Fuzzy Difference-in-Differences with Grouped Data

Clément de Chaisemartin* Xavier D'Haultfœuille† Félix Pasquier‡

February 14, 2022

Abstract

This paper considers the estimation of the effect of a binary treatment from a panel of groups in which the treatment rate may evolve over time. We assume common trends and consider fuzzy designs: within some groups, some units may be treated while others are not. In such contexts, the popular two-way fixed-effect regressions are not robust to heterogeneous treatment effects. Under some conditions allowing for such heterogeneity, we show that we can estimate average treatment effects with simple, linear estimators. Importantly, such estimators only rely on average outcomes and treatment rates at the group level. Thus, they can be used even if micro data with both the treatment status and outcome of units are not available. We apply our method to revisit the effect of radio programs on the rise of Nazism.

Keywords: differences-in-differences, panel data, time varying treatment effects, heterogeneous treatment effects, correlated random coefficient models.

JEL Codes: C21, C23

*Sciences Po Paris, clement.dechaisemartin@sciencespo.fr

†CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr.

‡CREST-ENSAE, felix.pasquier@ensae.fr

1 Introduction

Difference-in-differences is one of the most popular methods to identify causal effects of a binary treatment using observational data. This idea is very often implemented using a panel of groups exhibiting both cross-sectional and temporal variation in the treatment. Then one considers two-way fixed effect regressions, where the outcome of interest is regressed on group fixed effects, time period fixed effects and the treatment rate at the group level. However, even if the so-called common trends assumption holds, such regressions suffer from an important limitation: they are not robust to heterogeneous treatment effects (see de Chaisemartin and D'Haultfœuille, 2020, for general designs and Borusyak and Jaravel, 2017 and Goodman-Bacon, 2021 for staggered adoption and sharp designs). Specifically, the coefficient of the treatment rate in such regressions identifies a weighted average of treatment effects over all groups and periods of time, but with possibly negative weights. This implies that the coefficient may be negative even if all groups benefit from the treatment at all periods.

Realizing this has led to the development of several alternative estimators, robust to such heterogeneous treatment effects. We refer in particular to those of de Chaisemartin and D'Haultfœuille (2020) and de Chaisemartin and D'Haultfœuille (2021*b*) in a static framework and those of Callaway and Sant'Anna (2021), Sun and Abraham (2021), Borusyak et al. (2021) and de Chaisemartin and D'Haultfœuille (2021*a*) in a dynamic framework (for more details and a survey of this recent literature, see de Chaisemartin and D'Haultfœuille, 2022). However, these estimators only apply to sharp designs, namely designs for which the treatment status of all units belonging to the same group is the same. Yet, this requirement often fails to hold. For instance, consider a program that is in place in some districts but not in others. Only a fraction of the population in the first set of districts may benefit from the program.

Obviously, if individual data are available, and to the extent that the treatment is strictly exogenous, one can still rely on the aforementioned methods by defining "groups" as individuals. However, in many cases, only aggregated data at the group level are available, in the form of the average outcome and treatment rate at the group level. It could also be the case that the treatment is available at the individual level but the outcome is not, or conversely. For instance, many studies look at treatment effects on voting (see e.g. the study we revisit in our application in Section 5, or Enikolopov et al. (2011)) and voting is never available at the individual level The aim of this paper is to develop estimators robust to heterogeneous treatment effects for such contexts.

To this end, we consider an individual-level model and posit that both potential outcomes are additively separable into a group fixed effect, a time fixed-effect and idiosyncratic shocks. While this model implies some restrictions on average treatment at the group level, it still allows for heterogeneity in such effects across groups and over time. Not surprisingly given that we do not

rely on individual data, we also restrict selection into treatment within groups. Nevertheless, we show that this restriction is compatible with a generalized Roy model, provided that individuals do not have a rich information about their potential gains from being treated.

Under these restrictions, our individual-level model, once aggregated at the group level, leads to the correlated random coefficient model studied by Chamberlain (1992). Under a restriction on the design, this implies that average treatment effects on "movers", namely groups experiencing a change in their treatment rate, are identified. We show that the design restriction is weak with three periods or more and the estimator may still hold with two periods. Moreover, the identification strategy leads to elementary linear estimator, which are asymptotically normal as the number of groups tend to infinity.

We consider two important extensions. First, we show that it is straightforward to include time-varying covariates in our model, and modify the estimators accordingly. Such covariates are important in practice as they may render the common trend conditions more credible. Second, we show that we can quantify the importance of the heterogeneity of average treatment effects across groups sharing the same history of treatment rates.

Finally, we apply our results to revisit Adena et al. (2015) who study the impact of radio programs on the rise of Nazism during the Weimar Republic. This application is particularly suited to our methodology for at least three reasons. First, the design is fuzzy, and only aggregated data are available. Second, during the period under consideration, subscription to the radio increased substantially, with nonetheless important regional variations. Third, the content of the German radio programs changed substantially, from being first apolitical, then biased against Nazis and finally biased in favor of the Nazis. We thus expect important temporal variations in the treatment effects. Compared to the results of Adena et al. (2015), our results highlight in particular that the propaganda against Nazis did not have much impact, whereas the pro-Nazis slant of the 1933 radio programs had a large and positive impact on the votes for Nazis.

Our paper is related to de Chaisemartin and D'Haultfœuille (2018), which also considers fuzzy designs. An important difference is that the two estimators that are robust to heterogeneous treatment effects, their so-called "Wald-TC" and "Wald-CIC" estimators, cannot be computed with solely average outcomes and treatment rates at the group level. Another difference is that the asymptotic framework in de Chaisemartin and D'Haultfœuille (2018) is in the number of units rather than in the number of groups, as we do here. Adapting their result to a growing number of groups could be difficult, as they acknowledge (see Section 2.1 in their supplement). Our paper is also related to Chamberlain (1992). Specifically, we rationalize his model at the group level by a model on potential outcomes at the unit level. We thus show that estimators closely related to his are useful in the context of grouped panel data with heterogeneous treatment effects.

The paper is organized as follows. We display our basic set-up and main assumptions in Section 2.

Section 3 presents the identification results and our estimators for this basic set-up. Extensions are considered in Section 4. Section 5 is devoted to the application. All the proofs are gathered in the appendix.

2 Set-up and model

We consider a panel of G groups over T periods. For all $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, we let $N_{g,t}$ denote the number of units in this "cell" (g,t). For any random variable $A_{i,g,t}$ defined at the individual level, we let $\mathbf{A}_g = (A_{i,g,t})_{1\leq t\leq T, 1\leq i\leq N_{g,t}}$ be the vector collecting all the corresponding variables for group g. Then, we let $A_{g,t} = \sum_{i=1}^{N_{g,t}} A_{i,g,t}/N_{g,t}$ be the average of $A_{i,g,t}$ over the cell (g,t) and $\boldsymbol{A}_g^a = (A_{g,1},\dots,A_{g,T})'$. If $A_{g,t}$ is defined at the cell level, we define similarly $\mathbf{A}_g = (A_{g,t})_{1\leq t\leq T}$. We first impose that no group appears or disappears over time.

Assumption 1 (Balanced panel of groups) *For all* $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, $N_{g,t} > 0$.

We are interested in measuring the effect of a binary treatment on some outcome. For every (i,g,t), let $D_{i,g,t}$, $Y_{i,g,t}(0)$ and $Y_{i,g,t}(1)$ respectively denote the treatment status and the potential outcomes without and with treatment of the ith unit in cell (g,t). The observed outcome of the ith observation in group g at period t is $Y_{i,g,t} = Y_{i,g,t}(D_{i,g,t})$. Let $\Delta_{i,g,t} = Y_{i,g,t}(1) - Y_{i,g,t}(0)$ denote the treatment effect of ith unit in group g at period t. We consider the following static model. For all $(i,g,t) \in \{1,\dots,N_{g,t}\} \times \{1,\dots,G\} \times \{1,\dots,T\}$,

$$\begin{cases} Y_{i,g,t}(0) &= \alpha_{i,g} + \beta_t + \xi_{i,g,t} \\ \Delta_{i,g,t} &= \Lambda_{i,g} + \mu_t + \zeta_{i,g,t} \end{cases} \tag{1}$$

$\alpha_{i,g}$ and $\Lambda_{i,g}$ are unit-specific parameters while β_t and μ_t are common across units of all groups. Without loss of generality and for identification purposes, we suppose that $\beta_1 = \mu_1 = 0$. Let $U_{i,g,t} = (D_{i,g,t}, \alpha_{i,g}, \xi_{i,g,t}, \Lambda_{i,g}, \zeta_{i,g,t})$. We impose the following assumptions.

Assumption 2 (Independent groups and exchangeability within groups) *(i) The* G *random vectors* $(\boldsymbol{U}_g)_{g=1,\dots,G}$ *are independent and (ii) for all* $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, *the* $N_{g,t}$ *random vectors* $(U_{i,g,t})_{i=1,\dots,N_{g,t}}$ *are exchangeable conditional on* $\boldsymbol{D}_g^a$.

Assumption 2 requires that potential outcomes and treatment statuses of units in different groups be independent. Yet, treatment statuses and potential outcomes of units in the same group may be correlated over time. Also, units within the same group are assumed to exchangeable, which implies that they are identically distributed but allows for dependence between them.

Assumption 3 (Strong exogeneity) *For all* $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, $E[\xi_{1,g,t}|\mathbf{D}_g^a] = E[\zeta_{1,g,t}|\mathbf{D}_g^a] = 0$.

Assumption 3 requires that the shocks affecting a unit's potential outcome be mean-independent of the unit's treatment sequence and of the other units' belonging to the same group. It is related to the strong exogeneity condition for panel models, which is necessary to obtain the consistency of the fixed effect estimator. With Assumption 3, it is easy to see that the model implicitly rests on the following common trends assumption:

$$E[Y_{i,g,t}(0) - Y_{i,g,t-1}(0)] = \beta_t - \beta_{t-1} \tag{2}$$

$$E[Y_{i,g,t}(1) - Y_{i,g,t-1}(1)] = \beta_t - \beta_{t-1} + \mu_t - \mu_{t-1}. \tag{3}$$

Similar conditions are imposed in de Chaisemartin and D'Haultfœuille (2020, see Assumptions 4 and 9 therein) and in de Chaisemartin and D'Haultfœuille (2020, see Assumptions 4'M in the supplement). Finally, we consider the following assumption:

Assumption 4 (Sufficiency of the average treatment rates) *For all* $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, $E[\Lambda_{1,g}|D_{1,g,t}, \boldsymbol{D}_g^a] = E[\Lambda_{1,g}|\boldsymbol{D}_g^a]$.

Assumption 4 states that, conditional on group g's average treatment rates over time, all unit-specific treatment effects $\mathbf{\Lambda_g}$ are mean independent of the treatment status of units in the group (in particular of the unit's corresponding treatment status). Noteworthy, Assumption 4 is still compatible with some generalized Roy models. Specifically, assume that for all (i,g,t),

$$D_{i,g,t} = \mathbb{1}\left\{E[\Delta_{i,g,t}|\mathcal{I}_{i,g,t}] \geq C_{i,g,t}\right\}, \tag{4}$$

where $\mathcal{I}_{i,g,t}$ is the information set of the agent (technically, a sigma-algebra) at the time of his decision and $C_{i,g,t}$ represents the expected cost for i of being treated. Then, we have the following result:

Proposition 1 *Suppose that* (1) *and* (4) *hold,* $C_{i,g,t} = C_{g,t} + \eta_{i,g,t}$, $\mathcal{I}_{i,g,t}$ *is the sigma-algebra generated by* $(Z_{g,t}, C_{g,t}, \eta_{i,g,t})$ *for some variables* $Z_{g,t}$, $((\Lambda_{i,g})_i, \boldsymbol{Z}_g, \boldsymbol{C}_g) \perp\!\!\!\perp (\boldsymbol{\eta}_g, \boldsymbol{\zeta}_g)$ *and the variables* $\eta_{i,g,t}$ *are i.i.d. over* i *and* t. *Then, Assumption 4 holds.*

The main restriction we imposed on this Roy model is that the returns expected by individuals are all the same within a group g at t (and equal to $E[\Delta_{i,g,t}|Z_{g,t}, C_{g,t}]$. On the other hand, this model is compatible with self-selection of groups, where groups with the highest average treatment effect displaying also the highest treatment rates.

3 Identification and estimation

3.1 Identification

Our main insight is that under Assumptions 1-4, the average outcome and treatment variables $Y_{g,t}$ and $D_{g,t}$ satisfy the correlated random coefficient model of Chamberlain (1992). Then, identification of average treatment effects follow under a restriction on the design. Before presenting this restriction and the result, we introduce new notation. Let $\delta_0 = (\delta_{0,1}, \ldots, \delta_{0,2T})' = (\beta_2, \mu_2, \ldots, \beta_T, \mu_T)'$. We also define "movers" as groups experiencing at least one change in $D_{g,t}$:

$$M_g = \mathbb{1}\left\{\exists (t,t') : D_{g,t} \neq D_{g,t'}\right\}.$$

The reason why we introduce such movers is that in our model with heterogeneous treatment effects, we do not learn anything on the average treatment effects of "stable" groups (g such that $M_g = 0$). Thus, we focus hereafter on the average treatment effect on movers at date t:

$$\Delta_{0,t} := E[\Delta_{i,g,t} | M_g = 1].$$

Given our model, the identification of $\Delta_{0,t}$ will follow from that of δ_0 and $\phi_0 = (\phi_{0,1},\ \phi_{0,2}) = (E[\alpha_{1,g}|M_g = 1],\ E[\Lambda_{1,g}|M_g = 1])'$.

We also introduce useful vectors and matrices. First, we let $e_{n,k}$ denote the row vector of size n with 1 at coordinate k and 0 elsewhere, $\mathbf{0}_n$ (resp. $\mathbf{1}_n$) denote the row vector of size n with all coordinates equal to 0 (resp. 1). Then, let $X_g = (\mathbf{1}_T' \ \boldsymbol{D}_g^a)$ and

$$W_g = \begin{pmatrix} \mathbf{0}_{2(T-1)} \\ e_{T-1,1} \otimes (1, D_{g,2}) \\ \vdots \\ e_{T-1,T-1} \otimes (1, D_{g,T}) \end{pmatrix}.$$

Hence, with $T = 3$ for instance, we have

$$W_g = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & D_{g,2} & 0 & 0 \\ 0 & 0 & 1 & D_{g,3} \end{pmatrix}.$$

Similarly to ϕ_0, we also define $\phi_0(\boldsymbol{D}_g^a) = \left(E[\alpha_{1,g}|\boldsymbol{D}_g^a],\ E[\Lambda_{1,g}|\boldsymbol{D}_g^a]\right)'$. Finally, for any matrix A, let A^+ denote its Moore–Penrose inverse and let $\Pi(A) = I - AA^+$ be the orthogonal projector on the kernel of A'.

We can now present the design restriction and our main result.

Assumption 5 (Design restriction) *For all $g \in \{1, \ldots, G\}$, $E[W_g' \Pi(X_g) W_g]$ is non-singular.*

Theorem 1 *Suppose that Model* (1) *and Assumptions 1-4 hold. Then*

$$E[Y_g | \mathbf{D}_g^a] = W_g \delta_0 + X_g \phi_0(\mathbf{D}_g^a). \tag{5}$$

Moreover, if Assumption 5 holds, δ_0, ϕ_0 and $\Delta_{0,t}$ are respectively identified by

$$\delta_0 = E[W_g' \Pi(X_g) W_g]^{-1} E[W_g' \Pi(X_g) Y_g], \tag{6}$$

$$\phi_0 = E\left[(X_g' X_g)^{-1} X_g' (Y_g - W_g \delta_0) | M_g = 1 \right], \tag{7}$$

$$\Delta_{0,t} = \phi_{0,2} + \delta_{0,2(t-1)} \mathbb{1}\{t > 1\}. \tag{8}$$

Note that Equation (5) is a particular case of Model (4.4) in Chamberlain (1992). The identification of δ_0 presented here is the same as in Arellano and Bonhomme (2012), and extends that in Chamberlain (1992). Specifically, we need not focus on groups g such that X_g is full column rank (namely, the movers) and can also include "stayers", for which $\boldsymbol{D}_{g,1}^a = \cdots = \boldsymbol{D}_{g,t}^a$. In fact, stayers are necessary to achieve identification if $T = 2$, as shown below. Once δ_0 is known, identification of ϕ_0 and average treatment effects on movers follow easily from (7) and (8).

Assumption 5 is key for the identification of the model. To understand this condition better, Proposition 2 relates it to the distribution of $\boldsymbol{D}_g^a$.

Proposition 2 *Assumption 5 holds:*

1. *when $T = 2$, if and only if $P(M_g = 0) > 0$ and $V(D_{g,1} | M_g = 0) > 0$.*
2. *when $T \geq 3$, if for all t, there exists (t', t'') such that $D_{g,t'} - D_{g,t''}$, $D_{g,t}(D_{g,t'} - D_{g,t''})$, $D_{g,t} - D_{g,t''}$ and $D_{g,t'}(D_{g,t} - D_{g,t''})$ are not collinear.*

As already pointed out by Graham and Powell (2012), the two cases $T = 2$ and $T \geq 3$ are fundamentally distinct. With $T = 2$, $\Pi(X_g) = 0$ except if $M_g = 0$. This is why Assumption 5 requires that there is a positive fraction of stable groups. In order to identify μ_2, we also require variation in the treatment rates across such groups ($V(D_{g,1} | M_g = 0) > 0$). An example where such conditions hold is when some groups are untreated at both periods, while some other groups are fully treated at both periods. When $T \geq 3$, on the other hand, $\Pi(X_g)$ is never null. Then, a sufficient condition is that, roughly speaking, the treatment rates vary sufficiently from one period to another. The condition is actually weak: it may hold even if the support of $(D_{g,t}, D_{g,t'}, D_{g,t''})$ takes only four distinct values. As another example, in a sharp design where the support of $D_{g,t}$ is only $\{0, 1\}$ and $T = 3$, the condition above holds if the trajectories $(0, 0, 1)$, $(0, 1, 0)$, $(0, 1, 1)$ and $(1, 0, 1)$ are in the support of $(D_{g,1}, D_{g,2}, D_{g,3})$.

3.2 Estimation

We consider estimators based on the identification results in Theorem 1. As pointed out by Graham and Powell (2012), we have to regularize the empirical counterpart of (8). The reason behind is "quasi-stayers", for which $X_g'X_g$ is close to being singular: if the density of the variable $\det(X_g'X_g)$ is positive at 0, the plug-in estimator is inconsistent. To remedy this issue, we focus on movers exhibiting sufficient variation. Specifically, let

$$M_{g,h} = \mathbb{1}\left\{|\det(X_g'X_g)| > h\right\}.$$

We focus hereafter on $\Delta_{0,t}^h = E\left[\Delta_{i,g,t}|M_{g,h}=1\right]$. Let $G_h = \#\{g : M_{g,h}=1\}$. The estimator we consider is a variation of Chamberlain (1992)'s two-step GMM estimator:

$$\begin{aligned}\widehat{\delta} &= \left(\frac{1}{G}\sum_{g=1}^{G} W_g'\Pi(X_g)W_g\right)^{-1}\left(\frac{1}{G}\sum_{g=1}^{G} W_g'\Pi(X_g)Y_g\right)\\ \widehat{\phi}^h &= \frac{1}{G_h}\sum_{g:M_{g,h}=1}(X_g'X_g)^{-1}X_g'(Y_g - W_g\widehat{\delta}).\end{aligned}$$

Then, we let $\widehat{\Delta}_t^h = \widehat{\phi}_2^h + \widehat{\delta}_{2(t-1)}\mathbb{1}\{t>1\}$. As one could expect, this estimator is asymptotically normal.

Proposition 3 *Suppose that Assumptions 1-5 hold. Then*

$$\sqrt{G_h}\left(\widehat{\Delta}_t^h - \Delta_{0,t}^h\right) \xrightarrow{d} \mathcal{N}\left(0, V(\omega_{g,t})\right),$$

where $\omega_{g,t}$ is defined in Equation (20) *in the appendix.*

Estimating $\Delta_{0,t}$ rather than $\Delta_{0,t}^h$ is possible by letting h tend to 0 at an appropriate rate as G tends to infinity. Then, one faces a usual bias-variance trade-off: the bias decreases while the variance increases as h tends to 0. We refer to Graham and Powell (2012) for a thorough discussion of this issue.

4 Extensions

4.1 Including covariates

Our basic model implies the common trend conditions (2) and (2). Such conditions may not be credible unconditionally, but may hold conditional on some observed covariates. We show

here that our model can be simply extended to include time-varying covariates. First, we now assume that for all $(g,t) \in \times\{1,\ldots,G\} \times \{1,\ldots,T\}$ and $i \in \{1,\ldots,N_{g,t}\}$,

$$\left\{ \begin{array}{rl} Y_{i,g,t}(0) = & \alpha_{i,g} + \beta_t + Z^1_{i,g,t}{}'\lambda_{0,1} + \xi_{i,g,t} \\ \Delta_{i,g,t} = & \Lambda_{i,g} + \mu_t + Z^2_{i,g,t}{}'\lambda_{0,2} + \zeta_{i,g,t} \end{array} \right. . \tag{9}$$

We allow the vectors $Z^1_{i,g,t}$ and $Z^2_{i,g,t}$ to be empty by simply letting $Z^1_{i,g,t}{}'\lambda_{0,1} = 0$ or $Z^2_{i,g,t}{}'\lambda_{0,1} = 0$ in such cases. Though $Z^1_{i,g,t}$ and $Z^2_{i,g,t}$ are allowed to be identical, it is important to distinguish them for informational reasons. Our estimators will eventually rely on the averages $Z^1_{g,t}$ and $\sum_{i=1}^{N_{g,t}} D_{i,g,t}Z^2_{i,g,t}/N_{g,t}$, and we may not observe this latter average, for the same reasons we may not observe simultaneously $Y_{i,g,t}$ and $D_{i,g,t}$. Note that even if the vector $Z^2_{i,g,t}$ is empty, Model (9) still allows for heterogeneous treatment effects through the other terms in $\Delta_{i,g,t}$, in particular $\Lambda_{i,g}$.

Let $Z_{i,g,t} = (Z^1_{i,g,t}{}', D_{i,g,t}Z^2_{i,g,t}{}')$ and $\lambda_0 = (\lambda'_{0,1}, \lambda'_{0,2})'$, with, e.g., $Z_{i,g,t} = Z^1_{i,g,t}$ and $\lambda_0 = \lambda_{0,1}$ if $Z^2_{i,g,t}$ is empty. We modify the previous assumptions as follows:

Assumption 2' (Independence and exchangeability with covariates) *(i) The G random vectors $(\mathbf{U}_g, \mathbf{Z}_g)_{g=1,\ldots,G}$ are independent; (ii) the $N_{g,t}$ random vectors $(U_{i,g,t}, Z_{i,g,t})_{i=1,\ldots,N_{g,t}}$ are exchangeable conditional on $(\boldsymbol{D}^a_g, \boldsymbol{Z}^a_g)$.*

Assumption 3' (Strong exogeneity with covariates) *For all $(g,t) \in \{1,\ldots,G\}\times\{1,\ldots,T\}$, $E[\xi_{1,g,t}|\mathbf{D}_g, \mathbf{Z}_g] = E[\zeta_{1,g,t}|\mathbf{D}_g, \mathbf{Z}_g] = 0$.*

Assumption 4' (Sufficient statistics with covariates) *For all $g \in \{1,\ldots,G\}$, $E[\Lambda_{1,g}|\mathbf{D}_g, \mathbf{Z}^a_g] = E[\Lambda_{1,g}|\boldsymbol{D}^a_g, \mathbf{Z}^a_g]$.*

Assumption 5' (Design restriction with covariates) *For all $g \in \{1,\ldots,G\}$, the matrix $E[(W_g\, \boldsymbol{Z}^a_g)'\Pi(X_g)(W_g\, \boldsymbol{Z}^a_g)]$ is non-singular.*

Note that the generalized Roy model (4) remains compatible with Assumption 4' as Proposition 1 still applies. The only difference is that some components of $\boldsymbol{Z}_g$ are observed by the econometrician. Our main identifying result also extends in a straightforward way to this set-up with covariates.

Theorem 2 *Suppose that Model (9) and Assumptions 1, 2'-4' hold. Then*

$$E[Y_g|\mathbf{D}^a_g, \boldsymbol{Z}_g] = W_g\delta_0 + \boldsymbol{Z}^a_g\lambda_0 + X_g\phi_0(\mathbf{D}^a_g, \boldsymbol{Z}_g). \tag{10}$$

Moreover, if Assumption 5' holds, δ_0, λ_0, ϕ_0 and $\Delta_{0,t}$ are respectively identified by

$$(\delta_0'\, \lambda_0')' = E[(W_g\, \boldsymbol{Z}^a_g)'\Pi(X_g)(W_g\, \boldsymbol{Z}^a_g)]^{-1}E[(W_g\, \boldsymbol{Z}^a_g)'\Pi(X_g)Y_g], \tag{11}$$

$$\phi_0 = E\left[(X_g'X_g)^{-1}X_g'(Y_g - (W_g\, \boldsymbol{Z}^a_g)(\delta_0'\, \lambda_0')')|M_g = 1\right], \tag{12}$$

$$\Delta_{0,t} = \phi_{0,2} + \delta_{0,2(t-1)}\mathbb{1}\{t > 1\}. \tag{13}$$

Based on these constructive identification results, we obtain estimators of average treatment effects on the movers that are very similar to those in Section 3.2. Details are omitted.

4.2 Group variation in average treatment effects

We have focused so far on average treatment effects. Actually, Equation (5) shows that for all movers g, we directly identify $\phi_0(\boldsymbol{D}_g^a)$ by

$$\phi_0(\boldsymbol{D}_g^a) = (X_g'X_g)^{-1}E\left[Y_g - W_g\delta_0|\boldsymbol{D}_g^a\right]. \tag{14}$$

Recall that the second component of $\phi_0(\boldsymbol{d}^a)$ is $E\left[\Lambda_{1,g}|\boldsymbol{D}_g^a = \boldsymbol{d}^a\right]$, the average treatment effect of groups with a trajectory of treatment rates equal to $\boldsymbol{d}^a$. Equation (14) implies that we identify these conditional average treatment, and thus their distribution. In particular, we identify

$$V\left[\phi_0(\boldsymbol{D}_g^a)|M_g = 1\right] = V\left[(X_g'X_g)^{-1}E\left(Y_g - W_g\delta_0|\boldsymbol{D}_g^a\right)|M_g = 1\right].$$

This variance is appealing for at least two reasons. First, we have

$$V\left[\Lambda_{1,g}|M_g = 1\right] \geq V\left[E\left(\Lambda_{1,g}|\boldsymbol{D}_g^a = \boldsymbol{d}^a\right)\right].$$

Therefore, $V\left[\phi_0(\boldsymbol{D}_g^a)|M_g = 1\right]$ is a lower bound on the variance of group-level average treatment effects. Second, apart from the time trend μ_t in the average treatment effects, it is the heterogeneity of $\phi_0(\boldsymbol{D}_g^a)$, rather than that of $\Lambda_{1,g}$ conditional on $\boldsymbol{D}_g^a$, that is responsible for the bias of the estimator of Δ_t obtained by the two-way fixed effect regression (see de Chaisemartin and D'Haultfœuille, 2020).

5 Application to the impact of radio programs on the rise of Nazism

5.1 Set-up and descriptive statistics

In this section, we revisit Adena et al. (2015), who investigate the impact of biased radio programs in Germany on votes for the Nazi party in the period 1928-1933. To this end, Adena et al. (2015) use votes and radio subscription rate at the district $\times$ election level and two-way fixed effect regressions. They exploit the fact that subscription rates are quite heterogeneous over districts and increase substantially during the period 1928-1933.

As Adena et al. (2015) stress, the political content of radio broadcasts changed dramatically over time. At first, radio programs were completely apolitical (see Panel A, Figure 1 in Adena et al., 2015). But in 1929, the Weimar government decided to include political news with a pro-government slant. The Nazis and the communists were denied airtime, unlike other political

parties. This all changed in January 1933, when Hitler was appointed chancellor and gained control over radio. Radio broadcast turned from having no Nazi messages to airing pro-Nazi propaganda. To account for these changes, Adena et al. (2015) define their treatment as, basically "radio broadcast with a pro-Nazi inclination". Then, $D_{i,g,t}$ is equal to 1 if i in district g at t has access to radio with a pro-Nazi slant (as in 1933), 0 if she does not have access to radio or the programs are politically neutral (as in 1928) and -1 if she has access to radio and programs are anti-Nazi (as between 1929 and 1932).

This definition implies strong restrictions on treatment effects. If having access to radio reduces the probability of voting for the Nazi party by 5% in 1929-1932, say, the model predicts that it should increase the probability of voting for that party by the same 5% in 1933. As our model allows for unconstrainted time-varying treatment effects, we simply define $D_{i,g,t}$ to be 1 if i in district g has access to radio at t, 0 otherwise.

The subscription rates at the district level are not observed precisely on the election dates: they are only available in the month of April for 1931, 1932 and 1933, whereas election dates are May 1928, September 1930, July and November 1932 and March 1933. Adena et al. (2015) construct a predicted value of $D^a_{g,t}$ based in particular on local radio signal strength. However, their predicted value does not match well the overall increase in subscription rates observed over the period; see the second and third columns of Table 1 and Figure II in Adena et al. (2015). This prediction also leads to an important reduction in the standard deviation of subscription rates between districts.

We thus compute another predicted subscription rate as follows. First, we interpolate the radio subscription rates for the elections of July 1932, November 1932 and March 1933 based on the district-level quadratic model of radio subscription rates on time that fits the 3 observed rates of April 1931, April 1932 and April 1933. When only two observations are available, we interpolate these subscription rates using the district-level linear model that fits the two known values. Second, as the national subscription rate exhibits a linear time trend, we extrapolate the radio subscription rates for the elections of May 1928 and September 1930 from the linear model $D^a_{g,t} = \alpha_g + \beta_g t + u_t$ at the district level, using all observed subscription rates.

Table 1 presents some descriptive statistics on the observed subscription rates, the predictions obtained by Adena et al. (2015) and our own predictions. Our predicted subscription rate in March 1933 is much closer than Adena et al. (2015)s prediction to the actual subscription rate in April 1933. Table 1 also shows the average and dispersion of the outcome, which is the vote share for the Nazi party at the five parliamentary elections between 1928 and 1933.

Table 1: Descriptive Statistics

Date	Radio Subscription Rates			Vote Share for the Nazi Party
	Observed	Adena et al. (2015)	Ours	
May 1928		18.51	8.09	3.16
		(3.60)	(10.37)	(3.99)
September 1930		18.85	16.21	18.96
		(3.75)	(8.61)	(8.81)
April 1931	18.74			
	(8.25)			
April 1932	22.18			
	(8.27)			
July 1932		22.25	23.18	39.30
		(2.96)	(8.30)	(14.35)
November 1932		22.22	24.58	35.19
		(2.52)	(8.42)	(13.34)
March 1933		22.94	26.05	47.20
		(2.68)	(8.74)	(12.20)
April 1933	26.43			
	(8.87)			

G = 850. Notes: April 1931, 1932 and 1933 are the three dates where subscription rates are observed at the district level. The other five dates are election dates. Adena et al. (2015) estimate subscription rates on these dates using a fitting of subscription rates by local radio signal strength. We use district-specific models with linear time trends. Standard deviations over district are under parentheses.

We also slightly depart from Adena et al. (2015) in the way we include covariates. Adena et al. (2015) consider a rich specification by including 115 covariates in their TWFE regressions (see Appendix B for the whole list of such variables). As a result, the matrix

$$\sum_{g=1}^{G} (W_g \, \boldsymbol{Z}_g^a)' \Pi(X_g) (W_g \, \boldsymbol{Z}_g^a)/G$$

is singular and the identification condition for our set-up with covariates (Assumption 5') does not hold on our sample. To solve this issue, we select the most relevant covariates using the dou-

ble selection procedure of Belloni et al. (2014). Basically, the idea is to keep the most important control variables (in terms of their correlation with either the outcome or the treatment), while ensuring that the matrix above is non-singular. We refer to Appendix B for more details on the procedure and the subset of covariates that are selected by it.

5.2 Results

We first present the results of different TWFE regressions that we consider in Table 2. First, we replicate the results of Adena et al. (2015) in Column (I).[1] The treatment coefficient implies that having access to radio with anti-Nazi propaganda decreases the probability of voting for the Nazi party by 12.3% from 1929 to 1932, and then increases this probability by the same amount in 1933. This specification assumes away any effect in 1928. We then consider the same specification, but controlling for the radio subscription rate in 1928. This may be seen as a placebo test: given the absence of political programs at the radio that year, we expect the coefficient to be 0. The coefficient is not significant but note that it is quite large, around 1.5 as large as the coefficient of the treatment. Next, we consider the same two regressions, but with our predicted treatment rates, and with the subset of covariates obtained through the double selection procedure mentioned above rather than those used by Adena et al. (2015). We again find a positive and significant coefficient, though its magnitude is smaller. Also, the coefficient of the subscription rate in 1928 is now significant at the 10% level in Column (IV). Note that with these new specifications, only 850 districts are included, rather than 959. This is because the radio subscription rates are missing for 109 districts. As Adena et al. (2015) rely on a model based on radio signal strength, which is known for every district from 1928 to 1933, to predict subscription rates, they can still include these districts in their analysis, unlike us. We checked that the specification of Adena et al. (2015) gives similar results on this subsample: Colum (V) shows that results are very similar to the original ones.

[1] We were able to replicate their point estimate, though we obtain a larger standard error.

Table 2: Effects of radio on voting for the Nazis - TWFE regressions.

	Adena et al. (2015)	(I)	(II)	(III)	(IV)	(V)
Radio subscr. rates	0.123***	0.123***	0.092*	0.054**	0.046**	0.115***
	(0.027)	(0.042)	(0.047)	(0.018)	(0.020)	(0.043)
Radio subscr. rates in 1928			0.137		0.057	
			(0.111)		(0.034)	
Subscr. rates and cov.	Adena et al.	Adena et al.	Adena et al.	Ours	Ours	Adena et al.
G	959	959	959	850	850	850
R^2	0.972	0.991	0.991	0.992	0.992	0.992
Share of negative weights		63%				
Sum of negative weights		-3.145				

Notes: the first column is from Adena et al. (2015), Table 3. (I) replicates their specification. (II) adds to that specification subscription rates in 1928. Column (III) (resp. (IV)) is like (I) (resp. (II)) but with our own subscription rates and subset of covariates. Column (V) is Adena et al. (2015)'s specification using our restricted set of districts. Standard errors (under parentheses) are clustered at the electoral region level. * p < .1, ** p < .05, *** p < .01.

However, the results above may be biased, because of spatial or temporal heterogeneity in the treatment effects. We follow de Chaisemartin and D'Haultfœuille (2020) and compute the weights that the TWFE estimator in Adena et al. (2015)'s specification assigns to each district × elections average treatment effects. We estimate that 63% of district × election cell receive negative weights and negative weights sum to -3.145. The risk of a bias is thus a major concern here. There is no reason to assume that the effect of radio subscription would be exactly the opposite between 1930-1932 and 1933. We thus consider our model, where heterogeneity in the treatment effects is much less restricted. The results are displayed in Table 3. None of the treatment effects are significant. However, we do observe a large and statistically significant evolution in favor of Nazis of the average treatment effects in 1933 compared to 1928, with a point estimate for $\widehat{\mu}_{03/1933}$ around 19.3%. Another interesting finding is that the anti-Nazi

propaganda that took place between 1929 and 1932 did not seem to have much impact: none of the $\widehat{\mu}_t$ for t before 1933 are statistically significant. Finally, and reassuringly, the coefficient for 1928 is not significant at all usual level. But note that its sign is negative, contrary to what we obtained in Table 2. Accounting for possible heterogeneity in treatment effects thus had an important impact on the results.

Table 3: Effects of radio on voting for the Nazis - our estimates

Parameter	1928	1930	07/1932	11/1932	1933
$\widehat{\Delta}_t$	-0.093	-0.087	-0.027	-0.039	0.100
	(0.118)	(0.139)	(0.154)	(0.153)	(0.160)
$\widehat{\mu}_t$	0	0.006	0.067	0.054	0.193**
	–	(0.038)	(0.066)	(0.070)	(0.084)

Notes: $G = 850$. We use our estimator with the covariates selected by Belloni et al. (2014)'s double selection procedure. Standard errors, under parentheses, are clustered at the electoral region level. * p < .1, ** p < .05, *** p < .01.

References

Adena, M., Enikolopov, R., Petrova, M., Santarosa, V. and Zhuravskaya, E. (2015), 'Radio and the rise of the nazis in prewar germany', *The Quarterly Journal of Economics* **130**(4), 1885–1939.

Arellano, M. and Bonhomme, S. (2012), 'Identifying distributional characteristics in random coefficients panel data models', *The Review of Economic Studies* **79**(3), 987–1020.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**(2), 608–650.

Borusyak, K. and Jaravel, X. (2017), Revisiting event study designs. Working Paper.

Borusyak, K., Jaravel, X. and Spiess, J. (2021), Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419.

Callaway, B. and Sant'Anna, P. H. (2021), 'Difference-in-differences with multiple time periods', *Journal of Econometrics* **225**(2), 200–230.

Chamberlain, G. (1992), 'Efficiency bounds for semiparametric regression', *Econometrica: Journal of the Econometric Society* pp. 567–596.

de Chaisemartin, C. and D'Haultfœuille, X. (2021*a*), Difference-in-differences estimators of intertemporal treatment effects. arXiv preprint arXiv:2007.04267.

de Chaisemartin, C. and D'Haultfœuille, X. (2021*b*), Two-way fixed effects regressions with several treatments. arXiv preprint arXiv:2012.10077.

de Chaisemartin, C. and D'Haultfœuille, X. (2022), Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. arXiv preprint arXiv:2112.04565.

de Chaisemartin, C. and D'Haultfœuille, X. (2018), 'Fuzzy differences-in-differences', *The Review of Economic Studies* **85**(2), 999–1028.

de Chaisemartin, C. and D'Haultfœuille, X. (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**, 2964–2996.

Enikolopov, R., Petrova, M. and Zhuravskaya, E. (2011), 'Media and political persuasion: Evidence from russia', *American Economic Review* **101**(7), 3253–3285.

Goodman-Bacon, A. (2021), 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics* pp. 254–277.

Graham, B. S. and Powell, J. L. (2012), 'Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models', *Econometrica* **80**(5), 2105–2152.

Sun, L. and Abraham, S. (2021), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* **225**(2), 175–199.

A Proofs of the results

A.1 Proposition 1

Let us define $\Delta^n_{g,t} = E[\Delta_{i,g,t}|Z_{g,t}, C_{g,t}] - C_{g,t}$, so that

$$D_{i,g,t} = \mathbb{1}\left\{\Delta^n_{g,t} - \eta_{i,g,t} \geq 0\right\}. \tag{15}$$

Since the $(\eta_{i,g,t})_{i,t}$ are independent of $\boldsymbol{\Delta}^n_g$ and i.i.d., we have

$$P\left(D_{1,g,1} = d_{1,1}, \ldots, D_{N_{g,T},g,T} = d_{N_{g,T},T}|\boldsymbol{\Delta}^n_g\right) = \prod_{t=1}^{T}\left[F(\Delta^n_{g,t})^{\sum_{i=1}^{N_{g,t}} d_{i,t}}\left(1 - F(\Delta^n_{g,t})\right)^{N_{g,t} - \sum_{i=1}^{N_{g,t}} d_{i,t}}\right],$$

where F is the cumulative distribution function of $\eta_{i,g,t}$. This implies that

$$\boldsymbol{D}_g \perp\!\!\!\perp \boldsymbol{\Delta}^n_g|\boldsymbol{D}^a_g. \tag{16}$$

Now, conditional on $(\boldsymbol{Z}_g, \boldsymbol{C}_g)$, $\boldsymbol{D}_g$ is a function of $\boldsymbol{\eta}_g$. Thus, by the independence assumption,

$$\Lambda_{i,g} \perp\!\!\!\perp \boldsymbol{D}_g|\boldsymbol{Z}_g, \boldsymbol{C}_g. \tag{17}$$

As a result,

$$\begin{aligned}
E\left[\Lambda_{i,g}|\boldsymbol{D}_g\right] &= E\left[E\left[\Lambda_{i,g}|\boldsymbol{Z}_g, \boldsymbol{C}_g\right]|\boldsymbol{D}_g\right] \\
&= E\left[\Delta^n_{g,t} - \mu_t - E[\zeta_{i,g,t}|\boldsymbol{Z}_g, \boldsymbol{C}_g]|\boldsymbol{D}_g\right] \\
&= E\left[\Delta^n_{g,t} - \mu_t|\boldsymbol{D}_g\right] \\
&= E\left[\Delta^n_{g,t} - \mu_t|\boldsymbol{D}^a_g\right].
\end{aligned}$$

The first equality follows by the law of iterated expectations and (17). The second equality follows by Model (1). The third holds since $\zeta_{i,g,t}$ is independent of $(\boldsymbol{Z}_g, \boldsymbol{C}_g)$. The fourth follows by (16). The last equality implies that $E\left[\Lambda_{i,g}|\boldsymbol{D}_g\right]$ only depends on $\boldsymbol{D}^a_g$. Therefore, $E\left[\Lambda_{i,g}|\boldsymbol{D}_g\right] = E\left[\Lambda_{i,g}|\boldsymbol{D}^a_g\right]$, which proves that Assumption 4 holds.

A.2 Theorem 1

First, we have for all $(i, g, t) \in \{1, \ldots, N_{g,t}\} \times \{1, \ldots, G\} \times \{1, \ldots, T\}$,

$$\begin{aligned}
E[Y_{i,g,t}|\boldsymbol{D}^a_g] =& \beta_t + E[\alpha_{i,g}|\boldsymbol{D}^a_g] + E[D_{i,g,t}(\Lambda_{i,g} + \mu_t)|\boldsymbol{D}^a_g] \\
=& \beta_t + E[\alpha_{1,g}|\boldsymbol{D}^a_g] + E[D_{1,g,t}(\Lambda_{1,g} + \mu_t)|\boldsymbol{D}^a_g] \\
=& \beta_t + E[\alpha_{1,g}|\boldsymbol{D}^a_g] + E[D_{1,g,t}|\boldsymbol{D}^a_g]\left(E[\Lambda_{1,g}|\boldsymbol{D}^a_g] + \mu_t\right) \\
=& \beta_t + E[\alpha_{1,g}|\boldsymbol{D}^a_g] + D_{g,t}\left(E[\Lambda_{1,g}|\boldsymbol{D}^a_g] + \mu_t\right).
\end{aligned}$$

The first equality follows by Model (1) and the exchangeability condition in Assumption 3. The second uses Assumption 2. The third follows by Assumption 4. The last follows again by exchangeability. Hence,

$$E[Y_{g,t}|\boldsymbol{D}_g^a] = \beta_t + D_{g,t}\mu_t + E[\alpha_{1,g}|\boldsymbol{D}_g^a] + D_{g,t}E[\Lambda_{1,g}|\boldsymbol{D}_g^a].$$

We obtain (5) by stacking the equations above over t.

Next, consider (6). We have

$$\begin{aligned} E[W_g'\Pi(X_g)Y_g|\mathbf{D}_g] =& W_g'\Pi(X_g)W_g\delta_0 + W_g'\Pi(X_g)X_g\phi_0(\boldsymbol{D}_g) \\ =& W_g'\Pi(X_g)W_g\delta_0. \end{aligned}$$

The first equality follows by (5) and since $W_g'\Pi(X_g)$ is a function of $\boldsymbol{D}_g$. The second equality holds because $\Pi(X_g)X_g = 0$. Equation (6) follows by the law of iterated expectations and the fact that $E[W_g'\Pi(X_g)W_g]$ is nonsingular. Equation (7) then follows from (5), (6) and $\phi_0 = E[\phi_0(\boldsymbol{D}_g)|M_g = 1]$. Finally, Equation (8) follows by Model (1) and the normalization $\mu_1 = 0$.

A.3 Proof of Proposition 2

1. If $M_g = 1$, X_g is invertible and thus $\Pi(X_g) = 0$. Hence,

$$E\left[W_g'\Pi(X_g)W_g\right] = E\left[W_g'\Pi(X_g)W_gM_g\right].$$

Now, some algebra shows that if $M_g = 1$,

$$\Pi(X_g) = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Then, we obtain

$$\begin{aligned} E\left[W_g'\Pi(X_g)W_gM_g\right] =& \frac{1}{2}\begin{pmatrix} E[1-M_g] & E[D_{g,1}(1-M_g)] \\ E[D_{g,1}(1-M_g)] & E[D_{g,1}^2(1-M_g)] \end{pmatrix} \\ =& \frac{P(M_g=0)}{2}\begin{pmatrix} 1 & E[D_{g,1}|M_g=0] \\ E[D_{g,1}|M_g=0] & E[D_{g,1}^2|M_g=0] \end{pmatrix}. \end{aligned}$$

This matrix is nonsingular if and only if $P(M_g = 0) > 0$ and the determinant of the last matrix is not 0, which is equivalent to $V(D_{g,1}|M_g = 0) > 0$.

2. Let's suppose that $\det(X_g'X_g) \neq 0$ so that $\Pi(X_g) = I_T - X_g(X_g'X_g)^{-1}X_g$. Then, one can show

that for all $(i,j)\{1,\ldots,T\}^2$, $i \neq j$,

$$\Pi(X_g)_{i,i} = \sum_{\substack{k=2 \\ k\neq i}}^{T} \sum_{\substack{l=1 \\ l\neq i}}^{l-1} (D_l - D_k)^2 / \det(X_g' X_g)$$

$$\Pi(X_g)_{i,j} = \sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{T} (D_i - D_k)(D_j - D_k) / \det(X_g' X_g)$$

where $\Pi(X_g)_{i,j}$ denotes the component on the ith line and jth column of matrix $\Pi(X_g)$. Then, since $W_g'\Pi(X_g)W_g = (\Pi(X_g)_{.,2}\, D_2\Pi(X_g)_{.,2} \ldots \Pi(X_g)_{.,T}\, D_T\Pi(X_g)_{.,T})$ where for all $i \in \{2,\ldots,T\}$, $\Pi(X_g)_{.,i}$ is the ith column of matrix $\Pi(X_g)$, the sufficient condition arises straightly. When $\det(X_g'X_g) = 0$, the condition always holds unless $D = 1$ or $D = 0$.

A.4 Proposition 3

First, it follows from (6) that

$$\widehat{\delta} = \delta_0 + \left(\frac{1}{G}\sum_{g=1}^{G} W_g'\Pi(X_g)W_g\right)^{-1} \left(\frac{1}{G}\sum_{g=1}^{G} W_g'\Pi(X_g)\varepsilon_g\right),$$

where $\varepsilon_g = Y_g - W_g\delta_0 - X_g\phi_0(\boldsymbol{D}_g)$. Thus, by the central limit and Slutsky's theorems,

$$\sqrt{G}\left(\widehat{\delta} - \delta_0\right) = \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\psi_g + o_p(1), \tag{18}$$

where $\psi_g = (\psi_{1,g},\ldots,\psi_{2(T-1),g})' := E\left(W_g'\Pi(X_g)W_g\right)^{-1} W_g'\Pi(X_g)\varepsilon_g$. Next, let

$$\widetilde{\phi}^h = \frac{1}{G_h}\sum_{g:M_{g,h}=1} (X_g'X_g)^{-1}X_g'(Y_g - W_g\delta_0),$$

so that

$$\widehat{\phi}^h = \widetilde{\phi}^h - \left[\frac{1}{G_h}\sum_{g:M_{g,h}=1} (X_g'X_g)^{-1}X_g'W_g\right]\left(\widehat{\delta} - \delta_0\right). \tag{19}$$

The delta method and some algebra show that

$$\sqrt{G}\left(\widetilde{\phi}^h - \phi_0^h\right) = \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\chi_g + o_p(1),$$

where $\chi_g = M_{g,h}\left[(X_g'X_g)^{-1}X_g'(Y_g - W_g\delta_0) - \phi_0^h\right]/E[M_{g,h}]$. Combined with (18), (19) and Slutsky's theorem, this yields

$$\sqrt{G}\left(\widehat{\phi}^h - \phi_0^h\right) = \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\nu_g + o_p(1),$$

where $\nu_g = (\nu_{1,g}, \nu_{2,g})' := \chi_g - E[(X_g'X_g)^{-1}X_g'W_g|M_{g,h} = 1]\psi_g$. Finally, we obtain

$$\sqrt{G}\left(\widehat{\Delta}_t^h - \Delta_{0,t}^h\right) = \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\omega_{g,t} + o_p(1),$$

where

$$\omega_{g,t} = \nu_{2,g} + \psi_{2(t-1),g}\mathbb{1}\left\{t > 1\right\}. \tag{20}$$

The proposition follows by the central limit and Slutsky's theorems.

A.5 Theorem 2

First, we have for all $(i, g, t) \in \{1, \ldots, N_{g,t}\} \times \{1, \ldots, G\} \times \{1, \ldots, T\}$,

$$\begin{aligned}
E[Y_{i,g,t}|\boldsymbol{D}_g^a, \mathbf{Z}_g] =& \beta_t + Z_{g,t}\lambda_0 + E[\alpha_{i,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + E[D_{i,g,t}(\Lambda_{i,g} + \mu_t)|\boldsymbol{D}_g^a, \mathbf{Z}_g] \\
=& \beta_t + Z_{g,t}\lambda_0 + E[\alpha_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + E[D_{1,g,t}(\Lambda_{1,g} + \mu_t)|\boldsymbol{D}_g^a, \mathbf{Z}_g] \\
=& \beta_t + Z_{g,t}\lambda_0 + E[\alpha_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + E[D_{1,g,t}|\boldsymbol{D}_g^a, \mathbf{Z}_g]\left(E[\Lambda_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + \mu_t\right) \\
=& \beta_t + Z_{g,t}\lambda_0 + E[\alpha_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + D_{g,t}\left(E[\Lambda_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + \mu_t\right).
\end{aligned}$$

The first equality follows by Model (9) and the exchangeability condition in Assumption 3'. The second uses Assumption 2'. The third follows by Assumption 4'. The last follows again by exchangeability. Hence,

$$E[Y_{g,t}|\boldsymbol{D}_g^a, \mathbf{Z}_g] = \beta_t + D_{g,t}\mu_t + Z_{g,t}\lambda_0 + E[\alpha_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g] + D_{g,t}E[\Lambda_{1,g}|\boldsymbol{D}_g^a, \mathbf{Z}_g].$$

We obtain (10) by stacking the equations above over t.

Next, consider (11). We have

$$\begin{aligned}
E[(W_g\, Z_g)'\Pi(X_g)Y_g|\mathbf{D}_g, \mathbf{Z}_g] =& (W_g\, \boldsymbol{Z}_g^a)'\Pi(X_g)(W_g\, \boldsymbol{Z}_g^a)(\delta_0'\, \lambda_0')' + (W_g\, \boldsymbol{Z}_g^a)'\Pi(X_g)X_g\phi_0(\boldsymbol{D}_g, \mathbf{Z}_g) \\
=& (W_g\, \boldsymbol{Z}_g^a)'\Pi(X_g)(W_g\, \boldsymbol{Z}_g^a)(\delta_0'\, \lambda_0')'.
\end{aligned}$$

The first equality follows by (10) and since $(W_g\, \boldsymbol{Z}_g^a)'\Pi(X_g)$ is a function of $(\boldsymbol{D}_g, \mathbf{Z}_g)$. The second equality holds because $\Pi(X_g)X_g = 0$. Equation (11) follows by the law of iterated expectations and the fact that $E[(W_g\, Z_g)'\Pi(X_g)(W_g\, Z_g)]$ is nonsingular. Equation (12) then follows from (10), (11) and $\phi_0 = E[\phi_0(\boldsymbol{D}_g, \mathbf{Z}_g)|M_g = 1]$. Finally, Equation (13) follows by Model (9) and the normalization $\mu_1 = 0$.

B Additional details on the application

The control variables considered by Adena et al. (2015) include first socio-demographic characteristics. These are: a fifth order polynomial of population, the share of Jewish and Catholic

people in 1925, the share of workers in white- and blue-collar occupations in 1925, the share of unemployed and partially employed people in 1933, the number of World War I participants per 1,000 inhabitants in 1925, the number of social housing renters per 1,000 inhabitants in 1925, the number of welfare recipients per 1,000 inhabitants in 1925 and the logarithm of the average property tax in 1930. Second, variables related to preexisting political preferences are included: the shares of votes for the DNVP and the NSFB nationalistic parties, the shares of votes for the Zentrum and SPD non-nationalistic parties in the 1924 Parliamentary election and turnout. Finally, Adena et al. (2015) control for the determinants of radio transmitters location by adding the average altitude of the district, a dummy for city status of the district and the distance to the closest city with at least 50,000 inhabitants. All the variables above are interacted with time.

As the total number of covariates amounts to 115, we implement the double selection procedure from Belloni et al. (2014) to select the most relevant ones. In a first step, the treatment (radio subscription rate) is regressed on the whole set of control variables (including time and district dummies) using a Lasso. The variables whose estimated coefficient is different from zero are kept. In a second step, the outcome (vote share for the Nazi party) is regressed on the whole set of control variables (including time and district dummies) also using a Lasso. Once again, the variables whose associated estimated coefficient is different from zero are kept. Finally, vote share for the Nazi party is regressed on radio subscription rate and all the selected variables, that is to say variables that were kept in at least one of the two previous steps, using OLS. We adapt this last step by keeping all time and electoral districts dummies in the regression. The selected variables are presented in Table 4. Preexisting political preferences and socio-economic characteristics seem to be the most relevant determinants of support for the Nazi Party and/or radio exposure.

Table 4: Selected Covariates from Belloni et al. (2014)'s Procedure

Elections	**Selected Control Variables**
May 1928	Vote share for the Zentrum party in the 1924 parliamentary election Vote share for the DNVP party in the 1924 parliamentary election Share of blue-collar workers in 1925, Turnout
September 1930	Vote share for the Zentrum party in the 1924 parliamentary election Vote share for the DNVP party in the 1924 parliamentary election Vote share for the NSFB party in the 1924 parliamentary election Share of the Catholic population in 1925 Share of the blue-collar workers in 1925 Altitude, Turnout
July 1932	Vote share for the NSFB party in the 1924 parliamentary election Vote share for the DNVP party in the 1924 parliamentary election Share of the Catholic population in 1925, Share of white-collar workers in 1925 Logarithm of the average property tax in 1930, Altitude
November 1932	Vote share for the Zentrum party in the 1924 parliamentary election Vote share for the NSFB party in the 1924 parliamentary election Vote share for the DNVP party in the 1924 parliamentary election Share of the Catholic population in 1925 Share of white-collar workers in 1925, City dummy Share of unemployed people in 1933 Logarithm of the average property tax in 1930
March 1933	Vote share for the Zentrum party in the 1924 parliamentary election Vote share for the NSFB party in the 1924 parliamentary election Vote share for the DNVP party in the 1924 parliamentary election Share of the Catholic population in 1925 Share of white-collar workers in 1925 City dummy, Altitude Share of unemployed people in 1933 Logarithm of the average property tax in 1930