Generalized Random Forest Event Studies

Chris Liao[*]

December 1, 2025

Abstract

This note introduces a procedure for estimating heterogeneous dynamic treatment effects in event studies with staggered adoption, allowing heterogeneity to vary by treatment adoption date and with observation-level time-invariant covariates. The approach reformulates the problem as estimating a conditional linear model with binary regressors and estimates all cohort–event time conditional treatment effects jointly using a single generalized random forest, with propensity scores adapted to staggered treatment adoption in panel settings.

1 Motivation

Event studies are widely used to estimate the causal effects of policies over multiple periods. A survey by Roth (2022) found 70 papers in three leading economics journals between 2014 and 2018 that employ event study plots. Researchers may also be interested in treatment effect heterogeneity within event study designs. The literature has developed methods to estimate treatment effects that are robust to heterogeneity arising from variation in treatment timing (Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Athey and Imbens, 2022), but largely omits potential heterogeneity arising from other sources. A separate literature leverages machine learning methods for heterogeneous treatment effect estimation (Wager and Athey, 2018; Athey, Tibshirani and Wager, 2019; Nie and Wager, 2021; Chernozhukov et al., 2025), but these methods are primarily designed for cross-sectional settings rather than panel data. Chernozhukov et al. (2019) develops estimators of

heterogeneous treatment effects that vary across units and over time using a low-rank factor structure but their framework is not designed for causal inference in staggered treatment adoption settings.

I present a procedure for estimating heterogeneous dynamic treatment effects in staggered treatment adoption event studies with unit-level time-invariant covariates. I show that this procedure is equivalent to estimating a special case of the conditional linear model with binary regressors and propensity scores adapted for staggered treatment adoption and panel settings.

I build on existing work combining estimation on event studies and heterogeneous treatment effects using machine learning. Wang (2022) and Miao et al. (2023) both embed the causal forest from Athey, Tibshirani and Wager (2019) within a differences-in-difference framework but only consider a single pre- and post-treatment period, so their analysis omits treatment effect dynamics. Miller (2020) use causal forests to estimate heterogeneous and time-varying policy effects but rely on conditional exogeneity for identification, whereas my procedure relies on conditional parallel trends.

The most related work is Gavrilova, Langørgen and Zoutman (2025). Both this paper and Gavrilova, Langørgen and Zoutman (2025) are interested in estimating the effect of unit-level time-invariant covariates on treatment effects in an event-study framework with potentially multiple pre- and post-treatment periods and staggered treatment adoption using generalized random forests (Athey, Tibshirani and Wager, 2019). Generalized random forests estimate conditional average treatment effects using locally weighted averages of observation-level treatment effects, with weights determined by forest-based similarity in covariates. Gavrilova, Langørgen and Zoutman (2025) allows these weights to vary by covariate values, treatment cohorts, and event time, since they estimate separate forests for each cohort–event time effect. My approach restricts variation to the covariate value level, since I only use one forest. I argue in the body that this enhances interpretability.

2 Setup

I observe units $i = 1, \ldots, n$, and each unit i is observed from time $t = s_i, \cdots, e_i$ where $1 \le s_i < e_i \le T$ represent the first and last period i is observed, respectively. Observations are assumed independent across units but may be serially correlated within units. Each unit i is either treated in period $G_i \in \{2, \ldots, T-1\} \equiv G$ or never treated ($G_i = \infty$). Treatments are irreversible and binary. Units are separated into cohorts $G_i = g$ based on treatment timing. Outcomes are denoted by $Y_{i,t}$ and $Y_{i,t}^\infty$ is

the potential outcome for unit i in period t if it were never treated.

Sun and Abraham (2021) define the cohort-specific average treatment effect on the treated (CATT) k periods after treatment. For a treated cohort $g \in G$,

$$CATT_{g,k} = \mathbb{E}[Y_{i,g+k} - Y_{i,g+k}^{\infty} \mid G_i = g]. \tag{1}$$

$CATT_{g,k}$ captures the average treatment effect at event time $k = t - G_i$ for units in treatment cohort g.

Define the cohort–event time treatment indicator as

$$Z_{i,t}^{g,k} = \mathbf{1}\{G_i = g\} \cdot \mathbf{1}\{t = g + k\}$$

where $Z_{i,t}^{g,k} = 1$ if and only if unit i belongs to cohort g and at time t is at event time k. I define the reference period as $k = -1$. Sun and Abraham (2021) show that when the saturated fixed-effects regression below is estimated,

$$Y_{i,t} = \alpha_i + \gamma_t + \sum_{g \in G} \sum_{k \neq -1} \delta_{g,k} Z_{i,t}^{g,k} + \varepsilon_{i,t} \tag{2}$$

$\delta_{g,k}$ identifies the $CATT_{g,k}$ from Equation 1. α_i and γ_t are unit and time effects, respectively.

For ease of exposition, I define

- $\delta = (\delta_{g,k})_{g \in G, \, k \neq -1}$ as the vector of cohort–event time coefficients, with one entry for each cohort g and event time k,

- $Z_{i,t} = (Z_{i,t}^{g,k})_{g \in G, \, k \neq -1}$ as the corresponding vector of cohort–event time treatment indicators.

I can then rewrite Equation 2 as

$$Y_{i,t} = \alpha_i + \gamma_t + \langle \delta, Z_{i,t} \rangle + \varepsilon_{i,t} \tag{3}$$

3 Heterogeneous Treatment Effects

3.1 The Model

Each unit i has a set of time-invariant covariates $X_i \in \mathcal{X}$. I am interested in how X_i affects the CATT, which motivates my estimand of inference: the conditional CATT,

defined as

$$CATT_{g,k}(x) = \mathbb{E}[Y_{i,g+k} - Y_{i,g+k}^\infty \mid G_i = g, X_i = x]. \tag{4}$$

This is the cohort–event time equivalent of Equation 1 in Gavrilova, Langørgen and Zoutman (2025).

The model I will estimate is

$$Y_{i,t} = \alpha_i + f_t(X_i) + \langle \delta(X_i), Z_{i,t} \rangle + \varepsilon_{i,t}. \tag{5}$$

The estimated equation differs from Equation 3 in two ways:

1. I replace time fixed effects with flexible, time-varying functions of $f_t(x)$ that capture baseline heterogeneity in the outcome,

2. I define conditional cohort–event time coefficients $\delta_{g,k}(x)$ as functions of covariates. Furthermore, define the vector $\delta(X_i) = \big(\delta_{g,k}(X_i)\big)_{g \in G,\, k \neq -1}$

I retain the linear fixed effects from Equation 2 because my covariates are time-invariant. See Johannemann et al. 2021 for work on implementing alternative, lower-dimensional representations of categorical variables into generalized random forests. Equation 5 embeds the possibility of a model with time fixed effects if $f_t(x)$ is constant across all x and Appendix Section A extends derivations to this setting,

3.2 Quasi-Treatment Dates and First Differences

My goal is to rewrite Equation 5 in the centered regression form of Robinson (1988). This requires partialing out the unit fixed effect, so that the (residualized) outcome is a function of the dot product of the (residualized) treatment vector and $\delta(x)$.

I will first-difference all observations with respect to the reference period to eliminate the unit-level fixed effects. Since the notion of a reference period is not well-defined for never-treated observations, to enable first-differencing, I introduce the notion of the *quasi-treatment date*.

Each never-treated unit first appears at time s_i. Define

$$F_G(g \mid s, x) = \frac{\sum_{j:s_j=s} \mathbb{1}\{G_j = g,\, G_j < \infty\} K(x_j, x)}{\sum_{j:s_j=s} \mathbb{1}\{G_j < \infty\} K(x_j, x)} \tag{6}$$

which is the conditional cumulative distribution function of treatment dates among treated units conditional on first appearance date s and covariates x. $K(\cdot, \cdot)$ is a weighting function over covariates. I assign all never-treated units i a quasi-treatment

date Q_i, drawn from $F_G(\cdot \mid s_i, x_i)$. For treated units, the quasi-treatment date equals the actual treatment date.

I will next show how first-differencing eliminates the unit fixed effect and the resulting first-differenced reformulation of Equation 5. The first-differenced outcomes, baseline heterogeneity function and error are

$$Y_{i,t}^{FD} = Y_{i,t} - Y_{i,Q_i-1}$$
$$f_t^{FD}(X_i) = f_t(X_i) - f_{Q_i-1}(X_i)$$
$$\varepsilon_{i,t}^{FD} = \varepsilon_{i,t} - \varepsilon_{i,Q_i-1}$$

First-differenced treatment indicators are unchanged because the reference period is already omitted. The first-differenced form of Equation 5 is hence

$$\begin{aligned} Y_{i,t}^{FD} &= f_t(X_i) - f_{Q_i-1}(X_i) + \langle \delta(X_i), Z_{i,t} \rangle + \varepsilon_{i,t}^{FD} \\ &= \langle f(X_i), T_{i,t} \rangle + \langle \delta(X_i), Z_{i,t} \rangle + \varepsilon_{i,t}^{FD}. \end{aligned} \tag{7}$$

For ease of exposition,

- $f(X_i) = (f_s(X_i))_{s=1}^T$ is the vector of baseline heterogeneity functions

- $T_{i,t} = \left(T_{i,t}^s\right)_{s=1}^T$ is the vector of indicators, where row $T_{i,t}^s = 1$ if $s = t$, $T_{i,t}^s = -1$ if $s = Q_i - 1$ and 0 otherwise.

Henceforth, the observation for event time -1 is omitted.

The consequence of my first differencing procedure is that units will only ever be compared to other units with the same quasi-treatment date. Gavrilova, Langørgen and Zoutman (2025) also difference the outcome relative to a reference period, but do not need to use a quasi-treatment date. This is because Gavrilova, Langørgen and Zoutman (2025) estimate each cohort–event time parameter using a separate forest. The treated observations for that forest is a single treatment cohort observed at the relevant event time and the control observations are the never-treated cohort observed at that same calendar time. Since each forest includes only one treatment cohort, all treated units share a single reference period, allowing outcomes to be differenced relative to that period without assigning quasi-treatment dates.

3.3 Identifying the Conditional CATT

The following assumptions enable $\delta_{g,k}$ to identify the $CATT_{g,k}(x)$. The proof is in Appendix Section B.

Assumption 1 (Conditional Parallel Trends). *For all time periods t, $x \in \mathcal{X}$, and treatment cohorts $g \in \{G \cup \{\infty\}\}$*

$$\mathbb{E}[Y_{i,t}^{\infty} - Y_{i,Q_i-1}^{\infty} \mid G_i = g, X_i = x] = f_t(X_i) - f_{Q_i-1}(X_i)$$

It then follows that $\mathbb{E}[\epsilon_{i,t}^{FD} \mid X_i = x] = 0$

Assumption 2 (Conditional No Anticipation). *For all $k < 0$ and $x \in \mathcal{X}$,*

$$\mathbb{E}[Y_{i,g+k} \mid X_i = x] = \mathbb{E}[Y_{i,g+k}^{\infty} \mid X_i = x] \quad \text{for all } g \in G.$$

I also require that the overlap assumption and regularity conditions from Athey, Tibshirani and Wager (2019) hold.

3.4   Centered Regression Form

Define the conditional means

$$\begin{aligned}
e(x) &= \mathbb{E}[Z_{i,t} \mid X_i = x] \\
t(x) &= \mathbb{E}[T_{i,t} \mid X_i = x] \\
m(x) &= \mathbb{E}[Y_{i,t}^{FD} \mid X_i = x] = \langle f(x), t(x) \rangle + \langle \delta(x), e(x) \rangle
\end{aligned}$$

Subtracting the conditional means from the model in Equation 7 yields the centered regression from Robinson (1988).

$$Y_{i,t}^{FD} - m(X_i) = \langle f(X_i), T_{i,t} - t(X_i) \rangle + \langle \delta(X_i), Z_{i,t} - e(X_i) \rangle + \varepsilon_{i,t}^{FD} \qquad (8)$$

Define the centered treatment, baseline heterogeneity assignment and outcome as

$$\begin{aligned}
\widetilde{Z}_{i,t} &= Z_{i,t} - e(X_i) \\
\widetilde{T}_{i,t} &= T_{i,t} - t(X_i) \\
\widetilde{Y}_{i,t}^{FD} &= Y_{i,t}^{FD} - m(X_i)
\end{aligned}$$

The key insight from Robinson (1988) and Athey, Tibshirani and Wager (2019) is that if there is some neighborhood S where for all $x \in S$, $\delta(x)$ and $f(x)$ are constant, then we can estimate $\delta(x)$ and $f(x)$ by estimating Equation 8 on the subsample of observations j where $X_j \in S$.

Hence, Athey, Tibshirani and Wager (2019) show that for any covariate values x,

$\delta(x)$ is identified by the solution to

$$(\delta(x), f(x)) = \text{Var}[(\widetilde{Z}_{i,t} + \widetilde{T}_{i,t}) \mid X_i \in S]^{-1} \text{Cov}[(\widetilde{Z}_{i,t} + \widetilde{T}_{i,t}), \widetilde{Y}_{i,t}^{FD} \mid X_i \in S] \quad (9)$$

3.5 Estimation Framework

The finite-sample equivalent to Equation 9 is

$$(\hat{\delta}(x), \hat{f}(x)) = \left(\sum_{i,t} \alpha_i(x) \left(Z_{i,t} - e^{(-i)}(x) + T_{i,t} - t^{(-i)}(x) \right)^{\otimes 2} \right)^{-1} \cdot$$
$$\sum_{i,t} \alpha_i(x) \left(Z_{i,t} - e^{(-i)}(x) + T_{i,t} - t^{(-i)}(x) \right) \left(Y_{i,t} - m^{(-i)}(x) \right).$$

Here, $\alpha_i(x)$ are generalized random forest kernel weights measuring the relevance of observation i for estimating the $\delta(x)$ and are the observation-specific sample analog of the neighborhood S (Athey, Tibshirani and Wager, 2019).

Since Gavrilova, Langørgen and Zoutman (2025) estimate a separate forest for each cohort–event time estimate, each cohort–event time estimate will rely on upon cohort–event time kernel weights $\alpha_i^{g,k}(x)$. In the shared setting of this note and Gavrilova, Langørgen and Zoutman (2025), unit-level covariates are time-invariant. Because each $\delta_{g,k}(x)$ is a weighted average of other observations i's treatment effects, the setting of Gavrilova, Langørgen and Zoutman (2025) implies that even for identical covariate values x, the same observation i may be weighted differently depending on the treatment cohort or event time of the parameter being calculated.

Since my approach uses a single forest to estimate all parameters, the weighting structure $\alpha_i(x)$ is shared across all treatment cohort and event time combinations. This ensures that unit i's contribution to the conditional treatment effect remains consistent conditional on covariate value x. Note that my approach does not restrict treatment effects to be the same across cohort–event times conditional on covariates - it only restricts the weights used in the weighted average of other observations to be the same.

Suppose that the data are divided into k equally sized folds. $e^{(-i)}(\cdot)$, $t^{(-i)}(\cdot)$, and $m^{(-i)}(\cdot)$ are the treatment, baseline heterogeneity and outcome function estimated via k-fold cross-fitting, using the $k-1$ folds that omit observation i. Cross-fitting enables valid inference on $\delta(X_i)$ and $f(X_i)$ even when the estimated outcome and propensity functions have small errors because the conditional means functions are not trained on unit i (Chernozhukov et al., 2018).

Estimation of the propensity scores $e(x)$ and $t(x)$ requires modifying existing ap-

proaches from the literature because of the panel data setting. Since treatment is at the cohort–event time level, in a given time period t, if an observation belongs to treatment cohort g, it is immediately known that for all cohort–event time pairs that $Z_{i,t}^{g,k} = 0$, for $k \neq t - g$. A similar phenomenon occurs with $T_{i,t}$. Appendix Section C describes the mathematical formulation to estimate $e(x)$ and $t(x)$. The outcome function $m(x)$ can be estimated following standard procedure. Since Gavrilova, Langørgen and Zoutman (2025) use separate forests for each cohort–event time estimate, their treatment vectors, and hence propensity scores, do not run into the co-determinism problem above.

Appendix Section D shows how doubly robust scores for $\hat{\delta}(x)$ are estimated.

3.6 Discussion of Weighting Differences

The primary advantage of my weighting approach is interpretability, whereas Gavrilova, Langørgen and Zoutman (2025) prioritize modeling flexibility. Restricting the weights to depend only on covariate values, as I do, ensures that treatment effect heterogeneity varies only with covariates and does not additionally depend on treatment cohort or event time. This greatly simplifies interpretation: two observations with the same covariate profile contribute to the conditional treatment effect in the same way, regardless of when they adopt treatment. Such a restriction is particularly appealing in settings where treatment timing is plausibly idiosyncratic or otherwise uninformative about treatment effect heterogeneity. This restriction also makes it easier to deploy interpretability tools, such as the policy tree Athey and Wager (2021).

By contrast, allowing weights to vary with covariates, treatment cohort, and event time–as in Gavrilova, Langørgen and Zoutman (2025)–provides the most flexible way to model how time-invariant covariates influence outcomes. One benefit of this flexibility is that can illuminate differences how covariates affect short-run versus long-run treatment effects, or across cohorts that adopt a policy at different times. However, it also introduces an interpretability challenge: when treatment effects are allowed to vary arbitrarily across treatment cohorts, it becomes unclear what one learns from aggregating these effects into event-time profiles. These aggregated parameters may represent mixtures of fundamentally different groups, reducing their relevance for policy. For example, if two treated units exhibit very different covariate–treatment effect relationships, the average of their effects may be uninformative for predicting the effect on a third unit; it may be more informative to identify which of the two units the third more closely resembles and base predictions on that individual unit's effect instead.

References

Athey, Susan, and Guido W. Imbens. 2022. "Design-based analysis in Difference-In-Differences settings with staggered adoption." *Journal of Econometrics*, 226(1): 62–79. Annals Issue in Honor of Gary Chamberlain.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning With Observational Data." *Econometrica*, 89(1): 133–161.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics*, 47(2): 1148–1178.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Chernozhukov, Victor, Christian Hansen, Yuan Liao, and Yinchu Zhu. 2019. "Inference for heterogeneous effects using low-rank estimation of factor slopes." *arXiv preprint arXiv:1812.08089*.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Causal Parameters." *The Econometrics Journal*, 21(1): C1–C68. Publisher: Royal Economic Society.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2025. "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India." *Econometrica*, 93(4): 1121–1164.

Gavrilova, Evelina, Audun Langørgen, and Floris T. Zoutman. 2025. "Difference-in-Difference Causal Forests With an Application to Payroll Tax Incidence in Norway." *Journal of Applied Econometrics*.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277. Themed Issue: Treatment Effect 1.

Johannemann, Jonathan, Vitor Hadad, Susan Athey, and Stefan Wager. 2021. "Sufficient Representations for Categorical Variables." _eprint: 1908.09874.

Miao, Wei, Yiting Deng, Wei Wang, Yongdong Liu, and Christopher S. Tang. 2023. "The Effects of Surge Pricing on Driver Behavior in the Ride-Sharing Market: Evidence from a Quasi-Experiment." *Journal of Operations Management*, 69(5): 794–822.

Miller, Steve. 2020. "Causal forest estimation of heterogeneous and time-varying environmental policy effects." *Journal of Environmental Economics and Management*, 103: 102337.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle Estimation of Heterogeneous Treatment Effects." *Biometrika*, 108(2): 299–319.

Robinson, Peter M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica*, 56(4): 931–954. Publisher: Econometric Society.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights*, 4(3): 305–322.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, 225(2): 175–199. Themed Issue: Treatment Effect 1.

Uysal, S. Derya. 2015. "Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling." *Journal of Applied Econometrics*, 30(5): 763–786.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wang, Guihua. 2022. "The Effect of Medicaid Expansion on Wait Time in the Emergency Department." *Management Science*, 68(9): 6648–6665. Publisher: INFORMS.

A   The linear time fixed effects model

A special case of Section 3 arises when $f_t(X_i)$ affects outcomes uniformly within each period t, implying a time fixed effect. The model from Section 5 then simplifies to

$$Y_{i,t} = \alpha_i + \gamma_t + \langle \delta(X_i), Z_{i,t} \rangle + \varepsilon_{i,t}.$$

with first difference

$$Y_{i,t}^{FD} = \gamma_t - \gamma_{Q_i-1} + \langle \delta(X_i), Z_{i,t} \rangle + \varepsilon_{i,t}^{FD}.$$

Note that the average of the first-differenced outcome across all never-treated individuals at time t belonging to the quasi-treatment cohort $Q_i = g$ is $\gamma_t - \gamma_{g-1}$. Hence, I can define the demeaned outcome and residual as

$$\widetilde{Y}_{i,t}^{FD} = Y_{i,t}^{FD} - \frac{1}{|j \in Q_i, G_j = \infty|} \sum_{j \in Q_i, G_j = \infty} Y_{j,t}^{FD}$$

$$\widetilde{\varepsilon}_{i,t}^{FD} = \varepsilon_{i,t}^{FD} - \frac{1}{|j \in Q_i, G_j = \infty|} \sum_{j \in Q_i, G_j = \infty} \varepsilon_{j,t}^{FD}$$

and rewrite the regression as

$$\widetilde{Y}_{i,t}^{FD} = \langle \delta(X_i), Z_{i,t} \rangle + \widetilde{\varepsilon}_{i,t}^{FD}.$$

The framework in Section 3.4 can then be used to estimate $\delta(x)$, except that the baseline heterogeneity functions are omitted.

B   Proof of conditional CATT identification

From Equation 7, taking expectations conditional on cohort membership G_i and covariates $X_i = x$ gives

$$\mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = g, X_i = x] = f_{g+k}(x) - f_{g-1}(x) + \delta_{g,k}(x)$$

$$\mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = \infty, X_i = x] = f_{g+k}(x) - f_{g-1}(x).$$

Then,

$$\mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = g, X_i = x] - \mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = \infty, X_i = x]$$
$$= \delta_{g,k}(x)$$

Finally,

$$
\begin{aligned}
\delta_{g,k}(x) &= \mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = g, X_i = x] - \mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = \infty, X_i = x] \\
&= \mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = g, X_i = x] - \mathbb{E}[Y_{i,g+k}^{FD,\infty} \mid Q_i = g, G_i = \infty, X_i = x] \\
&\quad \text{by the definition of being untreated} \\
&= \mathbb{E}[Y_{i,g+k}^{FD} \mid Q_i = g, G_i = g, X_i = x] - \mathbb{E}[Y_{i,g+k}^{FD,\infty} \mid Q_i = g, G_i = g, X_i = x] \\
&\quad \text{by the parallel trends assumption in Assumption 1} \\
&= \mathbb{E}[Y_{i,g+k}^{FD} \mid G_i = g, X_i = x] - \mathbb{E}[Y_{i,g+k}^{FD,\infty} \mid G_i = g, X_i = x] \\
&= \mathbb{E}[Y_{i,g+k}^{FD} - Y_{i,g+k}^{FD,\infty} \mid G_i = g, X_i = x] \\
&= \mathbb{E}[Y_{i,g+k} - Y_{i,g-1} - (Y_{i,g+k}^{\infty} - Y_{i,g-1}^{\infty}) \mid G_i = g, X_i = x] \\
&= \mathbb{E}[Y_{i,g+k} - Y_{i,g-1}^{\infty} - (Y_{i,g+k}^{\infty} - Y_{i,g-1}^{\infty}) \mid G_i = g, X_i = x] \\
&\quad \text{by the no anticipation assumption in Assumption 2} \\
&= \mathbb{E}[Y_{i,g+k} - Y_{i,g+k}^{\infty} \mid G_i = g, X_i = x] \\
&= CATT_{g,k}(x)
\end{aligned}
$$

C   Propensity scores in panel data settings

Estimating the treatment propensity $E[Z_{i,t} \mid X_i]$ is a nontrivial task because of the panel data setting and staggered treatment adoption. Recall $Z_{i,t}^{g,k} = \mathbf{1}\{G_i = g\}\mathbf{1}\{t = g + k\}$. Now, for some (g,k) any for observation (i,t), conditional on covariates X_i,

$$\mathbb{E}[Z_{i,t}^{g,k} \mid X_i] = \mathbb{E}\big[\mathbf{1}\{G_i = g\}\mathbf{1}\{t = g + k\} \mid X_i = x\big] = \Pr(G_i = g, t = g + k \mid X_i).$$

By the chain rule for probabilities,

$$\Pr(G_i = g, t = g + k \mid X_i) = \Pr(G_i = g \mid X_i) \cdot \Pr(t = g + k \mid G_i = g, X_i).$$

$\Pr(G_i = g \mid X_i)$, is the probability of treatment time conditional on covariates. For $\Pr(t = g + k \mid G_i = g, X_i)$, notice that conditional on g and k, for any observation

(i, t), we immediately know whether $t = g + k$. Hence,

$$\Pr(G_i = g \mid X_i) \cdot \mathbf{1}\{t = g + k\}.$$

There are many known procedures for estimating $\Pr(G_i = g \mid X_i)$, including using a logit or probability forest.

The estimation of $\mathbb{E}[T_{i,t} \mid X_i = x]$ follows a similar logic. Note that the s-th row of $T_{i,t}$ can be redefined as

$$T_{i,t}^s = \mathbf{1}\{s = t\} + \mathbf{1}\{s = Q_i - 1\}.$$

The two indicators will never both be 1 because all observations at event time -1 are omitted. Hence,

$$
\begin{aligned}
\mathbb{E}[T_{i,t}^s \mid X_i] &= \mathbb{E}[\mathbf{1}\{s = t\} \mid X_i] + \mathbb{E}[\mathbf{1}\{s = Q_i - 1\} \mid X_i] \\
&= \mathbb{E}[\mathbf{1}\{t = s\} \mid X_i] + \mathbb{E}[\mathbf{1}\{Q_i = s + 1\} \mid X_i] \\
&= \mathbf{1}\{t = s\} + \Pr(Q_i = s + 1 \mid X_i).
\end{aligned}
$$

since the time period t and row s are known by definition.

$\Pr(Q_i = s + 1 \mid X_i)$ is also already known. I can integrate over the conditional empirical distribution of treatment dates conditional on an observation's first appearance in the data to obtain the conditional distribution of quasi-treatment dates.

D   Calculating Doubly Robust Scores

Since all my treatment variables are indicators and for all treated units i, $\sum_{g,k} Z_{i,t} = 1$, I can reframe staggered treatment adoption treatment as a single multivalued treatment, where each treatment arm corresponds to a treatment cohort–event time. I can then estimate doubly robust treatment effects using existing work on calculating doubly robust scores when treatments are multi-valued (Uysal 2015).

Define the average baseline outcome as

$$\widehat{Y}_{i,t}^{\text{baseline}} = m^{(-i)}(X_i) - \sum_{g,k} \widehat{P}(G_i = g \mid X_i)\mathbf{1}\{t = g + k\}\widehat{\delta}_{g,k}(X_i).$$

The predicted baseline outcome, accounting for which units are treated and untreated

is hence

$$
\widehat{\mu}_{i,t}^{g,k} = \begin{cases} \widehat{Y}_{i,t}^{FD,\text{baseline}} + \widehat{\delta}_{g,k}(X_i), & \text{if } G_i \neq \infty, \\ \widehat{Y}_{i,t}^{FD,\text{baseline}}, & \text{if } G_i = \infty. \end{cases}
$$

Define the outcome residual as

$$
\widetilde{Y}_{i,t}^{g,k} = Y_{i,t}^{FD} - \widehat{\mu}_{i,t}^{g,k}.
$$

For a given column (g,k) of observation (i,t), the inverse propensity estimate score is

$$
\text{IPW}_{i,t}^{g,k} = \frac{\mathbf{1}\{G_i = g\}\mathbf{1}\{t = g+k\}}{P(G_i = g \mid X_i)} - \frac{\mathbf{1}\{G_i = \infty\}}{P(G_i = \infty \mid X_i)}.
$$

Hence, the doubly robust score for $\hat{\delta}_{g,k}(X_i)$ is

$$
\begin{aligned}
\widehat{\varphi}_{i,t}^{g,k} &= \widehat{\delta}_{g,k}(X_i) + \text{IPW}_{i,t}^{g,k} \cdot \widetilde{Y}_{i,t}^{g,k} \\
&= \widehat{\delta}_{g,k}(X_i) + \left(\frac{\mathbf{1}\{G_i = g\}\mathbf{1}\{t = g+k\}}{P(G_i = g \mid X_i)} - \frac{\mathbf{1}\{G_i = \infty\}}{P(G_i = \infty \mid X_i)} \right) \cdot \left(Y_{i,t}^{FD} - \widehat{\mu}_{i,t}^{g,k} \right).
\end{aligned}
$$

E   Notes on Implementation

The procedure allows for consistent and unbiased estimation of CATT(x), following Athey, Tibshirani and Wager 2019. One can use the conditionally linear model forest (lm_forest) from the R package grf and should cluster by observations when using panel data. Alternatively, one can use the multi-arm causal forest, which is the multivariate extension of the R-learner from Nie and Wager 2021 if they partial out the baseline heterogeneity functions from Equation 7.[1] The literature on event study and difference-in-differences also provides weights that can be used to consistently aggregate $CATT(x)$ into more aggregate quantities of interest (Callaway and Sant'Anna 2021, Sun and Abraham 2021). By default, for each observation, lm_forest will produce $CATT(x)$ values for many treatment cohorts. Only the treatment cohort–event time estimate pertaining to that (i,t) should be used when estimating doubly robust scores.

---

[1]Note that Nie and Wager 2021 solves a global minimization function to estimate $\delta(x)$, as opposed to estimating a local parametric model of $\delta(x)$.