Supplementary Appendix: Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment

Brantly Callaway* Pedro H. C. Sant'Anna†

March 23, 2018

This supplementary appendix contains results for (1) the case where a researcher has access to repeated cross sections data rather than panel data; (2) additional details on group-time average treatment effects under an unconditional parallel trends assumption, paying particular attention to the possibilities of using regressions to estimate group-time average treatment effects; and (3) extending our results to using "not yet treated" observations as an additional control group. Appendix A (contained in the main text of the paper) contains the proofs of the main results, so we start the supplementary appendix with Appendix B.

Appendix B: Additional Results for Repeated Cross Sections

It is also fairly straightforward to extend our approach to the case with repeated cross sections data instead of panel data. Here we assume that for each individual we observe $(Y, G_1, \ldots, G_{\mathcal{T}}, C, T, X)$ where $T \in \{1, \ldots, \mathcal{T}\}$ denotes the time period when that individual is observed. We also define T_t to a dummy variable that is equal to 1 for observations in period t and 0 otherwise, and let $\lambda_t = P(T_t = 1)$ – which is the probability of a draw being in period t in the repeated cross sections data. Also, let $\lambda = 1/\mathcal{T}$ – the fraction of individuals in each period in the population. Then, our

*Department of Economics, Temple University. Email: brantly.callaway@temple.edu
†Department of Economics, Vanderbilt University. Email: pedro.h.santanna@vanderbilt.edu

sample consists of random draws from the mixture distribution

$$F_M(y, g_1, \ldots, g_{\mathcal{T}}, c, t, x) = \sum_{t=1}^{\mathcal{T}} \lambda_t F_{Y,G_1,\ldots,G_{\mathcal{T}},C,X|T}(y, g_1, \ldots, g_{\mathcal{T}}, c, x|t)$$

Thus, the sample sizes can be different across periods. And notice that, once one conditions on the time period, then expectations under the mixture distribution correspond to population expectations. Also, because X, G_g, and C do not change over time $P(G_g = 1|X, G_g + C = 1) = P_M(G_g = 1|X, G_g + C = 1)$ (i.e. one can directly use draws from the mixture distribution to estimate the population version of the generalized propensity score even in the case where the sample sizes change across periods). Using similar arguments, $E[G_g] = E_M[G_g]$ and $E[(p_g(X)C)/(1-p_g(X))] = E_M[(p_g(X)C)/(1-p_g(X))]$. Notice, however, that this argument does not hold for expectations involving Y. Then, using the similar arguments as in Theorem 1, one can show that

$$ATT\left(g,t\right) = \mathbb{E}_M\left[\lambda\left(\frac{T_t}{\lambda_t} - \frac{T_{g-1}}{\lambda_{g-1}}\right)\left(\frac{G_g}{\mathbb{E}_M\left[G_g\right]} - \frac{\frac{p_g\left(X\right)C}{1-p_g\left(X\right)}}{\mathbb{E}_M\left[\frac{p_g\left(X\right)C}{1-p_g\left(X\right)}\right]}\right)Y\right]$$

We do not provide a formal proof but the idea for the above result is the following. Let

$$\rho = \left(\frac{G_g}{\mathbb{E}_M\left[G_g\right]} - \frac{\frac{p_g\left(X\right)C}{1-p_g\left(X\right)}}{\mathbb{E}_M\left[\frac{p_g\left(X\right)C}{1-p_g\left(X\right)}\right]}\right)$$

Note that under the repeated cross sections sampling scheme, these weights are exactly the same as before. Then, we can write the RHS of the above equation as

$$\begin{aligned} E_M\left[\lambda\left(\frac{T_t}{\lambda_t} - \frac{T_{g-1}}{\lambda_{g-1}}\right)\rho Y\right] &= E_M[\lambda\rho Y_t|T_t = 1] - E_M[\lambda\rho Y_{g-1}|T_{g-1} = 1] \\ &= E[\lambda\rho Y_t|T_t = 1] - E[\lambda\rho Y_{g-1}|T_{g-1} = 1] \\ &= E[\rho(Y_t - Y_{g-1})] \\ &= ATT(g,t) \end{aligned}$$

Based on the above results, estimation is relatively straightforward with repeated cross sections and similar to what we did in the case with panel data. One simply needs to adjust the weights slightly as above. One could also use a similar multiplier boostrap procedure for inference though the influence function we need to be derived again and would be somewhat different than in the panel case. We omit those details. Another option that seems likely to work without requiring developing any theory would be to use the empirical boostrap. The main cost of this approach

would be additional computational time.

Appendix C: Analysis with "Not yet Treated" as a Control Group

In this appendix, we discuss the case where one considers the "not yet treated" instead of the "never treated" as a control group. This case is particularly relevant in applications when eventually (almost) all units are treated, though the timing of the treatment differs across groups. To carry this analysis, we make the following assumptions.

Assumption C.1. $\{Y_{i1}, Y_{i2}, \dots Y_{i\mathcal{T}}, X_i, D_{i1}, D_{i2}, \dots, D_{i\mathcal{T}}\}_{i=1}^n$ *is independent and identically distributed (iid).*

Assumption C.2. *For all* $t = 2, \dots, \mathcal{T}$, $g = 2, \dots, \mathcal{T}$ *such that* $g \leq t$,

$$\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, D_t = 0] \ a.s..$$

Assumption C.3. *For* $t = 2, \dots, \mathcal{T}$,

$$D_t = 1 \text{ implies that } D_{t+1} = 1$$

Assumption C.4. *For all* $t = 2, \dots, \mathcal{T}$, $g = 2, \dots, \mathcal{T}$, $P(G_g = 1) > 0$ *and* $P(D_t = 1|X) < 1$ *a.s..*

Assumptions C.1 and C.3 are the same as Assumptions 1 and 3 in the main text. Assumptions C.2 and C.4 are the analogue of Assumptions 2 and 4, but using those "not yet treated" ($D_t = 0$) as a control group instead of the "never treated" ($C = 0$ or $D_{\mathcal{T}} = 0$). Note that Assumption C.4 rule out the case in which eventually everyone is treated; in these time periods, there is no "control group" available, and therefore the data itself is not informative about the average treatment effect when $D_t = 1$ *a.s.*. In these cases, one should concentrate their attention only to the time periods such that $P(D_t = 1|X) < 1$ *a.s.*.

Remember that

$$ATT_X(g, t) = \mathbb{E}[Y_t(1) - Y_t(0)|X, G_g = 1].$$

Next lemma states that, under Assumptions C.1-C.4, we can identify $ATT_X(g, t)$ for $2 \leq g \leq t \leq \mathcal{T}$. This is the analogue of Lemma A.1.

Lemma C.1. *Under Assumptions C.1-C.4, and for* $2 \leq g \leq t \leq \mathcal{T}$,

$$ATT_X(g, t) = \mathbb{E}[Y_t - Y_{g-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0] \ a.s..$$

Proof of Lemma C.1: In what follows, take all equalities to hold almost surely (a.s.). Notice that for identifying $ATT_X(g,t)$, the key term is $E[Y_t(0)|X, G_g = 1]$. And notice that for $h > s$, $E[Y_s(0)|X, G_s = 1] = E[Y_s|X, G_h = 1]$, which holds because in time periods before an individual is first treated, their untreated potential outcomes are observed outcomes. Also, note that, for $2 \leq g \leq t \leq \mathcal{T}$,

$$\begin{aligned}\mathbb{E}[Y_t(0)|X, G_g = 1] &= \mathbb{E}[\Delta Y_t(0)|X, G_g = 1] + \mathbb{E}[Y_{t-1}(0)|X, G_g = 1] \\ &= \mathbb{E}[\Delta Y_t|X, D_t = 0] + \mathbb{E}[Y_{t-1}(0)|X, G_g = 1],\end{aligned} \tag{C.1}$$

where the first equality holds by adding and subtracting $E[Y_{t-1}(0)|X, G_g = 1]$ and the second equality holds by Assumption C.2. If $g = t-1$, then the last term in the final equation is identified; otherwise, one can continue recursively in similar way to (C.1) but starting with $\mathbb{E}[Y_{t-1}(0)|X, G_g = 1]$. As a result,

$$\begin{aligned}\mathbb{E}[Y_t(0)|X, G_g = 1] &= \sum_{j=0}^{t-g} \mathbb{E}[\Delta Y_{t-j}|X, D_t = 0] + \mathbb{E}[Y_{g-1}|X, G_g = 1] \\ &= \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0] + \mathbb{E}[Y_{g-1}|X, G_g = 1].\end{aligned} \tag{C.2}$$

Combining (C.2) with the fact that, for all $g \leq t$, $\mathbb{E}[Y_t(1)|X, G_g = 1] = \mathbb{E}[Y_t|X, G_g = 1]$ (which holds because observed outcomes for group g in period t with $g \leq t$ are treated potential outcomes), implies the result. □

With the result of Lemma C.1 at hands, we proceed to show that the $ATT\,(g,t)$ is nonparametrically identified under Assumptions C.1 - C.4 and for $2 \leq g \leq t \leq \mathcal{T}$. The following Theorem is the analogue Theorem 1 .

Theorem 1. *Under Assumptions C.1 - C.4 and for $2 \leq g \leq t \leq \mathcal{T}$, the group-time average treatment effect for group g in period t is nonparametrically identified, and given by*

$$ATT\,(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\,[G_g]} - \frac{\frac{P\,(D_g = 1|X, D_{g-1} = 0)\,(1 - D_t)}{1 - P\,(D_t = 1|X)}}{\mathbb{E}\left[\frac{P\,(D_g = 1|X, D_{g-1} = 0)\,(1 - D_t)}{1 - P\,(D_t = 1|X)}\right]}\right)(Y_t - Y_{g-1})\right]. \tag{C.3}$$

Proof of Lemma 1: Given the result in Lemma C.1,

$$\begin{aligned}ATT(g,t) &= \mathbb{E}[ATT_X(g,t)|G_g = 1] \\ &= \mathbb{E}\left[\mathbb{E}[Y_t - Y_{g-1}|X, G_g = 1] - \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0]\middle| G_g = 1\right] \\ &:= \mathbb{E}[A_X|G_g = 1] - \mathbb{E}[B_X^{n.yet}|G_g = 1],\end{aligned}$$

and we consider each term separately. For the first term

$$\begin{aligned}\mathbb{E}[A_X|G_g=1] &= \mathbb{E}[Y_t - Y_{g-1}|G_g=1] \\ &= \mathbb{E}\left[\frac{G_g}{\mathbb{E}[G_g]}(Y_t - Y_{g-1})\right]. \end{aligned} \tag{C.4}$$

For the second term, by repetition of the law of iterated expectations, and noticing that under Assumption C.3,

$$P(G_g=1|X) = P(D_g=1|X, D_{g-1}=0) \ \ a.s.,$$

we have

$$\begin{aligned}\mathbb{E}[B_X^{n.yet}|G_g=1] &= \mathbb{E}\left[\mathbb{E}[Y_t - Y_{g-1}|X, D_t=0]\middle|G_g=1\right] \\ &= \mathbb{E}\left[G_g\mathbb{E}[(1-D_t)(Y_t - Y_{g-1})|X, D_t=0]\middle|G_g=1\right] \\ &= \mathbb{E}\left[G_g\mathbb{E}\left[\frac{(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\middle|X\right]\middle|G_g=1\right] \\ &= \mathbb{E}[G_g]^{-1}\mathbb{E}\left[G_g\mathbb{E}\left[\frac{(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\middle|X\right]\right] \\ &= \mathbb{E}[G_g]^{-1}\mathbb{E}\left[\mathbb{E}[G_g|X]\,\mathbb{E}\left[\frac{(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\middle|X\right]\right] \\ &= \mathbb{E}[G_g]^{-1}\mathbb{E}\left[\mathbb{E}\left[\frac{P(G_g=1|X)(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\middle|X\right]\right] \\ &= \mathbb{E}[G_g]^{-1}\mathbb{E}\left[\frac{P(G_g=1|X)(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\right] \\ &= \mathbb{E}[G_g]^{-1}\mathbb{E}\left[\frac{P(D_g=1|X, D_{g-1}=0)(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\right] \qquad \text{(C.5)} \\ &= \frac{\mathbb{E}\left[\frac{P(D_g=1|X, D_{g-1}=0)(1-D_t)}{1-P(D_t=1|X)}(Y_t - Y_{g-1})\right]}{\mathbb{E}\left[\frac{P(D_g=1|X, D_{g-1}=0)(1-D_t)}{1-P(D_t=1|X)}\right]}, \qquad \text{(C.6)}\end{aligned}$$

where (C.6) follows from

$$\begin{aligned}\mathbb{E}\left[\frac{P(D_g=1|X, D_{g-1}=0)(1-D_t)}{1-P(D_t=1|X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{P(G_g=1|X)(1-D_t)}{1-P(D_t=1|X)}\middle|X\right]\right] \\ &= \mathbb{E}\left[\frac{P(G_g=1|X)}{1-P(D_t=1|X)}\mathbb{E}\left[(1-D_t)\middle|X\right]\right] \\ &= \mathbb{E}\left[\frac{P(G_g=1|X)}{1-P(D_t=1|X)}(1-P(D_t=1|X))\right] \\ &= \mathbb{E}[P(G_g=1|X)] \\ &= \mathbb{E}[\mathbb{E}[G_g|X]]\end{aligned}$$

$$= \mathbb{E}\left[G_g\right].$$

The proof is completed by combining (C.4) and (C.6). □

Once we have establish nonparametric identification of $ATT(g,t)$, we can follow a similar two-step estimation strategy as described in Section 3 . More precisely, under Assumptions C.1 - C.4 and for $2 \leq g \leq t \leq \mathcal{T}$, one can estimate $ATT(g,t)$ by

$$\widehat{ATT}_{n.yet}(g,t) = \mathbb{E}_n\left[\left(\frac{G_g}{\mathbb{E}_n\left[G_g\right]} - \frac{\frac{\hat{p}_{D_g}\left(X, D_{g-1}=0\right)\left(1-D_t\right)}{1-\hat{p}_{D_t}\left(X\right)}}{\mathbb{E}_n\left[\frac{\hat{p}_{D_g}\left(X, D_{g-1}=0\right)\left(1-D_t\right)}{1-\hat{p}_{D_t}\left(X\right)}\right]}\right)\left(Y_t - Y_{g-1}\right)\right],$$

where $\hat{p}_{D_g}\left(X, D_{g-1}=0\right)$ is an estimate of $P\left(D_g = 1 | X, D_{g-1}=0\right)$, and $\hat{p}_{D_t}\left(X\right)$ is an estimate of $P\left(D_t = 1|X\right)$.

Following similar steps as in Theorems 2 and 3, one can show that under suitable regularity conditions akin to those in Assumption 5, $\widehat{ATT}_{n.yet}(g,t)$ is consistent and asymptotically normal, and that one can use a multiplier bootstrap similar to the one described in Algorithm 1 to conduct asymptotically valid inference. We omit the details for conciseness.

Appendix D: Additional Results for the Case without Covariates

Panel Data

The case where the DID assumption holds without conditioning on covariates is of particular interest. In this appendix, we briefly consider whether or not it is possible to obtain $ATT(g,t)$ using a regression approach like the two period - two group case. A natural starting point is the model

$$Y_{igt} = \alpha_t + c_i + \gamma_{gt} G_{igt} + u_{igt}$$

where α_t is a vector of time period fixed effects (we normalize α_1 and γ_{g1} to be equal to 1), c_i is time invariant unobserved heterogeneity that can be distributed differently across groups, and G_{igt} is is a dummy variable indicating whether or not individual i is a member group g and the time period is t. Differencing the model across time periods results in

$$\Delta Y_{igt} = \alpha_t + \gamma_{gt} G_{igt} + \Delta u_{igt}$$

Notice that this is a fully saturated model in group and time effects. It is straightforward to show that

$$\gamma_{gt} = E[\Delta Y_t | G_g = 1] - E[\Delta Y_t | C = 1]$$

When $g = t$, this is exactly the DID estimator. Under the unconditional version of the parallel trends assumption, γ_{gt} should be equal to 0 for all $g > t$, and it is straightforward to test this using output from standard regression software (e.g. Wald test). For $g < t$, the long difference estimate of $ATT(g,t)$ can be constructed by

$$\begin{aligned} ATT(g,t) &= E[Y_t - Y_{g-1} | G_g = 1] - E[Y_t - Y_{g-1} | C = 1] \\ &= \sum_{s=g}^{t} (E[\Delta Y_s | G_g = 1] - E[\Delta Y_s | C = 1]) \\ &= \sum_{s=g}^{t} \gamma_{gs} \end{aligned}$$

This implies that, under the unconditional parallel trends assumption, $ATT(g,t)$ can be recovered using a regression approach. However, combining the estimates of the parameters in this way does not seem to offer much convenience relative to simply computing the estimates directly using the main approach suggested in the paper. Thus, unlike the 2-period case, it does not appear that there is as exact of a mapping from a regression coefficient to a group-time average treatment effect.

Common Approaches to Pre-Testing in the Unconditional Case

Finally in this section, we consider the most common approach to pre-testing the Unconditional DID assumption is to run the following regression (see Autor et al. (2007) and Angrist and Pischke (2008)).

$$Y_{it} = \alpha_t + \theta_g + \beta_0 D_{it} + \sum_{j=1}^{q} \beta_j \Delta D_{it,t+j} + \epsilon_{ist} \tag{B.1}$$

where D_{it} is a dummy variable for whether or not individual i is treated in period t (notice that this is not whether they are *first treated* in period t but whether or not they are treated at all; it is a post-treatment dummy variable), $\Delta D_{it,t+j}$ is a j period lead for individual i who is first treated in period $t + j$. For example, when $t = 2$, $\Delta D_{i2,4} = 1$ (for $j = 2$) for individuals who are first treated in period 4, which indicates that the group of individuals first treated in period 4 will be treated 2 periods from period t.

Then, one can test the unconditional parallel trends assumption by testing if $\beta_j = 0$ for

$j = 1, \ldots, q$. Under the Unconditional DID Assumption, each β_j will be 0. One advantage of this approach is that it allows simple graphs of pre-treatment trends. However, it is possible for this approach to miss departures from the unconditional parallel trends assumption that our test would not miss.

Consider the case with four periods and three groups – the control group, a group first treated in period 4, and a group first treated in period 3. Also, consider the case with $q = 1$. It is straightforward to show that $\beta_1 = \mathbb{E}[\Delta Y_3 | G_4 = 1] - \mathbb{E}[\Delta Y_3 | C = 1]$ and $\beta_1 = \mathbb{E}[\Delta Y_2 | G_3 = 1] - \mathbb{E}[\Delta Y_1 | C = 1]$ so that the estimate of β_1 will be a weighted average of these two pre-trends. Thus, the unconditional parallel trends assumption could be violated in ways that offset each other leading to β_1 being equal to 0. Our approach, on the other hand, is robust to these types of violations of the unconditional parallel trends assumption.

References

Angrist, J. D., and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, : Princeton University Press.

Autor, D. H., Kerr, W. R., and Kugler, A. D. (2007), "Does Employment Protection Reduce Productivity? Evidence From US States," *The Economic Journal*, 117(521), F189–F217.