Simple Approaches to Nonlinear Difference-in-Differences with Panel Data

Jeffrey M. Wooldridge
Department of Economics
Michigan State University

This version: August 5, 2022

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 Jaap Abbring for helpful comments on an earlier draft and to the participants of the session for comments and questions on my presentation.

Abstract: I derive simple, flexible strategies for difference-in-differences settings where the nature of the response variable may warrant a nonlinear model. In addition to covering the case of common treatment timing, I allow for 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 (QMLEs) in the linear exponential family (LEF) extend the pooled ordinary least squares (POLS) estimation in Wooldridge (2021). By using the conditional mean associated with the canonical link function, imputation and estimation pooled across the entire sample produce identical results. Moreover, using the canonical link 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. A small simulation study shows the estimators work well when the mean function is correctly specified; they also have some resiliency to misspecification.

1. Introduction

Difference-in-differences 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 the fixed effects and related 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 two-way fixed effects (TWFE) can lead to an estimated effect that is difficult to interpret. Sun and Abraham (2021) study the properties of event study estimators and propose an estimator that allows for heterogeneous treatment effects. Borusyak, Jaravel, and Spiess (2021) [BJS (2021)] derive imputation estimators, based on two-way fixed effects, 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 (2022) provides a nice discussion of these papers, and more.

In Wooldridge (2021), I argued that for staggered interventions under the commonly imposed assumptions of no anticipation and parallel trends, one can recover average treatment effects on the treated (ATTs) that vary by cohort and calendar time using linear models and standard regression techniques. Moreover, using an expanded equation with cohort and calendar time interactions, pooled ordinary least squares (POLS), random effects, and TWFE pooled across all of the data are equivalent. I also showed that the natural imputation approach

leads to the same ATT estimates. Covariates are allowed so that the parallel trends assumption only needs to hold after conditioning on covariates. The conclusion is that there is nothing inherently wrong with two-way fixed effects estimation: One simply needs to apply the method to a suitably flexible equation. Nevertheless, consistent estimation of the ATTs by POLS (TWFE) requires that the parallel trends assumption applies to changes over time in the mean of the outcomes; in this paper, I refer to this as the *linear parallel trends* (LPT) assumption and define it formally in the sections that follow.

In the current paper, I allow for a version of the parallel trends assumption that may be more attractive when the outcome variable, Y_t, is limited in range. For example, when Y_t is binary, the usual parallel trends assumption is potentially suspect. If Y_t is a fractional response, the PT assumption is also questionable – especially when that outcome regularly takes on values near or at the endpoints of zero or one. Another common situation where the linear PT 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. A perhaps better assumption is that PT holds in the ratio of means, which is implied by an exponential conditional mean function.

In the two-period case, Roth and Sant'Anna (2021) study the problem of invariance of the parallel trends assumption to strictly monotonic transformations of the outcome variable. They show that invariance requires a strong form of parallel trends 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 parallel trends 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 the common choices of $G(\cdot)$, $G^{-1}(Y_t(0))$ is not even defined. For a nonnegative variable $Y_t(0)$ with $P(Y_t(0) = 0) > 1$, I argue that the PT assumption can hold for $\log(E[Y_t(0)|D])$ but not for $E[Y_t(0)|D]$. With a corner at zero, $\log(Y_t(0))$ is not well defined, and so it cannot be under consideration in the Roth-Sant'Anna setting. In this paper I recognize that, in most cases, 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 this outcome. We are not interested in some possibly strange transformation of the outcome variable. Because of the special features of $Y_t(0)$, it may be more realistic to impose the PT assumption on a transformation of $E[Y_t(0)|D]$; but the goal of estimating ATTs in terms of the potential outcomes $Y_t(1)$ and $Y_t(0)$ does not change.

When $T = 2$ and Y_t is continuous, the analysis here is less general than that in 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. Moreover, it is not clear how to apply the AI (2006) approach to staggered interventions and when one introduces covariates.

From a practical perspective, my goal in this paper is to provide simple strategies that allow empirical researchers to compare results from linear DiD analyses with sensible nonlinear alternatives. I impose the minimal assumptions on distributions, requiring only specification of a conditional mean function. All other conditional moments, for both the marginal and joint distributions, 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 restrictions on serial dependence over time. A special case is pooled OLS estimation of a linear equation that I proposed in Wooldridge (2021). Consequently, in settings with limited dependent variables, the methods here allow one to assess robustness of findings by combining a linear analysis with a suitable nonlinear analysis.

It is important to remember that, in panel data settings with a small number of time periods, fixed effects strategies do not generally produce reliable estimators of parameters or partial effects. Including a full set of unit dummy variables usually results in an incidental parameters problems unless the number of time periods is sufficiently large. The exceptions are rare: A linear model estimated by fixed effects and an exponential model estimated by Poisson fixed effects. In the former case, Wooldridge (2021) shows that a carefully formulated pooled OLS regression and TWFE produce identical estimates. In the Poisson case, identification of ATTs on the level of the outcome is generally not possible when FE estimation is used, and so the FE Poisson estimator has limitations for estimating ATTs on levels. That TWFE coincides with pooled OLS in the linear case suggests that pooled methods in nonlinear cases may have good properties. This turns out to be true: I formally demonstrate that pooled QMLE in the LEF identifies the ATTs under weak assumptions.

The identification of the ATTs is based on random variables representing an underlying population of units that we observe over time periods $t \in \{1, 2, \ldots, T\}$. When I describe estimation, it is easiest to think of random sampling N units from the cross section. 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 confidence intervals would have to be treated in more detail. I do not present regularity conditions needed for pooled quasi-MLE methods to have the usual asymptotic properties because the required moment and smoothness assumptions are almost always satisfied in the difference-in-differences (DiD) setting.

The remainder of 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 considers any number of time periods with common intervention. I also show how to include covariates in the general common timing case. Section 4 turns to the general case of staggered interventions, allowing for covariates. Although I initially assume the existence of a never treated group, I show how to relax that assumption in Section 8. In addition, I formally state an equivalence result between imputation and pooled QMLE when using the canonical link in the LEF.

Section 5 discusses testing and correcting for violations of parallel trends. In Section 6, I present findings from a small simulation study. In Section 7 I apply the results for Poisson regression with an exponential mean function to the car thefts data, at the block level, in Di Tella and Schargrodsky (2004). Section 8 discusses two extensions: How to allow for all units being eventually treated and how to handle (staggered) exit from treatment. Section 9 contains concluding remarks and suggests some directions for future research.

2. The $T = 2$ Case

To understand the nature of the parallel trends assumption I impose in the general staggered case, it is helpful to begin with the $T = 2$ case as in Heckman, Ichimura, and Todd (1997), Abadie (2005), Athey and Imbens (2006), and Sant'Anna and Zhou (2020), among others. 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, meaning treatment after the first period and prior to the second. Therefore, $Y_1(0)$ 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 ($t = 2$):

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

There are two assumptions that serve to identify τ_2. The first is a *no anticipation* (NA) assumption. The strongest form of the assumption says that $Y_1(1) = Y_1(0)$; this is the version used by Heckman, Ichimura, and Todd (1997), Abadie (2005), and others. See also Abbring and Van den Berg (2003) for a discussion of the no anticipation assumption in the context of duration analysis.

A weaker version of no anticipation 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 in the linear case in two parts:

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

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

In the absence of (2.4), (2.3) has no content because it is the same as simply defining

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

However, combining (2.3) and (2.4) gives

$$\begin{aligned} E[Y_2(0)|D] - E[Y_1(0)|D] &= E[Y_2(0) - Y_1(0)|D] \\ &= (\alpha + \beta D + \gamma_2) - (\alpha + \beta D) = \gamma_2 \end{aligned} \tag{2.5}$$

In other words, on average, the trend $Y_2(0) - Y_1(0)$ 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 is not possible 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 program, the

PT assumption allows that selection may be based on differences in pre-treatment earnings but not on how the earnings are trending in the absence of the intervention.

When $Y_t(0)$ 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 $G(\cdot)$,

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

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

The key restriction is that $G(\cdot)$ is strictly increasing. For simple use of standard asymptotic analysis, we add that $G(\cdot)$ is sufficiently smooth. Just as when $G(\cdot)$ is the identify function, (2.6) imposes no restriction in isolation because it simply implies, definitionally, $\alpha = G^{-1}(E[Y_1(0)|D = 0])$ and $\beta = G^{-1}(E[Y_1(0)|D = 1]) - G^{-1}(E[Y_1(0)|D = 0])$. The key is that the same function $G(\cdot)$ appears in (2.7). Combining (2.6) and (2.7) gives

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

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

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

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

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

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

Assume further that

$$U_1, U_2 \text{ are continuous and independent of } D$$
$$U_1, U_2 \text{ are identically distributed with CDF } F(\cdot)$$

We do not restrict the dependence between U_1 and U_2 so that general serial correlation is allowed. Then, for $t = 1, 2$ with $\gamma_1 \equiv 0$, we have

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

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

$$E[Y_1^*(0)|D] = \alpha + \beta D \tag{2.9}$$
$$E[Y_2^*(0)|D] = \alpha + \beta D + \gamma_2, \tag{2.10}$$

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

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 $G(\cdot) = \exp(\cdot)$:

$$E[Y_t(0)|D] = \exp(\alpha + \beta D) \tag{2.11}$$
$$E[Y_2(0)|D] = \exp(\alpha + \beta D + \gamma_2) \tag{2.12}$$

These equations imply that the parallel trends assumption is in terms of the growth in the mean in the untreated state, not the change in the mean. In particular,

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

does not depend on D. Equivalently, the difference in logs of the mean is

$$\log\{E[Y_2(0)|D]\} - \log\{E[Y_1(0)|D]\} = \gamma_2$$

In this $T = 2$ setting we can never determine the function $G(\cdot)$, if any, such that assumptions (2.6) and (2.7) hold. For example, if PT holds for $G(\cdot)$ logistic then the linear PT assumption will fail. We have to take a stand on which function $G(\cdot)$ is most realistic. As we will see, our choice of $G(\cdot)$ definitely affects how we estimate the ATT, τ_2. Whether the estimates are sensitive to the choice of $G(\cdot)$ is something one would explore in an application.

2.1. Identification and Estimation

We can easily see that, under no anticipation (NA) (2.2) and the PT assumption in (2.6) and (2.7), the ATT τ_2 is identified. To see how, 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,

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

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

The second part of (2.13), $E[Y_2(0)|D = 1]$, is the one that requires use of (2.2), (2.6), and (2.7). By (2.7) we have

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

and, given that $G(\cdot)$ is assumed known, identification of $E[Y_2(0)|D = 1]$ rests on identifying α, β, and γ_2. From (2.6),

$$E(Y_1|D = 0) = E[Y_1(0)|D = 0] = G(\alpha),$$

and so

$$\alpha = G^{-1}(E(Y_1|D = 0))$$

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, we use (2.2) along with (2.6):

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

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

$$\alpha + \beta = G^{-1}(E(Y_1|D = 1))$$

or

$$\beta = G^{-1}(E(Y_1|D = 1)) - \alpha$$

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 by using the sample averages over the control and treated units, respectively:

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

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

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,

$$\hat{\alpha} = G^{-1}(\bar{Y}_{01}) \xrightarrow{p} \alpha$$

and

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

Of course, 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),

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

and so

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

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

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

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

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

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

When $G(\cdot)$ is the identify function – the linear DiD case – the estimator is

$$\begin{aligned} \hat{\tau}_2 &= (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01}) \\ &= (\bar{Y}_{12} - \bar{Y}_{02}) - (\bar{Y}_{11} - \bar{Y}_{01}), \end{aligned} \tag{2.18}$$

which is the basic DiD estimator. Clearly, (2.17) will generally change as the choice of $G(\cdot)$

changes. For example, when $G(\cdot) = \exp(\cdot)$,

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

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

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

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,

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

thereby providing an (approximate) proportional effect. When $G(z) = \exp(z)/[1 + \exp(z)]$, 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

$$\hat{\delta}_2 = G^{-1}(\bar{Y}_{12}) - \left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_2\right);$$

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

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, define the time-varying treatment indicator

$$W_{it} = D_i \cdot f2_{it}, t = 1,2, \tag{2.21}$$

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

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

Then impute a treatment effect for the treated units in $t = 2$:

$$\widehat{TE}_{i2} \equiv Y_{i2} - \hat{\alpha} - \hat{\beta} - \hat{\gamma}_2 \text{ if } D_i = 1$$

Next, average across the treated units in the second time period to get

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

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

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

Often the regression is written with $D_i \cdot f2_t$ in place of W_{it}; later, explicitly using W_{it} pays dividends for simplifying the calculation of estimates and standard errors.

A simple imputation strategy also applies in the general case. Further, for certain combinations of $G(\cdot)$ and the objective function used to estimate the parameters, methods that pool across all observations are available. Rather than cover the equivalences in the $T = 2$ case, I now turn to the general common timing case.

3. Common Timing with General T

3.1. No Covariates

Now consider the case where there are a total of $T \geq 2$ total time periods, indexed from $t = 1,\ldots,T$. An intervention occurs at $t = q$ and stays in effect through period T. Because there are only two groups to consider – the control group and the (eventually) treated group – it suffices to denote the potential outcomes as $Y_t(0)$ and $Y_t(1)$.

The treatment effects of interest are the ATTs in the treated time periods:

$$\tau_r = E[Y_r(1) - Y_r(0)|D = 1], r = q, q+1,\ldots,T \tag{3.1}$$

Again, we can identify the τ_r under two population assumptions.

Assumption NAC (**No Anticipation**, **Common**): For $1 \leq t \leq q-1$,

$$E[Y_t(1) - Y_t(0)|D = 1] = 0. \square \tag{3.2}$$

A version of NAC that is easier to work with but (technically) more restrictive is

$$Y_t = Y_t(0) = Y_t(1), 1 \leq t \leq q - 1,$$

which means that, for each population unit, the potential outcomes are the same prior to the intervention.

In cases where anticipation is a concern, one can modify the assumption and estimation methods to allow for anticipation in periods near the intervention period. There are two ways to think about this, with the first requiring no change in notation. We simply allow the time spaces between the index $t \in \{1,2,\ldots,T\}$ to vary, and so the jump from $q - 1$ to q could indicate, say, a three-year period. Probably a bit more natural is to replace $q - 1$ with, say, $p \leq q - 1$, and allow violation of NAC for $p \leq t \leq q - 1$. In terms of estimation, this simply means that we would drop periods $t \in \{p,p + 1,\ldots,q - 1\}$ in any analysis that follows. As we will see shortly, we could get by with imposing NAC only for $t = 1$ and dropping periods $\{2,\ldots,q - 1\}$. Naturally, this could be very costly in terms of precision when pre-treatment periods are dropped unnecessarily. To keep the notation and discussion relatively simple, I use the NAC assumption as stated above and one can think of varying lengths of time between periods.

Generally, we can write the observed outcomes as

$$Y_t = (1 - D)Y_t(0) + D \cdot Y_t(1), t = 1,\ldots,T \tag{3.3}$$

For now, we observe the treatment status, D, the sequence of observed outcomes, $\{Y_t : t = 1,2,\ldots,T\}$, and we know the data of the intervention, q.

In the linear case, the parallel trends assumption can be written as

$$E[Y_t(0)|D] = \alpha + \beta D + \gamma_t, t = 1,2,\ldots,T$$

where $\gamma_1 \equiv 0$ is a normalization without loss of generality; see, for example, Wooldridge (2021). It follows that

$$E[Y_t(0)|D] - E[Y_{t-1}(0)|D] = \gamma_t - \gamma_{t-1}, t = 2,\ldots,T,$$

and so the change in the mean outcome in the untreated state does not differ across the control and treatment groups. As discussed in Section 2, this PT assumption may not be natural when $Y_t(0)$ is discrete (such as binary or a count) or its range is limited in some substantive way (such as being in the unit interval). The next assumption is the parallel trends assumption with a general index function.

Assumption IPTC (**Index Parallel Trends**, **Common**): For a known, strictly increasing, continuously differentiable function $G(\cdot)$ and parameters α, β, and γ_2, ..., γ_T,

$$E[Y_t(0)|D] = G(\alpha + \beta D + \gamma_t), t = 1,2,\ldots,T, \tag{3.4}$$

where $\gamma_1 \equiv 0$. □

The binary response example in Section 2 extends immediately to the general case of common intervention. Namely, let

$$Y_t(0) = 1[Y_t^*(0) > 0]$$

$$Y_t^*(0) = \alpha + \beta D + \gamma_t + U_t, t = 1,\ldots,T$$

where the U_t, $t = 1,\ldots,T$ are independent of D, identically distributed across t, have zero mean, and have continuous cumulative distribution function $F(\cdot)$. Then

$$\begin{aligned} E[Y_t(0)|D] &= P[Y_t(0) = 1|D] = P[\alpha + \beta D + \gamma_t + U_t > 0|D] \\ &= 1 - F[-(\alpha + \beta D + \gamma_t)] \equiv G(\alpha + \beta D + \gamma_t) \end{aligned}$$

Notice how the usual PT assumption for a linear model holds for the latent variable $Y_t^*(0)$:

$$E[Y_t^*(0)|D] = \alpha + \beta D + \gamma_t$$

The exponential mean example also immediately extends. If

$$E[Y_t(0)|D] = \exp(\alpha + \beta D + \gamma_t) \tag{3.5}$$

(again with $\gamma_1 = 0$) then for $t = 2, \ldots, T$, the ratio

$$\frac{E[Y_t(0)|D]}{E[Y_1(0)|D]} = \exp(\gamma_t) \tag{3.6}$$

does not depend on D.

The general restriction is that, for the chosen, invertible function $G(\cdot)$,

$$G^{-1}(E[Y_t(0)|D]) - G^{-1}(E[Y_1(0)|D]) = \gamma_t, t = 2, \ldots, T. \tag{3.7}$$

When $G(z) = \exp(z)/[1 + \exp(z)]$, $G^{-1}(\cdot)$ is the log-odds transformation of the conditional mean (which applies to fractional as well as binary responses).

Given Assumption IPTC and the fact that $Y_t = Y_t(1)$ when $D = 1$, we can write the ATTs as

$$\tau_r = E(Y_r|D = 1) - E[Y_r(0)|D = 1] = E(Y_r|D = 1) - G(\alpha + \beta + \gamma_r),\ r = q, \ldots, T \tag{3.8}$$

As usual, the first term is always identified, and consistently estimated using the sample average of the treated units in time $t = r$:

$$\bar{Y}_{1r} = N_1^{-1} \sum_{i=1}^{N} D_i Y_{ir} \xrightarrow{p} E(Y_r|D = 1) \tag{3.9}$$

For the second term, it suffices to consistently estimate α, β, and γ_r, $r = q, \ldots, T$.

As in the $T = 2$ case, α and β are identified using observables when $t = 1$ because

$$\alpha = G^{-1}(E(Y_1|D = 0))$$

$$\beta = G^{-1}(E(Y_1|D = 1)) - \alpha$$

Then, γ_r, $r = q, \ldots, T$ are identified by the PT assumption (3.4) because

$$E(Y_r|D = 0) = G(\alpha + \gamma_r), r = q, \ldots, T$$

or

$$\gamma_r = G^{-1}(E(Y_r|D = 0)) - \alpha, r = q, \ldots, T \tag{3.10}$$

The means $E(Y_r|D = 0)$ are consistently estimated using the average of the period r outcomes for the control units:

$$\bar{Y}_{0r} \equiv N_0^{-1} \sum_{i=1}^{N} (1 - D_i) Y_{ir} \tag{3.11}$$

The previous argument that shows identification of α, β, γ_2, ..., γ_T only uses the PT assumption along with no anticipation in $t = 1$. In fact, we only need to identify α, β, and γ_r, $r = q, \ldots, T$, which means, technically, we can drop the NA and PT assumptions for periods $t \in \{2, \ldots, q - 1\}$. But that is the same as just entirely ignoring those time periods in estimation. We can always make that choice at the cost of precision. By convention, I assume that one wants to include all the listed periods in the analysis; if not, one simply thinks of $q - 1$ as the latest period before the intervention where NA and PT are assumed to hold. By allowing uneven spacing between time periods we need not even change notation; we simply allow an extended gap between the periods labeled $q - 1$ and q.

Given the NA and PT assumptions, rather than separately estimating the parameters it is more efficient (and convenient) to estimate the parameters at once using all untreated observations. To this end, define the time-varying treatment indicator as

$$W_{it} = D_i \cdot (fq_t + \cdots + fT_t) \equiv D_i \cdot p_t, \tag{3.12}$$

where $f2_t$, ..., fT_T is a set of mutually exclusive time dummies and $p_t = fq_t + \cdots + fT_t$ is a post-treatment period indicator. In the common timing setting, $\{W_{it} : t = 1,2,\ldots,T\}$ can have only two patterns: all zeros or $q-1$ zeros followed by $T-q+1$ ones. Using the NA and PT assumptions it is easy to see that

$$E(Y_{it}|D_i, W_{it} = 0) = G(\alpha + \beta D_i + \gamma_t), t = 1,\ldots,q-1 \quad (3.13)$$

$$= G(\alpha + \gamma_t), t = q,\ldots,T \quad (3.14)$$

Therefore, the parameters can be estimated using a pooled estimation method on the untreated observations, $W_{it} = 0$, to obtain $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}_t$, $t = 2,\ldots,T$. Given consistent parameter estimators, consistent estimators of the ATTs are

$$\hat{\tau}_r = \bar{Y}_{1r} - G\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_r\right), r = q,\ldots,T \quad (3.15)$$

This estimator can be viewed as an imputation estimator, where we use the untreated observations to obtain $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}_r$ and then impute $Y_{ir}(0)$ as

$$\hat{Y}_{ir}(0) = G\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_r\right) \quad (3.16)$$

In terms of specific estimation methods, it is natural to use a quasi-MLE approach using a log likelihood in the linear exponential family; see Gourieroux, Monfort, and Trognon (1984) for the cross-sectional case and Wooldridge (2010, Section 13.11.4) for a discussion of pooled estimation with panel data. When $G(z) = z$, it is natural to choose the normal log likelihood, which then leads to the pooled OLS estimators in Wooldridge (2021). For common nonlinear choices of $G(\cdot)$, the normal QMLE is not very attractive because it has essentially no chance of being efficient – 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 contenders for $G(\cdot)$, and then it is natural to use the Bernoulli quasi-log likelihood function (QLLF). For the latter

case, see Papke and Wooldridge (1996, 2008). Because the Bernoulli distribution is in the linear exponential family (LEF), the pooled QMLE is fully robust for estimating $(\alpha, \beta, \gamma_2, \ldots, \gamma_T)$ provided the mean is correctly specified, which holds under Assumptions NAC and IPTC. Estimation and inference are straightforward; see, for example, Wooldridge (2010, Section 13.11.4).

For obtaining estimators using the imputation approach, the choice of $G(\cdot)$, subject to being logically consistent with the nature of Y_{it}, is largely a matter of taste. For example, in the binary or fractional case, one would rarely have a compelling argument to choose logit over probit or vice versa. Nevertheless, in the current setting, there are some practical advantages to choosing $G^{-1}(\cdot)$ to be the *canonical link* function associated with the chosen density in the LEF. With such a choice, there is a simple equivalence between imputation and pooled QMLE estimation, something pointed out in Wooldridge (2021) for the linear case.

The idea is that we postulate that the conditional mean function holds for all draws, untreated and treated. That is, we act *as if*

$$E(Y_{it}|D_i, W_{it}) = G(\alpha + \beta D_i + \gamma_2 f2_t + \cdots + \gamma_T fT_t + \delta_q(W_{it} \cdot fq_t) + \cdots + \delta_T(W_{it} \cdot fT_t)\big), \tag{3.17}$$

where δ_q, ..., δ_T is a set of additional parameters. As in the $T = 2$ case, using $W_{it} \cdot fs_t$ rather than $D_i \cdot fs_t$ is convenient for obtaining the $\hat{\tau}_r$ along with valid standard errors. If $G(\cdot)$ is mean function associated with the canonical link for the chosen LEF density, the pooled QMLE, obtained using all observations, produces exactly the same estimates of α, β, and the γ_t as the pooled QMLE restricted to $W_{it} = 0$. Moreover, if one computes the average partial effects (APEs) of W_t using (3.17) and then evaluates these at $D = 1$ and $fr_t = 1$ with $fs_t = 0$ for $s \neq r$, the result is the estimates in (3.15). In other words, given the pooled QMLE estimators,

denoted with a "~," it can be shown that $\tilde{\alpha} = \hat{\alpha}$, $\tilde{\beta} = \hat{\beta}$, $\tilde{\gamma}_r = \hat{\gamma}_r$, and

$$\hat{\tau}_r = G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_r + \tilde{\delta}_r\right) - G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_r\right), r = q, \ldots, T \tag{3.18}$$

This follows because because $\tilde{\delta}_r$ is such that $\bar{Y}_{1r} = G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_r + \tilde{\delta}_r\right), r = q, \ldots, T$. A more general result with staggered entry and covariates is stated and discussed in Section 4.

The pooled method is convenient because popular software packages, such as Stata, produce average partial effects for different settings of the covariates and provides valid standard errors in the presence of serial correlation and misspecification of the chosen LEF density. In this simple case, the delta method is applied to (3.18) along with clustering. This equivalence means that, when $G^{-1}(\cdot)$ is the canonical link, the function $E(Y_{it}|D_i, W_{it})$ is only assumed correct when $W_{it} = 0$ – keeping with the spirit of estimators of ATTs when the LPT assumption is imposed. I will show a more general equivalence in the staggered case with covariates.

The virtues 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 3.1.

Table 3.1. Canonical Link and Log Likelihood Pairings		
Conditional Mean	LEF Density	Comments
Linear	Normal	Any Response. Leads to Pooled 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, No Natural Upper Bound

Anticipating the general recommendations below, in applying DiD to limited dependent variables it will make sense to compare DiD estimates from linear regression to a nonlinear method that exploits the special nature of Y_{it} – provided, of course, that the ATTs are properly computed for the nonlinear model.

3.2. Adding Covariates

It is straightforward to modify the index PT assumption to allow for covariates. First, we specify the no anticipation assumption conditional on the covariates.

Assumption CNAC (**Conditional NA**, **Common**): For a $1 \times K$ vector of time-constant covariates $\mathbf{X}$,

$$E[Y_t(1)|D = 1, \mathbf{X}] = E[Y_t(0)|D = 0, \mathbf{X}], \; t \in \{1, \ldots, q-1\}. \; \square \tag{3.19}$$

With the addition of covariates, we assume that the pre-intervention ATTs are zero for each subpopulation indexed by $\mathbf{X}$.

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

$$E[Y_t(0)|D, \mathbf{X}] = G(\alpha + \beta D + \mathbf{X}\boldsymbol{\kappa} + (D \cdot \mathbf{X})\boldsymbol{\eta} + \gamma_t + \mathbf{X}\boldsymbol{\pi}_t), \; t = 1, 2, \ldots, T, \tag{3.20}$$

where D is the binary treatment indicator and where we take $\gamma_1 \equiv 0$, $\boldsymbol{\pi}_1 \equiv \mathbf{0}$ as normalizations to define α and $\boldsymbol{\kappa}$. $\square$

Unlike in the case without covariates, when $t = 1$ the conditional mean in (3.20) now imposes functional form restrictions (unless $\mathbf{X}$ includes mutually exlusive and exhaustive binary indicators that partition the population). For flexibility, we can choose $\mathbf{X}$ to be functions of underlying control variables, such as in the common practice of including squares and interactions of underlying variables. Once we settle on the functional form

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

the substantive restriction imposed on the linear index in (3.20) is that the coefficients on terms involving D are time invariant. In the linear case the restriction becomes

$$E[Y_t(0)|D,\mathbf{X}] - E[Y_1(0)|D,\mathbf{X}] = \gamma_t + \mathbf{X}\boldsymbol{\pi}_t,\ t \in \{2,\ldots,T\}, \tag{3.21}$$

which does not depend on D. Wooldridge (2021) used this assumption to derive POLS estimators of the τ_r.

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

$$\frac{E[Y_t(0)|D,\mathbf{X}]}{E[Y_1(0)|D,\mathbf{X}]} = \exp(\gamma_t + \mathbf{X}\boldsymbol{\pi}_t),\ t = 2,\ldots,T,$$

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

Generally, we can exploit the conditional index PT assumption to identify the τ_r, which are still given by (3.1). Nothing changes for estimating $E[Y_r(1)|D = 1]$ because $Y_r = Y_r(1)$ for $D = 1$. For $E[Y_r(0)|D = 1]$, we first apply iterated expectations:

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

Next, by CIPTC,

$$E[Y_r(0)|D = 1, \mathbf{X}] = G(\alpha + \beta + \gamma_r + \mathbf{X}(\boldsymbol{\kappa} + \boldsymbol{\eta} + \boldsymbol{\pi}_r)) \tag{3.22}$$

and so

$$E[Y_r(0)|D = 1] = E[G(\alpha + \beta + \gamma_r + \mathbf{X}(\boldsymbol{\kappa} + \boldsymbol{\eta} + \boldsymbol{\pi}_r))|D = 1],\ r = q, \ldots, T \tag{3.23}$$

Given that $G(\cdot)$ is assumed known and we observe $\mathbf{X}$, (3.23) shows that $E[Y_r(0)|D = 1]$ is identified if the parameters

$$(\alpha, \beta, \gamma_r, \boldsymbol{\kappa}, \boldsymbol{\eta}, \boldsymbol{\pi}_r)$$

are identified. Provided the distribution of $\mathbf{X}$ is not degenerate, α, β, $\boldsymbol{\kappa}$, and $\boldsymbol{\eta}$ are identified using $t = 1$. The argument is a simple extension from the earlier one. First,

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

and, by CNAC with $t = 1$,

$$\begin{aligned} E(Y_1|D = 1, \mathbf{X}) &= E[Y_1(1)|D = 1, \mathbf{X}] = E[Y_1(0)|D = 1, \mathbf{X}] \\ &= G(\alpha + \beta + \mathbf{X}(\boldsymbol{\kappa} + \boldsymbol{\eta})) \end{aligned}$$

Combined, we simply have

$$E(Y_1|D, \mathbf{X}) = G(\alpha + \beta D + \mathbf{X}\boldsymbol{\kappa} + D \cdot \mathbf{X}\boldsymbol{\eta}), \tag{3.24}$$

where all variables are observed and $G(\cdot)$ 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 (2.24) under weak regularity conditions.

With $(\alpha, \beta, \boldsymbol{\kappa}, \boldsymbol{\eta})$ identified, all that remains for τ_r is identification of γ_r and $\boldsymbol{\pi}_r$. But, under CIPTC,

$$E(Y_r|D = 0, \mathbf{X}) = E[Y_t(0)|D, \mathbf{X}] = G(\alpha + \gamma_r + \mathbf{X}(\boldsymbol{\kappa} + \boldsymbol{\pi}_r)), \tag{3.25}$$

which shows that $\alpha + \gamma_r$ and $\boldsymbol{\kappa} + \boldsymbol{\pi}_r$ are identified using the period r observed outcomes for the

control group ($D = 0$). Given identification of α and $\boldsymbol{\kappa}$ from $t = 1$, γ_r and $\boldsymbol{\pi}_r$ are identified. Notice how we need not even obtain γ_s, $s < q$ (pre-intervention periods).

With the parameters in (3.22) identified, (3.23) is obtained by averaging over the distribution of $\mathbf{X}$ given $D = 1$. Technically, because we have a parametric function $G(\cdot)$, we do not need an overlap assumption. Nevertheless, the problems with extrapolation in treatment effects estimation are well-known. In practice, we should ensure the support condition

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

or, equivalently, $P(D = 1|\mathbf{X} = \mathbf{x}) < 1$ for all $\mathbf{x} \in \text{Supp}(\mathbf{X})$. The support condition 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_r and $\boldsymbol{\pi}_r$.

As before, with multiple pre-treatment periods, operationalizing the previous identification argument discards useful information in $t \in \{2, \ldots, q-1\}$ and is generally inefficient. Instead, we can use all untreated observations across all time periods by noting that, for a random draw i from the population,

$$\begin{aligned} E(Y_{it}|D_i, \mathbf{X}_i, W_{it} = 0) = G(&\alpha + \beta D_i + \mathbf{X}_i\boldsymbol{\kappa} + (D_i \cdot \mathbf{X}_i)\boldsymbol{\eta} \\ &+ \gamma_2 f2_t + \cdots + \gamma_T fT_t + (f2_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_2 + \cdots + (fT_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_T\big) \end{aligned} \tag{3.26}$$

where, again, W_{it} is the time-varying treatment indicator and the fs_t are time dummies. Equation (3.26) follows because $W_{it} = 0$ if and only if $t \leq q-1$ or $q \leq t \leq T$ and $D_i = 0$. Given (3.26) and a random sample of size N, the parameters $(\alpha, \beta, \boldsymbol{\kappa}, \boldsymbol{\eta}, \gamma_2, \ldots, \gamma_T, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_T)$ can be jointly estimated using any number of methods that identify parameters in conditional mean functions.

For the reasons discussed previously – robustness, familiarity, ease of estimation, and convenient inference – pooled QMLE using QLLFs in the LEF are attractive, with the

canonical link function providing extra computational benefits. Nevertheless, one may use any $\sqrt{N}$-consistent, asymptotically normal estimation method. In fact, other modeling approaches are allowed. For example, if Y_{it} is a corner solution response, we could use for the function $G(\cdot)$ that which is derived from the standard (homoskedastic) Tobit model. I do not pursue such possibilities here, preferring to specify only the conditional mean.

Given the parameter estimators, the τ_r are consistently estimated (with fixed T, $N \to \infty$) by

$$\hat{\tau}_r = \bar{Y}_{1r} - N_1^{-1} \sum_{i=1}^{N} D_i \cdot G\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_r + \mathbf{X}_i(\hat{\boldsymbol{\kappa}} + \hat{\boldsymbol{\eta}} + \hat{\boldsymbol{\pi}}_r)\right), \; r = q, \ldots, T. \tag{3.27}$$

Moreover, by applying Wooldridge (2010, Problem 12.17), a standard error for $\hat{\tau}_r$ can be obtained. In order to combine or test joint hypotheses about the τ_r, an asymptotic variance for $\hat{\boldsymbol{\tau}}$, $(T - q + 1)$ vector of ATT estimators, can be obtained. Alternatively, the panel bootstrap can be used. Even in fairly large sample sizes the bootstrap should be computionally feasible because pooled QMLEs in the LEF are not computationally time consuming.

In the linear case, (3.27) becomes

$$\hat{\tau}_r = \bar{Y}_{1r} - \left[\hat{\alpha} + \hat{\beta} + \hat{\gamma}_r + \bar{\mathbf{X}}_1(\hat{\boldsymbol{\kappa}} + \hat{\boldsymbol{\eta}} + \hat{\boldsymbol{\pi}}_r)\right]$$

where $\bar{\mathbf{X}}_1 = N_1^{-1} \sum_i D_i \cdot \mathbf{X}_i$ is the average of the treated subsample. Equation (3.27) extends the observation in Wooldridge (2021) for the linear case that $\hat{\tau}_r$ can be viewed as an imputation estimator. For each treated unit i in a treated time period r, define an imputed treatment effect as

$$\widehat{TE}_{ir} = Y_{ir} - G\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_r + \mathbf{X}_i(\hat{\boldsymbol{\kappa}} + \hat{\boldsymbol{\eta}} + \hat{\boldsymbol{\pi}}_r)\right), \tag{3.28}$$

where the second part is how $Y_{ir}(0)$ is imputed for a treated unit. Then we can write

$$\hat{\tau}_r = N_1^{-1} \sum_{i=1}^{N} D_i \cdot \widehat{TE}_{ir} \tag{3.29}$$

As in the $T = 2$ case without covariates, one can use a pooled QMLE using all of the data. Specifically, act as if the following conditional mean function holds for treated as well as untreated observations:

$$\begin{aligned} E(Y_{it}|D_i, \mathbf{X}_i, W_{it}) = G[&\alpha + \beta D_i + \mathbf{X}_i\boldsymbol{\kappa} + (D_i \cdot \mathbf{X}_i)\boldsymbol{\eta} \\ &+ \gamma_2 f2_t + \cdots + \gamma_T fT_t + (f2_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_2 + \cdots + (fT_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_T \\ &+ \delta_q(W_{it} \cdot fq_t) + \cdots + \delta_T(W_{it} \cdot fT_t) \\ &+ (W_{it} \cdot fq_t \cdot \dot{\mathbf{X}}_i)\boldsymbol{\xi}_q + \cdots + (W_{it} \cdot fT_t \cdot \dot{\mathbf{X}}_i)\boldsymbol{\xi}_T\Big], \end{aligned} \tag{3.30}$$

where the conditioning on W_{it} is redundant but useful for understanding the flexibility of the approach and also in obtaining proper standard errors. In forming the interactions $W_{it} \cdot fr_t \cdot \dot{\mathbf{X}}$, the controls $\dot{\mathbf{X}}$ have been demeaned using $E(\mathbf{X}|D = 1)$. The reason for demeaning the covariates is so that the δ_r have a useful meaning. In particular,

$$\delta_r = G^{-1}(E[Y_r(1)|D = 1]) - G^{-1}(E[Y_r(0)|D = 1]),$$

which is a treatment effect that we discussed earlier in the $T = 2$ case. It may be just as interesting to have an ATT stated, say, in terms of a percentage difference as on the level when $Y_{it} \geq 0$.

To obtain the parameter estimates, decorated with "~" and including the $\tilde{\delta}_r$, and the estimated ATTs in each treated period, use pooled QMLE with covariates $\dot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}_1$. Extending regression terminology, across all *t* and *i* use pooled QMLE of

$$\begin{aligned} Y_{it} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot \mathbf{X}_i, f2_t, ..., fT_t, f2_t \cdot \mathbf{X}_i, ..., fT_t \cdot \mathbf{X}_i, \\ W_{it} \cdot fq_t, ..., W_{it} \cdot fT_t, W_{it} \cdot fq_t \cdot \dot{\mathbf{X}}_i, ..., W_{it} \cdot fT_t \cdot \dot{\mathbf{X}}_i \end{aligned}$$

Then, $\tilde{\tau}_r$ is the average partial effect of W for the subpopulation $D = 1, fr_t = 1$ (setting all

other time dummies to zero):

$$\tilde{\tau}_r = N_1^{-1} \sum_{i=1}^{N} D_i \left[G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_r + \mathbf{X}_i(\tilde{\boldsymbol{\kappa}} + \tilde{\boldsymbol{\eta}} + \tilde{\boldsymbol{\pi}}_r) + \tilde{\delta}_r + \dot{\mathbf{X}}_i \tilde{\boldsymbol{\xi}}_r \right) - G\left(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma}_r + \mathbf{X}_i(\tilde{\boldsymbol{\kappa}} + \tilde{\boldsymbol{\eta}} + \tilde{\boldsymbol{\pi}}_r)\right) \right] \quad (3.3$$

In Section 4, I will show that, when $G^{-1}(\cdot)$ is the canonical link, all of the estimated coefficients in common with the imputation methods are the same. Moreover,

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

which means (3.31) is the same as (3.27).

A practical benefit of using (3.31) to obtain $\tilde{\tau}_r$ is that commonly used software packages – including Stata – provide a valid standard error for using the delta method. One can condition on the covariates or account for sampling variation in the $\mathbf{X}_i$. The online appendix shows Stata commands for obtaining both $\tilde{\tau}_r$ and its standard error.

Another potential benefit from using QMLE pooled across all observations is that it permits estimation of an average treatment effect (ATE) in each time period, not just an ATT. In particular, the estimated ATE in period $r \in \{q, \ldots, T\}$ would be

$$\begin{aligned} \tilde{\tau}_{r,ate} = N^{-1} \sum_{i=1}^{N} \Big[& G\left(\tilde{\alpha} + \tilde{\beta} D_i + \mathbf{X}_i \tilde{\boldsymbol{\kappa}} + (D_i \cdot \mathbf{X}_i)\tilde{\boldsymbol{\eta}} + \tilde{\gamma}_r + \mathbf{X}_i \tilde{\boldsymbol{\pi}}_r + \hat{\delta}_r + \dot{\mathbf{X}}_i \tilde{\boldsymbol{\xi}}_r \right) \\ & - G\left(\tilde{\alpha} + \tilde{\beta} D_i + \mathbf{X}_i \tilde{\boldsymbol{\kappa}} + (D_i \cdot \mathbf{X}_i)\tilde{\boldsymbol{\eta}} + \tilde{\gamma}_r + \mathbf{X}_i \tilde{\boldsymbol{\pi}}_r \right) \Big], \end{aligned} \quad (3.32)$$

which is obtained by computing the APE with respect to W for time period r. Here we average across both the treated and control units. The quantity $\hat{\tau}_{r,ate}$ is easily estimated using standard software packages and standard errors are obtained using the delta method. The difference with the ATT is that we average across treated and control units in time period r, not just the treated units.

In order for (3.32) to consistently estimate

$$\tau_{r,ate} = E[Y_r(1) - Y_r(0)],\ r = q, \ldots, T,$$

we have to strengthen the no anticipation and parallel trends assumption. In the latter case we would assume that the CIPT assumption holds in the treated state, $Y_t(1)$, in addition to the control state. I omit the details because the focus is usually on the ATTs.

4. Staggered Interventions

I now consider the case of staggered interventions, where different units are subjected to a treatment or intervention in different periods. This rolling out of the treatment generates different treatment cohorts.

4.1. 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 effectively being different treatment intensities. In any post-intervention period, units treated initially at time $t = q$ 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 8 I discuss relaxing this restriction.

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

$$Y_t(g),\ g \in \{q, \ldots, T, \infty\},\ t \in \{1, 2, \ldots, T\}, \tag{4.1}$$

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

$$\tau_{gr} = E[Y_r(g) - Y_r(\infty)|D_g = 1], r = g,\ldots,T;\ g = q,\ldots,T \tag{4.2}$$

For each (eventually) treated cohort g, τ_{gr}, $r = g,\ldots,T$ are the ATTs in all subsequent time periods.

We present the assumptions 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 $\mathbf{X}$,

$$E[Y_t(g)|D_g = 1,\mathbf{X}] = E[Y_t(\infty)|D_g = 1,\mathbf{X}]. \square \tag{4.3}$$

Assumption CIPTS (**Conditional Index PT**, **Staggered**): For $t = 1,2,\ldots,T$,

$$E[Y_{it}(\infty)|D_{iq},\ldots,D_{iT},\mathbf{X}_i] = G\left(\alpha + \sum_{g=q}^{T}\beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T}(D_{ig}\cdot\mathbf{X}_i)\boldsymbol{\eta}_g + \gamma_t + \mathbf{X}_i\boldsymbol{\pi}_t\right) \tag{4.4}$$

with $\gamma_1 \equiv 0$, $\boldsymbol{\pi}_1 \equiv \mathbf{0}$. $\square$

When $G(z) = z$ is the identify function, CIPTS implies, with $\mathbf{D}_i = (D_{iq},\ldots,D_{iT})$,

$$E[Y_{it}(\infty)|\mathbf{D}_i,\mathbf{X}_i] - E[Y_{i1}(\infty)|\mathbf{D}_i,\mathbf{X}_i] = \gamma_t + \mathbf{X}_i\boldsymbol{\pi}_t,$$

which allows the trend to depend on the covariates but not on the treatment cohort. This is the conditional parallel trends assumption used in Wooldridge (2021). When $t = 1$ and without covariates, (4.4) imposes no restrictions. With covariates, a functional form assumption is maintained:

$$E[Y_{i1}(\infty)|D_{iq},\ldots,D_{iT},\mathbf{X}_i] = G\left(\alpha + \sum_{g=q}^{T}\beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T}(D_{ig}\cdot\mathbf{X}_i)\boldsymbol{\eta}_g\right)$$

For $t \geq 2$, (4.4) imposes that the change in the index does not depend on $(D_{iq},\ldots,D_{iT})$.

The ATTs can be written as

$$\tau_{gr} = E(Y_r|D_g = 1) - E\left[G\left(\alpha + \beta_g + \gamma_r + \mathbf{X}_i\left(\boldsymbol{\kappa} + \boldsymbol{\eta}_g + \boldsymbol{\pi}_r\right)\right)\middle|D_g = 1\right], \tag{4.5}$$

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

The time-varying treatment indicator, W_{it}, now can be written as

$$\begin{aligned} W_{it} &= D_{iq}\cdot pq_t + \cdots + D_{iT}\cdot fT_t \\ &= D_{iq}\cdot(fq_t + \cdots + fT_t) + \cdots + D_{iT}\cdot pT_t \end{aligned} \tag{4.6}$$

where

$$pg_t = fg_t + \cdots + fT_t$$

is an indicator for the post treatment period if a unit is initially treated in period g. 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, $W_{it} = 0$, $t = 1,\ldots,T$.

The identification argument for the parameters extends the common timing case. Assumptions CNAS and CIPTS imply

$$E(Y_{i1}|D_{iq},\ldots,D_{iT},\mathbf{X}_i) = G\left[\alpha + \sum_{g=q}^{T}\beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T}(D_{ig}\cdot\mathbf{X}_i)\boldsymbol{\eta}_g\right], \tag{4.7}$$

and so, with strictly increasing $G(\cdot)$, population units in each treatment cohort, and no perfect collinearity in $\mathbf{X}_i$ in the population, α, β_q, ..., β_T, $\boldsymbol{\kappa}$, and $\boldsymbol{\eta}_q$, ..., $\boldsymbol{\eta}_T$ are all identified using the

first time period only. The CIPTS assumption implies that, for $r \geq q$,

$$E(Y_{ir}|D_{iq} = 0, \ldots, D_{iT} = 0|\mathbf{X}) = G[\alpha + \gamma_r + \mathbf{X}_i(\boldsymbol{\kappa} + \boldsymbol{\pi}_r)], \tag{4.8}$$

which verifies one can use the never treated units starting with the first intervention period to identify γ_r and $\boldsymbol{\pi}_r$, $r = q, \ldots, T$. With a never treated group, we need not use any of the eventually treated units in treated time periods – a feature of the CS (2021) approach. Using all control combinations (i,t) to estimate all parameters uses all of the conditional mean implications of the no anticipation and parallel trends assumption.

As in the common timing case, all parameters, including the γ_s, $\boldsymbol{\pi}_s$ for $s < q$, can be estimated at once using pooled estimation on the untreated observations. In particular, assumptions CNAS and CIPTS imply that

$$E(Y_{it}|D_{iq}, \ldots, D_{iT}, \mathbf{X}_i, W_{it} = 0) = G\left[\alpha + \sum_{g=q}^{T} \beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T} (D_{ig} \cdot \mathbf{X}_i)\boldsymbol{\eta}_g + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_s \right] \tag{4.9}$$

Implicitly, all of the valid control observations are used in each time period, which makes this approach generally more efficient than other approaches, such as Callaway and Sant'Anna (2021), that use long differences relative to a particular control group.

The above discussion essentially proves the following result.

Theorem 4.1: Assume that Assumptions CNAS and CIPTS hold for a strictly increasing function $G(\cdot)$. Assume that $\rho_g \equiv P(D_g = 1) > 0, g \in \{q, q+1, \ldots, T, \infty\}$. If $\mathbf{X}$ has a nondegenerate distribution then $\left(\alpha, \beta_q, \ldots, \beta_T, \boldsymbol{\kappa}, \boldsymbol{\eta}_q, \ldots, \boldsymbol{\eta}_T, \gamma_2, \ldots, \gamma_T, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_T\right)$ are identified and so are the ATTs. □

The following procedure uses all conditional mean implications of Assumptions CNAS and

CIPTS to estimate the parameters in the in equation (4.4).

Procedure 4.1 (**Imputation Estimation**):

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

$$\left(\alpha, \beta_q, \ldots, \beta_T, \boldsymbol{\kappa}, \boldsymbol{\eta}_q, \ldots, \boldsymbol{\eta}_T, \gamma_2, \ldots, \gamma_T, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_T\right)$$

by pooled QMLE in the LEF. The explanatory variables are

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

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

$$\hat{Y}_{igr}(\infty) \equiv G\left(\hat{\alpha} + \hat{\beta}_g + \mathbf{X}_i\hat{\boldsymbol{\kappa}} + \mathbf{X}_i\hat{\boldsymbol{\eta}}_g + \hat{\gamma}_r + \mathbf{X}_i\hat{\boldsymbol{\pi}}_r\right), r = g, \ldots, T \tag{4.10}$$

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

$$\begin{aligned}\hat{\tau}_{gr} &= N_g^{-1} \sum_{i=1}^{N} D_{ig}\left[Y_{ir} - \hat{Y}_{igr}(\infty)\right] \\ &= \bar{Y}_{gr} - N_g^{-1} \sum_{i=1}^{N} D_{ig} G\left(\hat{\alpha} + \hat{\beta}_g + \mathbf{X}_i\hat{\boldsymbol{\kappa}} + \mathbf{X}_i\hat{\boldsymbol{\eta}}_g + \hat{\gamma}_r + \mathbf{X}_i\hat{\boldsymbol{\pi}}_r\right). \square\end{aligned} \tag{4.11}$$

Procedure 4.1 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, $\hat{\boldsymbol{\tau}}$. The panel bootstrap – where units are resampled – is also valid and allows for any kind of serial dependence and model misspecification.

For robustness and simplicity, I recommend using pooled QMLE with a likelihood function in the LEF. By appropriately choosing the LLF, the pooled QMLE should also have satisfying efficiency properties. Nevertheless, the estimators from Procedure 4.1 are generally inefficient

– for a couple of reasons. First, the variances associated with the chosen LLF might differ from the actual conditional variances, $Var(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i,W_{it}=0)$. 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).

Rather than use an LLF in the LEF, one could use other models and estimation methods. 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. In equation (4.10), the function $G(\cdot)$ would be replaced with the mean function for the Tobit model; see, for example, Wooldridge (2010,Section 17.2).

As an alternative to the imputation approach, and to simplify calculation of standard errors, a pooled estimation method is convenient. The pooled estimation is nominally based on a conditional expectation for any time period t (and random draw i):

$$\begin{aligned} E(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i,\mathbf{W}_i) = G\Bigg[& \alpha + \sum_{g=q}^{T}\beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T}(D_{ig}\cdot\mathbf{X}_i)\boldsymbol{\eta}_g \\ & + \sum_{s=2}^{T}\gamma_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{X}_i)\boldsymbol{\pi}_s \\ & + \sum_{g=q}^{T}\sum_{s=g}^{T}\delta_{gs}(W_{it}\cdot D_{ig}\cdot fs_t) \\ & + \sum_{g=q}^{T}\sum_{s=g}^{T}(W_{it}\cdot D_{ig}\cdot fs_t\cdot\dot{\mathbf{X}}_{ig})\boldsymbol{\xi}_{gs}\Bigg] \end{aligned} \tag{4.12}$$

where

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

are the cohort-specific means of the covariates. The reason for centering the covariates in the quadruple interaction term is that it makes the δ_{gs} easier to interpret. 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. Alternatively, for each (i,t) we can define treatment indicators

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

and write the conditional mean as

$$\begin{aligned} E(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i,\mathbf{W}_i) = G\Bigg[& \alpha + \sum_{g=q}^{T} \beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T} (D_{ig} \cdot \mathbf{X}_i)\boldsymbol{\eta}_g \\ & + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_s \\ & + \sum_{g=q}^{T}\sum_{s=g}^{T} \delta_{gs} W_{itgs} \\ & + \sum_{g=q}^{T}\sum_{s=g}^{T} (W_{itgs} \cdot \dot{\mathbf{X}}_{ig})\boldsymbol{\xi}_{gs} \Bigg] \end{aligned} \tag{4.13}$$

In other words, one defines a different treatment indicator for each (g,s) pair with $s \in \{g,\ldots,G\}, g \in \{q,\ldots,T\}$.

If we only maintain Assumptions CNAS and CIPTS then mean function in (4.12) might be misspecified when $W_{it} = 1$ for some t. If we strengthen these assumptions – for example, CIPTS would apply to the potential outcomes $Y_t(g)$ as well as to $Y_t(\infty)$ – then we can identify the ATTs defined in (4.5) with pooled estimation across all observations.

Procedure 4.2 (**Pooled Estimation**):

1. Using all of the data, use pooled QMLE in the LEF to estimate

$$\begin{aligned}&\big(\alpha, \beta_q, \ldots, \beta_T, \boldsymbol{\kappa}, \boldsymbol{\eta}_q, \ldots, \boldsymbol{\eta}_T, \gamma_2, \ldots, \gamma_T, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_T,\\ &\delta_{qq}, \delta_{q,q+1}, \ldots, \delta_{qT}, \delta_{q+1,q+1}, \ldots, \delta_{q+1,T}, \ldots, \delta_{TT},\\ &\xi_{qq}, \xi_{q,q+1}, \ldots, \xi_{qT}, \xi_{q+1,q+1}, \ldots, \xi_{q+1,T}, \ldots, \xi_{TT}\big)\end{aligned}$$

The explanatory variables are

$$\begin{aligned}&1,\ D_{iq},\ \ldots,\ D_{iT},\ \mathbf{X}_i,\ D_{iq} \cdot \mathbf{X}_i,\ \ldots,\ D_{iT} \cdot \mathbf{X}_i,\\ &\quad f2_t,\ \ldots,\ fT_t,\ f2_t \cdot \mathbf{X}_i,\ \ldots,\ fT_t \cdot \mathbf{X}_i,\\ &\quad W_{it} \cdot D_{iq} \cdot fq_t,\ W_{it} \cdot D_{iq} \cdot f(q+1)_t,\ \ldots,\ W_{it} \cdot D_{iq} \cdot fT_t\\ &\quad W_{it} \cdot D_{i,q+1} \cdot f(q+1)_t,\ \ldots,\ W_{it} \cdot D_{i,q+1} \cdot f(q+1)_t,\ \ldots,\ W_{it} \cdot D_{iT} \cdot fT_t\\ &\quad W_{it} \cdot D_{iq} \cdot fq_t \cdot \dot{\mathbf{X}}_{iq},\ W_{it} \cdot D_{iq} \cdot f(q+1)_t \cdot \dot{\mathbf{X}}_{iq},\ \ldots,\ W_{it} \cdot D_{iq} \cdot fT_t \cdot \dot{\mathbf{X}}_{iq}\\ &\quad W_{it} \cdot D_{i,q+1} \cdot f(q+1)_t \cdot \dot{\mathbf{X}}_{i,q+1},\ \ldots,\ W_{it} \cdot D_{i,q+1} \cdot f(q+1)_t \cdot \dot{\mathbf{X}}_{i,q+1},\ \ldots,\ W_{it} \cdot D_{iT} \cdot fT_t \cdot \dot{\mathbf{X}}_{iT}\end{aligned} \tag{4.14}$$

where now $\dot{\mathbf{X}}_{ig} = \mathbf{X}_i - \bar{\mathbf{X}}_g$ are centered 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

$$\begin{aligned}\tilde{\tau}_{gr} = N_g^{-1} \sum_{i=1}^{N} D_{ig} \big[&G\big(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{X}_i\tilde{\boldsymbol{\kappa}} + \mathbf{X}_i\tilde{\boldsymbol{\eta}}_g + \hat{\gamma}_r + \mathbf{X}_i\tilde{\boldsymbol{\pi}}_r + \tilde{\delta}_{gr} + \dot{\mathbf{X}}_{ig}\boldsymbol{\xi}_{gr}\big)\\ &G\big(\tilde{\alpha} + \tilde{\beta}_g + \mathbf{X}_i\tilde{\boldsymbol{\kappa}} + \mathbf{X}_i\tilde{\boldsymbol{\eta}}_g + \hat{\gamma}_r + \mathbf{X}_i\tilde{\boldsymbol{\pi}}_r\big) \big].\end{aligned} \tag{4.15}$$

Alternatively, with the equation written as in (4.13), obtain the APEs of each indicator W_{itgr} and average across the subpopulation $W_{itgr} = 1$. □

This procedure has some benefits compared with Procedure 4.1. 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 $\hat{\delta}_{gr}$ are provided, and these are often of interest themselves. Generally, we can think of $\tilde{\delta}_{gr}$ as estimating

$$G^{-1}(E[Y_r(g)|D_g = 1]) - G^{-1}(E[Y_r(\infty)|D_g = 1]), \tag{4.16}$$

which is a treatment effect defined in terms of the linear index. We also obtain the $\tilde{\boldsymbol{\xi}}_{gr}$, which allows us to study whether and how the treatment effects vary with observed covariates. One can easily test whether such variables need to be interacted with the treatment indicators as allowed in the most general specification.

Another benefit of the pooled method is that one can, if desired, obtain overall average treatment effects for each treated period. Now, instead of setting a specific cohort dummy to unity (and setting the rest to zero), and 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 parallel trends assumption. In effect, we would need conditional PT to hold, with the same function $G(\cdot)$, for each $Y_t(g)$, $g \in \{q, q+1, \ldots, T\}$ 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. Obtaining analytical standard errors can be tricky in general, but the panel bootstrap can be applied to functions such as $(\tau_{gg} + \tau_{g,g+1} + \cdots + \tau_{gT})/(T - g + 1)$.

4.2. Equivalence Between Imputation and Pooled Estimation

For general choices of $G(\cdot)$, the $\tilde{\tau}_{gr}$ in (4.15) differ from the $\hat{\tau}_{gr}$ in (4.11), with the latter requiring fewer assumptions 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 linear exponential family, as discussed previously. I now formally state a result, which is proven in the appendix.

Proposition 4.1. (**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 in 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 shows that, for estimating the ATTs, we can use a pooled QMLE in the LEF without adding additional assumptions and consistently estimate the ATTs. In addition, we can, if desired, estimate ATEs for each period – as described earlier. It is also worth noting that including covariates **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 **X**, so that one could use the pooled estimator without the covariates. Including **X**, by effectively reducing the noise, can lead to more precise estimates of the τ_{gr}.

One might argue that the canonical link requirement is restrictive, but it actually serves a practical function. Namely, it limits the kinds of nonlinear models and estimation methods used by an empirical researcher. If a researcher goes beyond a linear model then, for the vast majority of cases, the mean function will be logistic or exponential. The method of estimation is set as the pooled QMLE using the paired log likelihood, as described in Table 3.1.

4.3. Examples

When $Y_{it} \in [0,1]$, including the binary case, a logistic function and estimation by pooled Bernoulli QMLE is most convenient because estimation and standard errors can be done using the pooled QMLE. The mean function is specified as

$$
\begin{aligned}
E(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i) = \Lambda\Bigg[& \alpha + \sum_{g=q}^{T} \beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T} (D_{ig} \cdot \mathbf{X}_i)\boldsymbol{\eta}_g \\
& + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{X}_i)\boldsymbol{\pi}_t \\
& + \sum_{g=q}^{T}\sum_{s=g}^{T} \delta_{gs}(W_{it} \cdot D_{ig} \cdot fs_t) \\
& + \sum_{g=q}^{T}\sum_{s=g}^{T} (W_{it} \cdot D_{ig} \cdot fs_t \cdot \dot{\mathbf{X}}_{ig})\boldsymbol{\xi}_{gs} \Bigg],
\end{aligned} \tag{4.17}
$$

where the centering of the covariates – in practice, replacing the means $E(\mathbf{X}|D_g = 1)$ with $\bar{\mathbf{X}}_g$ – ensures the estimates $\hat{\delta}_{gr}$ can be interpreted as the ATT on the log-odds ratio; see (4.16). 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 and then one sets the appropriate cohort/year dummy combinations to unit with the others set to zero.

If Y_{it} has a bound that can possibly change over i and t, say $Y_{it} \in [0, B_{it}]$, then the mean function in (4.17) 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.

When $Y_{it} \geq 0$ without a natural upper bound, the exponential mean makes sense as an alternative to a linear mean:

$$E(Y_{it}|D_{iq},\ldots,D_{iT},\mathbf{X}_i) = \exp\Bigg[\alpha + \sum_{g=q}^{T}\beta_g D_{ig} + \mathbf{X}_i\boldsymbol{\kappa} + \sum_{g=q}^{T}(D_{ig}\cdot\mathbf{X}_i)\boldsymbol{\eta}_g + \sum_{s=2}^{T}\gamma_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{X}_i)\boldsymbol{\pi}_t + \sum_{g=q}^{T}\sum_{s=g}^{T}\delta_{gs}(W_{it}\cdot D_{ig}\cdot fs_t) + \sum_{g=q}^{T}\sum_{s=g}^{T}(W_{it}\cdot D_{ig}\cdot fs_t\cdot\dot{\mathbf{X}}_{ig})\boldsymbol{\xi}_{gs}\Bigg], \tag{4.18}$$

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.

In the previous examples, and many others – including the linear case – one may want to impose restrictions on the parameters. A natural restriction would be to replace $W_{it}\cdot D_{ig}\cdot fs_t$ 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" indicators, say Z_{ith}, $h\in\{1,2,\ldots,T-q+1\}$, can greatly conserve on estimated parameters with many post-intervention periods. These also 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.

5. 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 parallel trends assumption. As a robustness check, it is helpful to combine nonlinear

models with simple models that allow heterogenous trends in the never treated state.

5.1. Testing the Parallel Trends 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 – a special case of the pooled QMLE approach in Section 4.1. Moreover, the tests are the same whether based only on the $W_{it} = 0$ observations or on pooled OLS using all observations – provided full flexibility is allowed in the treatment indicators, as in Section 4.1. The algebraic equivalence of the tests holds in the nonlinear case provided the canonical link function is used in the linear exponential family. 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 to 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. This is analogous 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 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 carried out a cluster-robust Wald statistic of joint significance of these pre-intervention indicators. This 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. This is in addition to the six dummies indicating legitimate treated observations. The event-study-type test is a cluster-robust joint test that the nine coefficients on the pre-intervention treatment indicators are all zero.

Another attractive 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 $D_{ig} \cdot t^2$, 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 trend test because the later 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 (4.12). Also, a full set of time dummies should be included, and interacted with the $\mathbf{X}_i$, if possible. This allows for an unrestricted aggregate trend that may also differ by observed heterogeneity, $\mathbf{X}_i$, in the never treated state, $Y_{it}(\infty)$.

5.2. Correcting for Violation of Parallel Trends

As discussed in Wooldridge (2021), the event-study-type test is inappropriate as a correction for pre-trends, as it would require that violation of parallel trends 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 the difference-in-difference-in-differences estimator of the singe ATT, τ_3.

Including heterogenous trends, even linear ones, 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, 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 bias but it can result in a severe loss of precision. Using a pre-test for heterogeneous trends is not ideal, but one also does not want to unnecessarily control for irrelevant variables that induce large standard errors.

As in the case without heterogeneous trends, the simplest analysis is obtained using pooled OLS, pooled logit (or fractional logit), and pooled Poisson estimation (with an exponential mean function). Then, the imputation and pooled estimators are identical, and the pooled method is very convenient for obtaining valid standard errors (at a minimum, clustering at the unit level). In equation (4.12), one simply adds the terms

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

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}, $D_g = 1, D_h = 0$ for $h \neq g, fr = 1, fs = 0$ for $s \neq r$, and $t = r$.

6. Simulations

I now report on Monte Carlo simulations when Y_{it} is binary or nonnegative with mass at zero. The simulations were performed in Stata 17 and the code is available upon request from

the author.

6.1. Binary Response with Common Timing

In this case the potential outcomes, $Y_t(0)$ and $Y_t(1)$, are binary. There are $T = 6$ time periods with a common intervention time at $q = 4$. In generating the data, the difference in the means of the underlying latent variables, $E[Y_t^*(1)] - E[Y_t^*(0)]$, changes across $t = 4$, 5, and 6. [Because of the nonlinearity of the model, the ATTs defined in terms of $Y_t(0)$ and $Y_t(1)$ would change over time even if $E[Y_t^*(1)] - E[Y_t^*(0)]$ were constant across t.] The single (time-constant) covariate, X, is generated as the average, over the six time periods, of independent exponential random variables with unit means. Treatment status, D, follows a logit:

$$D = 1[-0.5 + (X - 1) + V > 0], \tag{6.1}$$

where V is independent of X with a logistic distribution. This mechanism is used for all simulations and implies $P(D = 1) \approx 0.383$. For the first simulation, the potential outcomes are generated as

$$Y_t(0) = 1[(X - 1)/2 - 2D + (X - 1) \cdot D/4 + U_t(0) > 0], t = 1, \ldots, T, \tag{6.2}$$

where the $U_t(0)$ follow logistic distributions with a common effect C and independent idiosyncratic shocks. This mechanism ensures the potential outcome $Y_t(0)$ is correlated with X as well as with treatment status, D, and that $U_t(0)$ has substantial serial correlation. In the treated state, $Y_t(1) = Y_t(0)$ for $t \leq 3$ to ensure no anticipation. Then

$$Y_t(1) = 1[0.5 + (X - 1) - 2D + 0.2 \cdot f5_t + 0.3 \cdot f6_t + U_t(1) > 0], t = 4, \ldots, T, \tag{6.3}$$

where $U_t(1)$ depends on the same heterogeneity as $U_t(0)$ with separate independent shocks and also has a logistic distribution. This formulation implies that the index inside the logit

function has a time-varying treatment effect, although that effect does not vary with X. The treatment effects on the binary outcomes do vary with X because of the nonlinearity. The parallel trends assumption, conditional on X, is in force for the linear index.

The data generating mechanism implies $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.315$, $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.405$, and so there is reasonable balance in both potential outcomes between zero and one. The ATTs – the target parameters – are obtained by averaging across the 1,000 Monte Carlo replications. The logit conditional mean is estimated using pooled logit and the linear mean is estimated using pooled OLS. The R-squared from the linear POLS estimation is about 0.205, which seems realistic given that D, year dummies, and the covariate are controlled for flexibly.

The method proposed by Callaway and Sant'Anna (2021), where the comparison group is the never treated group (the default in Stata), is also included in the simulation . The two tests for pre-trends are computed for the pooled logit and pooled OLS estimation. The findings are reported in Table 6.1.

Table 6.1: **Binary Response**, **Common Timing**, **Logistic Errors**

1,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.061	0.060	0.022	0.062	0.037	0.061	0.042
τ_5	0.094	0.093	0.024	0.094	0.037	0.093	0.044
τ_6	0.113	0.113	0.026	0.113	0.038	0.112	0.044
Event Study Rejection Rate (2 df)	——	0.060		0.051		——	
Heterogeneous Trend Test (1 df)	——	0.053		0.042		——	

In this first scenario, the relative performance of the estimators is clear. Even with only 500 cross-sectional units, all three approaches show little bias when compared with the sample ATTs (which are averaged across the 1,000 simulations and should be close to the population

ATTs). In terms of precision, the correctly specified logit is much preferred, having Monte Carlo standard deviations no more than 69% of the linear model SDs. Not surprisingly, the CS (2021) estimators, which use only the single period prior to the intervention as the control period, are the least precise. Given that the logit model is correctly specified we hope that the parallel trends tests in the logit estimation have rejection frequencies close to the nominal size, 0.05, and that is the case. Because the linear model is misspecified, we might expect that the parallel trends tests should reject, but they also reject right around a 5% rate. In this scenario, that is not a bad thing because the linear model estimates show essentially no bias.

In other scenarios, the approaches based on the linear PT assumption can be badly biased. Table 6.2 reports on a simulation where the potential outcomes are generated as

$$Y_t(0) = 1[0.4 \cdot f4_t + 0.5 \cdot f5_t + 0.6 \cdot f6_t + (X-1)/2 - 2D + U_t(0) > 0], t = 1, \ldots, T \quad (6.4)$$

$$Y_t(1) = 1[0.9 \cdot f4_t + 1.1 \cdot f5_t + 1.3 \cdot f6_t + (X-1)/2 - 2D + U_t(1) > 0], t = 4, \ldots, T \quad (6.5)$$

The ATTs in this case are time varying and increasing. Again, the index satisfies the conditional PT assumption. In this case the average response probabilities are slightly larger than before: $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.384$ and $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.459$.

The simulation findings are striking. As it should, the pooled logit estimation is essentially unbiased. (Simulation results for the probit estimator show a slight downward bias, but well within the acceptable range.) Unfortunately, not only are the pooled OLS and CS (2021) estimates badly biased, they are, on average, actually negative, whereas the true ATTs are positive and practically large. To compound matters, the tests of parallel trends have no power for detecting the misspecification apparent in using linear methods. A topic for future research is to obtain diagnostics that would allow one to reject the misspecified linear model so that one is led to a nonlinear model.

Table 6.2: **Binary Response, Common Timing, Logistic Errors**							
1,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.079	0.078	0.028	–0.049	0.036	–0.049	0.043
τ_5	0.114	0.114	0.031	–0.037	0.038	–0.037	0.045
τ_6	0.159	0.158	0.038	–0.012	0.042	–0.013	0.047
Event Study Rejection Rate (2 df)		0.043		0.055			
Heterogeneous Trend Test (1 df)		0.054		0.042			

In a third scenario, the errors $U_t(0)$ and $U_t(1)$ are generated to have $Uniform(-2,2)$ distributions, where each depends on the common heterogeneity, C, as before. The potential outcomes are generated as in (6.3) and (6.4). In this specification, $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.319$ and $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.397$. When the support in the uniform distribution is sufficiently wide, the linear model is essentially correctly specified. Therefore, we might expect the the linear model to perform relatively well in this setting. The findings are reported in Table 6.3.

Table 6.3: **Binary Response, Common Timing, Uniform Errors**							
1,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.112	0.111	0.026	0.111	0.039	0.111	0.044
τ_5	0.159	0.159	0.029	0.159	0.041	0.159	0.047
τ_6	0.184	0.183	0.031	0.183	0.042	0.182	0.047
Event Study Rejection Rate (2 df)		0.038		0.046			
Heterogeneous Trend Test (1 df)		0.053		0.049			

Interestingly, while all methods show little bias, the logit estimator is substantially more efficient than the linear model or CS estimates. Evidently, the ramp function is well approximated by the logistic function, leading to better precision (and essentially no bias). As

before, the parallel trends tests reject right around 5% of the time even though, technically, both the linear and logit models are misspecified. And, again, this is a good outcome because the methods are doing very well for uncovering the ATTs.

In the final simulation for the binary response case, the potential outcomes are generated as in (6.4) and (6.5) with the errors having $Uniform(-2,2)$ distributions. The summary statistics in this simulation do not seem unusual: $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.374$, $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.449$, and the R-squared from the pooled LPM estimation is about 0.207. Because D is generated as before, $P(D = 1) \approx 0.383$. And yet, as shown in Table 6.4, the logit model now fares considerably worse than the linear model and CS approach, exhibiting a severe upward bias in each of the three ATTs. Both the linear regression and CS estimates have slight downward biases.

Table 6.4: **Binary Response**, **Common Timing**, **Uniform Errors**

1,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.126	0.199	0.032	0.112	0.042	0.111	0.047
τ_5	0.150	0.243	0.034	0.135	0.043	0.135	0.048
τ_6	0.175	0.287	0.035	0.159	0.043	0.158	0.048
Event Study Rejection Rate (2 df)		0.032		0.048			
Heterogeneous Trend Test (1 df)		0.054		0.050			

Simulations are necessarily special and not always realistic. For example, a uniform distribution for the composite error term, where the response probability can reach zero and one, does not seem particularly plausible. The purpose of the small study here is to show that one can obtain very different estimates depending on functional form. With at least one continuous covariate one can at least explore goodness-of-fit as a possible way of choosing among different conditional mean models. Unfortunately, the parallel trends tests do not

provide guidance in the simulation settings reported in Tables 6.2 and 6.4.

In some sense, the 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) when $T = 2$. In empirical practice, one can try a linear analysis along with a sensible nonlinear model, such as logit, and hopefully draw robust conclusions.

6.2. Count Outcome with Staggered Intervention

I now present simulations for nonnegative outcomes with mass at zero. In the first case, $Y_t(\infty)$ has a Poisson distribution conditional on unobserved heterogeneity. In particular,

$$Y_t^*(\infty) = 2 + 0.2 \cdot f2_t + 0.3 \cdot f3_t + 0.4 \cdot f4_t + 0.5 \cdot f5_t + 0.6 \cdot f6_t + X/5 - (D_4 + D_5 + D_6) + C$$

$$Y_t(\infty) \sim Poisson(Y_t^*(\infty))$$

where X is generated as before and $C \sim Normal(0,1)$. The distribution of $Y_t^*(\infty)$ conditional only on the cohort indicators is a mixture of a Poisson and lognormal random variable. The outcomes in the treated stages are generated as

$$\begin{aligned} Y_t(4) &\sim Poisson(Y_t^*(\infty) + (X-1)/5 + 0.4 \cdot f4_t + 0.8 \cdot f5_t + f6_t),\ t \geq 4 \\ Y_t(5) &\sim Poisson(Y_t^*(\infty) + (X-1)/5 + 0.6 \cdot f5_t + f6_t),\ t \geq 5 \\ Y_t(6) &\sim Poisson(Y_t^*(\infty) + (X-1)/5 + 0.4 \cdot f6_t),\ t = 6 \end{aligned}$$

The treatment cohorts are generated using an ordered probit, resulting in the following (approximate) shares:

$$\begin{aligned} P(D_\infty = 1) \approx 0.356,\ P(D_4 = 1) \approx 0.291 \\ P(D_5 = 1) \approx 0.226,\ P(D_6 = 1) \approx 0.127 \end{aligned}$$

The population R-squared in the pooled OLS estimation is about 0.126 and $P(Y_t(\infty) = 0) \approx 0.058$ (so zero is not a dominant outcome). The results are reported in Table

6.5.

Table 6.5: **Count Outcome, Staggered Intervention**							
1,000 Replications	Sample ATT	Exponential (Pooled Poisson)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_{44}	4.10	4.09	0.58	2.67	0.81	2.75	0.91
τ_{45}	11.15	11.14	1.38	8.24	1.78	8.36	1.75
τ_{46}	17.27	17.26	2.04	11.92	2.68	12.87	2.56
τ_{55}	8.18	8.18	1.38	5.80	1.75	6.72	1.65
τ_{56}	18.53	18.54	2.84	13.50	3.44	15.44	3.21
τ_{66}	6.34	6.33	1.36	1.59	2.24	4.62	2.08
Event Study Rejection Rate (9 df)	—	0.089		0.995		—	
Heterogeneous Trend Test (3 df)	—	0.064		0.999		—	

The clear-cut winner in this simulation is the exponential mean function (which is correctly specified) estimated using pooled Poisson regression (where the Poisson distribution is misspecified). The pooled QMLE is essentially unbiased whereas the pooled OLS estimators have severe downward biases. The CS estimators show less bias but are still very different, on average, from the sample ATTs. Moreover, the precision of the Poisson regression estimates is much better than either POLS or CS.

Compared with the binary case, the outcomes of the parallel trends tests are much more promising. The event-study-type test rejects in the linear case 99.5% of the time, and the heterogeneous trends test rejects in 999 out of 1,000 replications. Therefore, one would conclude 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 8.9% and 6.4% of the time, with the test using the three heterogeneous linear trends having particularly good size properties.

In the final simulation, I generate $Y_{it}(\infty)$ as a corner solution outcome: $P(Y_{it}(\infty) = 0) > 0$ with $Y_{it}(\infty)$ is continuous over strictly positive values. Unobserved heterogeneity is allowed by setting

$$Y_t^*(\infty) = 0.2 + X/5 - (D_4 + D_5 + D_6) + C,$$

where $C \sim Normal(0, 1)$. The corner solution outcome is generated with idiosyncratic variation over time as

$$Y_t(\infty) = R_t(\infty) \cdot \exp(Y_t^*(\infty)),$$

where the $R_t(\infty)$ are independent $Poisson(1)$ random variables. In other words, $Y_t(\infty)$ is the product of a Poisson random variable with unit mean and a random variable with a lognormal distribution conditional on X. The other cohort potential outcomes, with no anticipation imposed, are generated to have their own multiplicative idiosyncratic shocks as

$$\begin{aligned}
Y_t(4) &= R_t(4) \cdot \exp(Y_t^*(\infty) + (X - 1)/5 + 1.2 \cdot f4_t + 1.6f5_t + f6_t), \; t \geq 4 \\
Y_t(5) &= R_t(5) \cdot \exp(Y_t^*(\infty) + (X - 1)/5 + 1.2 \cdot f5_t + 1.8 \cdot f6_t), \; t \geq 5 \\
Y_t(6) &= R_t(6) \cdot \exp(Y_t^*(\infty) + (X - 1)/5 + f6_t), \; t = 6
\end{aligned}$$

where the $R_t(g)$ are mutually independent and distributed as $Poisson(1)$. The R-squared from the POLS regression is about 0.083. The proportion of zero outcomes in the population is fairly large, $P(Y_{it}(\infty) = 0) \approx 0.368$. The shares of the treatment cohorts are virtually the same as in Table 6.5. The simulation results are shown in Table 6.6.

Table 6.6: Corner Solution Outcome, Staggered Intervention							
1,000 Replications	Sample ATT	Exponential (Pooled Poisson)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_{44}	2.14	2.13	0.52	2.14	0.54	2.15	0.74
τ_{45}	3.62	3.61	0.77	3.61	0.80	3.61	0.91
τ_{46}	1.59	1.58	0.43	1.59	0.57	1.60	0.69
τ_{55}	2.28	2.28	0.63	2.28	0.68	2.26	0.99
τ_{56}	4.90	4.89	1.13	4.91	1.23	4.91	1.33
τ_{66}	1.87	1.83	0.75	1.87	0.97	1.90	1.53
Event Study Rejection Rate (9 df)	—	0.159		0.044		—	
Heterogeneous Trend Test (3 df)	—	0.090		0.046		—	

Interestingly, even though a zero outcome is much more likely compared with the simulation for count data in Table 6.5, the POLS estimation of the linear model [and CS (2021)] show no evidence of bias. (This may be partly due to the treatment effects being much smaller in magnitude.) The exponential mean is correctly specified so the absence of bias for the Poisson QMLE is not surprising. 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) SDs relative to the pooled QMLE are about 1.33 and 1.60, respectively. As before, CS (2021) is the least precise by some margin because it uses only the never treated group as the control group for each cohort and time period.

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: 15.9% for a nominal 5% test and the heterogeneous trend test rejects slightly too often. 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 suggest itself a priori as an attractive alternative or supplement to the linear model.

Other simulation scenarios suggest themselves. For example, one could generate the corner solutions to follow, say, Tobit models, or two-part models. Then, the exponential mean would also 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.

7. Empirical Application

I apply nonlinear DiD 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 centers. 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 through 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 through December as the post-intervention months. I take that approach here.

The outcome variable, *thefts*, is reported monthly, measured as the weekly average number of automobile thefts. More than 78% of the outcomes are zero, with increments of 0.25 up through the 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 estimate a linear model and an exponential model allowing for five different treatment effects by month.

Without covariates, the linear estimation is a standard DiD where *thefts* is averaged over the three pre-intervention months, and this is subtracted from the outcome in a particular treated month. This adjusted outcome is averaged across the treated and control blocks. As a comparison, I use an exponential model – which has a different underlying PT assumption – estimated by pooled Poisson regression. Also, the Callaway and Sant'Anna (2021) estimates are 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 through December as treated periods. In other words, the data on April and May are ignored. The estimates without covariates are given in Table 7.1.

Table 7.1: **No Covariates**

	(1)	(2)	(3)	(4)	(5)
	Single Effect (Linear)	Single Effect (Exponential)	Separate Effects (Linear)	Separate Effects (Exponential)	Separate Effects (CS, 2021)
τ	–0.087 (0.030)	–0.089 (0.033)	—	—	—
τ_8	—	—	–0.081 (0.041)	–0.084 (0.046)	–0.116 (0.056)
τ_9	—	—	–0.103 (0.030)	–0.103 (0.032)	–0.137 (0.048)
τ_{10}	—	—	–0.065 (0.047)	–0.067 (0.050)	–0.099 (0.062)
τ_{11}	—	—	–0.091 (0.031)	–0.092 (0.033)	–0.126 (0.049)
τ_{12}	—	—	–0.096 (0.031)	–0.098 (0.035)	–0.131 (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	—

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. The estimate 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 TE in each month. Again, there are no practical differences between the linear and exponential estimates, with the latter having slightly larger standard errors. The parallel trends test have large *p*-values, providing no evidence of the PT assumption in the linear or exponential means. The CS (2021) estimates are uniformly larger in magnitude and also less precise. Using only a single pre-treatment period when two more are available is difficult to justify, especially when there is no evidence against parallel trends.

Table 7.2 includes three binary covariates in additive form. These are binary indicators of whether the block houses a bank, public building, or gas 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 described in equation (3.30).

Table 7.2: **With Covariates**					
	(1)	(2)	(3)	(4)	(5)
	Single Effect (Linear)	Single Effect (Exponential)	Separate Effects (Linear)	Separate Effects (Exponential)	Separate Effects (CS, 2021)
τ	–0.087 (0.030)	–0.087 (0.035)	—	—	—
τ_8	—	—	–0.085 (0.042)	–0.090 (0.046)	–0.119 (0.056)
τ_9	—	—	–0.105 (0.030)	–0.104 (0.031)	–0.138 (0.047)
τ_{10}	—	—	–0.064 (0.047)	–0.066 (0.050)	–0.098 (0.062)
τ_{11}	—	—	–0.091 (0.031)	–0.095 (0.033)	–0.125 (0.049)
τ_{12}	—	—	–0.097 (0.032)	–0.104 (0.037)	–0.131 (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	—

Including the controls has very little impact on the estimates or standard errors for pooled estimation of the linear and exponential models. Thes estimators are more precise than the CS estimators, which show a similar pattern but are somewhat larger in magnitude. One advantage of the linear and exponential regressions is that one can see the coefficients on the interactions $W_{it} \cdot D_i \cdot fr_t \cdot \dot{\mathbf{X}}_{i1}$ to determine whether there are moderating effects. It turns out that the presence of a bank on a block drives the effects to essentially zero, a finding that may be explainable by the idea that car thieves may 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 time – the exponential model reaffirms the linear model estimates. In my view, this is not a bad thing. It demonstrates resiliency to the parallel trends assumption in that imposing it on two fairly different functional

forms leads to very similar estimates. The simulations in Section 6 demonstrate that in some cases one will get very similar estimates (on average) and, at least in principle, it is possible to obtain very different estimates.

8. Extensions

I now discuss some extensions that are straightforward to implement in the current framework.

8.1. All Units Eventually Treated

The discussion in the previous sections assumed the existence of a never treated group. This is not necessary, as the methods go through with little change if all units are treated by $t = T$. As discussed in Wooldridge (2021) in the linear without an never treated group, initially the ATTs are defined relative to the potential outcome $Y_t(T)$, which means the average gain $E[Y_t(g) - Y_t(T)|D_g = 1]$ for $g < T$ is is for being first treated in period g rather than 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 $g = T$. If we 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 $E[Y_t(T) - Y_t(\infty)|D_g = 1] = 0$ for $g < T$. Therefore, $E[Y_t(g) - Y_t(T)|D_g = 1] = E[Y_t(T) - Y_t(\infty)|D_g = 1]$ 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 identifying $E[Y_T(g) - Y_T(T)|D_g = 1]$.

As a practical matter, in Procedure 4.2 one drops any term involving D_{iT} because $D_{ig} + D_{i,g+1} + \cdots + D_{iT} = 1$; consequently, the terms involving D_{iT} are perfectly collinearity with other variables in the regression – as makes logical sense. In effect, $D_{iT} = 1$ acts as a

control group in each of the treated periods.

8.2. A Strategy with Exit

It is also possible to extend the pooled estimation methods to the case where the intervention turns off for at least 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 two periods, or only one period. The cohort first treated at $g = 5$ 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 $D_{g,h}$ 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. 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

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

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 $D_{g,h} \cdot fr_t$, $h = g, \ldots, T$, $r = g, \ldots, T$. 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 $W_{t,g,h,r} \equiv D_{g,h}fr_t$, $r = g, \ldots, T$, and then obtain the APEs with respect to each of these, averaging over the subsample corresponding to $W_{t,g,h,r} = 1$.

9. Concluding Remarks and Future Directions

I have proposed a simple yet flexible framework for estimating average treatment effects in staggered DiD settings when the (conditional) parallel trends 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.

One can use an imputation approach, based on first estimating the nonlinear model using the control observations, or a pooled method. In the leading cases mentioned above, the two methods are the same, with the pooled method being more convenient for obtaining standard errors and conducting inference.

I also 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 (with fewer degrees of freedom) that includes cohort-specific trends. I argue that the later

approach also leads to a sensible correction when PT is thought to be violated.

The simulations show that the nonlinear methods have essentially no bias when the conditional mean is correctly specified, and seem to work well under certain misspecifications. 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 $G(\cdot)$. 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. However, as discussed in Section 4, restricting oneself to canonical link functions based on the nature of Y_{it} is attractive.

As discussed in Section 8, it is essentially trivial to handle the situation where all units are treated by time period T. Moreover, with enough units in the treatment cohorts, one can handle staggered exit by defining a richer set of cohort dummy variables. How this method works when the cohort cells might be small remains to be seen – especially if used with nonlinear models.

Several additional directions suggest themselves. For example, sometimes the intervention has more than two levels (control and treatment). Allowing for multiple treatment levels seems relatively straightforward, provided one has enough data. Now one has to allow for multiple treatment levels as well as different treatment cohorts. Mechanically, it is easy to replace the binary indicator, W_{it} in (4.12), with a set of indicators for the different treatment levels. This would not be completely general because, in principle, one can define cohorts based on first period treated and first treatment level, but seems like a sensible way to allow cohort heterogeneity with multiple treatment levels. One could even replace the binary treatment with

a continuous treatment and estimate the average partial effects with respect to W_t across different treatment levels and averaging by cohort and time period. One word of caution: In the linear case, it seems unlikely that such strategies are the same as applying TWFE to an expanded equation – a desirable feature as discussed in Wooldridge (2021). I leave for the future an analysis of the precise treatment effects being recovered.

Another important practical extension is to replace the time constant covariates, $\mathbf{X}_i$, with time varying covariates, $\mathbf{X}_{it}$. To use time-varying covariates in the current setting, at a minimum they should not be influenced by the policy intervention. In panel data jargon, they should be "strictly exogenous." Then, the conditional expectation in (4.13) can be understood to be conditional on the entire history of the covariates, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT})$. The PT assumption is conditional on this history. In principle, with a large enough cross section, one could include the entire history in the conditional mean at time t – reminiscent of the Chamberlain (1980) device. It would be natural to conserve on degrees of freedom and replace $\mathbf{X}_i$ with $\mathbf{X}_{it}$, or maybe include the unit-specific time averages, $\bar{\mathbf{X}}_i = T^{-1} \sum_{s=1}^{T} \mathbf{X}_{is}$, as in the so-called Mundlak (1978) device. The τ_{gr} would be estimated as the APEs with respect to the indicators W_{itgr} averaged over the subpopulations $W_{itgr} = 1$.

Finally, the basic approach here should extend to repeated cross sections – as in Abadie (2005) and Callaway and Sant'Anna (2021) – but the details need to be worked out. For the identification derivations in Section 4 to extend to the case of repeated cross sections, the covariates $\mathbf{X}$, being time-constant, are assumed to be determined prior to the intervention – even though for some units the data are collected after the intervention. For example, in studying the effects of a staggered rollout of a job training program on older adults, where a new random sample is obtained each year, $\mathbf{X}$ would include variables that are determined prior

to the intervention – such as family background variables and highest grade achieved by age 25. Strictly speaking, a stationarity assumption would be imposed on the distribution of **X** so that sample in each time period are draw from the same distribution; Abadie (2005) and CS (2021) use assumptions that imply such a stationarity condition. In calculating the treatment effects, one would compute these as APEs with respect to $W_{gr} = 1$, where W_{gr} is a treatment indicator defined for being in treatment cohort *g* and time period *r*, averaged over the subsample with $W_{gr} = 1$.

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, 1491–1517.

Ai, C. and E.C. Norton (2003), "Interaction Terms in Logit and Probit Models," *Economics Letters* 80, 123–129.

Athey, S. and G.W. Imbens (2006), "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica* 74, 431–497.

Athey, S. and G.W. Imbens (2022), "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics* 226, 62-79.

Callaway, B. and P.H.C. Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics* 225, 200-230.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47, 225–238.

de Chaisemartin, C., and X. D'Haultfœuille (2020), "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review* 110, 2964–2996.

de Chaisemartin, C., and X. D'Haultfœuille (2022), "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey," forthcoming, *Econometrics Journal.*

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, 115-133.

Goodman-Bacon, A. (2021), "Difference-in-Differences with Variation in Treatment Timing," forthcoming, *Journal of Econometrics*.

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, 605-654.

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, 504-534.

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, 619-632.

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, 121-133.

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 *Advances in Econometrics*, Volume 39 (The Econometrics of Complex Survey Data). Kim Huynk, David Jacho-Chavez, and Gautam Tripathi (eds.), 109-135. Bingley, UK: Emerald Publishing, 2019.

Roth, J., and P.H.C. Sant'Anna (2021), "When is Parallel Trends Sensitive to Functional

Form?" working paper. https://arxiv.org/abs/2010.04814

Sant'Anna, P.H.C. and J. Zhao (2020), "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics* 219, 101-122.

Sun, L. and S. Abraham (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," forthcoming, *Journal of Econometrics* 225, 175–199.

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, 385-390.

Wooldridge, J.M. (2007), "Inverse Probability Weighted M-Estimation for General Missing Data Problems," *Journal of Econometrics* 141, 1281-1301.

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

Wooldridge, J.M. (2021), "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," working paper. https://www.researchgate.net/publication/353938385_Two-Way_Fixed_Effects_the_Two-Way_

Appendix A. Proof of Proposition 4.1

For this appendix, the notation is easier if 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 in Section 4 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 $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ be unique solutions to the equations

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

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

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

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

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

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}. \square \tag{a.4}$$

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

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

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

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 FOCs as

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

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

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

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

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

To use Proposition A.1 to proof Proposition 4.1, in the general setting with covariates, choose

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

and

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

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. 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 $\hat{\boldsymbol{\beta}}$, from the pooled estimation using all of the data, equals $\tilde{\boldsymbol{\beta}}$, the pooled QMLE from the $w_{it} = 0$ estimation.

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

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

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

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

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

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

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

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

or

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

Dividing by N_1 gives the result and Proposition 4.1 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 the 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 $d_{ig} \cdot fg_t, \ldots,$ $d_{ig} \cdot fT_t$, 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 $d_{ig} \cdot fg_t \cdot \dot{\mathbf{x}}_{ig}, \ldots, d_{ig} \cdot fT \cdot \dot{\mathbf{x}}_{ig}$.