NBER WORKING PAPER SERIES

A FLEXIBLE, HETEROGENEOUS TREATMENT EFFECTS DIFFERENCE-IN-DIFFERENCES ESTIMATOR FOR REPEATED CROSS-SECTIONS

Partha Deb
Edward C. Norton
Jeffrey M. Wooldridge
Jeffrey E. Zabel

Working Paper 33026
http://www.nber.org/papers/w33026

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024, Revised May 2025

The authors thank conference participants at the New York Camp Econometrics XVIII, at the American Society of Health Economists Annual Meeting, the 31st European Workshop on Econometrics and Health Economics, the 2024 Annual Health Econometrics Workshop, the Chinese Economists Society, the 2025 North American Winter Meeting of the Econometrics Society, and seminar participants at Oregon Health Sciences University and Universitat Pompeu Fabra. We also thank Anjelica Gangaram, Govert Bijwaard and Irene Botosaru for very helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Flexible, Heterogeneous Treatment Effects Difference-in-Differences Estimator for Repeated Cross-Sections

Partha Deb, Edward C. Norton, Jeffrey M. Wooldridge, and Jeffrey E. Zabel

NBER Working Paper No. 33026
October 2024, Revised May 2025
JEL No. C01, C21, C50

ABSTRACT

This paper proposes a method to estimate treatment effects in difference-in-differences designs in which the treatment start is staggered over time and treatment effects are heterogeneous by group, time, and covariates, and when the data are repeated cross-sections. We show that a linear-in-parameters regression specification with a sufficiently flexible functional form consisting of group-by-time treatment effects, two-way fixed effects, and interaction terms yields consistent estimates of heterogeneous treatment effects under general conditions. We also show that our method is identical to an imputation estimator. The estimates are efficient and aggregation of treatment effects and inference are straightforward. We call it FLEX, because it is a flexible linear model estimated by OLS with covariates (X). We illustrate the use of FLEX with an empirical example and provide comparisons to other recently derived estimators.

Partha Deb
Hunter College of the City University of New York
Department of Economics
and NBER
partha.deb@hunter.cuny.edu

Edward C. Norton
University of Michigan
School of Public Health
Department of Health Management and Policy,
and Department of Economics
and NBER
ecnorton@umich.edu

Jeffrey M. Wooldridge
Michigan State University
wooldri1@msu.edu

Jeffrey E. Zabel
Tufts University
jeff.zabel@tufts.edu

1 Introduction

The difference-in-differences (DID) study design is an important tool for causal inference in economics. In recent years, there has been an extraordinary number of new theoretical papers on how to obtain DID estimates in regression-based implementations of these study designs. In particular, the discovery that the two-way (group and time) fixed effects estimator of a model with a constant treatment effect can produce a biased estimate of the overall treatment effect when there is staggered treatment timing and heterogeneous treatment effects (e.g., de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021) has led to new approaches for dealing with both staggered timing and heterogeneous treatment effects (e.g., Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

What remains unknown is how the recent proposed methods to address heterogeneous treatment effects and staggered treatment timing work with repeated cross-sectional data that are common in applied research, as opposed to balanced panel data. The theoretical results showing consistent estimates of heterogeneous treatment effects with staggered treatment timing have been proven with panel data (Wooldridge, 2021; Borusyak et al., 2024), but not yet with repeated cross-sections. Proofs with repeated cross-sections are more complicated because covariates that are inherently time invariant are instead time varying when averaged at the group level (also true at the individual level) because of the changing composition of individuals over time. Furthermore, it is not possible to include only pre-treatment values of potentially endogenous time-varying variables, because these data are not observed for individuals who only appear once in the cross-sectional sample.

This paper addresses theoretical and empirical issues for difference-in-differences methods with repeated cross-sectional data with heterogeneous treatment effects and staggered treatment timing. We extend the methods proposed in Wooldridge (2021) to derive an estimator for repeated cross-sections under clearly stated assumptions that allow control variables to appear in a flexible way. We propose a linear-in-parameters specification with sufficient generality to be valid under general circumstances. Given a sample of repeated cross-sections, this specification can be estimated using ordinary least squares (OLS), which is consistent and efficient under the usual assumptions of stable unit treatment value, no bad controls, no

anticipation, and parallel trends. We show that this approach is identical to the imputation method in Borusyak et al. (2024) when both estimators are specified at the group-year level.

Our proposed DID method estimates a regression with heterogeneous treatment effects using OLS. It is simple and transparent. It is easy to know how identification is achieved, which treatment observations are compared to which control observations, and how many parameters are estimated. Because our method uses OLS regression, it is efficient in the class of linear estimators. Our method is flexible in how to incorporate covariates, which can be included additively or in a way that allows the treatment effects to vary with covariates. We call it FLEX, because it is a flexible linear model estimated by OLS with covariates (X).

In addition to the theoretical results, we demonstrate the use of FLEX with an empirical example of how an immigration enforcement program named Secure Communities may have affected enrollment in a government nutrition program (SNAP). Effects of this policy on SNAP have been substantively examined by East et al. (2023) and Alsan and Yang (2024). This policy-relevant example uses individual-level data from repeated cross-sections. The policy was implemented across counties in the US in a staggered way over several years. The policy may not have constant treatment effects at the group-time (county-year) level, so we allow for heterogeneous treatment effects. We compare the estimated average treatment effects on the treated (ATET) and their standard errors with those obtained using other popular methods (Cengiz et al., 2019; Callaway and Sant'Anna, 2021). The example demonstrates the features of our proposed FLEX estimator for repeated cross-sections with heterogeneous treatment effects, including being easy to estimate, transparent about the number of parameters estimated, allowing flexible controls for individual-level covariates, and having efficient standard errors. We also show how the results from our regressions can be displayed in a variety of commonly used graphical forms. The Stata code is available upon request.

2 Literature Review

2.1 Theoretical literature

Our paper is related to the extensive recent developments in econometrics about the estimation of treatment effects in difference-in-differences designs. De Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021) showed that difference-in-differences regressions that control for cohort and year fixed effects can identify treatment effects only when the treatment effect is constant over time. This common method, often called two-way fixed effects, uses observations in already treated groups as controls for other observations that start treatment later, in addition to using never-treated observations as controls. The overall average treatment effect is then a weighted average of comparisons that include earlier treated units to later treated units, instead of only treated to never-treated units plus not yet treated units.

Several other authors have proposed methods to avoid such unwanted comparisons. Callaway and Sant'Anna (2021) compare treated units in pre- and post-treatment periods to the period just prior to treatment to the same comparison for controls that can include the never- and not-yet treated units. Sun and Abraham (2021) use an event-study approach by exploiting both leads and lags. Borusyak et al. (2024) take a different approach and use what they call *imputation* to sweep out the effects of covariates and then use the residuals to estimate the treatment effects. Cengiz et al. (2019) propose a procedure that creates samples (stacks) of treated cohort observations pooled with never-treated observations, then combines those samples, which resolves the issue of unwanted comparisons. Cohort-level treatment effects are then estimated using a linear regression commonly referred to as a *stacked regression*. Wooldridge (2021) proposes an extended two-way fixed effects regression approach that includes interactions among dummy variables for treated cohorts, time periods, and covariates, allowing for estimation of heterogeneous treatment effects by cohort, time, and covariates. All these methods compare treated observations to not-yet-treated and never-treated observations, but never compare them to previously treated observations. There are recent summary papers by de Chaisemartin and D'Haultfœuille (2023), Roth et al. (2023), Freedman et al. (2023), and Baker et al. (2025).

2.2 Contribution

Our paper has several theoretical contributions. We propose FLEX—flexible linear model estimated by OLS with covariates (X)—for heterogeneous treatment effects in a difference-in-differences study design with repeated cross-sectional data and staggered treatment starting times. The linear-in-parameters specification can be estimated in a single linear regression. FLEX delivers consistent parameter estimates, allows for a flexible functional form with respect to covariates, and provides access to the OLS toolbox for inference and specification testing.

We prove that the FLEX treatment effect parameter estimates are consistent estimates of the group-time heterogeneous treatment effects that can be obtained in the repeated cross-section setting by an imputation method. To be precise, the FLEX estimates are identical to those from the Borusyak et al. (2024) imputation estimator when heterogeneity is specified at the cohort-year level. This imputation estimator extends both Borusyak et al. (2024) and Wooldridge (2021) to the group-time level for repeated cross-sectional data.

Given the simplicity of the underlying linear regression, aggregated treatment effects, such as the average treatment effect on the treated (ATET), can be derived easily from the estimated heterogeneous treatment effects. Inference on disaggregate and aggregate effects also follow naturally from the linear regression specification.

3 Theory

3.1 The Population Setting, Definitions, and Assumptions

Our goal is to set out a framework that focuses on the analysis of repeated cross sections where new units, i, belonging to one and only one group g, where $g = 1, 2, ..., G$, are (randomly) sampled from the population in every time period $t = 1, 2, ..., T$. Our framework also applies to panel data settings, most transparently when units i in groups g are followed over time t. In that case we are studying a stable population. When the data available are repeated cross sections, we assume the existence of a stable population from which units are randomly sampled in each time period. In defining parameters and stating assumptions, we assume that the populations are the same across t. But, in practice, there may be changes

in the population across t, a problem that can, at least partly, be overcome by controlling for observable characteristics.

We now turn to the policy setting. We assume that the intervention occurs at the group level indexed by $g = 1, 2, ..., G$. Each group, among those that receive treatment in the period of observation $t = 1, 2, ..., T$, receives the intervention for the first time in a time period from $t = 2, ..., T$ implying that there are no "always treated" groups. One or more groups ($< G$) receive the intervention for the first time in a particular time period. Let q denote the first time period in which treatment is received. We follow the existing literature by referring to the collection of groups associated with q as a cohort, $c = q$. Other subsets of untreated groups receive the intervention in subsequent time periods. Let $\bar{g}$ groups receive treatment by time period $Q \leq T$. Without loss of generality, let groups $g = 1, 2, ..., \bar{g}$ index the treated groups ordered by their cohort membership. Consequently, the never-treated groups can be indexed by $g = \bar{g} + 1, ..., G$. In what follows, it is convenient to refer to never-treated groups as being "treated" at ∞, i.e., $c = \infty$. Note that new entrant groups are not needed for each and every period. Also, note that groups, g, must map one-to-one to their cohort, i.e., the first time period in which the treatment is received.

Assumption 3.1 (SUTVA)**.** For each unit i in the population, the potential outcome depends only on unit i's assignment and not on the assignment of the other units. In other words,

$$Y_{it}\left(g_1, ..., g_{i-1}, g_i, g_{i+1}, ...\right) = Y_{it}\left(g_i\right), \, t = 1, ..., T,$$

where g_j is the group assignment for unit j. $\square$

The Stable Unit Treatment Value Assumption (SUTVA) [Rubin (1977,1986); Lechner (2011)] ensures that we observe one potential outcome for each unit i corresponding to i's treatment assignment. It also rules out spillover effects. Because we are assuming the treatment is an absorbing state, given SUTVA we need only index the potential outcomes by a single treatment assignment. Also, in stating Assumption 3.1, we allow for the possibility of an infinite or finite population; later, we assume independent sampling for each time period.

Let $R_g \in R_1, ..., R_{\bar{g}}, R_{\bar{g}+1}, ..., R_G$ denote a binary indicator for group membership. Let

$\{R_g\}^c$ denote the subset of groups in cohort c. The ATETs commonly of interest in staggered DID settings are the mean differences in the potential outcomes using $Y_t(\infty)$ as the reference outcome in a treated period t, i.e., the outcomes in the groups associated with the never-treated groups:

$$\tau_{gt} = E\left[Y_t(g) - Y_t(\infty) \,|R_g = 1\right], \; t = q, ..., T; \;\; g = 1, ..., \bar{g}. \tag{3.1}$$

For each (eventually) treated group g, τ_{gt}, $t = q, ..., T$ are the ATETs in all time periods including the first period in which treatment is received, q, and all following ones through the end of the observation period, T.

Assumption 3.2 (No Bad Controls, NBC)**.** For a $1 \times K$ vector of covariates $\{\mathbf{X}_t(g) : t = 1, ..., T; g = 1, ..., G\}$, which may be time-varying, the covariates are the same across all potential treatment assignments: $\mathbf{X}_t(g) = \mathbf{X}_t(\infty) = \mathbf{X}_t$ for all $g \in \{1, ..., G\}$ and $t = 1, ..., T$. □

Assumption 3.2 implies that the covariates vary exogeneous from the treatment status. Much of the literature on difference-in-differences in panel data contexts assumes the controls are dated prior to the first intervention date and do not change over time. This restriction helps ensure that one is not including 'bad controls' in the analysis—that is, elements in $\mathbf{X}_t$ that might be affected in the current or future periods by the intervention. Caetano et al. (2024) is an exception that formalizes the use of time-varying covariates, including to a situation we do not consider, i.e., to allow time-varying covariates to be affected by the treatment. We do not index the $\mathbf{X}_t$ using potential outcomes notation [such as $\mathbf{X}_t(g)$], and so we are maintaining that the covariates do not change with the treatment assignment. (This is different from saying that the treatment assignment cannot depend on $\mathbf{X}_t$—which, of course, we allow.) Allowing $\mathbf{X}_t$ to have time variation means that we can include predictors of the outcome whose paths are not influenced by treatment. For example, the outcome may be affected by local weather conditions, which are time-varying but not influenced by the treatment. More commonly, in the repeated cross-section case, the controls will necessarily be time varying—due to sampling variation across time periods.

Difference-in-differences analyses requires two additional key assumptions. The first assumption rules out anticipatory changes in the potential outcomes prior to the intervention

occuring for each eventually treated group. The second, parallel trends assumption, is stated conditional on group indicators that are less coarse than the cohort indicators. Consequently, because cohorts typically include more than one group and because the groups within a cohort may be heterogeneous, by conditioning on group indicators the assumptions are more general. Note that it is always possible to treat cohorts as "groups" in our framework.

Assumption 3.3 (Conditional No Anticipation, CNA)**.** For groups $g \in 1, 2, ..., \bar{g}$ and $t \in \{1, ..., c-1\}$,

$$E\left[Y_t(g)|R_1, ..., R_G, \mathbf{X}_t\right] = E\left[Y_t\left(\infty\right)|R_1, ..., R_G, \mathbf{X}_t\right]. \qquad \square$$

This simply means that, in any time period before the intervention occurs for any of the eventually treated groups, the potential outcomes are the same as the potential outcomes in the never treated state. This assumption can be violated if units within groups that are eventually treated anticipate the intervention and change their behavior prior to the first period of intervention. This formulation is adapted from Wooldridge (2021) (also in Wooldridge, 2023) which is stated for the panel data case.

Let $P_t \in P_1, P_2, ..., P_T$ denote binary indicators for observations in time periods $t = 1, 2, ..., T$. Then, in the staggered intervention case without exit, the time-varying treatment indicator is

$$W_t = \{R_g\}^q \cdot (P_q + ... + P_T) + \{R_g\}^{q+1} \cdot (P_{q+1} + ... + P_T) + ... + \{R_g\}^Q \cdot (P_Q + ... + P_T).$$

The observed outcome in every period is

$$Y_t = \{R_g\}^q \cdot Y_t\left(q\right) + \{R_g\}^{q+1} \cdot Y_t\left(q+1\right) + ...\{R_g\}^Q \cdot Y_t\left(Q\right) + \{R_g\}^\infty \cdot Y_t\left(\infty\right)$$

We state the parallel trends assumption assuming linearity of the conditional expectations and conditional on covariates and groups:

Assumption 3.4 (Conditional Parallel Trends, CPT)**.** For $t = 1, 2, ..., T$,

$$E\left[Y_t\left(\infty\right)|R_1, ..., R_G, \mathbf{X}_t\right] = \sum_{g=1}^{G} \beta_g R_g + \sum_{g=1}^{G} \left(R_g \cdot \mathbf{X}_t\right)\gamma_g + \mathbf{X}_t\pi_t + \eta_t. \qquad \square$$

Note that the Conditional Parallel Trends assumption in 3.4 also relaxes the analogous assumption in Callaway and Sant'Anna (2021) and other specifications where conditioning on cohort identity is replaced by conditioning on group identity, g.

Technically, we need not condition on the entire history of the covariates, $\{\mathbf{X}_t, t = 1, ..., T\}$ in assumption 3.4, and so the covariates need not satisfy a strict exogeneity assumption (see Wooldridge, 2010, Chapter 10). Nevertheless, if we think the treatment assignment influences the covariates in the future, Assumption CPT would generally fail. For a recent discussion of 'bad controls' in the DID setting, see Wooldridge (2024). Also, we require a sufficient number of observations per stratum in order to get precise estimates of the β_g and η_g in assumption 3.4.

Even if the covariates do not change over time, the terms $\mathbf{X}_t\pi_t$ and $(R_g \cdot \mathbf{X}_t)\,\gamma_g$ allow relaxation of the usual parallel trends assumption. Their inclusion plays the same role as in Callaway and Sant'Anna (2021), who apply standard treatment effects estimators when covariates are available; see also Wooldridge (2021, 2023). The presence of $(R_g \cdot \mathbf{X}_t)\,\gamma_g$ allows for substantial heterogeneity in how the average potential outcome changes with the groups (and therefore with the treatment groups).

Under assumptions 3.3 and 3.4, the parameters in 3.4 are identified using the untreated observations. In the subpopulation of untreated units at time t, we can write

$$E\left(Y_t | R_1, ..., R_G, \mathbf{X}_t, W_t = 0\right) = \sum_{g=1}^{G} \beta_g R_g + \sum_{g=1}^{G} \left(R_g \cdot \mathbf{X}_t\right)\gamma_g + \mathbf{X}_t\pi_t + \eta_t. \tag{3.2}$$

Equation 3.2 shows that all of the parameters are identified using the untreated observations, provided we have some units in every group.

The identification argument is easier to see in the simple 2×2 case, i.e., let $T = 2$ and $G = 2$. Let W denote the (only) treatment indicator and, to frame it like a typical 2×2 DID specification, let α denote the intercept for the value of the outcome for $G = 1$ and $T = 1$. We also remove group and time subscripts from all coefficients for simplicity. Then, equation 3.2 can be written as

$$E\left[Y_1\left(\infty\right)|W, \mathbf{X}_1\right] = \alpha + \beta W + \left(W \cdot \mathbf{X}_1\right)\gamma + \mathbf{X}_1\zeta \tag{3.3}$$

$$E\left[Y_2\left(\infty\right)|W, \mathbf{X}_2\right] = \alpha + \beta W + \left(W \cdot \mathbf{X}_2\right)\gamma + \eta_2 + \mathbf{X}_2\pi + \mathbf{X}_2\zeta \tag{3.4}$$

Under CNA, the expectation in equation 3.3 is the same when we replace $Y_1(\infty)$ with $Y_1(2)$. Because $W = 0$ implies $Y_1 = Y_1(\infty)$ and $W = 1$ implies $Y_1 = Y_1(2)$,

$$E(Y_1|W, \mathbf{X}_1) = \alpha + \beta W + (W \cdot \mathbf{X}_1)\gamma + \mathbf{X}_1\zeta,$$

which shows that, provided there are some treated and control units and $\mathbf{X}_1$ does not have perfectly collinear elements, the parameters α, β, γ, and ζ are identified using the control and (eventually) treated units in $t = 1$. By equation 3.4,

$$E(Y_2|W = 0, \mathbf{X}_2) = \alpha + \eta_2 + \mathbf{X}_2\pi + \mathbf{X}_2\zeta = (\alpha + \eta_2) + \mathbf{X}_2(\pi + \zeta)$$

Again, ruling out perfect collinearity in $\mathbf{X}_2$, $(\alpha + \eta_2)$ and $(\pi + \zeta)$ are identified by the second period control units. Because α and ζ are identified, so are η_2 and π. Returning to equation 3.4, we have

$$E[Y_2(\infty)|W = 1, \mathbf{X}_2] = (\alpha + \beta + \eta_2) + \mathbf{X}_2(\gamma + \pi + \zeta)$$

and so

$$E[Y_2(\infty)|W = 1] = (\alpha + \beta + \eta_2) + E(\mathbf{X}_2|W = 1)(\gamma + \pi + \zeta)$$

Identification of $E(\mathbf{X}_2|W = 1)$ follows immediately because we observe the second-period covariates for the all units, including the treated units. Because the other parameters are identified by assumptions 3.3 and 3.4, $E[Y_2(\infty)|W = 1]$ is identified. Then

$$\tau_2 = E[Y_2(2) - Y_2(\infty)|W = 1] = E(Y_2|W = 1) - E[Y_2(\infty)|W = 1]$$

is identified. The argument in the general staggered case is similar, with $E[Y_t(g)|W_g = 1] = E(Y_t|W_g = 1)$ always identified and $E[Y_t(\infty)|W_g = 1]$ identified under assumptions 3.3 and 3.4.

3.2 Estimation by Imputation and Ordinary Least Squares

The identification argument in subsection 3.1 immediately suggests an imputation approach to estimation. We assume the availability of random samples from the population at each time period t. The draws, indicated by i, are independent, but not generally identically distributed because the population distribution of both the outcome and control variables may change across t. (This same kind of heterogeneity is allowed in panel data settings.

Even if the population is stable across time, we allow for changing distributions across t.) For unit i, $t(i)$ represents the time period. The observed data for unit i is $Y_{i,t(i)}$, $\mathbf{X}_{i(t)}$, and the group indicators, $R_{i,t(i),j}$, $j = 1, ..., J$. As discussed earlier, the treatment assignment is determined by the group. Typically, the group for a unit does not change over time—an individual lives in the same state, say, over the time periods in question—but even if that is true, the repeated cross sections setting means that the draws of group indicators over time are from different samples of units.

Estimation of the parameters in equation 3.2 can proceed by estimating an OLS regression of the outcome on the untreated observations in the repeated cross-sectional dataset. Then, one mimics iterated expectations in the sample by using an out-of-sample prediction for $Y_t(\infty)$ (for the treated observations). Equivalently, the out-of-sample residuals are averaged over the appropriate group-time period pair to produce group-time specific ATT estimates. The following procedure extends Wooldridge (2021) to allow repeated cross sections and time-varying covariates. It is also related to the imputation method in Borusyak et al. (2024), who mention applying imputation in the repeated cross sections case. Here, we have derived an imputation estimator for repeated cross sections under clearly stated assumptions that allow random control variables to appear in a flexible way.

Procedure 3.1. [Imputation Estimation]

1. Use the control observations to estimate the parameters

$$(\beta_1, ..., \beta_G, \eta_1, ..., \eta_G, \gamma_2, ..., \gamma_T, \pi_2, ..., \pi_T) \tag{3.5}$$

by OLS:

$$\begin{aligned} Y_{i,t(i)} \text{ on } & R_{i,t(i),1}, ..., R_{i,t(i),G},\ P_{2,t(i)}, ..., P_{T,t(i)},\ \mathbf{X}_{i,t(i)} \\ & R_{i,t(i),1} \cdot \mathbf{X}_{i,t(i)}, ..., R_{i,t(i),G} \cdot \mathbf{X}_{i,t(i)},\ P_{2,t(i)} \cdot \mathbf{X}_{i,t(i)}, ..., P_{T,t(i)} \cdot \mathbf{X}_{i,t(i)} \end{aligned} \tag{3.6}$$

2. For unit i, impute $Y_{i,t(i)}(\infty)$ as

$$\begin{aligned} \hat{Y}_{i,t(i)}(\infty) = & \sum_{g=1}^{G} \hat{\beta}_g R_{i,t(i),g} + \sum_{g=1}^{G} \left(R_{i,t(i),g} \cdot \mathbf{X}_{i,t(i)}\right) \hat{\gamma}_g \\ & + \sum_{t=2}^{T} \hat{\eta}_t P_{t(i)} + \sum_{t=2}^{T} \left(P_{t(i)} \cdot \mathbf{X}_{i,t(i)}\right) \hat{\pi}_t \end{aligned} \tag{3.7}$$

3. For treatment group g, in period t, obtain

$$\begin{aligned}\hat{\tau}_{gt} &= N_{gt}^{-1} \sum_{i=1}^{N} R_{i,t(i),g} \cdot 1\left[t\left(i\right)=t\right] \cdot \left[Y_{i,t(i)}\left(g\right) - \hat{Y}_{i,t(i)}\left(\infty\right)\right] \\ &\equiv N_{gt}^{-1} \sum_{i=1}^{N} R_{i,t(i),g} \cdot 1\left[t\left(i\right)=t\right] \cdot \widehat{TE}_{i,t(i)} \\ &= \bar{Y}_{gt} - N_{gt}^{-1} \sum_{i=1}^{N} R_{i,t(i),c} \cdot 1\left[t\left(i\right)=t\right] \cdot \hat{Y}_{i,t(i)}\left(\infty\right) \end{aligned} \tag{3.8}$$

where

$$N_{gt}^{-1} = \sum_{i=1}^{N} Q_{i,t(i),c} \cdot 1\left[t\left(i\right)=t\right]$$

is the number of units in treatment group g in time period t, $\widehat{TE}_{i,t(i)} = Y_{i,t(i)} - \hat{Y}_{i,t(i)}\left(\infty\right)$ is the unit-specific estimated treatment effect, and

$$\bar{Y}_{gt} = N_{gt}^{-1} \sum_{i=1}^{N} Q_{i,t(i),g} \cdot 1\left[t\left(i\right)=t\right] \cdot Y_{i,t(i)} \tag{3.9}$$

is the average of the observed outcomes for units in treatment group g in period t. □

With a sufficient number of observations in each (g,t) cell, $\hat{\tau}_{gt}$, will have good statistical properties by the law of large numbers and central limit theorem. Nevertheless, the multi-step nature of the estimation makes inference challenging. A similar issue arises in the panel data setting in Borusyak et al. (2024), where unit-specific fixed effects are included in the first imputation step. Fortunately, there is an algebraic equivalence between imputation and a longer regression that uses all of the data. To describe the longer regression, let $\bar{\mathbf{X}}_{gt}$ be the average value of the covariates for treatment group g and time period t. Specifically, to the list of regressors in equation 3.6 we add treatment indicators and treatment indicators interacted with demeaned covariates:

$$R_{i,t(i),g} \cdot P_{i,t(i)},\ R_{i,t(i),g} \cdot P_{i,t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right),\ t = q, ..., T;\ \ g = 1, 2, ..., \bar{g} \tag{3.10}$$

If $R_{i,t(i),c} \cdot P_{i,t(i)} = 1$ then unit i is receiving treatment in time period t. The interactions $R_{i,t(i),c} \cdot P_{i,t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right)$ allow for heterogeneity in the treatment effects as a function of the observed covariates.

Proposition 3.1. [Equivalence of Imputation and OLS regression]

Using all of the data, consider the regression that includes all regressors in equations 3.6 and 3.10:

$$
\begin{aligned}
&Y_{i,t(i)} \text{ on } \mathbf{X}_{i,t(i)}, \\
&\qquad R_{i,t(i),1},\ ...,\ R_{i,t(i),G},\ R_{i,t(i),1} \cdot \mathbf{X}_{i,t(i)},\ ...,\ R_{i,t(i),G} \cdot \mathbf{X}_{i,t(i)}, \\
&\qquad P_{2,t(i)},\ ..., P_{T,t(i)},\ P_{2,t(i)} \cdot \mathbf{X}_{i,t(i)},\ ...,\ P_{T,t(i)} \cdot \mathbf{X}_{i,t(i)} \\
&\qquad R_{i,t(i),q} \cdot P_{q,t(i)},\ ...,\ R_{i,t(i),q} \cdot P_{Q,t(i)},\ ...,R_{i,t(i),Q} \cdot P_{Q,t(i)}, \\
&\qquad R_{i,t(i),q} \cdot P_{q,t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right),\ ...,\ R_{i,t(i),Q} \cdot P_{Q,t(i)} \cdot \left(\mathbf{X}_{i,t(i)} - \bar{\mathbf{X}}_{gt(i)}\right)
\end{aligned}
$$

Let the coefficients be $\left\{\tilde{\beta}_g : j = 1, ..., G\right\}$, $\{\tilde{\eta}_g : j = 1, ..., G\}$, $\{\tilde{\gamma}_t : t = 2, ..., T\}$, $\{\tilde{\pi}_r : t = 2, ..., T\}$, $\{\tilde{\tau}_{gt} : t = c, ..., T; g = 1, ..., \bar{g}\}$, and $\left\{\tilde{\delta}_{gt} : t = c, ..., T; g = 1, ..., \bar{g}\right\}$. Then

(i) For all g and t, $\tilde{\beta}_g = \hat{\beta}_g$, $\tilde{\eta}_g = \hat{\eta}_g$, $\tilde{\gamma}_t = \hat{\gamma}_t$, and $\tilde{\pi}_t = \hat{\pi}_t$.

(ii) For all $g \in \{1, 2, ..., \bar{g}\}$ and $t \in \{c, ..., T\}$,

$$\tilde{\tau}_{gt} = \hat{\tau}_{gt} \qquad \square$$

The equivalences in Proposition 3.1 are practically useful. Standard errors for the treatment effects $\tilde{\tau}_{gt} = \hat{\tau}_{gt}$ from the regression in equation 3.1 are readily available, and issues of clustering can be resolved in a standard pooled OLS setting. In addition to providing the $\hat{\tau}_{gt}$ and their standard errors, the $\tilde{\delta}_{gt}$ can be studied to determine if there are heterogeneous treatment effects.

As a special case of the regression in equation 3.1, one might have a single binary indicator, X_{it} (probably not time-varying), separating units into one of two groups. The scalar coefficients $\tilde{\delta}_{gt}$ would be the difference-in-difference-in-differences (DIDID) estimator for the group represented by $X_{it} = 1$. Again, inference is straightforward.

4 Regression Specifications

As the literature on methods for estimation of models for data with staggered entry into treatment has grown rapidly, the nomenclature used to describe various techniques has also proliferated. In naming methods, we think it is a) important to distinguish between the

estimands and the estimation methods, and b) to name the various estimators recently developed for estimating models in ways that indicate functional distinctions. In terms of the characteristics of the treatment effects parameters, some estimators assume homogeneous effects across groups and time. Others assume heterogeneity in one or the other dimension but not both. Yet others allow for heterogeneous effects across both cohorts and time but not across groups and time.

On the dimension of the parallel trends assumption, some estimators impose the parallel trends assumption for all periods prior to treatment, i.e., they assume that the regression specifications have only lagged treatment parameters. Other estimators impose the parallel trends assumption only for one baseline period prior to treatment (typically the period just preceding treatment) and the regression specifications include lag and lead treatment parameters with the lag parameters being the coefficients of interest for the ATET. Such models are often referred to as *event-study* models but that terminology corrupts definitions of event-study models from other areas of the econometrics literature.

We clarify our own use of language and notation by formally specifying the regression specifications we estimate using OLS to implement equation 3.1. In equation 4.1 below, we present a FLEX specification that explicitly displays all the regression parameters (and associated variables). The coefficients in the line denoted *lags* are the treatment effects at the group-time level in post-treatment periods. The coefficients in the line denoted *leads* are the pre-treatment differences between treated groups and the never-treated groups, except in one "baseline" period (chosen, without loss of generality, to be the period prior to treatment initiation, $q-1$). When the parameters in *leads* are estimated, we refer to these as *lags and leads* specifications. This specification is commonly referred to as an *event study* in the recent literature, (e.g. Roth, 2024), but we prefer the term *lags and leads* to disambiguate from an older, distinct use of the term "event study" in the econometrics literature (e.g. MacKinlay, 1997). In *lags only* specifications, all *lead* coefficients are set equal to zero. In other words, *lags only* specifications assume that all pre-treatment differences between treated cohorts and the never-treated cohort are identically equal to zero. These specifications include group and year fixed effects. These specifications also allow covariates to enter additively and interacted with each lag, lead, and fixed effect indicators.

$$
\begin{aligned}
&E\left(Y_{i,t(i)}|\{R_{ig}\},\{P_{i,t(i)}\},\boldsymbol{X}_{i,t(i)}\right) = \\
&\sum_{g=1}^{\bar{g}}\sum_{t=q}^{T}\tau_{gt}R_{ig}P_{i,t(i)} + \sum_{g=1}^{\bar{g}}\sum_{t=q}^{T}R_{ig}P_{i,t(i)}\cdot(\boldsymbol{X}_{i,t(i)}-\bar{\boldsymbol{X}}_g)\boldsymbol{\kappa}_{gt} && \textit{lags} \\
&+\sum_{g=1}^{\bar{g}}\sum_{t=1}^{q-2}\tau_{gt}R_{ig}P_{i,t(i)} + \sum_{g=1}^{\bar{g}}\sum_{t=1}^{q-2}R_{ig}P_{i,t(i)}\cdot(\boldsymbol{X}_{i,t(i)}-\bar{\boldsymbol{X}}_g)\boldsymbol{\kappa}_{gt} && \textit{leads} \\
&+\sum_{g=1}^{G}\beta_g R_{ig} + \sum_{t=2}^{T}\eta_t P_{i,t(i)} + \sum_{g=1}^{G}R_{ig}\cdot\boldsymbol{X}_{i,t(i)}\boldsymbol{\gamma}_g + \sum_{t=1}^{T}P_{i,t(i)}\cdot\boldsymbol{X}_{i,t(i)}\boldsymbol{\pi}_t + \boldsymbol{X}_{i,t(i)}\boldsymbol{\zeta} && (4.1)
\end{aligned}
$$

where $i = 1, ..., N$, $g = 1, ..., G$, $t = 1, ..., T$, and q denotes the first period in which treatment is received by observations in group g. Note that when covariates are interacted with treatment indicators, they are specified as deviations from group means. Deviating the covariates from means when they are interacted only with the group and only with the time indicators is not necessary, but doing say generally makes the coefficients on R_{ig} and $P_{i,t(i)}$ identical to ATETs.

Generally, the *lags and leads* specifications can be expected to be less efficient than *lags only*, because the latter uses all implications of the conditional parallel trends assumption. Moreover, *lags and leads* specifications may have more resiliency against some violations of parallel trends, but they will be more sensitive to other violations, e.g., they will be especially sensitive to violations of parallel trends that occur just before the intervention.

4.1 Aggregating the Treatment Effects

Rather than report a full set of treatment effect estimates for each treatment group and year, it is common to aggregate the effects to the cohort level, to the time level, to the exposure-time level (for staggered start), or most commonly to the aggregate level. One could do this by simple averaging. For example, the immediate effects, $\hat{\tau}_{gt}$, can be averaged over time periods $t = q$, ..., T. The one-period dynamic effects, $\hat{\tau}_{t,t+1}$, can be averaged over $t = q$, ..., $T-1$; and so on for each of the exposure lengths. To be precise, the aggregated average treatment effect on the treated (ATET) corresponding to equation 4.1 is given by:

$$
\hat{\tau} = \frac{1}{N}\sum_{g=1}^{\bar{g}}\sum_{t=q}^{T}\sum_{i=1}^{n_{gt}}\left[\hat{\tau}_{gt}\left(R_{ig}P_{i,t(i)}\right) + \left(R_{ig}P_{i,t(i)}\cdot(\boldsymbol{X}_{i,t(i)}-\bar{\boldsymbol{X}}_g)\right)\hat{\boldsymbol{\kappa}}_{gt}\right] \tag{4.2}
$$

where $i = 1, ..., N$, $g = 1, ..., G$, $t = 1, ..., T$, and q denotes the first period in which treatment is received by observations in group g. Note that n_{gt} is the number of observations in the g^{th} treated group in the t^{th} time period and $N = \sum_{g=1}^{\bar{g}} \sum_{t=q}^{T} n_{gt}$. This simplifies to

$$\hat{\tau} = \frac{1}{N} \sum_{g=1}^{\bar{g}} \sum_{t=q}^{T} [n_{gt} \hat{\tau}_{gt}] \tag{4.3}$$

because $\bar{x}_{i,t(i)}$ is de-meaned so that part of the equation sums to zero.

Note that $\hat{\tau}$ is simply the average of the treatment effects over all treated observations in all treated time periods. In specifications with no covariates, or covariates entered only additively, or if the covariates are de-meaned (using the means of covariates for the treated sample), then $\hat{\tau}$ is also a weighted average of the group-time effects $\hat{\tau}_{gt}$, where each weight is the sample size associated with each group and year, n_{gt}. Standard errors for the unweighted or weight averages are easy to obtain using standard software packages.

4.2 Other Approaches

As we mentioned previously, the two-step imputation method of Borusyak et al. (2024) using the group dummies and the OLS regression on the group dummies using all of the data produce numerically identical results. Moreover, if we include the pre-treatment indicators then, without controls, we obtain a repeated cross-sections version of the Sun and Abraham (2021) leads and lags event-study estimator that allows full heterogeneity by group and calendar time.

There are some other approaches that have become popular in empirical research. Callaway and Sant'Anna (2021) propose what are effectively leads and lag estimators because their methods reduce to estimating many 2×2 DiDs using the never treated group as the control group and the period just prior to the intervention as the control period. Without covariates, the Callaway and Sant'Anna (2021) approach is identical to Sun and Abraham (2021). With covariates, Callaway and Sant'Anna (2021) implement a regression adjustment approach to estimation of the treatment effects.

A novel aspect of Callaway and Sant'Anna (2021) is that, with covariates, they allow estimation methods other than linear regression adjustment. One estimator with nice robustness properties combines propensity score weighting with linear regression adjustment

(a version of IPWRA). Without the propensity score and without covariates, the Callaway and Sant'Anna (2021) approach is the same as FLEX without covariates. When covariates are introduced, Callaway and Sant'Anna (2021) approach is equivalent to a fully saturated FLEX specification. Note that the IPWRA estimator can have some resiliency to the assumed linear functional form.

Another popular approach uses a stacked difference-in-differences procedure (Stacked DID) described in Cengiz et al. (2019) and Wing et al. (2024). In each, the observations for each treated cohort are pooled with the observations from the never-treated cohort to form a data set that are then appended to form one stack. Note that this includes separate time and cohort indicators for each cohort. By doing so, a constant treatment effect difference-in-differences regression specification with group and time fixed effects delivers unbiased estimates of the cohort-level treatment effect. The issues reported by Goodman-Bacon (2021) and others do not apply. Then each of these stacks is pooled and a regression that interacts each term in the standard regression specification with stack indicators produces all cohort-by-time treatment effects in one application of a regression procedure. The same never-treated observations are used multiple times (as many times as there are treated cohorts). If there are many untreated groups, then the sample size for Stacked DID can be very large which can substantially increase the time to estimate the parameters.

4.3 Heterogeneous Time Trends

Like the estimates obtained from procedure 3.1, the event-study estimates require a kind of parallel trends assumption for consistency (Roth, 2022; Dette and Schumann, 2024). One approach to accounting for violations of parallel trends that occur even after including controls is to assume relatively simple heterogeneous trends in the absence of the intervention. In particular, with at least two pre-intervention periods per treated group, one can include in equation 3.1 interactions between the group indicators, $R_{i,t(i),j}$ and a linear time trend, t. The coefficients on the trend terms can be used to test for pre-trends (Dette and Schumann, 2024). As shown in Wooldridge (2021) in the panel data case, such tests do not suffer from 'contamination bias', provided the covariates are included flexibly, as in equation 3.1. That is because the imputation and OLS approaches continue to be identical.

Including group-specific trends can be costly in terms of precision because their inclusion creates collinearity with the treatment indicators. Of course, using a pre-test to decide whether to drop these terms can be problematic—just as when using the event-study approach.

4.4 Practical Considerations

In this section we discuss several practical considerations when estimating difference-in-differences models for data with repeated cross-sections, heterogeneous treatment effects, and staggered timing. Estimating treatment effects in a difference-in-differences model is usually a two-step process. The first step is to estimate the treatment parameters of a model and the second step is to aggregate some of those treatment parameters into estimated average treatment effects. For example, estimating the average treatment effect on the treated, or treatment effects by group or by time since the start of treatment. Some DID software packages seamlessly combine these two steps into one, showing only a final average treatment effect on the treated but not the intermediate step of estimating coefficients from a regression model. Thinking about DID as a two-step process is useful because it clarifies that there are two steps to decision making by the researcher, each with its own set of questions.

The set of questions for the first step revolves around identification of treatment effects and model specification. How are the treatment effects identified? What level of heterogeneity is allowed, specifically, heterogeneity at the group or the cohort level? Should the model assume that parallel trends holds in the pre period or allow for heterogeneous effects across groups in each time period before treatment starts? To what extent should the model specification control for covariates?

The set of questions for the second step includes answering the research question, presenting the results, and describing the extent of the heterogeneity in the treatment effects. After estimating the regression model, how should the estimated treatment coefficients be combined to form an overall ATET? Should the treatment effects be aggregated to other levels, for example, to show an event-study graph?

4.4.1 Estimation step

Our proposed FLEX approach also provides a useful modeling framework for all difference-in-differences models. Our general approach to estimation is that OLS can be used to estimate any flexible model, whether it is a flexible model with heterogeneous treatment effects or a parsimonious model with homogeneous treatment effects. Treatment effects can be heterogeneous with respect to time, group (or cohort), or both, and there can additionally be heterogeneous treatment effects by covariates. Most existing methods can be estimated with an OLS regression in the FLEX framework.

In our empirical examples, we explain when FLEX estimates the same treatment coefficients as other estimators.

4.4.2 Basic modeling decisions

There are three main modeling decisions, which can be seen as whether to allow more or less flexibility in the estimation step. They can also be seen as variations on equations 4.1. The first decision is whether to allow heterogeneity at the group level or the cohort level, when there are multiple groups per cohort. In our empirical examples, there are multiple states (groups) that start the policy treatment in some years and so are in the same cohort. In general we think that it is restrictive to impose homogeneity across all states that happen to start their policy in the same year. Therefore, our model specifications are at the group level. However, if the number of groups is large (e.g., $3,000$ counties), then one needs to think harder about the tradeoffs between model flexibility and possible over-fitting.

A second decision is whether to model lags only or both lags and leads. One might think that the event-study approach, which allows for different effects for treatment observations in the pre-treatment periods that follow a trend that is not parallel to that for the control observations, is more flexible than the leads only model because it does not impose the conditional parallel trends assumptions in the pre-treatment periods. But if parallel trends is violated in the treated periods, then the OLS estimator applied to the leads and lags model can be worse than the OLS estimator applied to the lags only model as it will have different biases for different violations of parallel trends.

The third is about how to include covariates. Covariates can be included separately as

additional controls or interacted with treatment effects. Group fixed effects sweep away any time-invariant covariates. However, unlike covariates in a balanced panel, covariates in a repeated cross section can change either because they are time-varying or because the random subsample has a different composition of covariates than the full sample. Therefore, one can control for covariates in a cross-section even when the covariates are inherently time-invariant. Endogenous covariates should not be included in the model. For example, controlling for experience when predicting income is not appropriate if the model is measuring the effect of a job training program on income and that job training program endogenously changes experience.

Interacting covariates with treatment effects allows for further treatment heterogeneity. This is a decision about whether to include the third and fourth rows of equation 4.1 or to assume that some of those coefficients are zero. The potential downside of this flexibility includes over-fitting. Alternatively, one can lean on economic theory and knowledge of institutions to focus on just one or a few covariates to interact with the treatment effects instead of all covariates. In our experience, including interactions between treatment effects and covariates does not lead to larger standard errors.

Finally, there is a practical issue about whether to demean covariates or not. In the flexible specifications of the regressions, group-by-year treatment effects vary by individual- and group-specific covariates. This is accomplished by interacting the group-by-year indicator variables with the covariates. To interpret the coefficients on the group-by-year indicators as the treatment effects in such models, it is necessary to demean the covariates in the interaction terms. Demeaning requires subtracting the group-by-year means of each covariate from its individual-level values. In these flexible specifications, the group and year fixed effects are also interacted with the covariates. But for these indicators, the interactions are with the raw covariates, not the demeaned ones. In addition, the raw covariates themselves are entered into the regression specification additively. We want to emphasize that it is not necessary to de-mean the covariates to estimate the ATET. As long as one is careful in combining the weighted average of the coefficients for the demeaned variables, the final ATETs are identical.

4.4.3 Aggregation and graphing

After estimating treatment effects at the group-time level, researchers have several options for how to present the results. One common approach is to aggregate to the time-since-treatment level to create an event study plot. This is useful when estimating a lags and leads model because one can plot the estimated difference between ever-treated groups and control groups in each pre period. This provides a visual test of the parallel trends and no anticipation assumptions. If one suspects or wants to test for heterogeneity by calendar time or by group, one could aggregate treatment effects to those levels and plot the results.

Researchers often want an overall ATET. This is easy to do in the FLEX approach. The ATET is the weighted average of treatment effects, where the weights are the number of treated observations in each group-time set in the post periods. We provide Stata code to show how to do this after estimating the single regression model, including standard errors.

4.4.4 Collapsing the data to the group level

When there are many individual observations for each group (or cohort), it is worth considering whether to collapse the data to the group (or cohort) level. The advantage is that collapsed data run faster in weighted OLS while returning identical parameter estimates and standard errors. If the data set is large, collapsed data could run considerably faster than the original data set. However, if there are covariates, then collapsing the data may not be feasible. Another case where one might consider collapsing the data is when there are a large number of groups (e.g., all U.S. counties) that require a very large number of parameters to estimate.

The identity (identical covariates from WOLS with collapsed data to covariates from OLS with non-collapsed data) only holds if you collapse not to the group-time level, but for every possible combination of covariates for each group-time combination. Therefore, the covariates must be discrete. For example, if there are 50 groups and 10 years of data, then there are 500 group-time combinations. Now suppose there are three dummy variables, which can be combined eight ways, that means that the data would be collapsed to 4000 possible group-time-covariate combinations. That could still be significantly smaller than the original data set and faster to run, but the potential benefits of collapsing are small when

there are many covariates. When there are continuous covariates, one can still collapse the data by specifying discrete versions of these variables.

4.4.5 Computational time

Because FLEX uses a single least-squares regression, its computational time is fast. It does not require multiple regressions (e.g., the imputation method), numerous pairwise comparisons (e.g., Callaway and Sant'Anna (2021)), or replicating the data (e.g., stacked). The computational time does increase, in proportion to the number of parameters, in models with many covariates and interactions between covariates and treatment effects, but even then we have found the computational time to be reasonable.

5 Empirical example

We show an empirical example using a data set that has features typical of repeated cross-section data with a difference-in-differences study design. We test the effect of the Secure Communities immigration enforcement program on Supplemental Nutrition Assistance Program (SNAP) enrollment. The idea is that when immigration enforcement increases in a community, some people will be more reluctant to sign up for a government program, even if they are U.S. citizens and eligible for the benefits (Alsan and Yang, 2024). The data are collected at the individual level. Treatment happens at the county level and the timing of the start of treatment is staggered over several years. We measure the outcome and covariates using a number of years of the American Community Survey, which is not a panel data set but instead a repeated cross section. The example is typical of many difference-in-differences study designs and is appropriate to illustrate our theoretical results.

5.1 Prior related empirical literature

There is a large and rapidly expanding empirical literature that uses difference-in-differences study design to assess the effects of policy changes. Here we briefly mention two papers that are most relevant to our empirical example (East et al., 2023; Alsan and Yang, 2024). Both papers study the effect of the Secure Communities immigration enforcement policy using repeated cross sectional data with staggered treatment timing.

East et al. (2023) measure the effects of the Secure Communities immigration enforcement on labor market outcomes. They exploited variation in the timing of exposure to Secure Communities over commuting zones to measure its effects on employment and hourly wages. They estimated a difference-in-differences regression with a homogeneous treatment effect, the traditional two-way fixed effects model, using data from the American Community Survey. East and colleagues limited their analysis to working-age populations who were most likely to be affected by Secure Communities—low-educated foreign born persons and U.S.-born individuals. They found that Secure Communities immigration enforcement decreased the employment of people who were likely to be undocumented immigrants, an expected finding. However, they also found that U.S.-born persons had slightly lower employment and hourly wages as a result of Secure Communities, an unexpected finding, perhaps due to a decline in labor demand.

Alsan and Yang (2024) explore a different policy question related to Secure Community immigration enforcement. They ask whether noncitizen deportations affect co-ethnic citizen participation in means-tested social insurance programs, even if those citizens are not personally at risk of deportation. Persons who decide not to enroll in those benefit programs if they are concerned that providing information to the government could result in deportation of relatives or persons in their network. This is one example of a chilling effect of deportation policies. Alsan and Yang look at the effects of Secure Community on two means-tested social insurance programs: Supplemental Security Income and the Supplemental Nutrition Assistance Program (SNAP). They use data from both the American Community Survey and the Panel Survey of Income Dynamics. Their treatment is measured at the county-year level. In addition, instead of a difference-in-differences study design, they use a triple difference by comparing the effect of Secure Community for Hispanics compared to black and white non-Hispanics. They allow for heterogeneous treatment effects by racial/ethnic group and by year since start of treatment, but not across counties. They found evidence for a chilling effect on Hispanic-headed households.

Our example borrows elements from both East et al. (2023) and Alsan and Yang (2024). We also use the American Community Survey to study the effect of Secure Community immigration enforcement, which has a staggered implementation across US counties and

data from a repeated cross section. Like Alsan and Yang (2024) we examine the effect on SNAP enrollment and like East et al. (2023) we use a difference-in-differences study design.

5.2 Secure Communities Immigration Enforcement and ACS data

We use repeated cross-sectional data from the American Community Survey (ACS) to show how to use heterogeneous effects difference-in-differences methods to estimate the effect of Secure Communities immigration enforcement on SNAP enrollment. The U.S. Census Bureau conducts the ACS monthly to collect information on individuals in communities, including information on employment, education, and enrollment in government benefit programs. The ACS publicly reports the location of respondents using Public Use Microdata Areas (PUMAs). PUMAs are non-overlapping, statistical geographic areas with at least 100,000 people. PUMA boundaries respect state boundaries but are often collections of counties. There are 2,487 PUMAs, somewhat fewer than the 3,144 counties in the US. When a PUMA consists of multiple counties, we assign the earliest data of implementation of the program in the constituent counties to the PUMA.

We use the data archive of Alsan and Yang (2022) to construct our analysis sample, including a crosswalk between counties (the geographic unit at which the treatment was assigned) and PUMAs (the geographic unit available in the ACS). Following Alsan and Yang (2024), we limit our sample to Hispanic citizens (including naturalized citizens) who did not move between states and who were either heads of households or their spouses. We also drop observations from Illinois, Massachusetts and New York as these states resisted the activation of Secure Communities. Unlike Alsan and Yang who restricted their sample to those with less than a high school education, we include those with a high school diploma (but not those with additional education beyond high school). We also exclude PUMAs that have fewer than 10 observations in any year 2005-2016. Few observations make it hard to estimate treatment effects at the group-time level, especially for models that include interactions with covariates. For purposes of illustration, although not required for estimating FLEX, we further restrict our sample to individuals who live in 682 PUMAs observed in each sample year. Our sample includes observations for more than half a million people who live in 682 PUMAs across 32 states during the years 2005-2016. Therefore, this is an individual-level

data set of repeated cross sections of all 682 PUMAs for all 12 years.

Out of the 682 PUMAs in our sample, 259 began Secure Communities immigration enforcement during our time period while 423 did not and are the never-treated controls (see Table 1). During the years 2005-2008, no PUMAs in our data set started Secure Communities immigration enforcement. Beginning in 2009, at least 50 PUMAs in our data set started Secure Communities immigration enforcement each year. The Secure Communities immigration enforcement began in the South and West Census regions. Treatment did not start in the Northeast region until 2012. See Table 1 for the number of PUMAs that were treated in the years 2009-2012.

A cohort comprises all PUMAs that implemented the Secure Communities immigration enforcement program in a calendar year. Each of the four treated cohorts is observed for at least four periods before the start of treatment and for at least five years after the start of treatment. In addition, one cohort consists of never-treated PUMAs.

We define a group as being a subset of PUMAs in a cohort that belong to the same Census region. Table 1 shows that there are 11 treated groups along with 4 never-treated groups (one for each Census region of the US).

There are a total of 581,660 observations on individuals. The ACS has an average of 850 observations per PUMA per year. The majority of observations are in the West and the fewest are in the Midwest (see Table 2). Around 20 percent of the sample was enrolled in SNAP (see Table 3). Slightly more than half the sample is female, the average age is 50, the income to poverty ratio is around 240, and about 60 percent have either a high-school diploma or a GED.

5.3 Model specifications

We estimate several alternative FLEX specifications based on the general model specification in equation 4.1. Our preferred FLEX specifications, which are numerically equivalent to imputation estimators, estimate heterogeneous treatment effects over either groups and event time or over cohorts and event time. When multiple PUMAs begin Secure Community immigration enforcement in the same calendar year, those PUMAs are in the same cohort. Because this empirical example has more than one group per cohort, it likely matters whether

to allow heterogeneous treatment effects at the group level or only at the cohort level, so we estimate both model specifications.

The *lags-only* models estimate treatment effects only in the periods after the start of the policies, i.e., these specifications use indicators for treatment lags only. This means that there are no separate coefficients for treatment and control groups in each pre-treatment period. In *lags and leads* models, we allow all periods—except one reference baseline period, which we specify as the period preceding initiation of treatment—to have treatment indicators for each ever-treated group or cohort. There are cohort-by-time or group-by-time coefficients in the pre-treatment periods as well as the post-treatment periods. In the *lags and leads* models, the treatment effects are compared to the year prior to the start of treatment; in the *lags-only* models, the treatment effects are compared to the average of all the years prior to the start of treatment. Note that one obtains the *lags-only* model by imposing the restriction that all the pre-treatment effects equal zero.

We made different choices about how to include covariates for individual-level characteristics. As a reminder, because the data are repeated cross-sections, although the sample populations are reasonably stable, the exact mean of individual-level variables for any PUMA (or equivalently for each group and cohort) will change from year to year. We estimate specifications that exclude covariates or include covariates in a flexible way by interacting all covariates with each lead and lag treatment coefficient, each group (or cohort) fixed effect, and each time fixed effect.

We show the results for the empirical example in tabular (Table 4) and graphical form (Figure 1). The table of results lists the average treatment effect on the treated (ATET) for a variety of model specifications. The top panel of each table shows results for *lags-only* models, while the lower panel of each table shows results for *lags and leads* models. We compare our FLEX model specifications to several other model specifications, which differ in the estimand, estimator, sample, model specification, and whether and how covariates are included.

The estimand for treatment effects varies widely across the models. The simplest assumes that the treatment effects are constant (homogeneous), both across groups and over time. For the *lags-only* models, this specification is commonly referred to as the homogeneous

two-way fixed effects (TWFE) regression. It assumes a one-time homogeneous shift in the outcome due to treatment. Note that the use of TWFE to refer to the homogeneous effects specification is misleading because it does not distinguish among a variety of estimands, each of which involve regression specifications that include fixed effects along two dimensions, all of which are more general than the homogeneous effect specification. For the *lags and leads* models, the simplest version assumes a separate effect for each point in time since the event, as described in Sun and Abraham (2021), but is homogeneous across groups (see the rows labeled *Event time ES* for the treatment heterogeneity in the lags and leads models).

We also estimate treatment effects using two popular alternative techniques, each of which has a way to resolve the issues arising from staggered entry into treatment. One *lags and leads* model uses the method of Callaway and Sant'Anna (2021) which implements a flexible regression adjustment estimator. Another approach uses stacked samples of data, where the observations for each cohort are first paired with all never-treated control observations, and then the cohort-specific datasets are pooled (Cengiz et al., 2019; Wing et al., 2024). These *stacked data* regressions can be used to estimate models with *lags-only* or *lags and leads* specifications of treatment coefficients.

5.4 Results

The outcome variable is whether the individual enrolled in SNAP. The results in Table 4 show that the ATET estimates are always negative, indicating that Secure Communities immigration enforcement leads to lower SNAP enrollment. The FLEX estimates are that immigration enforcement lowers SNAP enrollment by about $1.3 - 1.5$ percentage points, compared to an overall mean of around 20 percent. The ATET estimate implies at least a 6.5-7.5% decrease in the probability of enrolling in SNAP. The estimates are negative and statistically significant when the homogenous (constant) specification (TWFE) is used without any of the four covariates, but not statistically significant with covariates (gender, age, income, and education) added to the specification.

In more general regressions with specifications in which treatment varies by cohort and year or by group and year, the ATET estimates are negative and statistically significant at conventional levels. The point estimates are a bit larger in the cohort-by-year specifications as

compared to those in the group-by-year specifications. But the standard errors in the group-by-year specification are consistently smaller than those in the cohort-by-year specifications. These results can be seen in lags only and lags and leads specifications. It appears that the additional generality implied in the group-by-year heterogeneous specifications produces treatment estimates with greater precision because there is substantial within-cohort (by year) heterogeneity in outcomes.

The Callaway and Sant'Anna estimates, which allow for general heterogeneity at the cohort-by-year level, are similar to those obtained using FLEX. These estimates are also statistically significant but the standard errors are slightly larger than those obtained using FLEX. The stacked data regressions, which allow for heterogeneous effects across event-time, produce estimates are also statistically significant.

Another interesting pattern is that the ATET shrinks towards zero when covariates are interacted with the heterogeneous treatment effects. While this is not a universal finding, we have noticed in numerous empirical examples that including covariates often changes the magnitude of the ATET, while the standard errors either stay about the same or shrink.

We use our preferred FLEX specification, a linear-in-parameters model with heterogeneous treatment coefficients at the group-by-year level and interacted with covariates, to calculate estimates of ATET at disaggregate levels of interest to researchers. In panels (a) and (b) of Figure 1, we show the event study plots, from *lags only* models in panel (a) and from *lags and leads* models in panel (b). The latter is what is often referred to as the *event study* figure. An eyeball check shows that the parallel trends assumption is good for five pre-periods. The negative treatment effects by exposure year in the treatment periods are quite similar in magnitude across the *lags only* specification and the *lags and leads* specification.

The ATET estimates in each calendar year in a selection of treated periods are shown in panels (c) and (d). These are similar across the *lags only* specification (c) and *lags and leads* specification (d). The ATET by cohorts are shown in panels (e) and (f). They show some heterogeneity by cohort, with the early cohorts generally having more negative effects and the later cohorts having either no or positive effects. This heterogeneity would be worth exploring further.

5.5 Transparency

One advantage of our FLEX approach is that it is transparent. The model specification is clear; there are no hidden estimated parameters. This transparency is useful for comparing different possible FLEX model specifications and for comparing FLEX with other estimators. To demonstrate this transparency, we created tables showing the number of parameters estimated across different model specifications and estimators, before any aggregation of the treatment effects. The number of parameters estimated for the Secure Communities example are shown in Table 5. The numbers of parameters are shown separately for the covariates, main treatment effects (lags and leads), main fixed effects, and interactions of the main effects with covariates. All model specifications include main fixed effects for the group and time periods (what is often called two-way fixed effects). The total number of parameters is the sum of all covariates, main effects, fixed effects, and interactions.

There are several important patterns apparent in Table 5. By definition, there are no lead effects estimated in the lags-only models. The lags-only models assume that the parallel trends assumption is correct in the pre-periods; those models do not allow separate parameters for treatment and control groups each pre period. This assumption, if correct, is more efficient. However, models with both lags and leads allow one to plot event study graphs and test whether the parallel trends and no anticipation assumptions are correct.

Another difference between models is whether the model uses groups or cohorts. Because cohorts are collections of at least one group, there are at least as many parameters in the group models as in the cohort models. The interactions in the FLEX models are between covariates and treatment effects. Despite adding additional covariates, in our experience this additional treatment effect heterogeneity does not generally increase the standard errors of the estimated average treatment effect on the treated. Also, the researcher has the option to only interact a small number of the covariates with the treatment effects, based on what economic theory predicts or the evidence from prior related research supports.

Comparing the parameters helps to show the relationship between the different models. One interesting comparison is that the estimated coefficients for two of the models are identical. The lags-and-leads FLEX model with cohort and time effect heterogeneity and

no covariates interacted is identical to the Callaway and Sant'Anna model with cohort and time effect heterogeneity and no covariates. Although the estimated coefficients are identical (not shown), the two estimated average treatment effect on the treated are slightly different because of different weights used to average those identical estimated treatment effects (see Table 4).

6 Conclusions

Our paper makes several theoretical and practical contributions to the literature on difference-in-differences with staggered treatments applied to the analysis of repeated cross-sectional data. On the theoretical dimension, we prove that a linear regression with a sufficiently flexible functional form consisting of group-by-time treatment effects, two-way fixed effects, and interaction terms yields consistent estimates of heterogeneous treatment effects. The estimates are efficient and aggregation of treatment effects and inference are straightforward. The results hold under standard assumptions of stable unit treatment value, no bad controls, no anticipation, and conditional parallel trends. We prove that FLEX with lags only and appropriate interaction terms, estimated by ordinary least squares, returns numerically identical results as the imputation method by Borusyak et al. (2024).

The theoretical result about repeated cross-sectional data is of importance to many applied researchers, because data are often not balanced panel data. Our FLEX model extends other well-known DID models in other ways. FLEX with leads and lags extends Sun and Abraham (2021) and Callaway and Sant'Anna (2021) to repeated cross sections and allows for covariates to be included more flexibly. FLEX also allows heterogeneous treatment effects at the group level, instead of only at the cohort level.

On the empirical side, we demonstrated our FLEX methods with one publicly available data set to answer a policy-relevant research question about the effect of Secure Communities immigration enforcement. The empirical example used individual-level cross-sectional data with staggered treatment at the PUMAs. Our FLEX method and the imputation method obtained the same numerical result. Our FLEX method generally has smaller standard errors than other popular estimators. In summary, FLEX has the advantage of being easy to implement, flexible, fast, and best among linear unbiased estimators.

References

Alsan, M. and Yang, C. (2022). Replication data for: Fear and the Safety Net: Evidence from Secure Communities.

Alsan, M. and Yang, C. S. (2024). Fear and the Safety Net: Evidence from Secure Communities. *The Review of Economics and Statistics*, 106(6):1427–1441.

Baker, Andrew C. Callaway, B., Scott, C., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2025). Difference-inDiferences Designs: A Practitioner's Guide. Technical report.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*, page ˙https://doi.org/10.1093/restud/rdae007.

Caetano, C., Callaway, B., Payne, S., and Rodrigues, H. S. (2024). Difference in Differences with Time-Varying Covariates. *Working paper*.

Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2):200–230.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The Effect of Minimum Wages on Low-Wage Jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996.

de Chaisemartin, C. and D'Haultfœuille, X. (2023). Two-way fixed effects and difference-indifferences with heterogeneous treatment effects: a survey. *Econometrics Journal*, 26:C1–C30.

Dette, H. and Schumann, M. (2024). Testing for Equivalence of Pre-Trends in Difference-in-Differences Estimation. *Journal of Business & Economic Statistics*, 1–13.

East, C. N., Hines, A. L., Luck, P., Mansour, H., and Velásquez, A. (2023). The Labor Market Effects of Immigration Enforcement. *Journal of Labor Economics*, 41(4):957–996.

Freedman, S. M., Hollingsworth, A., Simon, K. I., Wing, C., and Yozwiak, M. (2023). Designing Difference in Difference Studies With Staggered Treatment Adoption: Key Concepts and Practical Guidelines. NBER Working Paper 31842. https://www.nber.org/papers/w31842.

Goodman-Bacon, A. (2021). Difference-in-differences with Variation in Treatment Timing. *Journal of Econometrics*, 225(2):254–277.

MacKinlay, A. C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1):13–39.

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

Roth, J. (2024). Interpreting Event-Studies from Recent Difference-in-Differences Methods. *arXiv working paper*. https://doi.org/10.48550/arXiv.2401.12309.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2023). What's Trending in Difference-in-differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*, 235(2):2218–2244.

Sun, L. and Abraham, S. (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, 225(2):175–199.

Wing, C., Freedman, S. M., and Hollingsworth, A. (2024). Stacked Difference-in-Differences. https://www.nber.org/papers/w32054.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. SSRN Working Paper 3906345. https://papers.ssrn.com/abstract=3906345.

Wooldridge, J. M. (2023). Simple Approaches to Nonlinear Difference-in-differences with Panel data. *The Econometrics Journal*, 26(3):C31–C66.

Table 1: Public Use Microdata Area participation in the Secure Communities immigration enforcement program by cohort and Census region

	Northeast	Midwest	South	West	Total
Never treated	10	24	143	246	423
2009	0	0	49	20	69
2010	0	0	32	18	50
2011	0	9	20	24	53
2012	47	6	5	29	87
Total	57	39	249	337	682

Notes: The repeated cross-sectional data are from the American Community Survey (ACS) data for residents of 682 PUMAs in 32 states for 12 years from 2005–2016. A cohort comprises PUMAs that implemented the program in a particular year.

Table 2: Secure Communities observations by cohort and region

	Northeast	Midwest	South	West	Total
Never treated	5680	8256	132281	235625	381842
2009	0	0	54124	14349	68473
2010	0	0	30950	15589	46539
2011	0	3336	15837	19629	38802
2012	27092	1429	1346	16137	46004
Total	32772	13021	234538	301329	581660

Notes: The repeated cross-sectional data are from the American Community Survey (ACS) data for residents of 682 PUMAs in 32 states for 12 years from 2005–2016. A cohort comprises PUMAs that implemented the program in a particular year.

Table 3: Sample means by Secure Communities status

	Treated	Never treated
Enrolled in SNAP	0.204	0.166
Secure Communities is activated	0.557	0.000
Female	0.544	0.550
Age	50.582	51.403
Income to Poverty ratio	235.350	247.876
High school diploma or GED	0.607	0.631
Observations	199818	381842

Notes: The repeated cross-sectional data are from the American Community Survey (ACS) data for residents of 682 PUMAs in 32 states for 12 years from 2005–2016.

Table 4: ATET of participation in the Secure Communities immigration enforcement program on SNAP enrollment

Model	Effect heterogeneity	Covariates	ATET	Std. err.	p-value
Lags only models					
FLEX	Group & Time	None	-0.0145	0.0044	0.0011
		Interacted	-0.0109	0.0034	0.0014
	Cohort & Time	None	-0.0117	0.0052	0.0243
		Interacted	-0.0071	0.0040	0.0794
Constant Effect	Homogeneous	None	-0.0099	0.0047	0.0363
		Additive	-0.0034	0.0038	0.3711
Lags and leads models					
FLEX	Group & Time	None	-0.0163	0.0053	0.0022
		Interacted	-0.0156	0.0044	0.0005
	Cohort & Time	None	-0.0130	0.0057	0.0236
		Interacted	-0.0121	0.0049	0.0125
CSRA	Cohort & Time	None	-0.0133	0.0059	0.0235
		Flexible	-0.0122	0.0050	0.0153
Event Study	Time-since-event	None	-0.0153	0.0055	0.0052
		Additive	-0.0122	0.0046	0.0085
Stacked DID	Time-since-event	None	-0.0160	0.0056	0.0046
		Additive	-0.0126	0.0048	0.0089

Notes: For models with heterogeneous effects, ATET is a weighted average of the estimand. Standard errors of ATET are based on cluster (group) robust standard errors of the coefficients in the estimand. As summarized in Table 3, 4 covariates are used. Some regression specifications have no covariates (None), in some covariates enter only additively (Additive), while in others covariates enter additively and interacted with the estimand coefficients and with group and year indicators (Interacted). FLEX refers to the model developed in this paper. Constant Effect refers to the homogeneous effect two-way fixed effects estimator. Event Study refers to the heterogeneous over event-time but constant across groups two-way fixed effects estimator. Stacked DID refers to regressions on samples of data in which each cohort is first associated with never-treated controls and then the samples associated with each cohort are pooled (Cengiz et al., 2019). CSRA refers to the Callaway and Sant'Anna (2021) regression-adjustment estimator that uses a flexible regression specification to estimate the parameters of the conditional mean model and an influence function approach to calculate the standard errors of the estimates.

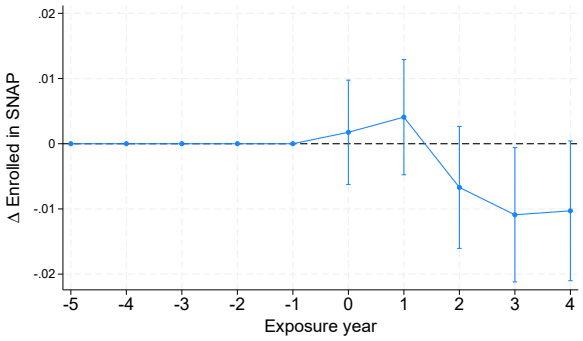
Table 5: Numbers of parameters estimated in alternative specifications

	Effect		Main effects			Interactions			
Model	heterogeneity	Covariates	Lags	Leads	FE	Lags	Leads	FE	Total
Lags only models									
FLEX	Group & Time	0	68	0	25	0	0	0	94
		4	68	0	25	272	0	100	470
	Cohort & Time	0	26	0	15	0	0	0	42
		4	26	0	15	104	0	60	210
Constant Effect	Homogeneous	0	1	0	25	0	0	0	27
		4	1	0	25	0	0	0	31
Lags and leads models									
FLEX	Group & Time	0	68	18	25	0	0	0	112
		4	68	53	25	272	212	100	735
	Cohort & Time	0	26	18	15	0	0	0	60
		4	26	18	15	104	72	60	300
CSRA	Cohort & Time	0	26	18	88	0	0	0	176
		4	26	18	88	0	0	0	352
Event Study	Time-since-event	0	12	12	25	0	0	0	50
		4	12	12	25	0	0	0	54
Stacked DID	Time-since-event	0	12	12	100	0	0	0	125
		4	12	12	100	0	0	0	129

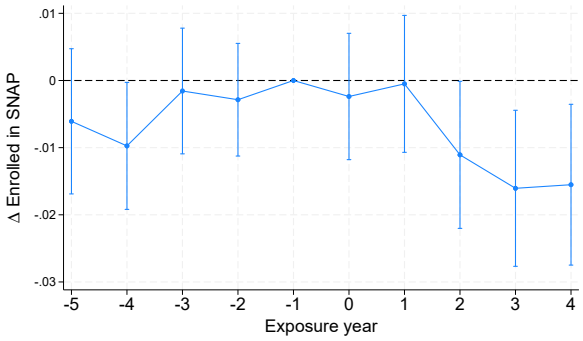
Notes: The repeated cross-sectional data are from the American Community Survey (ACS) data for residents of 682 PUMAs in 32 states for 12 years from 2005–2016. Some regression specifications have no covariates, in some covariates enter only additively, while in others covariates enter additively and interacted with the estimand coefficients and with group and year indicators. FLEX refers to the model developed in this paper. Constant Effect refers to the homogeneous effect two-way fixed effects estimator. Event Study refers to the heterogeneous over event-time but constant across groups two-way fixed effects estimator. Stacked DID refers to regressions on samples of data in which each cohort is first associated with never-treated controls and then the samples associated with each cohort are pooled (Cengiz et al., 2019). For each of these regression specifications, the column labeled "Total" also includes an intercept. CSRA refers to the Callaway and Sant'Anna (2021) regression-adjustment estimator that uses a flexible regression specification to estimate the parameters of the conditional mean model.

Figure 1: Heterogeneous ATET of participation in the Secure Communities immigration enforcement program on SNAP enrollment

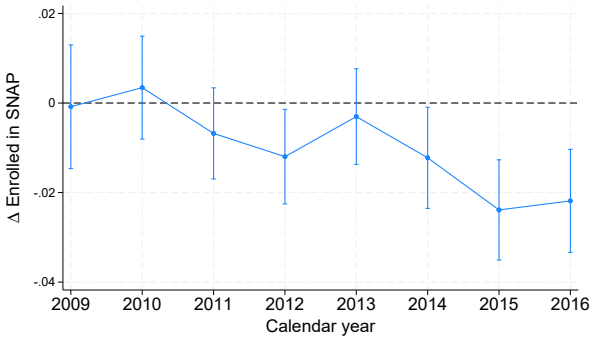
(a) ATET by exposure year: *lags only*



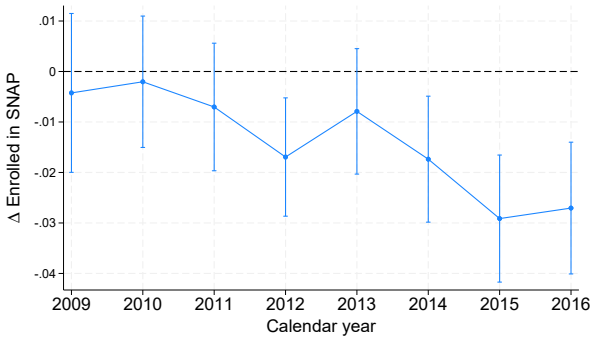
(b) ATET by exposure year: *lags and leads*



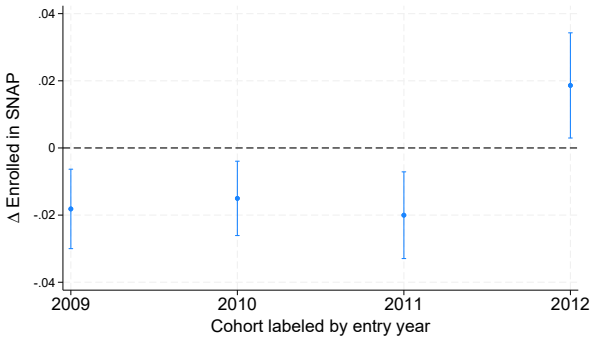
(c) ATET by calendar year: *lags only*



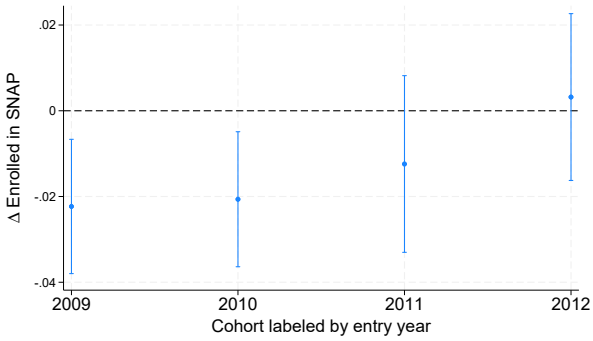
(d) ATET by calendar year: *lags and leads*



(e) ATET by cohort: *lags only*



(f) ATET by cohort: *lags and leads*



Notes: Regression models estimated with a fully interacted specification with estimands specified at the group by time level.