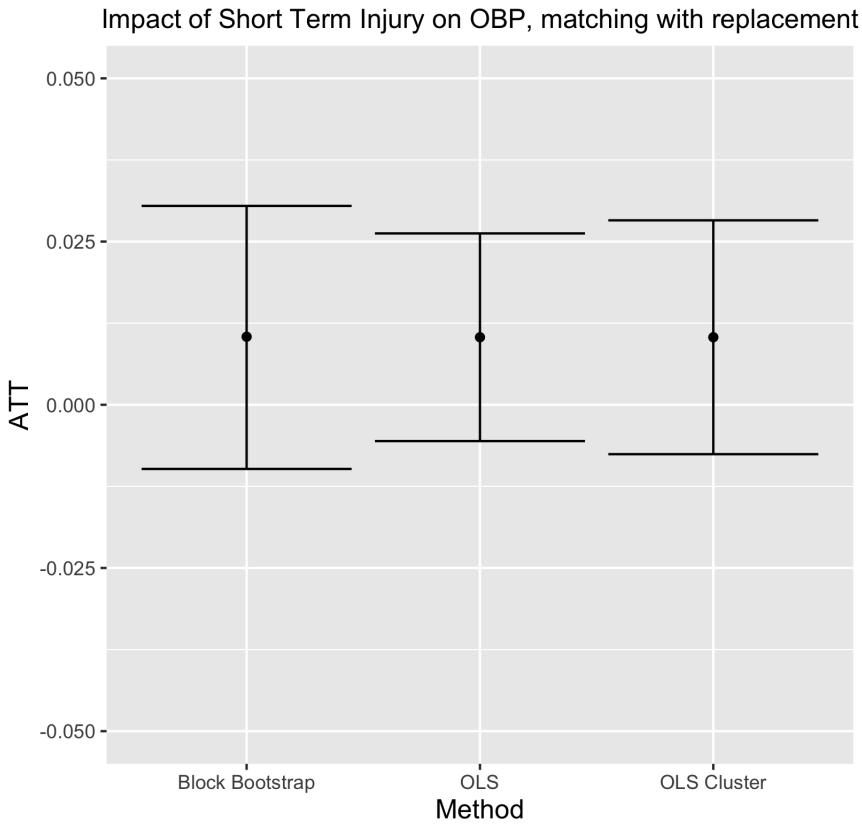

Figure 3: Plot comparing block bootstrap, OLS and Cluster OLS inference methods for the ATT in our 2017 baseball injury example.

avenue for extending Theorem 1 to cover the other two GroupMatch designs. Empirically, we have found that the block bootstrap performs well when matches are calculated using any of the three GroupMatch designs.

Setting aside the technical barriers associated with extending the theory to GroupMatch without replacement, our new approach provides a competitor method to the existing randomization inference framework described by Pimentel et al. (2020) available for GroupMatch without replacement. The randomization inference framework offers the advantage of finite sample validity and freedom from making assumptions about the sampling distribution of the response variables; on the other hand, the block bootstrap method avoids the need to assume a sharp null hypothesis. In general these same considerations arise in choosing between sampling-based inference and randomization-based inference for a cross-sectional matched study, although such choices have received surprisingly little direct and practical attention in the literature thus far.

The falsification test proposed in Section 6 is subject to several common criticisms levied at falsification tests, particularly their ineffectiveness in settings with low power. One possible approach is to reconfigure the test to assume violation of time agnosticism as a null hypothesis and seek evidence in the data to reject it; Hartman and Hidalgo (2018) recommend a similar change for falsification tests used to assess covariate balance. However, even in the absence of such a change the test may prove useful in concert with a sensitivity anal-

ysis. Sensitivity analysis, already widely studied in causal inference as a way to assess the role of ignorability assumptions, places a nonzero bound on the degree of violation of an assumption and reinterprets the study's results under this bound, often repeating the process for larger and larger values of the bound to gain insight. Such a procedure, which focuses primarily on assessing the impact of small or bounded violations of an assumption, naturally complements our falsification test, which (as shown in our simulations) can successfully rule out large violations but is more equivocal about minor violations.

Unfortunately no sensitivity analysis appropriate for block bootstrap inference has yet been developed, either for time agnosticism or other strong assumptions such as ignorability. The many existing methods for sensitivity analysis (developed primarily with ignorability assumptions in mind) are unsatisfying in our framework for a variety of different reasons: some rely on randomization inference (Rosenbaum, 2002b), others focus on weighting methods rather than matching (Zhao et al., 2019; Soriano et al., 2021), and others are limited to specific outcome measures (Ding and VanderWeele, 2016) or specific test statistics (Cinelli and Hazlett, 2020). We view the development of compelling sensitivity analysis approaches to be an especially important methodological objective for matching under rolling enrollment.

References

Abadie, A. and Imbens, G. W. (2006), "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74, 235–267.

— (2008), "On the failure of the bootstrap for matching estimators," *Econometrica*, 76, 1537–1557.

— (2011), "Bias-corrected matching estimators for average treatment effects," *Journal of Business & Economic Statistics*, 29, 1 – 11.

— (2012), "A Martingale representation for matching estimators," *Journal of the American Statistical Association*, 107, 833 – 843.

Abadie, A. and Spiess, J. (2021), "Robust post-matching inference," *Journal of the American Statistical Association*, 1–13.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019), "Democracy does cause growth," *Journal of Political Economy*, 127, 47–100.

Antonelli, J., Cefalu, M., Palmer, N., and Agniel, D. (2018), "Doubly robust matching estimators for high dimensional confounding adjustment," *Biometrics*, 74, 1171–1179.

Austin, P. C. and Small, D. S. (2014), "The use of bootstrapping when using propensity-score matching without replacement: a simulation study," *Statistics in medicine*, 33, 4306–4319.

Baumer, B. S. (2008), "Why on-base percentage is a better indicator of future performance than batting average: An algebraic proof," *Journal of Quantitative Analysis in Sports*, 4.

Begly, J. P., Guss, M. S., Wolfson, T. S., Mahure, S. A., Rokito, A. S., and Jazrawi, L. M. (2018), "Performance outcomes after medial ulnar collateral ligament reconstruction in Major League Baseball positional players," *Journal of Shoulder and Elbow Surgery*, 27, 282–290.

Bohl, A. A., Fishman, P. A., Ciol, M. A., Williams, B., LoGerfo, J., and Phelan, E. A. (2010), "A longitudinal analysis of total 3-year healthcare costs for older adults who experience a fall requiring medical care," *Journal of the American Geriatrics Society*, 58, 853–860.

Cinelli, C. and Hazlett, C. (2020), "Making sense of sensitivity: Extending omitted variable bias," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 39–67.

Conte, S., Camp, C. L., and Dines, J. S. (2016), "Injury trends in Major League Baseball over 18 seasons: 1998-2015," *Am J Orthop*, 45, 116–123.

Ding, P. and VanderWeele, T. J. (2016), "Sensitivity analysis without assumptions," *Epidemiology (Cambridge, Mass.)*, 27, 368.

Efron, B. and Morris, C. (1975), "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Fogarty, C. B. (2020), "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies," *Journal of the American Statistical Association*, 115, 1518–1530.

Frangiamore, S. J., Mannava, S., Briggs, K. K., McNamara, S., and Philippon, M. J. (2018), "Career Length and Performance Among Professional Baseball Players Returning to Play After Hip Arthroscopy," *The American Journal of Sports Medicine*, 46, 2588–2593.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

Hartman, E. and Hidalgo, F. D. (2018), "An Equivalence Approach to Balance and Placebo Tests," *American Journal of Political Science*, 62, 1000–1013.

Haviland, A., Nagin, D. S., Rosenbaum, P. R., and Tremblay, R. E. (2008), "Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data," *Developmental psychology*, 44, 422–436.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15, 199–236.

Holland, P. W. and Rubin, D. B. (1980), "Causal Inference in Prospective and Retrospective Studies," .

Imai, K., Kim, I. S., and Wang, E. (2020), "Matching methods for causal inference with time-series cross-section data," Tech. rep., Harvard University.

Imbens, G. W. (2011), "Experimental design for unit and cluster randomid trials," in *Conference International Initiative for Impact Evaluation, Cuernavaca*.

Keele, L. (2015), "The statistics of causal inference: A view from political methodology," *Political Analysis*, 313–335.

Li, X. and Ding, P. (2017), "General forms of finite population central limit theorems with applications to causal inference," *Journal of the American Statistical Association*, 112, 1759–1769.

Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001), "Balanced risk set matching," *Journal of the American Statistical Association*, 96, 870–882.

Lu, B. (2005), "Propensity score matching with time-dependent covariates," *Biometrics*, 61, 721–728.

Otsu, T. and Rai, Y. (2017), "Bootstrap inference of matching estimators for average treatment effects," *Journal of the American Statistical Association*, 112, 1720–1732.

Pesarin, F. and Salmaso, L. (2010), *Permutation Tests for Complex Data: Theory, Applications and Software*, Wiley.

Pimentel, S. D., Forrow, L. V., Gellar, J., and Li, J. (2020), "Optimal matching approaches in health policy evaluations under rolling enrollment," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183, 1411–1435.

Romano, J. P. (1989), "Bootstrap and randomization tests of some nonparametric hypotheses," *The Annals of Statistics*, 141–159.

Rosenbaum, P. R. (1999), "Choice as an alternative to control in observational studies," *Statistical science*, 259–278.

— (2002a), "Covariance adjustment in randomized experiments and observational studies," *Statistical Science*, 17, 286–327.

— (2002b), *Observational Studies*, New York, NY: Springer.

Rubin, D. B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 318–328.

Soriano, D., Ben-Michael, E., Bickel, P. J., Feller, A., and Pimentel, S. D. (2021), "Interpretable sensitivity analysis for balancing weights," *arXiv preprint arXiv:2102.13218*.

Stuart, E. (2010), "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25.

Stuart, E. A., King, G., Imai, K., and Ho, D. (2011), "MatchIt: nonparametric preprocessing for parametric causal inference," *Journal of statistical software*.

Wasserman, E. B., Abar, B., Shah, M. N., Wasserman, D., and Bazarian, J. J. (2015), "Concussions are associated with decreased batting performance among Major League Baseball players," *The American Journal of Sports Medicine*, 43, 1127–1133.

Witman, A., Beadles, C., Liu, Y., Larsen, A., Kafali, N., Gandhi, S., Amico, P., and Hoerger, T. (2019), "Comparison group selection in the presence of rolling entry for health services research: Rolling entry matching," *Health services research*, 54, 492–501.

Wu, J. and Ding, P. (2020), "Randomization tests for weak null hypotheses in randomized experiments," *Journal of the American Statistical Association*, 1–16.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019), "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

A Proof of Theorem

A.1 Assumptions

We begin by rewriting $\sqrt{N_1}U^*$ as follows:

$$\sqrt{N_1}U^* = \frac{1}{N_1}\sum_{i=1}^{N}\left(\widehat{\Delta}_i^* - \widehat{\Delta}_{adj}\right) = \sum_{i=1}^{N} W_i^*(\hat{\Delta}_i - D_i\hat{\Delta}_{adj})$$

Here the W_i^*, quantities we denote as the bootstrap weights, are random variables of the form Q_i/N_1 where Q_i is a count of the number of times observation i is selected to appear in the bootstrap sample. This new form enables us to work separately with stochasticity arising from the bootstrap and stochasticity arising from the original data-generating process, and it also makes it easy to generalize our results to other bootstrap approaches as discussed below.

Our proof relies on the assumptions on sampling described in Section 2.2 and the GroupMatch identification assumptions described in Section 2.3, with the exception of the exact matching assumption. In addition we invoke two additional sets of assumptions, which we denote M and R following similar labeling in Otsu and Rai (2017), from whom we adapt our proof strategy.

Assumption M. (Conditions for $\hat{\Delta}_{adj}$)

1. Let the population distribution of $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,T})$ be continuous on $\mathbb{R}^{kT}$ with compact and convex support $\mathbb{X}^T$. In addition, let the densiity of $\mathbf{X}_i$ be bounded, and bounded away from zero on its support.

2. For some $r \geq 1$, $\frac{N_1^r}{N_0} \to \theta$ for $\theta \in (0, \infty)$.

3. For $z = 0, 1$, let $\mu_z^t(\mathbf{X})$ be Lipschitz in $\mathbb{X}^L$ for all $t = L+1, \ldots, T$.

4. For all $t \in \{L+1, \ldots, T\}$, and $z = 0, 1$, $E[Y_{i,t}^4(z)|Z_{i,t-1} = z, \mathbf{X}_i]$ and $Cov(Y_{it}, Y_{it'}|D_i = d, \mathbf{X}_{i,t} = x, \mathbf{X}_{i,t'} = x')$ are bounded uniformly on $\mathbb{X}^T$, and $\text{Var}(Y_{i,t}(z)|Z_{i,t-1} = z, \mathbf{X}_i)$ is Lipschitz in $\mathbb{X}^T$ and bounded away from zero.

Assumption R. (Conditions for $\mu_d(x)$)
For $d = 0, 1$ and λ satisfying $\sum_{l=1}^{kL} \lambda_l = kL$, the derivative $\partial^{kL}\mu_d^t(x)$ exists and satisfies $\sup_{x \in \mathbb{X}} |\partial^{kL}\mu_d^t(x)| \leq R$ for some $R > 0$ and for all $t \in \{L+1, \ldots T\}$. Furthermore, $\hat{\mu}_0(x)$ satisfies $|\hat{\mu}_0(\cdot) - \mu_0(\cdot)|_{kL-1} = o_p(N^{-1/2+1/(kL)})$.

While not necessary to prove Theorem 1, we mention one additional set of conditions on the bootstrap weights W_i^*. These are satisfied trivially by the nonparametric bootstrap we adopt, but in fact the proof goes through for any bootstrap algorithm that can be represented by a set of bootstrap weights satisfying these conditions (for instance, the wild bootstrap). Again following Otsu and Rai (2017), we denote this set of assumptions as Assumption W and refer to it in our proof to make the path for generalization clear.

Assumption W. (Conditions for W_i^*)

1. $(W_1^*, ..., W_N^*)$ is exchangeable and independent of $\mathbf{O} = (\mathbf{Y}, \mathbf{D}, \mathbf{X})$.

2. $\sum_{i=1}^N (W_i^* - \bar{W}^*)^2 \xrightarrow{p} 1$ where $\bar{W}^* = \frac{1}{N}\sum_{i=1}^N W_i^*$

3. $\max_{i=1,...,N} |W_i^* - \bar{W}^*| \xrightarrow{p} 0$

4. $E[W_i^{*2}] = O(N^{-1})$ for all $i = 1, ..., N$

A.2 Lemmas

To bound the size of matching discrepancies, we compare those obtained by GroupMatch with instance replacement to nearest-neighbor matching at a fixed timepoint, in which only one instance from each control unit can potentially be used in a match. Nearest-neighbor matching matches each treated unit to the control instance that is most similar. While GroupMatch also uses nearest-neighbor matching, nearest-neighbor matching at a fixed timepoint considers a smaller pool of control instances since there is only one instance for each control unit that can potentially be used in a match. Comparing matching discrepancies between GroupMatch with instance replacement and nearest-neighbor matching is important to apply Lemma A.2, used to prove Theorem 2, in Abadie and Imbens (2011). For each treated unit i, let $j_m(i)$ and $j_m^{gm}(i)$ represent the mth instance used as a match for nearest-neighbors at a fixed timepoint and GroupMatch with instance replacement respectively. Let $U_{m,i} = \mathbf{X}_{j_m(i)} - \mathbf{X}_{i,T_i}$ and $U_{m,i}^{gm} = \mathbf{X}_{j_m^{gm}(i)} - \mathbf{X}_{i,T_i}$ be the matching discrepancies under nearest neighbors at a fixed timepoint and GroupMatch with instance replacement matching respectively.

Lemma 1. $||U_{m,i}^{gm}|| \leq ||U_{m,i}||$ for all m, i.

Proof. Because nearest neighbors (NN) matching at a fixed timepoint only considers one instance from each control trajectory whereas GroupMatch with instance replacement (GM) considers multiple, the set of control instances that can be used in a match in NN

matching, $\mathcal{C}$, is a subset of the set of control instances that can be used in a match in GM, $\mathcal{C}^{gm}$: $\mathcal{C} \subseteq \mathcal{C}^{gm}$. Both NN and GM match treated units to the control instance that minimizes $U_{m,i}$ and $U^{gm}_{m,i}$, so we have that

$$||U^{gm}_{m,i}|| = min_{m,j\in\mathcal{C}^{gm}}|\mathbf{X}_j - \mathbf{X}_{i,T_i}| \leq min_{m,j\in\mathcal{C}}|\mathbf{X}_j - \mathbf{X}_{i,T_i}| = ||U_{m,i}||$$

where min_m denotes the mth minimum.

Lemma 2. $E[K_M(i,t)^q]$ is bounded uniformly in N.

Proof. This proof follows closely with the proof of Lemma 3 in Abadie and Imbens (2006) with modifications described below. Let f^t be the density of $\mathbf{X}_{i,t}$ and define $\underline{f} = inf_{x,w,t} f^t_w(x)$ and $\overline{f} = sup_{x,w,t} f^t_w(x)$. We define the catchment area, $\mathbb{A}_M(i,t)$ as the subset of $\mathbb{X}$ such that control unit i at time t is matched to each observation j at time t' with $W_j = 1 - W_i$ and $\mathbf{X}_{j,t'} \in \mathbb{A}_M(i,t)$:

$$\mathbb{A}_M(i,t) = \{x | \sum_{l|W_l=W_i, l\neq i} 1\{(min_{t'}||\mathbf{X}_{l,t'} - x||) \leq ||\mathbf{X}_{i,t} - x||\} 1\{min_{t'}||\mathbf{X}_{i,t'} - x|| \geq ||\mathbf{X}_{i,t} - x||\} \leq M\}$$

Ultimately, we want to bound the volume of the catchment area. In order to do this, we need to bound the probability that the distance to a match exceeds some value. To derive this bound, we must account for our trajectory structure by showing the following inequality:

$$\begin{aligned}
&Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN^{-1/k}_{1-W_i} | W_1, ..., W_N, \mathbf{X}_{i,t'} = x, j \in \mathcal{J}_M(i)) \\
&\leq Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN^{-1/k}_{1-W_i} | W_1, ..., W_N, \mathbf{X}_{i,t'} = x, (j,t) = j_M(i,t')) \\
&= \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} Pr(min_t ||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN^{-1/k}_{1-W_i} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^{N_{1-W_i}-m} \times \\
&Pr(min_t ||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| \leq uN^{-1/k}_{1-W_i} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^m \\
&\leq \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| > uN^{-1/k}_{1-W_i} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^{N_{1-W_i}-m} \times \\
&Pr(||\mathbf{X}_{j,t} - \mathbf{X}_{i,t'}|| \leq uN^{-1/k}_{1-W_i} | W_1, ..., W_N, W_j = 1 - W_i, \mathbf{X}_{i,t'} = x)^m
\end{aligned}$$

The second inequality follows from the fact that the probability that the minimal distance over a trajectory is less than or equal to any particular instance. The rest of the proof follows directly from the proof of Abadie and Imbens (2006)'s Lemma 3 after substituting in our catchment area, this inequality, and indexing over time.

Lemma 3. Given that $E[K_M(i,t)^q]$ is uniformly bounded, $\sum_i \sum_t K_M(i,t) = N_1$, and $K_M(i,t) \geq 0$, we have that $\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t''')] \leq cN_1^{1+o(1)}$ for some constant c.

Proof. For all $\epsilon > 0$, we have that

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t''')]$$
$$= \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon] +$$
$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); max_{t_1=t,t',t'',t'''} K_M(i,t_1) > N_1^\epsilon]$$
$$\leq \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon] +$$
$$4 \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t_1) > N_1^\epsilon]$$

This upper bound is derived from the fact that

$$E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); max_{t_1=t,t',t'',t'''} K_M(i,t_1) > N_1^\epsilon]$$
$$\leq \sum_{t_1 \in \{t,t',t'',t'''\}} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t_1) > N_1^\epsilon]$$

and since we are summing over all t, t', t'', t''' we are able to replace the summation above with multiplying by four. We bound each of the terms separately. First,

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^\epsilon]$$
$$\leq N_1^{3\epsilon} \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)]$$
$$= T^3 N_1^{3\epsilon+1}$$

For the next term we use proof by contradiction to show

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] \leq N_1.$$

Suppose $\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] > N_1$, then:

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$\geq (N_1^\epsilon)^{\frac{r}{\epsilon}} \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$= N_1^r \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t)K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon]$$
$$> N_1^r N_1 = N_1^{r+1}$$
$$\implies NT^4 sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] > N_1^{r+1}$$
$$\implies sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t')K_M(i,t'')K_M(i,t'''); K_M(i,t) > N_1^\epsilon] > \frac{N_1^{r+1}}{NT^4} = cN_1$$

for some constant c. This follows from Assumption M.
So, then we have that:

$$
\begin{aligned}
&sup_{i,t} E[K_M(i,t)^{\frac{r}{\epsilon}+4}] \\
&\geq sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t') K_M(i,t'') K_M(i,t''')] \\
&> sup_{i,t,t',t'',t'''} E[K_M(i,t)^{\frac{r}{\epsilon}+1} K_M(i,t') K_M(i,t'') K_M(i,t'''); K_M(i,t) > N_1^{\epsilon}] \\
&> cN_1
\end{aligned}
$$

The first inequality holds by applying Cauchy-Schwarz twice. But then $E[K_M(i,t)^q]$ is not uniformly bounded for all q, so by contradiction

$$\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t'''); K_M(i,t) > N_1^{\epsilon}] \leq N_1.$$

So we have that:

$$
\begin{aligned}
&\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t''')] \\
&= \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t'''); \\
&K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^{\epsilon}] + \\
&\sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t'''); max_{t_1=t,t',t'',t'''} K_M(i,t_1) > N_1^{\epsilon}] \\
&\leq \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t'''); \\
&K_M(i,t), K_M(i,t'), K_M(i,t''), K_M(i,t''') \leq N_1^{\epsilon}] + \\
&4 \sum_i \sum_t \sum_{t'} \sum_{t''} \sum_{t'''} E[K_M(i,t) K_M(i,t') K_M(i,t'') K_M(i,t'''); K_M(i,t) > N_1^{\epsilon}] \\
&\leq T^3 N_1^{3\epsilon+1} + 4c' N_1 \leq c N_1^{1+o(1)}
\end{aligned}
$$

for some constant c.

A.3 Proof

We follow the proof of Theorem 1 in Otsu and Rai (2017) closely, adapting where necessary to address the possible presence of multiple correlated potential outcomes from the same trajectory multiple control instances in the matched design. In the case where every control unit has only one instance, our argument reduces to exactly that presented in Otsu and Rai (2017).

We decompose $\sqrt{N_1} U^*$ as follows:

$$
\begin{aligned}
\sqrt{N_1}U^* &= \sum_{i=1}^{N} W_i^*(\hat{\Delta}_i - D_i\hat{\Delta}_{adj}) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)(\hat{\Delta}_i - D_i\hat{\Delta}_{adj}) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)(D_i(\hat{\Delta}_i - \hat{\Delta}_{adj}) + (1 - D_i)\hat{\Delta}_i) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)[D_i(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i}) - \hat{\Delta}_{adj}) + (1 - D_i)\sum_{t=1}^{T} \frac{K_M(i,t)}{C}(Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t}))] \\
&= \sqrt{N_1}(T^* + R^*_{1N_1} + R^*_{2N_1})
\end{aligned}
$$

We define the following:

$$
\begin{aligned}
e_{i,t} &= Y_{i,t} - \mu_{D_i}(\mathbf{X}_{i,t}) \\
\xi_{i,t} &= (2D_i - 1)(\mu_{D_i}(\mathbf{X}_{i,t}) - \mu_{1-D_i}(\mathbf{X}_{i,t})) - \Delta
\end{aligned}
$$

We can now rewrite the three components as follows.

$$
\begin{aligned}
\sqrt{N_1}T^* &= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)(D_i(e_{i,t=T_i} + \xi_{i,t=T_i}) - (1 - D_i)\sum_{t=1}^{T} \frac{K_M(i,t)}{C} e_{i,t}) \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)[D_i((Y_{i,t=T_i} - \mu_1^{T_i}(\mathbf{X}_{i,t})) + (\mu_1^{T_i}(\mathbf{X}_{i,t}) - \mu_0(\mathbf{X}_{i,t}) - \Delta)) \\
&\quad - (1 - D_i)\sum_{t=1}^{T} \frac{K_M(i,t)}{C}(Y_{i,t} - \mu_0(\mathbf{X}_{i,t}))] \\
&= \sum_{i=1}^{N} (W_i^* - \bar{W}^*)[D_i(Y_{i,t=T_i} - \mu_0(\mathbf{X}_{i,t}) - \Delta) - (1 - D_i)\sum_{t=1}^{T} \frac{K_M(i,t)}{C}(Y_{i,t} - \mu_0(\mathbf{X}_{i,t}))]
\end{aligned}
$$

$$
\sqrt{N_1}R^*_{1N_1} = \sum_{i=1}^{N} (W_i^* - \bar{W}^*)(D_i(\mu_0(\mathbf{X}_{i,t}) - \hat{\mu}_0(\mathbf{X}_{i,t})) - (1 - D_i)\frac{K_M(i,t)}{C}\sum_{t=1}^{T}(\mu_0(\mathbf{X}_{i,t}) - \hat{\mu}_0(\mathbf{X}_{i,t})))
$$

$$
\sqrt{N_1}R^*_{2N_1} = \sum_{i=1}^{N} (W_i^* - \bar{W}^*)D_i(\Delta - \hat{\Delta}_{adj})
$$

We have that $Pr\{\sqrt{N_1}R^*_{1N_1} > \epsilon|\mathbf{O}\} \xrightarrow{p} 0$ and $Pr\{\sqrt{N_1}R^*_{2N_1} > \epsilon|\mathbf{O}\} \xrightarrow{p} 0$ for any $\epsilon > 0$, by the same argument as Otsu and Rai (2017) which utilizes our Assumptions W, R, Lemma 2

and the Markov Inequality. This part of the argument also relies on Lemma 1, although the reliance is not explicit in Otsu and Rai (2017); see Abadie and Imbens (2011) for a similar derivation where the role of Lemma 1 is more clear.

Next, to show that that $\sup_r |Pr\{\sqrt{N_1}T^* \leq r|\mathbf{O}\} - Pr\{\sqrt{N_1}(\hat{\Delta}_{adj} - \Delta) \leq r\}| \xrightarrow{p} 0$, we define:

$$\begin{aligned}\eta_i &= [D_i(Y_{i,t=T_i} - \mu_0(\mathbf{X}_{i,t}) - \Delta) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \mu_0(\mathbf{X}_{i,t}))]/\sqrt{N_1} \\ &= D_i(e_{i,t=T_i} + \xi_{i,t=T_i}) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}e_{i,t}\end{aligned}$$

We have the following:

$$\begin{aligned}\sigma_N^2 &= \sigma_{1N}^2 + \sigma_2^2 \\ \sigma_{1N}^2 &= \frac{1}{N_1}\sum_{i=1}^{N} Var(\sum_{t=1}^{T}(D_i 1\{t=T_i\} + \frac{K_M(i,t)}{C}(1-D_i))Y_{it}|\mathbf{D},\mathbf{X}) \\ &= \frac{1}{N_1}\sum_{i=1}^{N} Cov\Big\{\sum_{t=1}^{T}(D_i 1\{t=T_i\} + \frac{K_M(i,t)}{C}(1-D_i))Y_{it}, \\ &\quad \sum_{t=1}^{T}(D_i 1\{t=T_i\} + \frac{K_M(i,t)}{C}(1-D_i))Y_{it}|\mathbf{D},\mathbf{X}\Big\} \\ &= \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t=T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t'=T_i\} + \frac{K_M(i,t')}{C}(1-D_i))Cov(Y_{it},Y_{it'}) \\ \sigma_2^2 &= E[((\mu_1^{T_i}(\mathbf{X}_{i,t}) - \mu_0(\mathbf{X}_{i,t})) - \Delta)^2|D_i=1]\end{aligned}$$

Within σ_{1N}^2, note that $Var(\sum_{t=1}^{T}(D_i 1\{t=T_i\} + \frac{K_M(i,t)}{C}(1-D_i))Y_{it}|\mathbf{D},\mathbf{X})$ reduces to $Var(Y_{it})$ for treated units. However, for control units, we have this variance term plus extra covariance terms between that control unit instance and the other instances in its trajectory. From here, we are able to follow the same proof strategy as in Otsu and Rai (2017), with our modified assumptions, and some modifications (detailed below) to Lemmas (i)-(iii) in Otsu and Rai (2017) to account for the extra covariance terms resulting from the trajectory structure of the control units. Lemma(i)-(iii) in Otsu and Rai (2017) show that the sampling variance of the η_i's converge to the population variance, σ_N^2, and that other random variables converge in probability to 0. They show this by leveraging the boundedness of higher order moments of η_i and use of the Markov inequality. We are able to utilize the same arguments, with changes laid out more explicitly for lemma (i), below, and generalized for lemmas (ii) and (iii).

To apply Lemmas (i)-(iii) in Otsu and Rai (2017) we define:

$$\hat{\sigma}_{1N}^2 = \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t=T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t'=T_i\} + \frac{K_M(i,t')}{C}(1-D_i))e_{i,t}e_{i,t'}$$

Then note that:

$$
\begin{aligned}
&E[(\hat{\sigma}^2_{1N} - \sigma^2_{1N})^2] \\
&= E[\{(\frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i))e_{i,t}e_{i,t'}) - \\
&(\frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i))Cov(Y_{it}, Y_{it'}))\}^2] \\
&= \frac{1}{N_1^2}E[\{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i)) \\
&\times (e_{i,t}e_{i,t'} - Cov(Y_{it}, Y_{it'}))\}^2] \\
&= \frac{1}{N_1^2}E[\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{j=1}^{N}\sum_{t_1=1}^{T}\sum_{t_1'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i)) \\
&\times (D_j 1\{t_1 = T_j\} + \frac{K_M(j,t_1)}{C}(1-D_j))(D_j 1\{t_1' = T_j\} + \frac{K_M(j,t_1')}{C}(1-D_j)) \\
&\times (e_{i,t}e_{i,t'} - Cov(Y_{it}, Y_{it'}))(e_{j,t_1}e_{j,t_1'} - Cov(Y_{jt_1}, Y_{jt_1'}))] \\
&= \frac{1}{N_1^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{t_1=1}^{T}\sum_{t_1'=1}^{T}E[(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1-D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1-D_i)) \\
&\times (D_i 1\{t_1 = T_i\} + \frac{K_M(i,t_1)}{C}(1-D_i))(D_i 1\{t_1' = T_i\} + \frac{K_M(i,t_1')}{C}(1-D_i)) \\
&\times E[(e_{i,t}e_{i,t'} - Cov(Y_{it}, Y_{it'}))(e_{i,t_1}e_{i,t_1'} - Cov(Y_{it_1}, Y_{it_1'}))|D_i, \mathbf{X}_{i,t}, \mathbf{X}_{i,t'}, \mathbf{X}_{i,t_1}, \mathbf{X}_{i,t_1'}]] \\
&\leq \frac{1}{N_1^2}(N_1 + cN_1^{1+o(1)})\times \\
&3 \times max(sup_{d,x}E[e_{it}^4|D_i = d, \mathbf{X}_{i,t} = x], \\
&- 2sup_{d,x,x'}E[e_{it}^2|D_i = d, \mathbf{X}_{i,t} = x]sup_{d,x,x'}Cov(Y_{it}, Y_{it'}|D_i = d, \mathbf{X}_{i,t} = x, \mathbf{X}_{i,t'} = x'), \\
&sup_{d,x,x'}Cov(Y_{it}, Y_{it'}|D_i = d, \mathbf{X}_{i,t} = x, \mathbf{X}_{i,t'} = x')^2) \to 0
\end{aligned}
$$

The inequality follows from Lemma 3. The convergence follows from Assumption M(4) and Lemma 2.

A.3.1 Difference-in-differences Estimator

This proof extends easily to the difference-in-differences ATT estimator described in Section 4.3 with the modifications described below.

We replace $\hat{\Delta}_{adj}$ with $\hat{\Delta}_{DiD}$ and the outcome $Y_{i,t}$ with the difference-in-differences outcome $Y_{i,t} - Y_{i,t-1}$

We modify

$$e_{i,t} = Y_{i,t} - Y_{i,t-1} - (\mu_{D_i}(\mathbf{X}_{i,t}) - \mu_{D_i}(\mathbf{X}_{i,t-1}))$$

and

$$\xi_{i,t} = (2D_i - 1)((\mu_{D_i}(\mathbf{X}_{i,t}) - \mu_{1-D_i}(\mathbf{X}_{i,t})) - (\mu_{D_i}(\mathbf{X}_{i,t-1}) - \mu_{1-D_i}(\mathbf{X}_{i,t-1})))$$

.

We also update our variance formulas to reflect the additional covariance terms introduced by the difference-in-differences outcome. In particular, our formula for σ^2_{1N} includes additional terms:

$$\begin{aligned}
\sigma^2_{1N} &= \frac{1}{N_1}\sum_{i=1}^{N} Var(\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(Y_{it} - Y_{i,t-1})|\mathbf{D}, \mathbf{X}) \\
&= \frac{1}{N_1}\sum_{i=1}^{N} Cov\Big\{\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(Y_{it} - Y_{i,t-1}), \\
&\sum_{t=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(Y_{it} - Y_{i,t-1})|\mathbf{D}, \mathbf{X}\Big\} \\
&= \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}(D_i 1\{t = T_i\} + \frac{K_M(i,t)}{C}(1 - D_i))(D_i 1\{t' = T_i\} + \frac{K_M(i,t')}{C}(1 - D_i))\times \\
&(Cov(Y_{it}, Y_{it'}) - Cov(Y_{it}, Y_{i,t'-1}) - Cov(Y_{i,t-1}, Y_{it'}) + Cov(Y_{i,t-1}, Y_{i,t'-1}))
\end{aligned}$$

To show that $E[(\hat{\sigma}^2_{1N} - \sigma^2_{1N})^2] \to 0$ we use a similar argument to before, where we are able to bound the difference between the $e_{i,t}e_{i,t'}$ and $Cov(Y_{it}, Y_{it'})$ by a fourth order polynomial in e_{it} terms and use the assumption that the expectation of these fourth moments is bounded.

B Weighted Least Squares

Weighted least squares (WLS) is commonly used with matching weights to calculate the ATT and its corresponding confidence interval (Ho et al., 2007; Stuart et al., 2011). For WLS, we have the following estimator for β:

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

Where W is a diagonal matrix with entries corresponding to the matching weights. Note that $W = W^T$. We can compute the variance as follows:

$$Var(\hat{\beta}) = (X^T W X)^{-1} X^T W [Var(Y)]((X^T W X)^{-1} X^T W)^T$$

Software to compute WLS estimates, such as lm and Zelig (which calls on lm) in R, often assumes that $Var(Y) = \sigma^2 W^{-1}$. If this is true we get a cancellation that yields a nicer formula for variance:

$$
\begin{aligned}
Var(\hat{\beta}) &= (X^TWX)^{-1}X^TW\sigma^2W^{-1}((X^TWX)^{-1}X^TW)^T \\
&= (X^TWX)^{-1}X^TWW^{-1}W^TX\sigma^2(W^TWX)^{-1} \\
&= (X^TWX)^{-1}(X^TWX)\sigma^2(W^TWX)^{-1} \\
&= \sigma^2(W^TWX)^{-1}
\end{aligned}
$$

In R, lm (and Zelig) use this formula[3] in order to compute standard errors and confidence intervals for WLS regression. However, when this assumption is not true, as is the case in our simulations (and often in practice), this formula is incorrect.

Assume $Var(Y) = \sigma^2 I$, as in some of our simulations, then:

$$
\begin{aligned}
Var(\hat{\beta}) &= (X^TWX)^{-1}X^TW\sigma^2I((X^TWX)^{-1}X^TW)^T \\
&= \sigma^2(X^TWX)^{-1}X^TW^2X(X^TWX)^{-1}
\end{aligned}
$$

To get an unbiased estimator of σ^2 we note that:

$$
E[e^TIe] = tr(I\Sigma_{ee})
$$

Where $e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^TWX)^{-1}X^TWY$. Now,

$$
\begin{aligned}
\Sigma_{ee} &= Cov(Y - X(X^TWX)^{-1}X^TWY, Y - X(X^TWX)^{-1}X^TWY) \\
&= Var(Y) - 2Cov(Y, X(X^TWX)^{-1}X^TWY) + Var(X(X^TWX)^{-1}X^TWY) \\
&= \sigma^2I - 2X(X^TWX)^{-1}X^TWVar(Y) + (X(X^TWX)^{-1}X^TW)^T(X(X^TWX)^{-1}X^TW)Var(Y) \\
&= \sigma^2(I - 2X(X^TWX)^{-1}X^TW + WX(X^TWX)^{-1}X^TX(X^TWX)^{-1}X^TW)
\end{aligned}
$$

Thus, to get an unbiased estimator of σ^2 we must divide e^Te by the trace of $I - 2X(X^TWX)^{-1}X^TW + WX(X^TWX)^{-1}X^TX(X^TWX)^{-1}X^TW$.

When we use this formula to create confidence intervals for WLS, instead of using lm, our simulations cover in the linear DGP, uncorrelated errors case.

[3] Last verified 04-11-2022.