The Econometrics Journal

Econometrics Journal (2023), volume **26**, pp. C31–C66.
https://doi.org/10.1093/ectj/utad016

Simple approaches to nonlinear difference-in-differences with panel data

JEFFREY M. WOOLDRIDGE

Department of Economics Michigan State University.
E-mail: wooldri1@msu.edu

First version received: 19 August 2022; final version accepted: 6 February 2023.

Summary: I derive simple, flexible strategies for difference-in-differences settings where the nature of the response variable may warrant a nonlinear model. I allow for general staggered interventions, with and without covariates. Under an index version of parallel trends, I show that average treatment effects on the treated (ATTs) are identified for each cohort and calendar time period in which a cohort was subjected to the intervention. The pooled quasi-maximum likelihood estimators in the linear exponential family extend pooled ordinary least squares estimation of linear models. By using the conditional mean associated with the canonical link function, imputation and pooling across the entire sample produce identical estimates. Generally, pooled estimation results in very simple computation of the ATTs and their standard errors. The leading cases are a logit functional form for binary and fractional outcomes—combined with the Bernoulli quasi-log likelihood (QLL)—and an exponential mean combined with the Poisson QLL.

Keywords: *Difference-in-differences*, *staggered intervention*, *nonlinear model*, *logit model*, *Poisson regression*.

JEL codes: *C23*, *C54*.

1. INTRODUCTION

Difference-in-differences (DiD) methods have become very popular for intervention analysis, with recent emphasis on estimating average treatment effects when the intervention is implemented as a staggered rollout. In Wooldridge (2005), I studied properties of estimators that allow heterogeneous slopes, including the case of time-varying treatment interventions. But the analysis was incomplete, failing to investigate situations where allowing for effects to vary by treatment intensity identifies interesting average treatment effects. Work by de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021) showed that the common practice of assuming a constant treatment effect and using the two-way fixed effects (TWFE) estimator can lead to an estimated coefficient that is difficult to interpret. Sun and Abraham (2021) study the properties of event study estimators and propose weighted estimators that allow for heterogeneous treatment effects. Borusyak et al. (2022) [BJS (2022)] derive imputation estimators, based on TWFE, to obtain heterogenous treatment effects. Callaway and Sant'Anna (2021) [CS (2021)] apply treatment effects estimators to long differences, using both never treated and already treated units as control groups. The recent survey by de Chaisemartin and D'Haultfœuille (2023) provides a valuable discussion of these papers, and more.

In Wooldridge (2021) I showed that, for staggered interventions under the commonly imposed assumptions of no anticipation (NA) and parallel trends (PT), one can recover average treatment effects on the treated (ATTs) using a linear model that includes treatment cohort dummies, time period dummies, and a set of cohort-by-time treatment indicators. When estimated by pooled ordinary least squares (POLS), the coefficients on the treatment dummies are consistent for the cohort/time-specific ATTs. Moreover, the random effects and TWFE estimators are identical to the POLS estimators, with the latter equivalence establishing that controlling for a small number of cohort dummies is the same as controlling for a potentially large number of unit-specific dummies. I also showed that a natural two-step imputation approach leads to the same ATT estimates as pooling across all of the data. Time-constant covariates are allowed so that the PT assumption only needs to hold after conditioning on covariates, and the equivalences among the estimators continue to hold. Moreover, the POLS/TWFE estimators of the ATTs are identical to the BJS (2022) estimators. An implication of these equivalence results is that there is nothing inherently wrong with TWFE estimation: One simply needs to apply the method to a suitably flexible linear equation. Nevertheless, consistent estimation of the ATTs by POLS/TWFE requires that PT applies to the mean of the potential outcomes—a *linear parallel trends* (LPT) assumption.

In the current paper, I allow for versions of the PT assumption that may be more attractive when the outcome variable, Y_t, is limited in range. For example, when Y_t is binary, the LPT assumption is potentially suspect. This point was made by Blundell and Costa Dias (2009) with two periods and a binary outcome, and their proposed solution is a special case of the framework I propose here with two periods and no covariates. If Y_t is a fractional response, the LPT assumption is also questionable—especially when that outcome regularly takes on values near or at the endpoints of zero and one. Another common situation where the LPT assumption might fail is when Y_t is a count variable or a corner solution outcome that can take on the value zero with some regularity as well as large values. An arguably better assumption is that PT holds in the ratio of the means, which is implied by an exponential conditional mean function.

In the two-period case, Roth and Sant'Anna (2023) study the problem of invariance of the PT assumption to strictly monotonic transformations of the outcome variable. They show that invariance requires a strong form of PT stated in terms of the cumulative distribution function for the potential outcome in the untreated state. Even in their $T = 2$ setting without covariates, the Roth–Sant'Anna setting does not apply to the models I study here. Rather than studying transformations of $Y_t(0)$, the outcome in the untreated state, I propose nonlinear conditional mean functions that satisfy a kind of PT assumption. As is well known, the mean of the transformation and the transformation of the mean are not the same for nonlinear functions. To appreciate the difference, when the potential outcome in the untreated state, $Y_t(0)$, is binary, the Roth–Sant'Anna result is that PT either holds for any coding of the variable as $\{a, b\}, a < b$, or not at all. By contrast, in Section 2 I present an example where the PT assumption fails for $E[Y_t(0)|D]$, where D is the treatment indicator, but is satisfied for a strictly increasing transformation, $G^{-1}(E[Y_t(0)|D])$. For common choices of $G(\cdot)$, $G^{-1}(Y_t(0))$ is not even defined. For a nonnegative variable $Y_t(0)$ with $P(Y_t(0) = 0) > 1$, it is easily seen that the PT assumption can hold for $\log(E[Y_t(0)|D])$, but not for $E[Y_t(0)|D]$. With zero as a possible outcome, $\log(Y_t(0))$ is not well defined, and so it cannot be considered in the Roth–Sant'Anna setting. Here I assume that we have an outcome variable of interest—such as an employment indicator, the fraction of workers participating in a pension plan, the number of murders at the county level, or the amount of a new fertilizer used by a farmer—and we would like to estimate ATTs in terms of the averages of this outcome.

When $T = 2$ and $Y_t(0)$ is continuous, the analysis here is less general than the 'changes-in-changes' approach proposed by Athey and Imbens (2006) [AI (2006)]. AI (2006) allow for an

unknown function assumed to be strictly increasing in unobserved heterogeneity, and they derive an estimator of the average treatment effect on the treated (ATT) using estimated cumulative distribution functions. Unfortunately, point identification is lost when the response variable has discreteness unless additional assumptions are imposed. Melly and Santangelo (2015) extend the AI (2006) approach to allow for covariates, but it is unclear how to extend the approach to many time periods and, especially, to staggered interventions.

From a practical perspective, the current paper provides simple strategies that allow empirical researchers to compare results from linear DiD analyses with sensible nonlinear alternatives. The approach only requires specification of a conditional mean function, imposing no extra assumptions. All other conditional moments, for both the marginal and joint distributions across time, are unrestricted. Using pooled quasi-maximum likelihood (QMLE) in the class of linear exponential family (LEF) distributions ensures that the resulting estimators of the ATTs do not rely on other distributional assumptions, nor on restrictions in patterns of serial dependence across time. As often happens in 'nonlinear' settings, a special case is POLS estimation of a linear equation, which I proposed in Wooldridge (2021). Consequently, in settings with limited dependent variables, the methods here allow one to assess robustness of findings by comparing estimates from a linear model with those from a suitable nonlinear model.

With nonlinear models in a panel data setting with a small number of time periods, fixed effects strategies do not generally produce reliable estimators of parameters or partial effects. Technically, including a full set of unit dummy variables usually results in an incidental parameters problem. Therefore, the BJS (2022) imputation approach generally is not available for nonlinear conditional mean functions. The only known nonlinear model/estimation method where including unit-specific dummies does not lead to the incidental parameters problem is an exponential model estimated by Poisson fixed effects (FE). For nonnegative outcomes, I briefly discuss the possibility of estimating the ATTs using Poisson FE in Section 3.

In the linear case, the equivalence between TWFE and POLS suggests that pooled methods that control for cohort dummies in nonlinear conditional means may have good properties. As I demonstrate in Sections 2 and 3, this turns out to be the case: The ATTs are identified from a conditional expectation that involves cohort and time dummies, along with covariates, under a modified PT assumption.

Identification of the ATTs is based on random variables representing an underlying population of units present over time periods $t \in \{1, 2, ..., T\}$. When I describe estimation, it is easiest to think of random sampling N units from the cross section and collecting information on all T periods. Based on asymptotic analysis, inference is straightforward in a scenario where T is fixed and N grows without bound. The estimation methods are valid in other situations—for example, with clustering in the treatment assignment, with large-T panels, and even with spatial correlation—but issues regarding calculation of standard errors and handling potential spillover effects would have to be treated in more detail. The underlying regularity conditions are unremarkable and so I omit those details.

The remainder of this paper is organized as follows. I start with the two-period case in Section 2 in order to demonstrate identification in the simplest setting. Section 3 turns to the general case of staggered interventions over many time periods. The weakest assumptions for identification lead naturally to an imputation procedure. Nevertheless, I formally show that the pooled QMLE using the canonical link function in the LEF is numerically identical to the imputation procedure.

Section 4 discusses testing and correcting for violations of PT. In Section 5, I summarize simulation findings for binary and nonnegative outcomes. In Section 6, I apply the Poisson regression estimator with an exponential mean function to the car thefts data, at the block level, in Di Tella

and Schargrodsky (2004). Section 7 discusses four extensions that are relatively straightforward using my proposed framework: How to allow for all units being eventually treated; a strategy for handling (staggered) exit from treatment; how to incorporate time-varying covariates; and an approach to nonbinary treatments. Section 8 contains concluding remarks. An appendix contains the proof of the key algebraic equivalence results and the Online Appendix includes the details of the simulations.

2. THE $T = 2$ CASE

To understand the nature of the key assumptions used in the general staggered case, it is helpful to begin with the $T = 2$ case—the setting studied in Heckman et al. (1997), Abadie (2005), Athey and Imbens (2006), and Sant'Anna and Zhou (2020). The identification discussion is in terms of random variables representing a population. Estimation can proceed under different sampling schemes, but it is easiest to think of obtaining a random sample from the population, with unrestricted correlation allowed across the two time periods.

Index the potential outcomes with a time subscript, $t \in \{1, 2\}$, where the first period is the control period. The potential outcomes are denoted $Y_t(0)$ and $Y_t(1)$, where the value in $(\cdot)$ indicates the state of the world (untreated or treated). The time-constant binary treatment indicator is D, with $D = 1$ implying treatment after the first period and prior to the second. Consequently, $Y_1(0)$ and $Y_1(1)$ are the potential outcomes in the period before treatment status has been assigned.

The parameter of interest is the average treatment effect on the treated (ATT) in the second time period ($t = 2$):

$$\tau_2 = E[Y_2(1) - Y_2(0) \,|\, D = 1]. \tag{2.1}$$

Two assumptions serve to identify τ_2. The first is a *no anticipation* (NA) assumption. The strongest form of the assumption is $Y_1(1) = Y_1(0)$; this is the version used by Heckman et al. (1997), Abadie (2005), and others—sometimes only implicitly. See also Abbring and Van den Berg (2003) for a discussion of the NA assumption in the context of duration analysis.

A weaker version of NA is sufficient for the development here:

$$E[Y_1(1) - Y_1(0) \,|\, D = 1] = 0, \tag{2.2}$$

which means, on average, among the eventually treated group there are no anticipatory changes that affect the potential outcomes prior to the intervention.

The second assumption imposed to identify τ_2 is the *parallel trends* (PT) assumption, also called *common trends*. For the extension to the nonlinear case, it is useful to state the PT assumption for the linear case in two parts:

$$E[Y_1(0) \,|\, D] = \alpha + \beta D, \tag{2.3}$$

$$E[Y_2(0) \,|\, D] = \alpha + \beta D + \gamma_2. \tag{2.4}$$

By itself, (2.3) is unrestrictive because it is the same as simply defining

$$\alpha \equiv E[Y_1(0) \,|\, D = 0], \beta \equiv E[Y_1(0) \,|\, D = 1] - E[Y_1(0) \,|\, D = 0].$$

That $E[Y_1(0) \,|\, D]$ is allowed to depend on D means that assignment can depend systematically on the potential outcome in the control state in the first time period. Only when (2.3) is combined

with (2.4) is an assumption imposed, namely,

$$E\left[Y_2\left(0\right)|D\right]-E\left[Y_1\left(0\right)|D\right]=E\left[Y_2\left(0\right)-Y_1\left(0\right)|D\right]=\left(\alpha+\beta D+\gamma_2\right)-\left(\alpha+\beta D\right)=\gamma_2. \tag{2.5}$$

In other words, on average, the trend $Y_2\left(0\right)-Y_1\left(0\right)$ in the control state does not differ across the control and treated groups—a substantive restriction. I will call (2.5) the *linear parallel trends* (LPT) assumption. The representation in (2.3) with β unrestricted recognizes that the average level of the outcome in the first period can systematically change with D, allowing for the kind of selection that causes systematic bias if we were to observe only a single time period. The LPT assumption rules out the possibility that selection into treatment is based on the trend in the untreated state. If an administrator is choosing participants in a job training programme, the LPT assumption allows that selection may be based on differences in pre-treatment earnings, but not on how the earnings are expected to trend in the absence of the intervention.

When $Y_t\left(0\right)$ is limited in some important way—for example, it is binary, a fraction, or is restricted to be nonnegative—the LPT assumption can be unrealistic. Instead, assume that for a known, strictly increasing, continuously differentiable function $G\left(\cdot\right)$,

$$E\left[Y_1\left(0\right)|D\right]=G\left(\alpha+\beta D\right), \tag{2.6}$$

$$E\left[Y_2\left(0\right)|D\right]=G\left(\alpha+\beta D+\gamma_2\right). \tag{2.7}$$

The key restriction is that $G\left(\cdot\right)$ is strictly increasing, with smoothness assumed only so that standard asymptotic theory applies. Just as when $G\left(\cdot\right)$ is the identify function, (2.6) imposes no restriction in isolation because it simply implies, definitionally, $\alpha=G^{-1}\left(E\left[Y_1\left(0\right)|D=0\right]\right)$ and $\beta=G^{-1}\left(E\left[Y_1\left(0\right)|D=1\right]\right)-G^{-1}\left(E\left[Y_1\left(0\right)|D=0\right]\right)$. The key is that the same function $G\left(\cdot\right)$ appears in (2.7). Combining (2.6) and (2.7) gives

$$G^{-1}\left(E\left[Y_2\left(0\right)|D\right]\right)-G^{-1}\left(E\left[Y_1\left(0\right)|D\right]\right)=\gamma_2, \tag{2.8}$$

which shows that the PT assumption applies to a nonlinear transformation of the means $E\left[Y_t\left(0\right)|D\right]$. Equivalently, the linear PT assumption holds for the indices inside the function $G\left(\cdot\right)$.

As an example of where we can derive (2.6) and (2.7) from a more primitive model, suppose $Y_t\left(0\right)$ is binary and generated by a latent variable, $Y_t^*\left(0\right)$:

$$Y_t\left(0\right)=1\left[Y_t^*\left(0\right)>0\right],t=1,2,$$

where $1\left[\cdot\right]$ is the indicator function and

$$Y_1^*\left(0\right)=\alpha+\beta D+U_1,$$
$$Y_2^*\left(0\right)=\alpha+\beta D+\gamma_2+U_2.$$

Assume further that

U_1,U_2 are continuous and independent of D

U_1,U_2 are identically distributed with CDF $F\left(\cdot\right)$.

The dependence between U_1 and U_2 is unrestricted so that general serial correlation is allowed. Then, for $t = 1, 2$ with $\gamma_1 \equiv 0$, we have

$$E[Y_t(0)|D] = P[Y_t(0) = 1|D] = P[\alpha + \beta D + \gamma_t + U_t > 0|D]$$
$$= 1 - F[-(\alpha + \beta D + \gamma_t)] \equiv G(\alpha + \beta D + \gamma_t),$$

which is exactly as in (2.6) and (2.7). The usual linear PT assumption holds for the underlying latent variable,

$$E\left[Y_1^*(0)|D\right] = \alpha + \beta D, \tag{2.9}$$

$$E\left[Y_2^*(0)|D\right] = \alpha + \beta D + \gamma_2, \tag{2.10}$$

but LPT generally fails for $E[Y_t(0)|D]$.

The same conclusion follows if we start with a standard unobserved effects model, $Y_t^*(0) = C + \gamma_2 f2_t + V_t$, where C is unobserved heterogeneity, $f2_t$ is a second period time dummy, and V_t is independent of C. We can always write $C = \alpha + \beta D + A$ where $E(A|D) = 0$. If we assume A is actually independent of D—common in correlated random effects settings—then the previous example holds by defining $U_t = A + V_t$, $t = 1, 2$.

As a second leading example, suppose that $Y_t(0) \geq 0$ (without a natural upper bound). $Y_t(0)$ could be a continuous variable, a count variable, or even a mixed variable with a corner at zero. Assume that (2.6) and (2.7) hold with $G(\cdot) = \exp(\cdot)$:

$$E[Y_1(0)|D] = \exp(\alpha + \beta D), \tag{2.11}$$

$$E[Y_2(0)|D] = \exp(\alpha + \beta D + \gamma_2). \tag{2.12}$$

These equations imply that the PT assumption is in terms of the growth in the mean in the untreated state:

$$\frac{E[Y_2(0)|D]}{E[Y_1(0)|D]} = \exp(\gamma_2), \tag{2.13}$$

does not depend on D. Equivalently, $\log\{E[Y_2(0)|D]\} - \log\{E[Y_1(0)|D]\} = \gamma_2$ does not depend on D.

In this $T = 2$ setting, we cannot determine the function $G(\cdot)$—if such a function exists—such that assumptions (2.6) and (2.7) hold. We have to take a stand on which function $G(\cdot)$ is most realistic, and our choice of $G(\cdot)$ affects how we estimate the ATT, τ_2. How sensitive the estimates of τ_2 are to the choice of $G(\cdot)$ is something one should explore in an application.

2.1. *Identification and estimation*

Under the NA assumption (2.2) and the PT assumption in (2.6) and (2.7), the ATT τ_2 is identified. To show this, write

$$\tau_2 = E[Y_2(1)|D = 1] - E[Y_2(0)|D = 1]. \tag{2.14}$$

Now, because $Y_2 = Y_2(1)$ when $D = 1$, we can always estimate $E[Y_2(1)|D = 1] = E(Y_2|D = 1)$ using the sample average of the treated units in $t = 2$. Given a random sample of

size N,

$$\bar{Y}_{12} \equiv N_1^{-1} \sum_{i=1}^{N} D_i Y_{i2} = \left(\frac{N_1}{N}\right)^{-1} \left(N^{-1} \sum_{i=1}^{N} D_i Y_{i2}\right) \xrightarrow{p} E\left(Y_2 | D = 1\right),$$

where $N_1 = \sum_{i=1}^{N} D_i$ is the number of treated units. (Naturally, we assume $0 < P\left(D = 1\right) < 1$ so that there are some treated and some control units.)

The second part of (2.14), $E\left[Y_2\left(0\right) | D = 1\right]$, is the one that requires using (2.2), (2.6), and (2.7). By (2.7),

$$E\left[Y_2\left(0\right) | D = 1\right] = G\left(\alpha + \beta + \gamma_2\right),$$

and, given that $G\left(\cdot\right)$ is assumed known, identification of $E\left[Y_2\left(0\right) | D = 1\right]$ rests on identifying α, β, and γ_2. From (2.6), $E(Y_1 | D = 0) = E\left[Y_1\left(0\right) | D = 0\right] = G\left(\alpha\right)$, and so $\alpha = G^{-1}\left[E(Y_1 | D = 0)\right]$. Because $E(Y_1 | D = 0)$ is the mean of the observed outcomes for the control units in the first period, α is identified. For β, use (2.2) along with (2.6):

$$E(Y_1 | D = 1) = E\left[Y_1\left(1\right) | D = 1\right] = E\left[Y_1\left(0\right) | D = 1\right] = G\left(\alpha + \beta\right),$$

where the first equality holds because $Y_1 = Y_1\left(1\right)$ when $D = 1$ and the second equality is the NA assumption. Therefore, we can write

$$\alpha + \beta = G^{-1}\left[E(Y_1 | D = 1)\right] \text{ or } \beta = G^{-1}\left[E(Y_1 | D = 1)\right] - \alpha.$$

Because α is identified and $E(Y_1 | D = 1)$ is the mean of the observed outcomes for the treated units in the first period, β is also identified.

Given a random sample at time $t = 1$, $E(Y_1 | D = 0)$ and $E(Y_1 | D = 1)$ are consistently estimable using the sample averages over the control and treated units, respectively:

$$\bar{Y}_{01} \equiv N_0^{-1} \sum_{i=1}^{N} \left(1 - D_i\right) \cdot Y_{i1} \xrightarrow{p} E(Y_1 | D = 0)$$

$$\bar{Y}_{11} \equiv N_1^{-1} \sum_{i=1}^{N} D_i \cdot Y_{i1} \xrightarrow{p} E(Y_1 | D = 1),$$

where $\bar{Y}_{01}$ is the average for the control group in $t = 1$ and $\bar{Y}_{11}$ is the average for the (eventually) treated group in $t = 1$. Therefore, by Slutsky's theorem,

$$\hat{\alpha} = G^{-1}\left(\bar{Y}_{01}\right) \xrightarrow{p} \alpha \text{ and } \hat{\beta} = G^{-1}\left(\bar{Y}_{11}\right) - \hat{\alpha} \xrightarrow{p} \beta.$$

Because we can estimate population means under different sampling schemes, α and β are identified much more generally.

All that is left is to identify and estimate is γ_2. From (2.7),

$$E\left(Y_2 | D = 0\right) = E\left[Y_2\left(0\right) | D = 0\right] = G\left(\alpha + \gamma_2\right),$$

and so

$$\gamma_2 = G^{-1}\left[E\left(Y_2 | D = 0\right)\right] - \alpha. \tag{2.15}$$

Because α is identified from $t = 1$ and $E(Y_2|D = 0)$ is identified using the control units in $t = 2$, γ_2 is identified. The natural estimator is

$$\hat{\gamma}_2 = G^{-1}\left(\bar{Y}_{02}\right) - \hat{\alpha}, \tag{2.16}$$

and, again by the law of large numbers and Slutsky's theorem, $\hat{\gamma}_2 \xrightarrow{p} \gamma_2$.

Putting together all of the estimators, a consistent estimator of τ_2 is

$$\hat{\tau}_2 = \bar{Y}_{12} - G\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_2\right) = \bar{Y}_{12} - G\left(G^{-1}\left(\bar{Y}_{11}\right) + \left(G^{-1}\left(\bar{Y}_{02}\right) - G^{-1}\left(\bar{Y}_{01}\right)\right)\right), \tag{2.17}$$

where the last expression shows that $\hat{\tau}_2$ is a particular nonlinear transformation of the sample averages of the four different groups.

When $G(\cdot)$ is the identify function—the linear PT case—(2.17) becomes

$$\hat{\tau}_2 = \left(\bar{Y}_{12} - \bar{Y}_{11}\right) - \left(\bar{Y}_{02} - \bar{Y}_{01}\right) = \left(\bar{Y}_{12} - \bar{Y}_{02}\right) - \left(\bar{Y}_{11} - \bar{Y}_{01}\right), \tag{2.18}$$

which is the basic DiD estimator. Clearly, (2.17) will generally change as the choice of $G(\cdot)$ changes. For example, when $G(\cdot) = \exp(\cdot)$,

$$\begin{aligned}\hat{\tau}_2 &= \bar{Y}_{12} - \exp\left[\log\left(\bar{Y}_{11}\right) + \left(\log\left(\bar{Y}_{02}\right) - \log\left(\bar{Y}_{01}\right)\right)\right] \\ &= \bar{Y}_{12} - \bar{Y}_{11} \cdot \left(\frac{\bar{Y}_{02}}{\bar{Y}_{01}}\right).\end{aligned} \tag{2.19}$$

The term $\bar{Y}_{02}/\bar{Y}_{01}$ measures the growth from $t = 1$ to $t = 2$ in the average of the control group. Therefore, the second term in (2.19) can be viewed as starting with the average of the treated group in the first period, $\bar{Y}_{11}$, and adjusting it using the growth in the control unit average from $t = 1$ to $t = 2$. This imputed value is used as the comparison for the average outcome for the treated units in the second period, $\bar{Y}_{12}$.

We can define another parameter, hidden in the analysis so far, that may be of interest. Using the assumption that $G(\cdot)$ is strictly increasing, define

$$\begin{aligned}\delta_2 &\equiv G^{-1}\left(E\left[Y_2(1)|D = 1\right]\right) - G^{-1}\left(E\left[Y_2(0)|D = 1\right]\right) \\ &= G^{-1}\left(E\left[Y_2(1)|D = 1\right]\right) - (\alpha + \beta + \gamma_2),\end{aligned} \tag{2.20}$$

which is equivalent to defining δ_2 such that

$$\tau_2 = G(\alpha + \beta + \gamma_2 + \delta_2) - G(\alpha + \beta + \gamma_2).$$

Whereas τ_2 is the ATT in the second time period defined in terms of the means, δ_2 is a treatment effect obtained by applying $G^{-1}(\cdot)$ to the potential outcome means. In the case of the exponential model,

$$\delta_2 = \log\left(E\left[Y_2(1)|D = 1\right]\right) - \log\left(E\left[Y_2(0)|D = 1\right]\right),$$

thereby providing an (approximate) proportional effect. When $G(z) = \exp(z)/\left[1 + \exp(z)\right]$, δ_2 is the change in the log-odds of the expected values for the treated subpopulation. Generally, a consistent estimator of δ_2 is

$$\hat{\delta}_2 = G^{-1}\left(\bar{Y}_{12}\right) - \left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_2\right) = \left[G^{-1}\left(\bar{Y}_{12}\right) - G^{-1}\left(\bar{Y}_{02}\right)\right] - \left[G^{-1}\left(\bar{Y}_{11}\right) - G^{-1}\left(\bar{Y}_{01}\right)\right],$$

assuming the inverse function $G^{-1}(\cdot)$ is well defined at all sample averages. This is a kind of DiD estimator, but where the means are transformed. For example, when $G(\cdot) = \exp(\cdot)$,

$$\hat{\delta}_2 = \left[\log\left(\bar{Y}_{12}\right) - \log\left(\bar{Y}_{02}\right)\right] - \left[\log\left(\bar{Y}_{11}\right) - \log\left(\bar{Y}_{01}\right)\right],$$

provided each average is positive. As we see more generally below, $\hat{\delta}_2$ can be obtained by a pooled estimation method for certain choices of $G(\cdot)$ and estimation methods.

In the binary case with repeated cross sections, the difference between (2.17) and (2.18) is related to Puhani's (2012) investigation of the proper way to define average treatment effects in binary response models. In the $T = 2$ DiD setting, for any strictly increasing function $G(\cdot)$, the Ai and Norton (2003) effect of the intervention is always (2.18), whereas Puhani effectively argued for (2.17). Puhani's conclusion is not based on estimation of the ATT, but we now see that Puhani's definition is the correct one for identifying τ_2 when the PT assumption is stated in terms of the linear index.

For later reference, it is useful to return to the linear case and show how (2.18) can be obtained from an imputation approach. Define a second period dummy variable $f2_t = 1$ if $t = 2$, zero otherwise. Also, for any unit i, define the time-varying treatment indicator

$$W_{it} \equiv D_i \cdot f2_t, t = 1, 2, \tag{2.21}$$

so that $W_{i2} = 1$ means unit i is treated in period two (with $W_{i1} \equiv 0$ for all i and $W_{i2} = D_i$). In the imputation step, obtain $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}_2$ from the ordinary least squares (OLS) regression using only untreated observations:

$$Y_{it} \text{ on } 1, D_i, f2_t, \ t = 1, 2; \ i = 1, ..., N \text{ if } W_{it} = 0. \tag{2.22}$$

Then impute a treatment effect for the treated units in $t = 2$: $\widehat{TE}_{i2} \equiv Y_{i2} - \left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_2\right)$ if $W_{i2} = 1$. Next, average across the treated units in the second time period to get

$$\bar{Y}_{12} - \left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_2\right) = \left(\bar{Y}_{12} - \bar{Y}_{11}\right) - \left(\bar{Y}_{02} - \bar{Y}_{01}\right),$$

the estimator in (2.18).

Alternatively, it follows from Wooldridge (2021) that $\hat{\tau}_2$ is obtained as the coefficient on W_{it} in the POLS regression

$$Y_{it} \text{ on } 1, D_i, f2_t, W_{it}, t = 1, 2; \ i = 1, ..., N. \tag{2.23}$$

Often, the regression is written with $D_i \cdot f2_t$ in place of W_{it}. Even in simple settings, and especially in the general staggered setting, explicitly using W_{it} pays dividends for simplifying the calculation of estimates and standard errors.

2.2. *Adding covariates*

Often a researcher has access to (pre-intervention) control variables, $\mathbf{X}$. In this subsection, I formally discuss how to incorporate such variables into the analysis.

ASSUMPTION CNA (CONDITIONAL NA): *For a* $1 \times K$ *vector of time-constant covariates* $\mathbf{X}$,

$$E\left[Y_1(1) \mid D = 1, \mathbf{X}\right] = E\left[Y_1(0) \mid D = 1, \mathbf{X}\right]. \tag{2.24}$$

Assumption CNA says that the pre-intervention ATTs are zero for each subpopulation indexed by $\mathbf{X}$.

We also modify the index PT assumption.

ASSUMPTION CIPT (CONDITIONAL IPT): *For a* $1 \times K$ *vector of time-constant covariates* $\mathbf{X}$ *and a function* $G(\cdot)$ *satisfying the requirements in Assumption IPTC,*

$$E\left[Y_t(0) \mid D, \mathbf{X}\right] = G\left(\alpha + \beta D + \mathbf{X}\kappa + (D \cdot \mathbf{X})\eta + \gamma_t + \mathbf{X}\pi_t\right), t = 1, 2, \tag{2.25}$$

where D is the binary treatment indicator and where we take $\gamma_1 \equiv 0$, $\pi_1 \equiv \mathbf{0}$ as normalizations to define α and κ.

Unlike in the case without covariates, when $t = 1$ the conditional mean in (2.25) now imposes functional form restrictions (unless $\mathbf{X}$ consists only of mutually exclusive and exhaustive binary indicators that partition the population). With disaggregated data—say, for individuals, firms, or schools—we can include geographic indicators — say, for counties, states or regions — among the covariates if we think underlying trends vary by geographic location. For flexibility, we can choose $\mathbf{X}$ to be functions of underlying control variables, such as the common practice of including squares and interactions.

Once we settle on the functional form in $t = 1$,

$$E[Y_1(0)|D, \mathbf{X}] = G(\alpha + \beta D + \mathbf{X}\kappa + (D \cdot \mathbf{X})\eta),$$

the substantive restriction on the linear index in (2.25) is that the coefficients on terms involving D are time invariant. In the linear case, the restriction becomes

$$E[Y_2(0)|D, \mathbf{X}] - E[Y_1(0)|D, \mathbf{X}] = \gamma_2 + \mathbf{X}\pi_2, \tag{2.26}$$

which does not depend on D. Wooldridge (2021) used an extension of this assumption in the general common timing case and derived a POLS estimators of the ATTs. Generally, by conditioning on $\mathbf{X}$ we are effectively partitioning the population by different outcomes of $\mathbf{X}$ and then assuming that the index PT assumption holds within each partition separately (rather than across the entire population).

Assumption CIPT allows PT to be violated for $E[Y_t(0)|D, \mathbf{X}]$ provided it holds for the index. With $G(\cdot) = \exp(\cdot)$,

$$\frac{E[Y_2(0)|D, \mathbf{X}]}{E[Y_1(0)|D, \mathbf{X}]} = \exp(\gamma_2 + \mathbf{X}\pi_2),$$

which implies that the growth in $E[Y_t(0)|D, \mathbf{X}]$ can depend on an unrestricted constant and on $\mathbf{X}$ in a fairly flexible way, but not on D.

We can exploit the conditional index PT assumption to identify τ_2, still given by (2.14). Nothing changes for estimating $E[Y_2(1)|D = 1]$ because $Y_2 = Y_2(1)$ for $D = 1$. For $E[Y_2(0)|D = 1]$, first apply iterated expectations:

$$E[Y_2(0)|D = 1] = E\{E[Y_2(0)|D = 1, \mathbf{X}]|D = 1\}.$$

Next, by Assumption CIPT,

$$E[Y_2(0)|D = 1, \mathbf{X}] = G(\alpha + \beta + \gamma_2 + \mathbf{X}(\kappa + \eta + \pi_2)), \tag{2.27}$$

and so

$$E[Y_2(0)|D = 1] = E[G(\alpha + \beta + \gamma_2 + \mathbf{X}(\kappa + \eta + \pi_2))|D = 1]. \tag{2.28}$$

Given that $G(\cdot)$ is assumed known and we observe $\mathbf{X}$, (2.28) shows that $E[Y_2(0)|D = 1]$ is identified if the parameters $(\alpha, \beta, \gamma_2, \kappa, \eta, \pi_2)$ are identified. Provided $\mathbf{X}$ does not include variables that are perfectly collinear, α, β, κ, and η are identified using $t = 1$. The argument is a simple extension from the case without covariates. First,

$$E(Y_1|D = 0, \mathbf{X}) = E[Y_1(0)|D = 0, \mathbf{X}] = G(\alpha + \mathbf{X}\kappa),$$

and, by CNA,

$$E\left(Y_1|D=1,\mathbf{X}\right)=E\left[Y_1\left(1\right)|D=1,\mathbf{X}\right]=E\left[Y_1\left(0\right)|D=1,\mathbf{X}\right]=G\left(\alpha+\beta+\mathbf{X}\left(\kappa+\eta\right)\right).$$

Combined, we can write, for the control and (eventually) treated units,

$$E\left(Y_1|D,\mathbf{X}\right)=G\left(\alpha+\beta D+\mathbf{X}\kappa+D\cdot\mathbf{X}\eta\right), \tag{2.29}$$

where all variables are observed in the first time period and $G\left(\cdot\right)$ is a known, strictly increasing smooth function. With a random sample for $t=1$ (or some other sampling scheme that permits consistent estimation), we can consistently estimate all parameters in (2.29) under weak regularity conditions.

With $(\alpha,\beta,\kappa,\eta)$ identified, all that remains for τ_2 is identification of γ_2 and π_2. Now, under CIPT,

$$E\left(Y_2|D=0,\mathbf{X}\right)=E\left[Y_2\left(0\right)|D=0,\mathbf{X}\right]=G\left(\alpha+\gamma_2+\mathbf{X}\left(\kappa+\pi_2\right)\right), \tag{2.30}$$

which shows that $\alpha+\gamma_2$ and $\kappa+\pi_2$ are identified using the $t=2$ observed outcome for the control group $(D=0)$. Given identification of α and κ from $t=1$, γ_2 and π_2 are identified.

With the parameters in (2.27) identified, (2.28) is obtained by averaging over the distribution of $\mathbf{X}$ given $D=1$. Technically, because we have a parametric function $G\left(\cdot\right)$, we do not need an overlap assumption. Nevertheless, problems with extrapolation in treatment effect estimation with parametric models are well known. In practice, we should verify the support condition

$$\text{Supp}\left(\mathbf{X}|D=1\right)\subset\text{Supp}\left(\mathbf{X}|D=0\right),$$

or, equivalently, $P\left(D=1|\mathbf{X}=\mathbf{x}\right)<1$ for all $\mathbf{x}\in\text{Supp}\left(\mathbf{X}\right)$. The overlap assumption ensures that when we average over the treated units we are averaging over covariate values that are used in obtaining estimators of the parameters, particularly γ_2 and π_2.

It is convenient to implement the previous procedure by estimating $\alpha,\beta,\kappa,\eta,\gamma_2$, and π_2 at once using all of the control observations. To see how, for a random draw i from the population we can write

$$\begin{aligned}E\left(Y_{it}|D_i,\mathbf{X}_i,W_{it}=0\right)=G\left(\alpha+\beta D_i+\mathbf{X}_i\kappa+\left(D_i\cdot\mathbf{X}_i\right)\eta\right.\\ \left.+\ \gamma_2 f2_t+\left(f2_t\cdot\mathbf{X}_i\right)\pi_2\right),\ t=1,2,\end{aligned} \tag{2.31}$$

where $W_{it}=D_i\cdot f2_t$ is the time-varying treatment indicator. Equation (2.31) follows because $W_{it}=0$ if and only if $t=1$ or $t=2$ and $D_i=0$. Given a random sample of size N, the parameters $(\alpha,\beta,\kappa,\eta,\gamma_2,\pi_2)$ can be jointly estimated using the control observations by any estimation method that identifies parameters in conditional mean functions.

Given consistent parameter estimators, the ATT τ_2 is consistently estimated by

$$\hat{\tau}_2=\bar{Y}_{12}-N_1^{-1}\sum_{i=1}^{N}D_i\cdot G\left(\hat{\alpha}+\hat{\beta}+\hat{\gamma}_2+\mathbf{X}_i\left(\hat{\kappa}+\hat{\eta}+\hat{\pi}_2\right)\right). \tag{2.32}$$

By applying the delta method with averaging—see Wooldridge (2010, problem 12.17)—a standard error for $\hat{\tau}_2$ can be obtained. Alternatively, the panel bootstrap can be used. Even in large cross-sectional sample sizes the bootstrap is computionally feasible because pooled estimation of nonlinear models is not particularly time consuming.

In the linear case, (2.32) becomes

$$\hat{\tau}_2=\bar{Y}_{12}-\left[\hat{\alpha}+\hat{\beta}+\hat{\gamma}_2+\bar{\mathbf{X}}_1\left(\hat{\kappa}+\hat{\eta}+\hat{\pi}_2\right)\right],$$

where $\bar{\mathbf{X}}_1 = N_1^{-1} \sum_i D_i \cdot \mathbf{X}_i$ is the average over the treated subsample. Equation (2.32) extends the observation in Wooldridge (2021) for the linear case that $\hat{\tau}_2$ can be viewed as an imputation estimator, where, for each i, $G\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_2 + \mathbf{X}_i\left(\hat{\kappa} + \hat{\eta} + \hat{\pi}_2\right)\right)$ is the imputed value of $Y_{i2}(0)$ for the treated subsample ($D_i = 1$).

Concerning specific estimation methods, a quasi-MLE approach using a log likelihood in the LEF is appealing; see Gourieroux et al. (1984) for the cross-sectional case and Wooldridge (2010, Section 13.11.4) for a discussion of pooled estimation with panel data. When $G(z) = z$, a natural choice is the normal log likelihood, which then leads to the pooled OLS estimators in Wooldridge (2021) (without needing normality or any other restrictions on the distribution). For common nonlinear choices of $G(\cdot)$, the Gaussian (normal) QMLE is not very attractive because it has essentially no chance of being efficient due to likely heteroskedasticity—even if we make the assumption of no serial correlation across t. If Y_{it} is a binary or fractional response, the logit and probit functions are leading candidates for $G(\cdot)$, and then it is natural to use the Bernoulli quasi-log likelihood function (QLLF). See Papke and Wooldridge (1996, 2008) for the fractional case. Because the Bernoulli distribution is in the LEF, the QMLE pooled across the $W_{it} = 0$ observations is fully robust for estimating $(\alpha, \beta, \kappa, \eta, \gamma_2, \pi_2)$ provided the mean is correctly specified, which holds under Assumptions CNA and CIPT. Details on estimation and inference are covered in Wooldridge (2010, Section 13.11.4). For nonnegative responses, including count variables and corner solutions, an exponential mean estimated using the Poisson QMLE is very natural.

An alternative to the imputation method is pooled QMLE using all of the data. Specifically, act as if the following conditional mean function, for randomly drawn unit i, holds for treated as well as untreated observations:

$$E\left(Y_{it}|D_i, \mathbf{X}_i, W_{it}\right) = G\big[\alpha + \beta D_i + \mathbf{X}_i\kappa + \left(D_i \cdot \mathbf{X}_i\right)\eta + \gamma_2 f2_t + \left(f2_t \cdot \mathbf{X}_i\right)\pi_2 + \delta_2\left(W_{it} \cdot f2_t\right) + \left(W_{it} \cdot f2_t \cdot \dot{\mathbf{X}}_i\right)\xi_2\big], \; t = 1, 2, \tag{2.33}$$

where the conditioning on W_{it} is redundant, but useful for understanding the flexibility of the approach (and also for obtaining the ATTs and standard errors). In forming the interactions $W_{it} \cdot f2_t \cdot \dot{\mathbf{X}}_i$, the controls $\dot{\mathbf{X}}_i \equiv \mathbf{X}_i - E\left(\mathbf{X}_i|D_i = 1\right)$ have been centred. The reason for demeaning the covariates is so that δ_2 has a useful meaning, as shown in equation (2.20). It may be as interesting to have an ATT stated, say, in terms of a percentage difference in the mean as on the mean itself when $Y_{it} \geq 0$.

To obtain all parameter estimates, decorated with '$\sim$' (and including $\tilde{\delta}_2$ and $\tilde{\xi}_2$), use pooled QMLE with covariates $\dot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}_1$. Extending regression terminology, across all i and $t = 1, 2$, use pooled QMLE of

$$Y_{it} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot \mathbf{X}_i, f2_t, f2_t \cdot \mathbf{X}_i, W_{it} \cdot f2_t, W_{it} \cdot f2_t \cdot \dot{\mathbf{X}}_i.$$

Then, $\tilde{\tau}_2$ is the average partial effect of W_t at $f2_t = 1$, averaged across the subsample with $D = 1$:

$$\tilde{\tau}_2 = N_1^{-1} \sum_{i=1}^{N} \big[D_i G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_2 + \mathbf{X}_i\left(\tilde{\kappa} + \tilde{\eta} + \tilde{\pi}_2\right) + \tilde{\delta}_2 + \dot{\mathbf{X}}_i\tilde{\xi}_2\right) - G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_2 + \mathbf{X}_i\left(\tilde{\kappa} + \tilde{\eta} + \tilde{\pi}_2\right)\right)\big]. \tag{2.34}$$

The expression in (2.34) is computed routinely by econometrics packages that support standard models of binary, fractional, and nonnegative responses. Moreover, formulas for standard errors

that account for sampling error in the parameter estimates and in $\{\mathbf{X}_i : i = 1, ..., N\}$ are readily available in Wooldridge (2010, problem 12.17) and coded in popular software packages.

Another benefit of using QMLE pooled across all observations is that it permits estimation of the average treatment effect (ATE) in the second period. Specifically,

$$\tilde{\tau}_{2,ate} = N^{-1}\sum_{i=1}^{N}\left[G\left(\tilde{\alpha} + \tilde{\beta}D_i + \mathbf{X}_i\tilde{\kappa} + (D_i \cdot \mathbf{X}_i)\,\tilde{\eta} + \tilde{\gamma}_2 + \mathbf{X}_i\tilde{\pi}_2 + \hat{\delta}_2 + \dot{\mathbf{X}}_i\tilde{\xi}_2\right)\right.$$
$$\left. - \; G\left(\tilde{\alpha} + \tilde{\beta}D_i + \mathbf{X}_i\tilde{\kappa} + (D_i \cdot \mathbf{X}_i)\,\tilde{\eta} + \tilde{\gamma}_2 + \mathbf{X}_i\tilde{\pi}_2\right)\right], \tag{2.35}$$

which is obtained by computing the average partial effect (APE) with respect to W_t evaluated at $f2_t = 1$. The difference between $\tilde{\tau}_{2,ate}$ and $\tilde{\tau}_2$ is that (2.35) is averaged across the both the control and treated units. As with $\tilde{\tau}_2$, $\tilde{\tau}_{2,ate}$ is easily estimated using standard software packages, and standard errors are obtained using the delta method.

To ensure (2.35) consistently estimates $\tau_{2,ate} = E\left[Y_2\left(1\right) - Y_2\left(0\right)\right]$, we must strengthen the NA and PT assumption. In particular, the CNA assumption would have to also apply to the $D = 0$ subpopulation and the CIPT assumption would also have to hold in the treated state, $Y_t\left(1\right)$. For the most part, I focus on the ATT because it is the most common treatment effect parameter in DiD settings.

Generally speaking, even within the class of LEF distributions, the imputation and pooled approaches produce different estimates of the ATT, τ_2. Technically, the pooled method adds a functional form assumption not required by imputation: (2.33) must hold across the treated as well as the control observations. Practically, imposing (2.33) rather than the weaker assumption (2.31) seems relatively harmless, as we are already making a functional form assumption for $t = 1$ and combining it with a conditional PT assumption.

In some leading cases the imputation and pooled QMLE methods produce identical estimates of τ_2, that is, $\tilde{\tau}_2 = \hat{\tau}_2$. This is the case whenever $G^{-1}\left(\cdot\right)$ is the *canonical link* function associated with the chosen density in the LEF. With such a choice, it turns out that the parameter estimates of α, β, κ, η, γ_2, and π_2 are identical. Moreover,

$$\bar{Y}_{12} = N_1^{-1}\sum_{i=1}^{N} D_i G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_2 + \mathbf{X}_i\left(\tilde{\kappa} + \tilde{\eta} + \tilde{\pi}_2\right) + \tilde{\delta}_2 + \dot{\mathbf{X}}_i\tilde{\xi}_2\right),$$

which implies (2.32) and (2.34) are identical. Wooldridge (2021) showed this for the special case of a linear conditional mean estimated by POLS.

The benefits of the canonical link function have appeared previously in the causal effects literature. Wooldridge (2007) emphasizes its role in obtaining doubly robust estimators of average treatment effects that combine regression adjustment—generally interpreted as quasi-MLE in the LEF—and inverse propensity score weighting. Extending the case of linear regression, Negi and Wooldridge (2021) show that using the canonical link in the LEF preserves consistency of ATE estimators under general conditional mean misspecification with randomized controlled trials. The most relevant combinations of canonical link mean functions and LEF densities are given in Table 1.

Anticipating the general recommendations below, in applying DiD to limited dependent variables it will make sense to compare DiD estimates from linear regression to the ATT estimates from a nonlinear method that exploits the special nature of Y_{it}.

Table 1. Canonical link and log likelihood pairings.

Conditional Mean	LEF Density	Comments
Linear	Normal	Any response; leads to OLS.
Logistic	Bernoulli	Binary or fractional response.
Logistic	Binomial	Nonnegative response with known upper bound.
Logistic	Multinomial	Multinomial or multiple fractional response.
Exponential	Poisson	Nonnegative response (count, corner), no natural upper bound.

3. STAGGERED INTERVENTIONS

I now consider the case of staggered interventions, where different units are subjected to a treatment or intervention in different periods. Naturally, what follows includes interventions with common timing (including $T = 2$) as a special case.

3.1. Potential outcomes and parameters of interest

The first intervention occurs at time $q \in \{2, \ldots, T\}$ and then some additional units are treated for the first time in subsequent periods. As in Wooldridge (2021), I treat the staggered intervention as defining different treatment cohorts that are then subjected to different treatment intensities in time $t \geq q$. In any post-intervention period, units treated initially at time q will have been exposed to the intervention longer than units first treated in later periods. Initially, I focus on the case where there is a never treated group, so that at $t = T$ there are still units not subjected to the intervention. In Section 7 I discuss relaxing this restriction.

As in Wooldridge (2021) and Athey and Imbens (2022), the potential outcomes are denoted

$$Y_t(g), g \in \{q, \ldots, T, \infty\}, t \in \{1, 2, \ldots, T\}, \tag{3.1}$$

where g indicates the first time subjected to the intervention—it can be thought of as a 'group' or 'cohort'—and t is calendar time. The case $g = \infty$ indicates the potential outcome in the never treated state. In other words, $Y_t(\infty)$ is the potential outcome at time t when a unit is not subjected to the intervention over the observed stretch of time. Listing potential outcomes that vary only by cohort and calendar time reflects the assumption of no reversibility with staggered entry.

The ATTs of interest are

$$\tau_{gr} = E\left[Y_r(g) - Y_r(\infty) \,|\, D_g = 1\right], r = g, \ldots, T;\ g = q, \ldots, T, \tag{3.2}$$

where the $\left\{D_g : g = q, \ldots, T\right\}$ are cohort indicators defining the first period a unit is subjected to the intervention. For now, we assume there is a group with $D_\infty = 1$, which indicates never being treated over the observed time period. For each (eventually) treated cohort g, $\tau_{gr}, r = g, \ldots, T$ are the ATTs in all subsequent time periods.

The key assumptions are stated conditional on covariates, with a special case being when $\mathbf{X}$ is null.

ASSUMPTION CNAS (CONDITIONAL NO ANTICIPATION, STAGGERED): *For* $g \in \{q, \ldots, T\}$, $t \in \{1, \ldots, g-1\}$, *and covariates* $\mathbf{X}$,

$$E\left[Y_t(g) \,|\, D_g = 1, \mathbf{X}\right] = E\left[Y_t(\infty) \,|\, D_g = 1, \mathbf{X}\right]. \tag{3.3}$$

In equation (3.3), $t < g$, and so Assumption CNAS implies zero treatment effects, on average, prior to the intervention for cohort g.

ASSUMPTION CIPTS (CONDITIONAL INDEX PT, STAGGERED): *For* $t = 1, 2, \ldots, T$,

$$E\left[Y_t(\infty) \mid D_q, \ldots, D_T, \mathbf{X}\right] = G\left(\alpha + \sum_{g=q}^{T} \beta_g D_g + \mathbf{X}\kappa + \sum_{g=q}^{T} \left(D_g \cdot \mathbf{X}\right)\eta_g + \gamma_t + \mathbf{X}\pi_t\right), \tag{3.4}$$

with the normalizations $\gamma_1 \equiv 0$, $\pi_1 \equiv \mathbf{0}$.

Equation (3.4) makes it clear that treatment cohort assignment is allowed to be confounded with respect to $Y_t(\infty)$, even after conditioning on $\mathbf{X}$. When $\beta_g \neq 0$, the mean response in the never treated (NT) state can depend on the cohort assignment, D_g. When $\eta_g \neq \mathbf{0}$ this dependence can be heterogeneous in $\mathbf{X}$ in the linear index function. When $G(z) = z$, CIPTS implies, with $\mathbf{D} = \left(D_q, \ldots, D_T\right)$,

$$E\left[Y_t(\infty) \mid \mathbf{D}, \mathbf{X}\right] - E\left[Y_1(\infty) \mid \mathbf{D}, \mathbf{X}\right] = \gamma_t + \mathbf{X}\pi_t,$$

which allows unrestricted aggregate trends and allows those trends to vary with the covariates. The key restriction is that the trend in the conditional mean cannot vary by treatment cohort. In other words, cohort assignment is unconfounded, conditional on $\mathbf{X}$, with respect to the trend. This is the conditional PT assumption used in Wooldridge (2021). When applied to $t = 1$ and without covariates, (3.4) imposes no restrictions, and allows assignment to treatment cohort to be systematically related to the never treated potential outcome. With covariates, a functional form assumption is maintained:

$$E\left[Y_1(\infty) \mid D_q, \ldots, D_T, \mathbf{X}\right] = G\left(\alpha + \sum_{g=q}^{T} \beta_g D_g + \mathbf{X}\kappa + \sum_{g=q}^{T} \left(D_g \cdot \mathbf{X}\right)\eta_g\right).$$

The index can depend on the cohort assignments, D_g, and $\mathbf{X}$, in a flexible way. For $t \geq 2$, (3.4) imposes that the *change* in the index does not depend on $\left(D_q, \ldots, D_T\right)$—an extension of the usual linear conditional PT assumption.

Assumption CIPTS implies that the ATTs can be written as

$$\tau_{gr} = E\left(Y_r \mid D_g = 1\right) - E\left[G\left(\alpha + \beta_g + \gamma_r + \mathbf{X}\left(\kappa + \eta_g + \pi_r\right)\right) \middle| D_g = 1\right], \tag{3.5}$$

which shows the τ_{gr} are identified if the parameters in the linear index are identified.

For a unit i, the time-varying treatment indicator, W_{it}, can be written as

$$W_{it} = D_{iq} \cdot (fq_t + \cdots + fT_t) + \cdots + D_{iT} \cdot fT_t. \tag{3.6}$$

It is easy to see that $W_{it} \cdot D_{ig} \cdot fs_t = D_{ig} \cdot fs_t$ for all $s = g, \ldots, T$ (that is, time periods where cohort g is subjected to the intervention). The condition $W_{it} = 0$ means that if unit i is in cohort g then $t < g$. For a never treated unit, $W_{it} = 0, t = 1, \ldots, T$. With no reversibility, the sequence $\{W_{it} : t = 1, \ldots, T\}$ can only have a string of zeros followed by a string of ones.

The identification argument for the parameters extends the two-period case. Assumptions CNAS and CIPTS imply

$$E\left(Y_1 \mid D_q, \ldots, D_T, \mathbf{X}\right) = G\left[\alpha + \sum_{g=q}^{T} \beta_g D_g + \mathbf{X}\kappa + \sum_{g=q}^{T} \left(D_g \cdot \mathbf{X}\right)\eta_g\right], \tag{3.7}$$

and so, with strictly increasing $G(\cdot)$, population units in each treatment cohort, and no perfect collinearity in $\mathbf{X}$, α, β_q, ..., β_T, κ, and η_q, ..., η_T are all identified using the first time period only. The CIPTS assumption implies that, for $r \geq q$,

$$E\left(Y_r | D_q = 0, \ldots, D_T = 0 | \mathbf{X}\right) = G\left[\alpha + \gamma_r + \mathbf{X}\left(\kappa + \pi_r\right)\right], \tag{3.8}$$

which verifies that one can use the never treated units starting with the first intervention period to identify γ_r and $\pi_r, r = q, ..., T$. With a never treated group, we need not use any of the eventually treated units as controls. Nevertheless, using all control combinations $W_{it} = 0$ to estimate all parameters uses all of the implications of the CNAS and CIPTS assumptions. In the population, write

$$E\left(Y_t | D_q, \ldots, D_T, \mathbf{X}, W_t = 0\right) = G\left[\alpha + \sum_{g=q}^{T} \beta_g D_g + \mathbf{X}\kappa + \sum_{g=q}^{T} \left(D_g \cdot \mathbf{X}\right)\eta_g + \sum_{s=2}^{T} \gamma_s f s_t + \sum_{s=2}^{T} \left(f s_t \cdot \mathbf{X}\right)\pi_s\right]. \tag{3.9}$$

Estimation based on (3.9) is typically more efficient than approaches that use only a subset of the valid time periods and control units. An example is CS (2021), which uses long differences relative to the time period just prior to the intervention and, sometimes, adopts only a subset of the valid controls (the most common being the never treated group). The identification discussion here is an alternative to that in CS (2021), with the current analysis applying more broadly to settings where a nonlinear function $G(\cdot)$ is needed to satisfy Assumption CIPTS.

The above discussion essentially proves the following result.

THEOREM 3.1. *Assume that Assumptions CNAS and CIPTS hold for a strictly increasing function $G(\cdot)$. Assume that $\rho_g \equiv P\left(D_g = 1\right) > 0$, $g \in \{q, q+1, \ldots, T, \infty\}$. If $\mathbf{X}$ has a non-degenerate distribution—no element is constant and there is no perfect collinearity among the elements—then $\left(\alpha, \beta_q, \ldots, \beta_T, \kappa, \eta_q, \ldots, \eta_T, \gamma_2, \ldots, \gamma_T, \pi_2, \ldots, \pi_T\right)$ are identified and so are the ATTs.*

The following procedure follows naturally from the identification result. Sufficient for consistency (with fixed T and $N \to \infty$) is random sampling from the cross section.

PROCEDURE 1 (IMPUTATION ESTIMATION):

1. For the chosen function $G(\cdot)$, use the $W_{it} = 0$ observations to estimate the parameters

$$\left(\alpha, \beta_q, \ldots, \beta_T, \kappa, \eta_q, \ldots, \eta_T, \gamma_2, \ldots, \gamma_T, \pi_2, \ldots, \pi_T\right)$$

by pooled QMLE in the LEF. The explanatory variables are:

$$1,\ D_{iq},\ ...,\ D_{iT},\ \mathbf{X}_i,\ D_{iq} \cdot \mathbf{X}_i,\ ...,\ D_{iT} \cdot \mathbf{X}_i,$$
$$f2_t,\ ...,\ fT_t,\ f2_t \cdot \mathbf{X}_i,\ ...,\ fT_t \cdot \mathbf{X}_i.$$

2. For cohort $g \in \{q, \ldots, T\}$, impute $Y_{ir}(\infty)$ for $W_{ir} = 1$:

$$\hat{Y}_{igr}(\infty) \equiv G\left(\hat{\alpha} + \hat{\beta}_g + \mathbf{X}_i\hat{\kappa} + \mathbf{X}_i\hat{\eta}_g + \hat{\gamma}_r + \mathbf{X}_i\hat{\pi}_r\right), r = g, \ldots, T. \tag{3.10}$$

3. For $r = g, \ldots, T$, obtain the imputation estimator of τ_{gr}:

$$\begin{aligned}\hat{\tau}_{gr} &= N_g^{-1}\sum_{i=1}^{N} D_{ig}\left[Y_{ir} - \hat{Y}_{igr}(\infty)\right] \\ &= \bar{Y}_{gr} - N_g^{-1}\sum_{i=1}^{N} D_{ig}G\left(\hat{\alpha} + \hat{\beta}_g + \mathbf{X}_i\hat{\kappa} + \mathbf{X}_i\hat{\eta}_g + \hat{\gamma}_r + \mathbf{X}_i\hat{\pi}_r\right).\end{aligned} \tag{3.11}$$

Procedure 1 extends the linear imputation estimators studied in Wooldridge (2021). With random sampling across i (and treating T as fixed), one can apply the delta method to obtain standard errors of the $\hat{\tau}_{gr}$, and even an estimator of the asymptotic variance of the vector of all estimators, $\hat{\tau}$ (Wooldridge, 2010, problem 12.17). The panel bootstrap—where units are resampled—is also valid and allows for any kind of serial dependence and model misspecification. In order to rely on the estimates in (3.11), we should have sufficient overlap in the covariate distributions across treatment cohorts. The precise conditions are complicated because, in effect, different control groups are used in estimating the different treatment effects. Because the never treated group acts as a control for estimating all ATTs, sufficient is $\text{Supp}\left(\mathbf{X}|D_g = 1\right) \subset \text{Supp}\left(\mathbf{X}|D_\infty = 1\right)$ for all treated cohorts g. (And this overlap condition is essentially necessary to identify ATTs in $t = T$.)

For robustness and simplicity, I recommend using pooled QMLE with a likelihood function in the LEF. By appropriately choosing the log-likelihood function (LLF), the pooled QMLE should also have acceptable efficiency properties. Nevertheless, the estimators from Procedure 1 are generally inefficient—for a couple of reasons. First, we are allowing that the variances associated with the chosen LLF are different from the actual conditional variances, $Var\left(Y_{it}|D_{iq}, \ldots, D_{iT}, \mathbf{X}_i, W_{it} = 0\right)$. Probably more importantly, the pooled method ignores serial correlation in estimation. To exploit serial correlation along with a particular conditional variance function, one could use a nonlinear generalized least squares approach of the type described in Wooldridge (2010, Section 12.9.2), although one should first establish identification of the ATTs.

Rather than using an LLF in the LEF, one could use other models and estimation methods as dictated by the particular application. For example, if Y_{it} is a corner solution, one could use a Tobit model in step (1), estimating the parameters by by pooled (quasi-) MLE after specifying that the underlying latent variable, $Y_{it}^*(\infty)$, follows the linear version of the model. In equation (3.10), the function $G(\cdot)$ would be replaced with the mean function for the Tobit model; see, for example, Wooldridge (2010, Section 17.2). A two-part model is also a possibility where the conditional PT assumption is imposed in both parts of the model. Using an exponential mean function and Poisson regression is simple and robust, but it does rely on the exponential mean function being correct.

An alternative to the imputation approach, convenient because it simplifies calculation of standard errors, is a method that pools across all observations. Pooled estimation is nominally

based on the following conditional expectation for any time period t and random draw i:

$$E\left(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i,\mathbf{W}_i\right)=G\left[\alpha+\sum_{g=q}^{T}\beta_g D_{ig}+\mathbf{X}_i\kappa+\sum_{g=q}^{T}\left(D_{ig}\cdot\mathbf{X}_i\right)\eta_g\right.$$
$$+\sum_{s=2}^{T}\gamma_s fs_t+\sum_{s=2}^{T}\left(fs_t\cdot\mathbf{X}_i\right)\pi_s$$
$$+\sum_{g=q}^{T}\sum_{s=g}^{T}\delta_{gs}\left(W_{it}\cdot D_{ig}\cdot fs_t\right)$$
$$\left.+\sum_{g=q}^{T}\sum_{s=g}^{T}\left(W_{it}\cdot D_{ig}\cdot fs_t\cdot\dot{\mathbf{X}}_{ig}\right)\xi_{gs}\right], \tag{3.12}$$

where

$$\dot{\mathbf{X}}_{ig}=\mathbf{X}_i-E\left(\mathbf{X}_i|D_{ig}=1\right),$$

are the cohort-specific means of the covariates. The reason for centring the covariates in the quadruple interaction term is that it makes the δ_{gs} easier to interpret—say, as ATTs on the log-odds or the log of the conditional mean. One might want to center $\mathbf{X}_i$ in the other interactions in order to, say, obtain more easily interpretable coefficients on the D_{ig}, but these coefficients are not measuring treatment effects. Including W_{it} in the triple and quadruple interactions does not change the estimated coefficients, but its presence is useful for emphasizing the flexibility in the pattern of treatment effects and for computing standard errors of the ATTs. It is helpful to think of the variables in the first two lines of (3.12) as consisting of controls and then the variables in the third and fourth lines, those involving W_{it}, are the treatment variables of interest.

Alternatively, for each (i,t) we can define a set of treatment indicators

$$W_{itgs}=D_{ig}\cdot fs_t,\ g=q,\ldots,T;s=g,\ldots,T,$$

so that for each unit and each time period we have a treatment indicator for each (g,s) pair with $g\in\{q,\ldots,T\}$, $s\in\{g,\ldots,G\}$. We can then write the conditional mean for a random draw as

$$E\left(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i,\mathbf{W}_i\right)=G\left[\alpha+\sum_{g=q}^{T}\beta_g D_{ig}+\mathbf{X}_i\kappa+\sum_{g=q}^{T}\left(D_{ig}\cdot\mathbf{X}_i\right)\eta_g\right.$$
$$+\sum_{s=2}^{T}\gamma_s fs_t+\sum_{s=2}^{T}\left(fs_t\cdot\mathbf{X}_i\right)\pi_s$$
$$+\sum_{g=q}^{T}\sum_{s=g}^{T}\delta_{gs}W_{itgs}$$
$$\left.+\sum_{g=q}^{T}\sum_{s=g}^{T}\left(W_{itgs}\cdot\dot{\mathbf{X}}_{ig}\right)\xi_{gs}\right]. \tag{3.13}$$

Technically, under only Assumptions CNAS and CIPTS, the mean function in (3.12) could be misspecified when $W_{it}=1$ for some time periods. If the functional form also holds for $W_{it}=1$,

then we can identify the ATTs defined in (3.5) with pooled estimation across all observations. This is tantamount to assuming that $E\left[Y_t(g)\,|D_g = 1, \mathbf{X}\right]$ is the same as the first two lines in (3.12) with the additional term $\delta_{gt} + \dot{\mathbf{X}}_{ig}\xi_{gt}$ for $t = g, \ldots, T$ (all inside the function $G(\cdot)$). Because we are already imposing a functional form on $E[Y_t(\infty)\,|\mathbf{D}, \mathbf{X}]$, it seems a minor additional assumption to impose the natural functional form on $E\left[Y_t(g)\,|D_g = 1, \mathbf{X}\right]$.

PROCEDURE 2 (POOLED ESTIMATION): 1. Using all of the data, apply pooled QMLE in the LEF to estimate

$$\begin{aligned}&\big(\alpha, \beta_q, \ldots, \beta_T, \kappa, \eta_q, \ldots, \eta_T, \gamma_2, \ldots, \gamma_T, \pi_2, \ldots, \pi_T,\\ &\delta_{qq}, \delta_{q,q+1}, \ldots, \delta_{qT}, \delta_{q+1,q+1}, \ldots, \delta_{q+1,T}, \ldots, \delta_{TT},\\ &\quad \xi_{qq}, \xi_{q,q+1}, \ldots, \xi_{qT}, \xi_{q+1,q+1}, \ldots, \xi_{q+1,T}, \ldots, \xi_{TT}\big).\end{aligned}$$

The explanatory variables are

$$\begin{aligned}&1.\ D_{iq},\ \ldots,\ D_{iT},\ \mathbf{X}_i,\ D_{iq}\cdot\mathbf{X}_i,\ \ldots,\ D_{iT}\cdot\mathbf{X}_i,\\ &fq_t,\ \ldots,\ fT_t,\ fq_t\cdot\mathbf{X}_i,\ \ldots,\ fT_t\cdot\mathbf{X}_i,\\ &W_{it}\cdot D_{iq}\cdot fq_t, W_{it}\cdot D_{iq}\cdot f(q+1)_t, \ldots, W_{it}\cdot D_{iq}\cdot fT_t\\ &W_{it}\cdot D_{i,q+1}\cdot f(q+1)_t, \ldots, W_{it}\cdot D_{i,q+1}\cdot f(q+1)_t, \ldots, W_{it}\cdot D_{iT}\cdot fT_t\\ &W_{it}\cdot D_{iq}\cdot fq_t\cdot\dot{\mathbf{X}}_{iq}, W_{it}\cdot D_{iq}\cdot f(q+1)_t\cdot\dot{\mathbf{X}}_{iq}, \ldots, W_{it}\cdot D_{iq}\cdot fT_t\cdot\dot{\mathbf{X}}_{iq}\\ &W_{it}\cdot D_{i,q+1}\cdot f(q+1)_t\cdot\dot{\mathbf{X}}_{i,q+1}, \ldots, W_{it}\cdot D_{i,q+1}\cdot f(q+1)_t\cdot\dot{\mathbf{X}}_{i,q+1}, \ldots, W_{it}\cdot D_{iT}\cdot fT_t\cdot\dot{\mathbf{X}}_{iT},\end{aligned} \tag{3.14}$$

where now $\dot{\mathbf{X}}_{ig} = \mathbf{X}_i - \bar{\mathbf{X}}_g$ are centred around cohort sample averages.

2. For $\tilde{\tau}_{gr}$, obtain the average partial effect with respect to the binary variable W_t, evaluated at $D_g = 1$, $fr_t = 1$, and all other cohort and time dummies set to zero. Average across the subsample with $D_{ig} = 1$ to get

$$\begin{aligned}\tilde{\tau}_{gr} = N_g^{-1}\sum_{i=1}^{N} D_{ig}\big[&G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{X}_i\tilde{\kappa} + \mathbf{X}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{X}_i\tilde{\pi}_r + \tilde{\delta}_{gr} + \dot{\mathbf{X}}_{ig}\xi_{gr}\right)\\ &- G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{X}_i\tilde{\kappa} + \mathbf{X}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{X}_i\tilde{\pi}_r\right)\big].\end{aligned} \tag{3.15}$$

Alternatively, with the equation written as in (3.13), obtain the APEs of each indicator W_{itgr} and average across the subsample $W_{itgr} = 1$.

This procedure has some benefits compared with Procedure 1. First, most econometric software packages support pooled quasi-MLE in the LEF—often under the label of 'generalized linear models'— along with cluster-robust standard errors for average partial effects. Moreover, often a simple option can be used to obtain standard errors that account for sampling variation in the $\bar{\mathbf{X}}_g$. Second, the estimates $\tilde{\delta}_{gr}$ are provided, and these are often of interest themselves. Generally, we can think of $\tilde{\delta}_{gr}$ as estimating

$$\delta_{gr} \equiv G^{-1}\left(E\left[Y_r(g)\,|D_g = 1\right]\right) - G^{-1}\left(E\left[Y_r(\infty)\,|D_g = 1\right]\right), \tag{3.16}$$

which is a treatment effect defined in terms of the linear index. For example, when $G(\cdot) = \exp(\cdot)$, $\tilde{\delta}_{gr}$ is an estimate proportionate effect, easily turned into a percent by multiplying by 100. Third,

we also obtain the $\tilde{\xi}_{gr}$, which allows us to study whether and how the treatment effects vary with observed covariates. When we compute the ATTs at different settings of $\mathbf{X}$ we can see which factors amplify or dampen the treatment effects. One can easily test whether control variables need to be interacted with the treatment indicators as allowed in the most general specification.

Given the estimates $\tilde{\tau}_{gr}$ (or the imputation analogs), they can be aggregated in different ways. For example, in event study settings, it is common to report estimates by treatment intensity. The so-called static effect is a weighted average of the immediate effects, $\tilde{\tau}_{gg}$, $g = q, \ldots, T$, where the weights are the proportions of the treated cohorts relative to all eventually treated cohorts. Similarly, dynamic effects are weighted averages of $\tilde{\tau}_{g,g+h}$ for $h \geq 1$ for all cohorts where these effects are identified. Obtaining standard errors for these kinds of effects can be done analytically or via the panel bootstrap.

A fourth benefit of the pooled method is that one can, if desired, obtain overall average treatment effects for each treated period. Now, instead of averaging over the subsample with $D_{ig} = 1$, one computes the average partial effect with respect to W_t across the entire sample with $fr_t = 1$ for the chosen time period, r (and $fs_t = 0$ for $s \neq r$). The approach identifies the ATEs only if we strengthen the PT assumption. In effect, we would need conditional PT to hold, with the same function $G(\cdot)$, for each potential outcome $Y_t(g)$, $g \in \{q, q+1, \ldots, T\}$ in addition to $g = \infty$. Ideally, one obtains separate estimates of τ_{gr} and then aggregates them in a desirable way, such as by treatment cohort, intensity of treatment, or even into a single, average effect.

As discussed above with the imputation approach, if a particular model for the conditional mean, or even the entire conditional distribution, is more appealing than the relatively simple means that are typically used with the LEF family of distributions, one can apply any pooled estimation method and then compute the APEs with respect to W_t at the different settings of the D_g and fr_t.

The case of common timing, allowing many pre- and post-intervention periods, is a special case, and the formulas and calculations simplify. For the pooled estimation, the conditional mean function (3.12) becomes

$$E(Y_{it}|D_i, \mathbf{X}_i, \mathbf{W}_i) = G\left[\alpha + \beta D_i + \mathbf{X}_i\kappa + (D_i \cdot \mathbf{X}_i)\eta + \sum_{s=2}^{T}\gamma_s fs_t + \sum_{s=2}^{T}(fs_t \cdot \mathbf{X}_i)\pi_s + \sum_{s=g}^{T}\delta_s(W_{it} \cdot D_i \cdot fs_t) + \sum_{s=g}^{T}\left(W_{it} \cdot D_i \cdot fs_t \cdot \dot{\mathbf{X}}_{ig}\right)\xi_{gs}\right], \tag{3.17}$$

where D_i is the 'eventually treated' indicator and, as before, W_{it} is the time-varying treatment indicator. The variables in the first line of (3.17) serve as controls and the treatment variables are in the second line.

3.2. *Equivalence between imputation and pooled estimation*

For general choices of $G(\cdot)$, the $\tilde{\tau}_{gr}$ in (3.15) differ from the $\hat{\tau}_{gr}$ in (3.11), with the latter requiring a slightly weaker functional form assumption for consistency. Nevertheless, in some popular cases the estimators are numerically the same. Wooldridge (2021) showed the estimators are the same in the linear case. It turns out that an extension of the equivalence result holds when $G^{-1}(\cdot)$ is chosen to be the canonical link function in the LEF, as discussed previously when $T = 2$. I now state a formal result, which is proven in the Appendix.

PROPOSITION 3.1 (EQUIVALENCE OF IMPUTATION AND POOLED QMLE): *In the staggered intervention setting without reversibility, suppose that* $G^{-1}(\cdot)$ *is the canonical link function associated with the chosen LEF density. If the solution to the pooled QMLE is unique, the estimates of common parameters are the same as the imputation estimates and the estimated ATTs are the same as the imputation estimates:* $\tilde{\tau}_{gr} = \hat{\tau}_{gr}$, $r = g, \ldots, T$; $g = q, \ldots, T$.

This is a useful result, as it ties together imputation estimation and pooled estimation across the entire sample.

It is worth noting at this point that including covariates $\mathbf{X}$ can help with precision of the estimates if those covariates are useful predictors of $Y_t(\infty)$. This is particularly true if the PT assumption holds without conditioning on $\mathbf{X}$, so that one could use the pooled estimator without the covariates. Including $\mathbf{X}$, by effectively reducing the noise from unobservables, can lead to more precise estimates of the τ_{gr}.

As discussed previously, there is no need to restrict oneself to a canonical link function in the LEF, but that might serve a practical purpose. Namely, it limits the kinds of nonlinear models and estimation methods attempted by an empirical researcher. If one is to go beyond a linear model to exploit the special nature of Y_{it}, it seems prudent to limit the kinds of nonlinear models that one will attempt. As a check on data mining, the model and estimation method are determined by the conditional mean/density pairs listed in Table 1.

3.3. Examples

When $Y_{it} \in [0, 1]$ (binary or fractional), a logistic function using the Bernoulli QLL is convenient because estimation and standard errors can be done using the pooled QMLE without imposing the additional functional form assumption. The mean function is specified as in (3.12) with $G(\cdot) = \Lambda(\cdot) \equiv \exp(\cdot)/[1 + \exp(\cdot)]$ as the logistic function. Centring of the covariates [replacing $E(\mathbf{X}|D_g = 1)$ with $\bar{\mathbf{X}}_g$] ensures the estimates $\hat{\delta}_{gr}$ can be interpreted as the ATT on the log-odds ratio; see (3.16). Any statistical package that does pooled logit, or fractional logit, and allows calculation of average partial effects and their standard errors can be used for proper inference. The APE is computed with respect to W_t and then one sets the appropriate cohort/year dummy combinations to unit with the others set to zero. In many applications it seems likely that a probit mean function will deliver similar estimates.

If Y_{it} has a bound that can possibly change over i and t, say $Y_{it} \in [0, B_{it}]$, then the logistic mean function simply gets multiplied by B_{it}, and so do the calculations of the APEs. This can be accomplished by choosing the binomial QLLF with a logit link function and total number of 'trials' equal to B_{it}. Note that Y_{it} need not be an integer and it can have corners at zero, B_{it}, or both.

When $Y_{it} \geq 0$ without a natural upper bound, the exponential mean makes sense as an alternative to a linear mean. In (3.12), choose $G(\cdot) = \exp(\cdot)$. This mean function should be coupled with the Poisson QLL in pooled QMLE estimation. Because $\log(\cdot)$ is the canonical link function for the Poisson log likelihood, the estimates of parameters and APEs from the estimation pooled across all of the data are identical to the imputation estimates.

Other than the linear conditional mean, where pooled OLS and TWFE are equivalent, the exponential mean is the one other case where a TWFE estimator is available that does not suffer from an incidental parameters problem. In particular, the FE Poisson estimator, with time dummies included, can be used to estimate the parameters γ_s, π_s, δ_{gs}, and ξ_{gs} without imposing any distributional assumptions or restrictions on serial dependence (Wooldridge, 1999). How does

FE Poisson compare with the pooled Poisson QMLE with the mean in (3.13) and $G(\cdot) = \exp(\cdot)$? Based on thousands of simulated data sets, it is evident that, without covariates, the estimates of γ_s and δ_{gs} are numerically the same. With covariates, the pooled QMLE and FE Poisson approaches produce different parameter estimates. An important issue is how one would turn the FE Poisson estimates into ATTs on the levels—that is, how one would estimate the τ_{gr}. Even more challenging is to obtain valid inference on the estimated ATTs after Poisson FE estimation. The approach in Martin (2017) for estimating average partial effects, and obtaining valid standard errors, can be adapted to the current setting, but this is notably more difficult than the pooled QMLE approach with cohort dummies.

As before, given separate estimates of the τ_{gr}, one may want to aggregate these in interesting ways. Alternatively, it is straightforward to impose restrictions directly. For example, one could replace $W_{it} \cdot D_{ig} \cdot fr_t$ with indicators for amount of time subjected to the intervention. The implicit assumption is that cohort and calendar time matter only insofar as they imply different exposure lengths. Including $T - q + 1$ 'intensity' (or 'exposure') indicators, say Z_{ith}, $h \in \{1, 2, \ldots, T - q + 1\}$, can greatly conserve on estimated parameters with many post-intervention periods. These can be interacted with the covariates in the nonlinear model. In the extreme case, one simply includes W_{it} by itself, possibly interacted with covariates, and then computes the APE with respect to W_{it} averaged over the $W_{it} = 1$ subsample. As in the case of a linear model, restricting the model in this way may produce misleading estimates of the ATTs for general staggered interventions.

4. TESTING AND CORRECTING FOR VIOLATION OF PARALLEL TRENDS

Because of its important role in identifying the ATTs, it is desirable to have simple tests of the PT assumption. As a robustness check, it is helpful to combine nonlinear models with simple models that allow heterogenous trends in the never treated state.

4.1. Testing the PT assumption

In the linear case, Wooldridge (2021) has shown that tests of the PT assumption are easily carried out in the context of pooled OLS estimation. Moreover, the tests are the same whether based only on the $W_{it} = 0$ observations (imputation regression) or on pooled OLS using all observations—provided full flexibility is allowed in the treatment indicators, as in Section 3.1. In other words, tests obtained pooling over the entire sample are equivalent to the commonly used 'pre-trend' tests that use only the untreated observations. As discussed by Wooldridge (2021), this means the tests using post-treatment data are not 'contaminated' by using treated observations—provided that the treatment effects are allowed to be flexible. In other words, the tests will not reject due to misspecification of the model for treatment effect heterogeneity.

The algebraic equivalence of the pooled tests and pre-trends tests carries over to the nonlinear case provided the canonical link function is used in the LEF. Technically, if one uses a different mean function or different objective function, the test should be carried out using only the $W_{it} = 0$ observations (although it seems unlikely the difference would be important in practice).

There are two approaches that apply to the general index case covered in this paper. In the common timing case, the first adds the interactions $D_i \cdot fs_t$ for $s = 2, \ldots, q - 1$, and does a joint test in these terms. If $q = 2$ (only one pre-treatment period), there is nothing we can test. If $q = 3$, one adds $D_i \cdot f2_t$ and does a cluster robust t test. With $q > 3$, there are multiple restrictions to

test. In the general nonlinear case, one can implement the test as a cluster-robust Wald test of exclusion restrictions. This is similar to the kind of 'event study' tests that are common when including treatment indicators that vary only by the number of periods away from treatment, both before and after an intervention.

In the general staggered case, the dummies $D_{ig} \cdot fs_t$ are included for $g = q,\ q+1,...,\ T$ and for $s = \{2, \ldots, g-1\}$. In other words, for each cohort, include indicators for the pre-intervention periods. One carries out a cluster-robust Wald statistic of joint significance of these pre-intervention indicators. This approach can result in many restrictions to test. For example, with $T = 6$ and $q = 4$, one would add $D_{i4} \cdot f2_t$, $D_{i4} \cdot f3_t$, $D_{i5} \cdot f2_t$, $D_{i5} \cdot f3_t$, $D_{i5} \cdot f4_t$, $D_{i6} \cdot f2_t$, $D_{i6} \cdot f3_t$, $D_{i6} \cdot f4_t$, and $D_{i6} \cdot f5_t$—or nine dummy variables. These are included in addition to the six dummies indicating legitimate treated observations. This event-study-type test is a cluster-robust joint test that the nine coefficients on the pre-intervention treatment indicators are all zero.

A different approach, which conserves on degrees of freedom, is to add the cohort-specific linear trends, $D_{iq} \cdot t$, $D_{i,q+1} \cdot t$,..., $D_{iT} \cdot t$. This test will have as many degrees of freedom as there are treatment cohorts. In the common timing case, the significance of the single term $D_i \cdot t$ can be tested using a cluster-robust t statistic. With many pre-treatment periods, one could add more functions of time, such as $D_{ig} \cdot t^2$ for each cohort g, but it seems that if important differences in trends are present, a linear trend will pick those up in most cases.

Generally, there is a tradeoff between the event-study-type test and the heterogenous linear trend test because the latter has fewer degrees of freedom, but does not look in all directions where PT might be violated. Incidentally, if one has controls $\mathbf{X}_i$, they should be included flexibly as in (3.12). This includes interacting them with a full set of time dummies, if possible, to allow unrestricted aggregate trends in the never treated state, $Y_{it}(\infty)$, that may also differ by observed heterogeneity, $\mathbf{X}_i$.

4.2. Correcting for violation of parallel trends

As discussed in Wooldridge (2021), the event-study approach is generally inappropriate as a correction for pre-trends, as it would require that violation of PT disappears just when we need it to. By contrast, the assumption that each cohort has a separate linear trend in the absence of the intervention is a reasonable—albeit not completely general—model of heterogenous trends. As discussed in Wooldridge (2021), in the $T = 3$ case with intervention only in the final period, including $D_i \cdot t$ in the POLS estimation, produces a difference-in-difference-in-differences estimator of the single ATT, τ_3. This estimator was used by Bell et al. (1999) to account for heterogeneous trends.

Including heterogeneous linear trends can be costly in terms of precision. Because of the nature of the staggered intervention, where treatment dummies $W_{it} \cdot D_{ig} \cdot fs_t$ are included in a flexible way and turn on in later periods, the treatment dummies are collinear with the heterogenous trends $D_{ig} \cdot t$ (but not perfectly so with at least two pre-treatment periods). Of course, multicollinearity does not cause inconsistency, but it can result in a severe loss of precision. As in other situations that employ pre-testing, using a pre-test for heterogeneous trends, and then deciding on the next step based on the outcome of the test, is not ideal. The statistical properties of such procedures warrant further study in linear and nonlinear contexts. At the same time, one does not want to unnecessarily control for irrelevant variables (heterogeneous trends) that induce large standard errors.

The simplest analysis is obtained using pooled OLS, pooled logit (or fractional logit), or pooled Poisson estimation (with an exponential mean function). Then, the imputation and pooled estimators are identical, and the pooled method is convenient for obtaining valid standard errors (clustering at least at the unit level). In equation (3.12), one simply adds the terms

$$D_{iq} \cdot t, D_{i,q+1} \cdot t, ..., D_{iT} \cdot t. \tag{4.1}$$

When using the pooled QMLE with a canonical link function, and computing the ATTs using standard software for marginal effects, one must be careful to evaluate the linear trend, t, at the appropriate period. Specifically, for τ_{rg}, set $D_g = 1$, $D_h = 0$ for $h \neq g$, $fr = 1$, $fs = 0$ for $s \neq r$, and $t = r$.

5. SIMULATIONS

In this section I summarize two sets of simulations meant to determine how the nonlinear methods fair against methods based on linear PT. In all scenarios, the no anticipation assumption holds, and PT holds in the underlying linear index. The first competing method is the POLS procedure proposed in Wooldridge (2021), which produces the same ATT estimates as TWFE applied to the same flexible model; what Wooldridge (2021) calls 'extended' TWFE. POLS is also identical to the imputation estimates that uses cohort dummies (see Proposition 3.2) and the BJS (2022) imputation estimates based on TWFE. The other competitor is CS (2021), using the never treated group as the only control group (the default in the Stata user-written software `csdid`). The panel data are generated for six time periods ($T = 6$) with three periods prior to the first intervention, and $N = 500$ units are drawn from the population distribution. All simulations use 5,000 Monte Carlo replications. The precise data generating mechanisms and the tables are provided in the Online Appendix; here I present an overview of the findings. The simulations were performed in Stata 17 and the code is available upon request from the author.

5.1. Binary outcome with common timing

In the first set of simulations, I imposed common timing of the intervention at $q = 4$ and generated the potential outcomes, $Y_t(0)$ and $Y_t(1)$, to be binary responses. I used two error distributions in the index formulation: logistic (with mean zero, variance $\pi^2/3$) and $Uniform(-2, 2)$. I also generated the underlying linear indices, $Y_t^*(0)$ and $Y_t^*(1)$, in two different ways. In the first, $Y_t^*(0)$ exhibits no trend (but is correlated with the eventually treated indicator, D, and the time-constant continuous covariate, X). The index $Y_t^*(1)$ exhibits a mild upward trend. In the second case, both $Y_t^*(0)$ and $Y_t^*(1)$ have fairly strong (and different) upward trends. The fraction of treated units in both cases is $P(D = 1) \approx 0.382$.

First consider the scenario with logistic errors and mild trend. The logit model is correctly specified and so one would expect the logit pooled MLE (PMLE) to perform well. That is the case, as the estimator exhibits essentially no bias for all three sample ATTs, which are $SATT_4 = 0.060$, $SATT_5 = 0.094$, and $SATT_6 = 0.113$. Both pooled OLS and CS (2021) are also essentially unbiased when the aggregate trend is mild. However, POLS and CS (2021) are notably less efficient than PMLE, with CS (2021) being the least precise. The POLS standard deviations are at least 45% larger than the logit PMLE standard errors; the CS (2021) standard errors are at least 69% larger than the logit PMLE standard errors. In terms of the PT tests, both the event study test (two degrees of freedom) and the heterogeneous trend test (one degree of freedom) have rejection rates between 4.2% and 4.6% for a 5% significance level test. These

rejection rates are expected for the logit PMLE, as the index PT assumption holds. Rejection rates around 5% are not necessarily expected for POLS, as the linear PT assumption is violated. Nevertheless, the failure to reject the linear PT assumption is a reasonable outcome considering the POLS estimators show essentially no bias.

When $Y_t^*(0)$ and $Y_t^*(1)$ have substantial upward trends (but conditional PT holds), the story is very different. The logit PMLE continues to perform well with essentially no bias for estimating $SATT_4 = 0.078$, $SATT_5 = 0.114$, and $SATT_6 = 0.158$. By contrast, POLS and CS (2021) not only have large biases, but they average to nontrivial *negative* numbers for all three ATTs. The large, downward biases in the linear PT appear to be related to the strong aggregate trends. The average response probabilities in the control and treated states, $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.385$ and $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.459$, do not seem unusual. The population R-squared from the POLS regression is about 0.276, which seems reasonable given the regression includes a full set of time period dummies. In addition to showing no bias, the PMLE remains more precise than the other two estimators. Unfortunately, the PT tests are not useful diagnostics for determining that the linear model is flawed, as the rejection rates are very close to the nominal 5% level.

When the errors are generated as $Uniform(-2, 2)$, all estimation methods are based on misspecified response probabilities. (The response probability for the $Uniform(-2, 2)$ distribution, which hits the limit values of zero and one, is sometimes called the 'ramp function'.) Nevertheless, when $Y_t^*(0)$ has no trend and $Y_t^*(1)$ has a mild trend, all estimators have very small biases, which shows some resiliency to the precise functional form. Interestingly, the logit PMLEs are still most efficient by nontrivial amounts. In the case with strong trends in $Y_t^*(0)$ and $Y_t^*(1)$, the linear estimators now show notably less bias than the logit PMLE. The SATTs are $SATT_4 = 0.125$, $SATT_5 = 0.150$, and $SATT_6 = 0.174$ and the Monte Carlo means of the POLS estimators are 0.111, 0.135, and 0.160, respectively. By contrast, the logit PMLE has means 0.197, 0.243, and 0.287, which shows substantial upward bias. Unfortunately, the PT tests are again unhelpful for model selection, as the rejection percentages are all around 5%.

Simulations are necessarily special and not always realistic. The purpose of these simulations is to show that one can obtain very similar estimates in some cases and also very different estimates depending the true functional form and underlying aggregate trends. With at least one covariate with substantial variation, one can explore goodness-of-fit as a possible way of choosing among different conditional mean models. Unfortunately, the PT tests do not provide guidance about misspecification in the binary response simulations provided here.

In some sense, simulations over many different scenarios can be viewed as exploring the limits of the bounds on the treatment effects of the kind derived in Athey and Imbens (2006) (for the $T = 2$ case). In empirical practice, one can try a linear analysis along with a sensible nonlinear model, such as logit, and hopefully draw robust conclusions.

5.2. Nonnegative outcome with staggered intervention

Next, I report on two simulations for a staggered intervention when the observed outcome, Y_{it}, is a nonnegative outcome with mass at zero. The treatment occurs first for different units in periods four, five, and six, with the cohort indicators generated by an ordered probit model. The (approximate) shares are

$$P(D_\infty = 1) \approx 0.357, \;\; P(D_4 = 1) \approx 0.291$$

$$P(D_5 = 1) \approx 0.225, \;\; P(D_6 = 1) \approx 0.127,$$

and so there is a relatively large never treated group.

In the first case, $Y_t(\infty)$ has a Poisson distribution conditional on unobserved heterogeneity following an exponential conditional mean, but there is no trend in the linear index. The zero outcome is not particularly important, with $T^{-1}\sum_{t=1}^{T} P(Y_t(\infty)=0) \approx 0.058$. The potential outcomes $Y_t(4)$, $Y_t(5)$, and $Y_t(6)$ are generated similarly to $Y_t(\infty)$ except the linear indices have nontrivial, and fairly different, aggregate trends. No anticipation holds, and the conditional PT assumption holds on the linear index.

From the findings reported in the Online Appendix, the clear-cut winner is the exponential mean function estimated using Poisson pooled QMLE (where the Poisson distribution is not close to the true distribution). The pooled QMLE is essentially unbiased for each SATT (six of them), whereas the pooled OLS estimators have downward biases over 30% in all cases and over 50% in some. The CS estimators show less bias than POLS, but are still very different, on average, from the SATTs. Moreover, the precision of the Poisson regression estimates is notably better than either POLS or CS.

The outcomes of the PT tests are much more promising than in the binary case. The event-study-type test rejects in the linear model 99.5% of the time, and the heterogeneous trends test has a 99.9% rejection rate. Therefore, one would conclude almost certainly that the PT assumption is violated in the linear model. Because the observed Y_{it} is a count variable, the hope is that one would turn to pooled Poisson regression with an exponential mean. The PT tests for the exponential model reject only 9.9% and 6.9% of the time, with the latter test using the three heterogeneous linear trends having particularly good size with $N = 500$ and three treated cohorts.

In the second simulation for $Y_t \geq 0$, I generated $Y_t(\infty)$ as a corner solution outcome: $P(Y_t(\infty)=0) > 0$ with $Y_{it}(\infty)$ continuous over strictly positive values. The underlying linear index, $Y_t^*(\infty)$, does not have a trend, but it depends on treatment cohort, unobserved heterogeneity, and the covariate, X. Zero is a fairly common outcome for $Y_t(\infty)$, with $T^{-1}\sum_{t=1}^{T} P(Y_t(\infty)=0) \approx 0.368$. The potential outcomes $Y_t^*(4)$, $Y_t^*(5)$, and $Y_t^*(6)$ have pronounced and different aggregate trends, generating a flexible pattern in the six treatment effects. The population R-squared from the POLS regression is about 0.083.

The conditional mean still has an exponential form and so the Poisson QMLE is consistent. It also shows essentially no bias, which is encouraging because, with $N = 500$, the final treated cohort has, on average, about 63 units. Somewhat surprisingly, the POLS and CS (2021) methods do just as well in terms of bias. This may be partly due to the treatment effects being much smaller in magnitude than in the count case. Because the outcome variable is not close to having a Poisson distribution, and fairly strong serial correlation is present, there is no guarantee that the Poisson QMLE using the exponential mean is efficient. Nevertheless, in this scenario the Poisson QMLE is the most precise of the three estimators. For example, for τ_{46}, the POLS and CS (2021) Monte Carlo standard deviations relative to the pooled QMLE are about 1.31 and 1.57, respectively. CS (2021) is the least precise by some margin because it uses only one period prior to the intervention and, in this simulation, it adopts the never treated group as the only control group for each treated cohort/time period combination.

The linear model is not systematically rejected using the PT tests—a sensible outcome because the POLS estimators show essentially no bias. The event-study-type test for pooled Poisson appears to reject somewhat too often (15.7% for a nominal 5% test), whereas the heterogeneous trends test is better behaved. The findings suggest that, with this kind of data generating mechanism, one would expect the linear and exponential models to produce similar estimates, with the precision of the pooled Poisson estimator probably being better. Because the observed outcome

is nonnegative with a corner at zero, the exponential mean model estimated by Poisson QMLE suggests itself, a priori, as an attractive alternative, or at least supplement, to the linear model.

Other scenarios for nonnegative responses suggest themselves. For example, generate the corner solution to follow a Tobit model. Then the exponential mean would be misspecified. Preliminary simulations show that both linear and exponential model can well approximate the ATTs even with more than 50% of the outcomes at the corner.

6. EMPIRICAL APPLICATION

I apply DiD with an exponential mean to the data in Di Tella and Schargrodsky (2004), who study an intervention in July 1994 where more police were assigned to certain blocks in Buenos Aires, Argentina, after a terrorist attack on the main Jewish Center in Buenos Aires. Of 876 blocks, 37 were provided with extra police to protect Jewish centres. While the number of treated units is relatively small, it is still more than in policy interventions using, say, the 50 states in the United States.

The data are reported monthly, running from April to December of 1994. Because the terrorist attack occurred midway through July, a case can be made for dropping July and using April, May, and June as the pre-intervention months and August to December as the post-intervention months. I take that approach here.

The outcome variable, *thefts*, reported monthly, is measured as the weekly average number of car thefts. More than 78% of the outcomes are zero, with increments of 0.25 up to a maximum of 2.5. Di Tella and Schargrodsky (2004) used a standard linear DiD analysis with several robustness checks, such as including covariates and looking at spillover effects (for which they find no evidence). Here I begin with a linear model, first imposing a single effect and then estimating five different treatment effects that vary by month. Without covariates, the linear estimation is a standard DiD, where *thefts* is averaged over the three pre-intervention months (April, May, and June) and the average is subtracted from the outcome in a particular treated month. This adjusted outcome is averaged across the treated and control blocks with the difference being the estimated ATT for that month. I estimate comparable exponential models using Poisson regression. The Callaway and Sant'Anna (2021) estimates are also reported in the case of estimating separate effects. Without covariates and with common timing, the CS (2021) estimates are 2×2 DiDs using June as the only control period and using each of August to December as treated periods. The estimates without covariates are given in Table 2.

Column (1) effectively reproduces the constant effect specification of Di Tella and Schargrodsky (2004) when July is not included. It implies that the presence of more police reduced car thefts by about 0.087 per month, on average—a nontrivial effect compared with the overall average, 0.093. The estimate from the exponential mean in column (2) is obtained by mimicking the linear analysis, where the time-varying treatment indicator is included along with the indicator of being a treated unit and a single post-intervention period. The estimate from the exponential model is practically the same, and even slightly less precise.

Columns (3), (4), and (5) show estimates allowing for a different ATT in each month. Again, there are no practical differences between the linear and exponential estimates, with the latter having slightly larger standard errors. The PT tests all have p-values above 0.5, providing no evidence against the PT assumption in the linear or exponential means. The CS (2021) estimates show a similar pattern, but are uniformly larger in magnitude; they are also much less precise. The CS (2021) procedure is essentially an event study approach, which only uses the period before

Table 2. No covariates.

	(1) Single Effect (Linear)	(2) Single Effect (Exponential)	(3) Separate Effects (Linear)	(4) Separate Effects (Exponential)	(5) Separate Effects (CS, 2021)
τ	−0.087	−0.089	—	—	—
	(0.030)	(0.033)			
τ_8	—	—	−0.081	−0.084	−0.116
			(0.041)	(0.046)	(0.056)
τ_9	—	—	−0.103	−0.103	−0.138
			(0.030)	(0.032)	(0.048)
τ_{10}	—	—	−0.065	−0.067	−0.099
			(0.047)	(0.050)	(0.062)
τ_{11}	—	—	−0.091	−0.092	−0.126
			(0.031)	(0.033)	(0.049)
τ_{12}	—	—	−0.096	−0.098	−0.131
			(0.031)	(0.035)	(0.049)
Event Study p-value (2 df)	0.652	0.650	0.530	0.489	—
Heterogeneous Trend Test (1 df)	0.818	0.812	0.576	0.523	—

Table 3. With covariates.

	(1) Single Effect (Linear)	(2) Single Effect (Exponential)	(3) Separate Effects (Linear)	(4) Separate Effects (Exponential)	(5) Separate Effects (CS, 2021)
τ	−0.087	−0.087	—	—	—
	(0.030)	(0.035)			
τ_8	—	—	−0.085	−0.090	−0.119
			(0.042)	(0.046)	(0.056)
τ_9	—	—	−0.105	−0.104	−0.138
			(0.030)	(0.031)	(0.047)
τ_{10}	—	—	−0.064	−0.066	−0.098
			(0.047)	(0.050)	(0.062)
τ_{11}	—	—	−0.091	−0.095	−0.125
			(0.031)	(0.033)	(0.049)
τ_{12}	—	—	−0.097	−0.104	−0.131
			(0.032)	(0.037)	(0.050)
Event Study p-value (2 df)	0.652	0.650	0.542	0.518	—
Heterogeneous Trend Test (1 df)	0.818	0.812	0.609	0.571	—

the treatment as the control period. The two pre-treatment 'treatment effects' using CS (2021) are also statistically insignificant (with p-values 0.642 and 0.280).

Table 3 includes three binary control variables, indicating whether the block houses a bank, a public building, or a petrol station. In the case of a single effect, these covariates are only interacted with the treatment variable, $W_{it} = D_i \cdot post_t$. In columns (3) and (4), they are included in a fully flexible way, as shown in equation (3.17). CS (2021) also allows for full flexibility.

Including the controls has very little impact on the ATT estimates or their standard errors. Nevertheless, an advantage of the linear and exponential regressions is that one can study the coefficients on the interactions, $W_{it} \cdot D_i \cdot fr_t \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)$ (not reported here) to determine the presence of moderating effects. Interestingly, the presence of a bank on a block drives the treatment effects to essentially zero, a finding that may be explainable by the idea that car thieves already think the police presence and other security is high near blocks that house a bank. One caution is that the estimated effects are based on relatively few observations. The full set of results and Stata code are available on request from the author.

In this application—which has a common intervention timing—the exponential model reaffirms the linear model estimates. In my view, this is not a bad thing. It demonstrates resiliency to the PT assumption in that imposing it on two fairly different functional forms leads to very similar estimates. Probably the small variation in the outcome variable plays a role in the similarity of the estimates. The simulations in Section 5 demonstrate that one can get very similar estimates (on average) and, in certain scenarios, very different estimates.

7. EXTENSIONS

I now discuss some extensions that are straightforward in the current setting.

7.1. All units eventually treated

The development in the previous sections assumed the existence of a never treated group. This is not necessary with multiple intervention periods, as the methods go through with little change if all units are treated by $t = T$. As discussed in Wooldridge (2021) in the linear case, without a never treated group the ATTs are initially defined relative to the potential outcome $Y_t(T)$. The average gain $E\left[Y_t(g) - Y_t(T) \,|\, D_g = 1\right]$ for $g < T$ is for being treated first in period g rather than in the last period. If the PT assumption is stated for $Y_t(T)$ rather than $Y_t(\infty)$, all of the previous methods go through. Clearly, we cannot estimate a treatment effect for the final treated cohort because there is no control group at time $t = T$ for $g = T$. However, if thinking about the potential outcome $Y_t(\infty)$ makes sense, even though we never observe this potential outcome, for $t < T$ the no anticipation assumption implies $E\left[Y_t(T) - Y_t(\infty) \,|\, D_g = 1\right] = 0$ for $g < T$. Therefore, $E\left[Y_t(g) - Y_t(T) \,|\, D_g = 1\right] = E\left[Y_t(g) - Y_t(\infty) \,|\, D_g = 1\right]$ for $g < T$ and $t \in \{g, g+1, \ldots, T-1\}$, and one can interpret the ATTs just as when there is a never treated group. For $g < T$ and $t = T$, we are always identifying $E\left[Y_T(g) - Y_T(T) \,|\, D_g = 1\right]$.

As a practical matter, in Procedure 2 one simply drops any term involving D_{iT} because $D_{ig} + D_{i,g+1} + \cdots + D_{iT} = 1$. In effect, $D_{iT} = 1$ acts as a control group in each of the treated periods.

7.2. A strategy with exit

It is possible to extend the pooled estimation methods to the case where the intervention turns off for some units, possibly in a staggered way. I discuss the linear case in Wooldridge (2021). The idea is to expand the notation of a cohort to be indexed by the first and last treatment dates, with the assumption that the intervention is in force over the entire interval. For example, with $T = 6$ and the first intervention at $q = 4$, the first treated cohort can be treated for all three periods, the first two periods, or only for the first period. The cohort first treated in period five can be treated for one or two periods, and the final treated cohort is treated for the one period.

We can represent this situation generally by defining a set of cohort dummies D_{gh} for $g \leq h \leq T$, where g is the first period of treatment and h is the last period. The case $h = \infty$ is allowed and represents the case of treatment through time T (so $D_{g\infty} = D_g$ in the previous notation). Initially, assume a never treated group as indicated by $D_\infty = 1$. The potential outcomes are now $Y_t(g, h)$ and the ATTs of interest are

$$\tau_{ghr} \equiv E\left[Y_r(g,h) - Y_r(\infty) \,|\, D_{gh} = 1\right], r = g, g+1, \ldots, T,$$

where $Y_r(\infty)$ is the PO in the never treated state. Note that ATTs are defined even when $r > h$; that is, after the intervention has been removed. Estimating these ATTs allows one to determine whether an intervention has lasting effects even after it has been removed.

Estimation is straightforward. In place of the interactions $D_g \cdot fr_t, r = g, \ldots, T$, one includes $D_{gh} \cdot fr_t$ for $g \leq h$ and $r = g, \ldots, T$. Even with a modest number of treated periods this can result in many ATTs, especially if there is exit for each treated cohort. As before, imposing restrictions on the parameters of the mean functions, or aggregating the estimated effects, is easy in principle. If there is no never treated group, but there is a group treated only in $t = T$, this group plays the role of the control group—just as when there is no exit.

In order to obtain valid standard errors for the $\hat{\tau}_{ghr}$, it is easiest to define treatment indicators $W_{tghr} \equiv D_{gh} fr_t, r = g, \ldots, T$, and then obtain the APEs with respect to each of these, averaging over the subsample corresponding to $W_{tghr} = 1$.

It is important to know that this approach allows endogeneity of exit only through its correlation with time-constant observables and unobservables. In other words, by expanding the list of cohort indicators to account for starting and ending time, we are allowing the mean of $Y_t(\infty)$ to vary systematically with these expanded cohorts. What is not being allowed is for a shock to $Y_t(\infty)$ at time t to lead to exit in a future period. This is essentially a strict exogeneity assumption on the time-varying treatment indicator once unobserved heterogeneity has been accounted for, the same restriction required for the FE estimator to consistently estimate parameters in the presence of attrition from a panel. One diagnostic that can be used is the same that is used to test no anticipation: add the lead value of the treatment indicator, $W_{i,t+1}$, in the expanded equation and perform a robust t statistics (losing the last time period).

7.3. Time-varying covariates

I have focused on time-constant covariates because using pre-intervention covariates ensures that one is not using 'bad controls' (or 'overcontrolling'). Nevertheless, it is somewhat common to see time-varying covarates included in applications. At a minimum, time-varying covariates should not be influenced by the policy intervention. Then, one can use the conditional expectation in (3.12) with $\mathbf{X}_i$ replaced by $\mathbf{X}_{it}$. In the linear case, this would allow trends to differ by cohort based on how $\{\mathbf{X}_{it} : t = 1, \ldots, T\}$ is trending because $D_{ig} \cdot \mathbf{X}_{it} - D_{ig} \cdot \mathbf{X}_{i1} = D_{ig}(\mathbf{X}_{it} - \mathbf{X}_{i1})$. (With time-constant covariates, this term vanishes.) Generally, τ_{gr} would be estimated as the APEs with respect to W_t averaged over the subpopulation $D_g = 1, fr_t = 1$. The estimated δ_{gr} will be easiest to interpret if the covariates $\mathbf{X}_{it}$ are centred around the means $\bar{\mathbf{X}}_{gt}$ for cohort g in period t. With a large enough cross section, one might even add the time averages, $\bar{\mathbf{X}}_i = T^{-1} \sum_{s=1}^{T} \mathbf{X}_{is}$, as in the Mundlak (1978) device. See also Wooldridge (2019) for the nonlinear case.

7.4. Multiple treatment levels and continuous treatments

Sometimes an intervention of interest has more than two levels. Mechanically, allowing for multiple treatment levels is relatively straightforward, but there are issues of how much heterogeneity should be allowed. If one simply wants to control for the heterogeneity in different treatment timing, as before, then the cohort indicators, D_{ig}, are defined as before, and one can replace the binary indicator, W_{it}, in (3.12) with a set of indicators for the different treatment levels (with $W_{it} = 0$ indicating not treated). This straightforward approach seems sensible, although it would

not be completely general because, in principle, one can define cohorts based on first treatment level as well as first period treated. For example, define indicators D_{iga} where g is the initial treatment period and a is the initial treatment level. One could even replace the binary treatment in (3.12) with a continuous treatment and estimate average partial effects with respect to W_t across different treatment levels, also averaging by (g, t) pairs. I leave for the future an analysis of the precise treatment effects being recovered.

8. CONCLUDING REMARKS

I have proposed a simple yet flexible framework for estimating average treatment effects in staggered DiD settings when the (conditional) PT assumption holds for a known, strictly increasing transformation of the conditional mean of the response variable. I argued that logit, fractional logit, and Poisson regression (with an exponential mean) are particularly attractive pooled quasi-MLEs. One can estimate a full set of ATTs indexed by cohort/calendar time or impose restrictions, or the estimated effects can be aggregated in various ways. Covariates are easily accommodated to allow more flexible statements of the PT assumption.

One can use an imputation approach, based on first estimating the nonlinear model using the control observations, or a pooled method. In the cases mentioned above, the two methods produce identical estimates. Generally, whether or not one uses a canonical link function, the pooled QMLE is more convenient for obtaining standard errors and conducting inference; under stronger assumptions, one can also obtain ATEs.

I proposed simple tests of the conditional PT assumption. In the context of the equation with fully heterogeneous treatment effects, using only the control units or pooling across all observations leads to the same test—provided the canonical link function is used. This is true for an event-study type test that includes pre-treatment indicators as well as a test (usually with fewer degrees of freedom) that includes cohort-specific trends. I argue that the latter approach also leads to a sensible correction when PT is thought to be violated.

The simulations show that, although in some cases linear and nonlinear models produce similar ATT estimates on average, there are cases where one or both can exhibit bias. The possibility of using data-driven methods to choose among different transformation functions, $G(\cdot)$, should be further explored, although there is a limit to what the data can tell us. In simple cases, such as $T = 2$ without covariates, the data are silent on the choice of the function $G(\cdot)$. When we have multiple pre-treatment periods, covariates, or both, model selection tests of the kind discussed in Rahmani and Wooldridge (2019) might be useful.

ACKNOWLEDGEMENTS

Prepared for the special session 'The New Difference-in-Differences', sponsored by *The Econometrics Journal*, at the 2022 Meetings of the Royal Economic Society. I am grateful to managing editor Jaap Abbring and two anonymous referees for helpful comments on an earlier draft, to the data editor Joan Llull and an anonymous replicator, to the participants of the special session for comments on my presentation, and to various participants in short courses who have seen this material at various stages.

REFERENCES

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*, 1–19.

Abbring, J. H. and G. J. Van den Berg (2003). The nonparametric identification of treatment effects in duration models. *Econometrica 71*, 1491–517.

Ai, C. and E. C. Norton (2003). Interaction terms in logit and probit models. *Economics Letters 80*, 123–9.

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*, 431–97 AI (2006).

Athey, S. and G. W. Imbens (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics 226*, 62–79.

Bell, B., R. Blundell and J. van Reenen (1999). Getting the unemployed back to work: The role of targeted wage subsidies. *International Tax and Public Finance 6*, 339–60.

Blundell, R. and M. Costa Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources 44*, 565–640.

Borusyak, K., X. Jaravel and J. Spiess (2022). Revisiting event study designs: robust and efficient estimation. Working paper, University College London BJS (2022).

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*, 200–30 CS (2021).

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*, 2964–96.

de Chaisemartin, C. and X. D'Haultfœuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *Econometrics Journal 63*, C1–30.

Di Tella, R. and E. Schargrodsky (2004). Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review 94*, 115–33. Data available at https://www.aeaweb.org/aer/data/march2004_ditella_data.zip.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*, 254–77.

Gourieroux, C., A. Monfort and A. Trognon (1984). Pseudo-maximum likelihood methods: Theory. *Econometrica 52*, 681–700.

Heckman, J. J., H. Ichimura and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*, 605–54.

Martin, R. S. (2017). Estimation of average marginal effects in multiplicative unobserved effects panel models. *Economics Letters 160*, 16–19.

Melly, B. and G. Santangelo (2015). The changes-in-changes model with covariates. Working paper, Bern University.

Mundlak, Y. (1978). On the pooling of cross section and time series data. *Econometrica 46*, 69–85.

Negi, A. and J. M. Wooldridge (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews 40*, 504–34.

Papke, L. E. and J. M. Wooldridge (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics 11*, 619–32.

Papke, L. E. and J. M. Wooldridge (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics 145*, 121–33.

Puhani, P. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear 'difference-in-differences' models. *Economics Letters 115*, 85–87.

Rahmani, I. and J. M. Wooldridge (2019). Model selection tests for complex survey samples. In K. Huynk, D. Jacho-Chavez and G. Tripathi (Eds.), *Advances in Econometrics, Volume 39*, 109–35 (The Econometrics of Comple x Survey Data). Bingley, UK: Emerald Publishing.

Roth, J. and P. H. C. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*, 737–47.

Sant'Anna, P. H. C. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*, 101–22.

Sun, L. and Abraham S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*, 175–99.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, *90*, 77–97.

Wooldridge, J. M. (2005). Fixed effects and related estimators for correlated random-coefficient and treatment effect panel data models. *Review of Economics and Statistics 87*, 385–90.

Wooldridge, J. M. (2007). Inverse probability weighted m-estimation for general missing data problems. *Journal of Econometrics 141*, 1281–301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, second edition. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics 211*, 137–50.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906345.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Managing editor Jaap Abbring handled this manuscript.

APPENDIX

A.1. Proof of Proposition 3.1

For this Appendix, the notation is easier when variables are lower case. Plus, that emphasizes the results are algebraic in nature; they hold for any outcome of data provided there is no perfect collinearity.

The following proposition is useful for establishing the equivalence between the pooled QMLE formulation of the estimators and their imputation counterparts.

PROPOSITION A.1. *Consider a panel data set* $\{(y_{it}, \mathbf{h}_{it}, \mathbf{m}_{it}, w_{it}) : t = 1, \ldots, T; i = 1, \ldots, N\}$ *where* $\mathbf{h}_{it}$ *is* $1 \times K$, $\mathbf{m}_{it}$ *is* $1 \times L$, *and* w_{it} *is a binary indicator. Further, assume that* $w_{it}\mathbf{m}_{it} = \mathbf{m}_{it}$ *for all* i *and* t *[so that* $(1 - w_{it})\mathbf{m}_{it} = \mathbf{0}$*]. For a strictly increasing function* $G(\cdot)$ *defined on* $\mathbb{R}$, *let* $\tilde{\beta}$ *and* $\tilde{\gamma}$ *be unique solutions to the equations*

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{h}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta} + \mathbf{m}_{it}\tilde{\gamma}\right)\right] = \mathbf{0} \tag{A.1}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{m}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta} + \mathbf{m}_{it}\tilde{\gamma}\right)\right] = \mathbf{0}. \tag{A.2}$$

Let $\hat{\beta}$ be the unique solution to the equations

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - w_{it})\mathbf{h}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\hat{\beta}\right)\right] = \mathbf{0}. \tag{A.3}$$

If for some $L \times K$ matrix $\mathbf{A}$, $w_{it}\mathbf{h}_{it} = \mathbf{m}_{it}\mathbf{A}$ for all (i, t), then

$$\hat{\beta} = \tilde{\beta}. \tag{A.4}$$

Proof: Because $(1 - w_{it}) \cdot \mathbf{m}_{it} = \mathbf{0}$, conditions (A.1) and (A.2) can be written as

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - w_{it})\mathbf{h}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta}\right)\right] + \sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}\mathbf{h}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta} + \mathbf{m}_{it}\tilde{\gamma}\right)\right] = \mathbf{0}, \tag{A.5}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}\mathbf{m}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta} + \mathbf{m}_{it}\tilde{\gamma}\right)\right] = \mathbf{0}. \tag{A.6}$$

Because $w_{it}^2 = w_{it}$, we can substitute $w_{it}\mathbf{h}_{it} = w_{it}\mathbf{m}_{it}\mathbf{A}$ and use algebra to write the first order conditions (FOCs) as

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - w_{it})\mathbf{h}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta}\right)\right] + \mathbf{A}'\sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}\mathbf{m}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta} + \mathbf{m}_{it}\tilde{\gamma}\right)\right] = \mathbf{0}, \tag{A.7}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}\mathbf{m}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta} + \mathbf{m}_{it}\tilde{\gamma}\right)\right] = \mathbf{0}. \tag{A.8}$$

Plugging (A.8) into (A.7) gives

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - w_{it})\mathbf{h}_{it}'\left[y_{it} - G\left(\mathbf{h}_{it}\tilde{\beta}\right)\right] = \mathbf{0},$$

which is the same set of equations as (A.3) and implies $\tilde{\beta} = \hat{\beta}$ by uniqueness. □

To use Proposition A.1 to prove Proposition 3.1, in the general setting with covariates, choose

$$\mathbf{h}_{it} = \left(1, d_{iq}, \ldots, d_{iT}, f2_t, \ldots, fT_t, \mathbf{x}_i, d_{iq} \cdot \mathbf{x}_i, \ldots, d_{iT} \cdot \mathbf{x}_i, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i\right),$$

and

$$\mathbf{m}_{it} = \left(d_{iq} fq_t, \ldots, d_{iq} fT_t, \ldots, d_{iT} fT_t, d_{iq} fq_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, d_{iq} fT_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, d_{iT} fT_t \cdot \dot{\mathbf{x}}_{iT}\right).$$

The equations in (A.1) and (A.2) are known to hold for the first order conditions of the pooled QMLE in the LEF when $G^{-1}(\cdot)$ is the canonical link function; in fact, this is one common characterization of the canonical link. By definition of w_{it}, $w_{it} d_{ig} fr_t = d_{ig} fr_t$, and so $w_{it}\mathbf{m}_{it} = \mathbf{m}_{it}$. Moreover, as described in Wooldridge (2021) in the linear case, $w_{it}\mathbf{h}_{it} = \mathbf{m}_{it}\mathbf{A}$ for a suitably chosen matrix $\mathbf{A}$. It follows that $\hat{\beta}$, from the pooled estimation using all of the data, equals $\tilde{\beta}$, the pooled QMLE from the $w_{it} = 0$ estimation.

The imputation estimate, $\hat{\tau}_{gr}$, is

$$\begin{aligned}\hat{\tau}_{gr} &= N_g^{-1} \sum_{i=1}^{N} d_{ig} \left[y_{ir} - G\left(\hat{\alpha} + \hat{\beta}_g + \mathbf{x}_i\hat{\kappa} + \mathbf{x}_i\hat{\eta}_g + \hat{\gamma}_r + \mathbf{x}_i\hat{\pi}_r\right)\right] \\ &= N_g^{-1} \sum_{i=1}^{N} d_{ig} \left[y_{ir} - G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{x}_i\tilde{\kappa} + \mathbf{x}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{x}_i\tilde{\pi}_r\right)\right].\end{aligned}$$

For the pooled estimation, the APE with respect to w_t—that is, setting it to zero and one—and evaluating at $fr_t = 1$, $fs_t = 0$ for $s \neq r$, and averaging over the $d_{ig} = 1$ subsample, is

$$\begin{aligned}\tilde{\tau}_{gr} = N_g^{-1} \sum_{i=1}^{N} d_{ig} \big[&G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{x}_i\tilde{\kappa} + \mathbf{x}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{x}_i\tilde{\pi}_r + \tilde{\delta}_{gr} + \dot{\mathbf{x}}_{ig}\tilde{\xi}_{gr}\right) \\ &\times G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{x}_i\tilde{\kappa} + \mathbf{x}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{x}_i\tilde{\pi}_r\right)\big].\end{aligned}$$

It is clear that $\tilde{\tau}_{gr} = \hat{\tau}_{gr}$ if

$$N_g^{-1} \sum_{i=1}^{N} d_{ig} y_{ir} = N_g^{-1} \sum_{i=1}^{N} d_{ig} G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{x}_i\tilde{\kappa} + \mathbf{x}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{x}_i\tilde{\pi}_r + \tilde{\delta}_{gr} + \dot{\mathbf{x}}_{ig}\tilde{\xi}_{gr}\right).$$

But this holds by the FOCs for the pooled estimation problem. In particular, the FOC with respect to δ_{gr} is

$$\begin{aligned}0 = \sum_{i=1}^{N}\sum_{t=1}^{T} d_{ig} fr_t \Bigg[y_{it} - G\Bigg(&\tilde{\alpha} + \sum_{h=q}^{T} \tilde{\beta}_h d_{ih} + \mathbf{x}_i\tilde{\kappa} + \sum_{g=q}^{T} (d_{ih} \cdot \mathbf{x}_i)\,\tilde{\eta}_h \\ &+ \sum_{s=2}^{T} \tilde{\gamma}_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i)\,\tilde{\pi}_s \\ &+ \sum_{h=q}^{T}\sum_{s=h}^{T} \tilde{\delta}_{hs}\,(d_{ih} \cdot fs_t) + \sum_{h=q}^{T}\sum_{s=h}^{T} (d_{ih} \cdot fs_t \cdot \dot{\mathbf{x}}_{ih})\,\tilde{\xi}_{hs}\Bigg)\Bigg],\end{aligned}$$

or

$$\sum_{i=1}^{N} d_{ig} y_{ir} = \sum_{i=1}^{N} d_{ig} G\left(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{x}_i\tilde{\kappa} + \mathbf{x}_i\tilde{\eta}_g + \tilde{\gamma}_r + \mathbf{x}_i\tilde{\pi}_r + \tilde{\delta}_{gr} + \dot{\mathbf{x}}_{ig}\tilde{\xi}_{gr}\right).$$

Dividing by N_1 gives the result and Proposition 3.1 is proven.

This result applies to estimation of event-study-style equations, where $d_{ig} fs_t$ for $s < g$ are included in $\mathbf{h}_{it}$. By definition of w_{it}, $w_{it} d_{ig} fs_t = 0$ (and so these are trivial linear combinations of $\mathbf{m}_{it}$). The result also extends to when cohort-specific trends, $d_{ig} \cdot t$, are included, and even with the terms $d_{ig} \cdot t \cdot \mathbf{x}_i$. The

equivalence now follows from the fact that $d_{ig} \cdot t$, which is included in the vector $\mathbf{h}_{it}$, is such that $w_{it} \cdot d_{ig} \cdot t$ is a linear combination of $d_{ig} \cdot fg_t, ..., d_{ig} \cdot fT_t$, which are all in $\mathbf{m}_{it}$. (The coefficients in the linear combination are $g, g+1, ..., T$.) Similarly, $w_{it} \cdot d_{ig} \cdot t \cdot \mathbf{x}_i$ is a linear combination of $d_{ig} \cdot fg_t \cdot \dot{\mathbf{x}}_{ig}, ..., d_{ig} \cdot fT \cdot \dot{\mathbf{x}}_{ig}$. In fact, the proof goes through if $\mathbf{x}_i$ is replaced with $\mathbf{x}_{it}$, whether or not the covariates are demeaned using averages for each (g, t) pair. In other words, the imputation and pooled approaches with time-varying covariates are identical when using the canonical link function in the LEF.