INFERENCE WITH "DIFFERENCE IN DIFFERENCES" WITH A SMALL NUMBER OF POLICY CHANGES

Timothy G. Conley and Christopher R. Taber*

Abstract—In difference-in-differences applications, identification of the key parameter often arises from changes in policy by a small number of groups. In contrast, typical inference assumes that the number of groups changing policy is large. We present an alternative inference approach for a small (finite) number of policy changers, using information from a large sample of nonchanging groups. Treatment effect point estimators are not consistent, but we can consistently estimate their asymptotic distribution under any point null hypothesis about the treatment. Thus, treatment point estimators can be used as test statistics, and confidence intervals can be constructed using test statistic inversion.

I. Introduction

THIS paper presents a new method of inference for difference-in-differences type fixed-effect regression methods for circumstances in which only a small number of groups provide information about treatment parameters of interest. In the difference-in-differences methodology, identification of the treatment parameter typically arises when a group changes some particular policy. We use N_1 to denote the number of treatment groups that change their policy in the data and N_0 to denote the number of control groups that do not change their policy. The usual asymptotic approximations assume that both N_1 and N_0 are large. However, even when the total number of observations is large, the number of actual policy changes observed in the data is often very small. For example, often only a few states change a law within the time span (T) of the data. In such cases, we argue that the standard large-sample approximations used for inference are not appropriate.[1] We develop an alternative approach to inference under the assumption that N_1 is finite, using asymptotic approximations that let N_0 grow large (with T fixed). Point estimators of the treatment effect parameter(s) are not consistent since N_1 and T are fixed. However, we can use information from the N_0 control groups to consistently estimate the distribution of these point estimators up to the true values of the parameter. This allows us to use treatment parameter point estimators as test statistics for any hypothesized true treatment parameter values and to construct confidence intervals by inverting these test statistics.

Received for publication March 28, 2008. Revision accepted for publication May 4, 2009.

* Conley: Booth School of Business, University of Chicago; Taber: University of Wisconsin–Madison.

We thank Federico Bandi, Alan Bester, Phil Cross, Chris Hansen, Rosa Matzkin, Bruce Meyer, Jeff Russell, and Elie Tamer for helpful comments and Aroop Chaterjee and Nathan Hendren for research assistantship. All errors are our own. T.C. gratefully acknowledges financial support from the NSF (SES 9905720) and from the IBM Corporation Faculty Research Fund at the University of Chicago Graduate School of Business. C.T. gratefully acknowledges financial support from the NSF (SES 0217032). Stata and Matlab code to implement the methods here can be found at the authors' websites.

[1] Of course in some special cases, the classical linear model assumptions will be satisfied, enabling small sample inference (see, e.g., Donald & Lang, 2007). Here our methods remain useful as specification checks, but they will be most valuable when the classical model may not be applicable.

The following simple model illustrates our basic problem and approach to its solution:

$$Y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt}, \tag{1}$$

where d_{jt} is a policy variable whose coefficient α is the object of interest, θ_j is a time-invariant fixed effect for group j, γ_t is a time fixed effect that is common across all groups but varies across time $t = 1, \ldots, T$, and η_{jt} is a group $\times$ time random effect.

Suppose that only the $j = 1$ group experiences a treatment change and that it happens to be a permanent one-unit change at period t^*. All other groups have a time-invariant policy: $d_{j1} = \ldots = d_{jT}$. Consider estimating model (1) by using ordinary least squares (OLS), controlling for group and time effects using dummy variables. Let $\widehat{\alpha}_{FE}$ be this regression estimate of α. It is straightforward to show that $\widehat{\alpha}_{FE}$ can be written as a difference of differences:

$$\begin{aligned}\widehat{\alpha}_{FE} = \alpha &+ \left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{1t} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{1t}\right] \\ &- \left(\frac{1}{(N_0)}\sum_{j=2}^{N_0+1}\frac{1}{(T-t^*)}\sum_{t=t^*+1}^{T}\eta_{jt}\right. \\ &\left. - \frac{1}{(N_0)}\sum_{j=2}^{N_0+1}\frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}\right).\end{aligned}$$

Under the usual assumption that η_{jt} has mean zero conditional on the regressors, $\widehat{\alpha}_{FE}$ is unbiased. However, it is not consistent. As the number of groups grows, only the term in parentheses vanishes; the term in brackets remains unchanged as N_0 gets large (with T fixed), that is

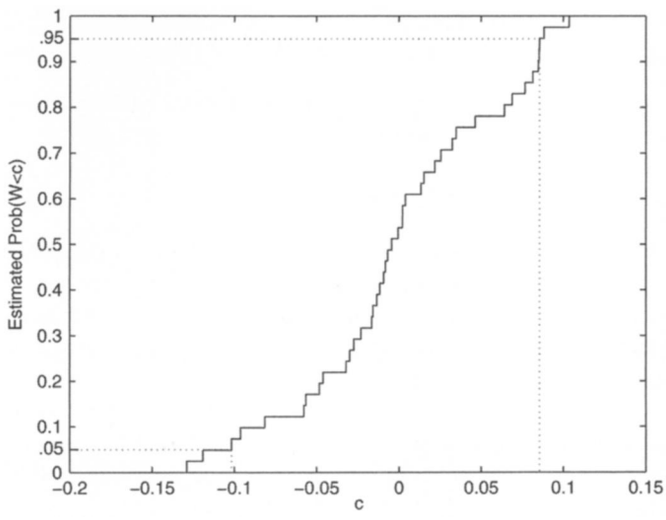
$$(\widehat{\alpha}_{FE} - \alpha) \xrightarrow{p} W \equiv \left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{1t} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{1t}\right].$$

In other words, the $\widehat{\alpha}_{FE}$ estimate is equal to the true parameter of interest α plus noise W. The key issue is that because T is fixed and the number of treatment groups is fixed at N_1, the noise W does not vanish as the total number of groups grows larger.

This problem is rarely acknowledged in empirical work, and researchers often ignore it when calculating standard errors. If the classical linear model were applicable, standard methods would yield the correct small sample inference (see Donald & Lang, 2007). However, for many applications, the classical model does not apply (e.g., due to nonnormal η_{jt} or serial correlation in η_{jt}). In such cases, classical inference can be misleading.

The Review of Economics and Statistics, February 2011, 93(1): 113–125

FIGURE 1.—EXAMPLE ESTIMATE OF CDF FOR W



In this paper, we show that although $\widehat{\alpha}_{FE}$ is not consistent, we can still conduct inference and construct confidence intervals for α with a general η_{jt} distribution. The key idea behind our approach is that although the control groups are uninformative regarding α, they can still contain information about the distribution of the noise W, a linear combination of η's. Thus, the large number of observations for the controls may allow consistent estimation of the W distribution. To be precise, a necessary condition for our approach is that the distribution of W can be identified from the population of controls. A sufficient condition is random assignment of treatment change conditional on group and time dummy variables, which implies common η distributions for treatments and controls. Under such an assumption, we can use the residuals from the control groups to learn about the limiting distribution of W. Let $\hat{\eta}_{jt}$ denote residuals and $\hat{W}_j$ denote the function of residuals that is analogous to W:

$$\hat{W}_j \equiv \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\widehat{\eta}_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\widehat{\eta}_{jt}.$$

As N_0 gets large, the $\hat{W}_j$ will have the same distribution as W. A test of the hypothesis that $\alpha = \alpha_0$ is easily conducted by comparing $(\widehat{\alpha}_{FE} - \alpha_0)$ with the empirical distribution of $\{\hat{W}_j\}_{j=2}^{N_0}$. The null hypothesis is rejected when $(\widehat{\alpha}_{FE} - \alpha_0)$ is a sufficiently unlikely (tail) event according to this distribution.

We illustrate this approach in figure 1 which is based on data from our empirical example in section IV. We present the empirical distribution of $\hat{W}_j$, a consistent estimate of the distribution of $(\widehat{\alpha}_{FE} - \alpha_0)$ under the null hypothesis that $\alpha = \alpha_0$. An acceptance region can be constructed by finding appropriate quantiles of this empirical distribution. For example, the interval $-.11$ to .09 in figure 1 corresponds to an approximately 90% acceptance region. If $(\widehat{\alpha}_{FE} - \alpha_0)$ does not fall within that range, the null hypothesis $\alpha = \alpha_0$ is rejected. The set of α_0 thats fails to be rejected provides an approximate 90% confidence interval for α. In this example, $\hat{\alpha}$ is approximately .08, which yields a 90% confidence interval for α of $-.01$ to .19.

Our approach is related to a large body of existing work on difference-in-differences models and inference in more general group effect models.[2] It is complementary to typical approaches focusing on situations where the numbers of treatment and control groups, N_1 and N_0, are both large (Moulton, 1990) or both small (Donald & Lang, 2007). It is also in the spirit of comparisons of changes in treatment groups to changes in control groups often done by careful applied researchers. For example, Anderson and Meyer (2000) examine the effect of changes in the unemployment insurance payroll in Washington State on a number of outcomes using a difference-in-differences approach with all other states representing the control groups. In addition to standard analysis, they compare the change in the policy in Washington State to the distribution of changes across other states during the same period of time in order to determine whether it is an outlier consistent with a policy effect.[3]

This approach is relevant for a wide range of applications. Examples include Gruber, Levine, and Staiger (1999) who use comparisons between the five treatment states that legalized abortion prior to *Roe* v. *Wade* versus the remaining states. Our results apply directly, with N_1 corresponding to the five initial movers. For expositional and motivational purposes, we focus on the difference-in-differences case, but our approach is appropriate more generally in treatment effect models with a large number of controls and a small number of treatments.[4] Hotz, Mullin, and Sanders

[2] There are so many examples of difference-in-differences style empirical work that we do not attempt to survey them. See Meyer (1995), Angrist and Krueger (1999), and Bertrand, Duflo, and Mullainathan (2004) for overviews of difference-in-differences methods. Wooldrige (2003) provides an excellent and concise survey of closely related group effect models.

[3] Though it does not appear in the published version, section 4.6 of Bertrand, Duflo, and Mullainathan (2002) describes a placebo laws experiment that is related to some aspects of our approach. They use simulation experiments under specific joint hypotheses about the policy and distribution of covariates to assess the size and power of typical tests (based on large-N_0 and large-N_1). Such experiments could also be used to recover the finite sample distribution of a treatment effect parameter under a particular null hypothesis.

Abadie, Diamond, and Hainmuller (2010) (ADH) is another related paper that uses placebo laws to do inference. However, their main focus is on how to choose the best comparisons for the treated units using combinations of untreated units, which they call synthetic controls. They provide theoretical justification for the use of synthetic controls and compare estimates obtained for the treated units to estimated placebo effects for untreated units to test the null of no treatment effect. In contrast, our paper focuses on inference for treatment parameters after the important choice of controls has been made by the researcher.

[4] One can also find many studies that use a small number of treatments and controls. However, if there exist group × time effects, the usual approach for inference is inappropriate. An alternative sample design is to collect many control groups (with the inherent cost of a reduction of match quality). One could then use our methods for appropriate inference. For example, Card and Krueger (1994) examine the impact of the New Jersey minimum wage law change on employment in the fast food industry. Their sample design has only one control group (eastern Pennsylvania), but they could have collected data from many control states to contrast with the available treatment state. We view this not as a substitute for the analysis that they perform, but rather a complement to check the robustness of the results.

(1997) provide a good example outside the difference-in-differences literature: they estimate the effect of teenage pregnancy on labor market outcomes of mothers. The key to their analysis is using miscarriage as an instrument for teenage motherhood. Of their sample of 980 women who had a teenage pregnancy, only 68 experienced miscarriages. Our basic approach could be extended to this type of application, with the 68 miscarriages taken as fixed like N_1 and the approximate distributions of estimators calculated treating only the nonmiscarried pregnancies as a large sample.

Our final example is the study of merit aid policies, which we use in section IV to illustrate our methods. Merit aid programs provide college tuition assistance to students who attend college in state and maintain a sufficiently high grade point average during high school. Some of the studies in the literature estimate the effect using only a single state that changed its law (Georgia), while newer studies make use of ten states.[5] We demonstrate our methodology and show that accounting for the small number of treatment states is important as the confidence intervals become substantially larger than those formed by the standard approach.

The closest analog to our inference method in econometrics is work on testing for end-of-sample structural breaks—in particular, work such as that by Dufour, Ghysels, and Hall (1994) and Andrews (2003) on the problem of testing for a structural break over a fixed and perhaps very short interval at the end of a sample. They develop tests that are asymptotically valid as the number of observations before the potential break point grows, holding fixed the number of time periods after the break. Their exact models, hypotheses of interest, and structure of proofs differ considerably from ours, but we both use the same basic idea for inference. This idea is to use the small number of observations after the break or N_1 changers as the basis for constructing a test statistic whose reference distribution can be well estimated using the large number of observations before the potential break or N_0 controls.

The remainder of this paper presents our approach in the simplest case of group × time data (e.g., collected at the state × year level) and a common treatment parameter in section II. Extensions to allow heterogeneity in treatment parameters across groups, individual-level data, and cross-sectional dependence and heteroskedasticity are described in section III. In section IV, we present an illustrative example of our approach by studying the effect of merit scholarships. Section V presents the results of a small simulation study of our estimator's performance, followed by a brief conclusion in section VI. Proofs of propositions 1 and 2 are contained in an appendix; all other material is contained in a Web appendix available at the *Review*'s Web site, http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00049.

II. Base Model

Our base model is for situations where data are available at a group × time level:

$$Y_{jt} = \alpha d_{jt} + X'_{jt}\beta + \theta_j + \gamma_t + \eta_{jt}, \tag{2}$$

where d_{jt} is the policy variable that need not be binary, X_{jt} is a vector of regressors with parameter vector β, θ_j is a time-invariant fixed effect for group j, γ_t is a time effect that is common across all groups but varies across time $t = 1, \ldots, T$, and η_{jt} is a group × time random effect. We take α to be the parameter of interest. We use the label "group" because in typical applications, j would index states, counties, or countries, though nothing precludes a group from being a single individual. This data could be either intrinsically group level or aggregates of individuals within a group. In section IIIB, we extend this framework to data with multiple individuals per group, retaining the feature that d_{jt} varies only across group-time cells not within them.

The key problem motivating our approach is that for many groups, there is no temporal variation in d_{jt}. We adopt the convention of indexing the N_1 groups whose value of d_{jt} changes during the observed time span with the integers 1 to N_1. The integers from $N_1 + 1$ to $N_1 + N_0$ then refer to the remaining groups for which d_{jt} is constant from $t = 1$ to T. We treat N_1 and T as fixed, taking limits as N_0 grows large. We assume throughout that at least one group changes its policy so that $N_1 \geq 1$.

It is convenient to partial out variation explained by indicators for groups and times and to have notation for averages across groups and time. Therefore, for generic variable Z_{jt}, we define $\overline{Z}_j = \frac{1}{T}\sum_{t=1}^{T} Z_{jt}$, $\overline{Z}_t = \frac{1}{N_0+N_1}\sum_{j=1}^{N_0+N_1} Z_{jt}$, and use the notation $\overline{Z}$ for the average of Z_{jt} across both groups and time periods. We define a variable $\widetilde{Z}_{jt}$ that equals the residual from a projection of Z_{jt} on group and time indicators: $\widetilde{Z}_{jt} = Z_{jt} - \overline{Z}_j - \overline{Z}_t + \overline{Z}$. The essence of difference in differences is that we can rewrite regression model (2) as

$$\widetilde{Y}_{jt} = \alpha\widetilde{d}_{jt} + \widetilde{X}'_{jt}\beta + \widetilde{\eta}_{jt}, \tag{3}$$

and we can then estimate α by regressing $\widetilde{Y}_{jt}$ on $\widetilde{d}_{jt}$ and $\widetilde{X}_{jt}$. Let $\widehat{\alpha}$ and $\widehat{\beta}$ denote the OLS estimates of α and β in equation (3).

We assume a set of regularity conditions stated as assumption 1, most of them routine. The conditions need to imply that changes in η_{jt} are uncorrelated with changes in regressors, and the usual moment and rank conditions hold. The only (slightly) unusual condition we use describes the cross-sectional dependence of our data. We generalize the standard independence assumption to allow the data to be cross-sectionally strong mixing (see Conley, 1999). This presumes the existence of a coordinate space in which our observations can be indexed. Mixing refers to observations approaching independence as their distance grows, a direct analog of the time series property with the same name. We omit an explicit notation for these coordinates for ease of exposition.

Assumption 1. *$((X_{j1}, \eta_{j1}), \ldots, (X_{jT}, \eta_{jT}))$ is strong mixing across groups; $(\eta_{j1}, \ldots, \eta_{jT})$ is expectation zero conditional on $(d_{j1}, \ldots, d_{jT})$ and $(X_{j1}, \ldots, X_{jT})$; all random variables have finite second moments. The regressors in equation (3),*

[5] Our specifications are motivated by Dynarski (2004).

$\widetilde{d}_{jt}, \widetilde{X}_{jt}$, are linearly independent. Finally, we assume that after the projection of X on time and group fixed effects, the residual regressors $\widetilde{X}_{jt}$ still have variation in the limit, which we state as

$$\frac{1}{N_0+N_1}\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{X}_{jt}\widetilde{X}_{jt}' \xrightarrow{p} \Sigma_x,$$

where Σ_x is finite and of full rank.

Assumption 1 is similar but weaker than the standard set of assumptions made in difference-in-differences applications. It is weaker in that we allow the data to be weakly dependent across groups rather than the usual assumption of independence across groups. The key difference between our setup and the usual setting is that we are assuming N_1 is small and fixed versus the usual assumption that it is large, and our corresponding assumption that there is temporal variation in d_{jt} only for N_1 observations. In proposition 1, we state that OLS yields a consistent estimator of β (as $N_0 \to \infty$, N_1, T fixed), and we derive the limiting distribution of $\widehat{\alpha}$:

Proposition 1. *Under assumption 1, $N_0 \to \infty : \widehat{\beta} \xrightarrow{p} \beta$ and $\widehat{\alpha}$ is unbiased and converges in probability to $\alpha + W$, with:*

$$W = \frac{\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)(\eta_{jt}-\overline{\eta}_j)}{\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)^2}. \tag{4}$$

Proof. See the appendix.

The proposition states that while $\widehat{\alpha}$ is unbiased, it is not consistent (as $N_0 \to \infty$, N_1, T fixed). Its limiting distribution is centered at α, with deviation from α given by W, a linear combination of $(\eta_{jt} - \overline{\eta}_j)$ for $j = 1$ to N_1 and $t = 1$ to T. The nice aspect of this result is that inference for α remains feasible if we can estimate relevant aspects of the distribution of W.

Our approach is to estimate the conditional distribution of W given the observable d_{jt} for the treatment groups. Thus, we need to identify the conditional distribution of $\{(\eta_{jt} - \overline{\eta}_j)\}$ for $j = 1$ to N_1 and $t = 1$ to T given the corresponding set of d_{jt} values. In order to do so, we assume that the distribution of $(\eta_{jt} - \overline{\eta}_j)$ given d_{jt} for the treatments is the same as that for the controls. The time-invariant d_{jt} for our controls cannot be informative about all forms of conditional η_{jt} distributions given the treatments' time-varying d_{jt} series. Thus for feasibility, we must restrict ourselves to a model that is estimable with time-invariant d_{jt}. Random assignment of d_{jt} conditional on X_{jt}, time dummies, and group dummies would be sufficient here, implying common $(\eta_{jt} - \overline{\eta}_j)$ distributions for treatments and controls. Assumptions implying common η distributions for treatments and controls are beyond what is necessary for difference-in-differences applications with large N_1. Large N_1 allows more heterogeneity in the distribution of η conditional on d_{jt} to be tolerated. Terms like W will vanish, and distribution approximations can exploit the large treatment sample size. However, in many cases, researchers justify their difference-in-differences approach by arguing that it is reasonable to think of d_{jt} as randomly assigned (conditional on group and time dummy variables). When this is the case, our approach imposes no further restrictions.

For ease of exposition, we first discuss estimation under a simple model in which the $(\eta_{j1}, \ldots, \eta_{jT})$ are independent of regressors and independent and identically distributed (i.i.d.) across groups, stated as assumption 2. This still allows arbitrary serial correlation in η_{jt}. It is important to note that assumption 2 is not necessary for our approach; it can be replaced by any model of cross-sectionally stationary data, with, for example, spatially correlated or conditionally heteroskedastic η_{jt}, that is, estimable given data from the controls.[6] In the Web appendix, we present an example model that allows temporal and spatial dependence in η_{jt} and heteroskedasticity depending on group population.[7]

Assumption 2. *$(\eta_{j1}, \ldots, \eta_{jT})$ is i.i.d. across j and independent of $(d_{j1}, \ldots, d_{jT})$ and $(X_{j1}, \ldots, X_{jT})$, with a bounded density.*

To see how the distribution of $(\eta_{jt} - \overline{\eta}_j)$ can be estimated under assumption 2, consider the residual for a member of the control group ($j > N_1$),

$$\begin{aligned}\widetilde{Y}_{jt} - \widetilde{X}_{jt}'\hat{\beta} &= \widetilde{X}_{jt}'(\hat{\beta}-\beta) + (\eta_{jt} - \overline{\eta}_j - \overline{\eta}_t + \overline{\eta}) \\ &\xrightarrow{p} (\eta_{jt} - \overline{\eta}_j).\end{aligned} \tag{5}$$

The term involving $\widetilde{X}_{jt}$ vanishes since $\hat{\beta}$ is consistent, and the η term simplifies because $\overline{\eta}_t$ and $\overline{\eta}$ vanish. Thus, if $\{(\eta_{jt} - \overline{\eta}_j)\}_{t=1}^{T}$ is i.i.d. across groups, its distribution for the treatment groups, $j \leq N_1$, is trivially identified using residuals for control groups $j > N_1$.

We first consider estimators implied by the sample analog estimator of the distribution of $\{(\eta_{jt} - \overline{\eta}_j)\}_{t=1}^{T}$, that is, the empirical distribution of residuals from control groups.[8] This implies an estimator of the conditional distribution of W given the d_{jt} for the treatment groups. Defining this distribution as $\Gamma(w) \equiv \Pr(W < w \mid \{d_{jt}, j = 1, \ldots, N_1, t = 1, \ldots, T\})$, its sample analog estimator is

$$\widehat{\Gamma}(w) \equiv \left(\frac{1}{N_0}\right)^{N_1} \sum_{\ell_1=N_1+1}^{N_1+N_0} \cdots \sum_{\ell_{N_1}=N_1+1}^{N_1+N_0} 1\left(\frac{\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)(\widetilde{Y}_{\ell_j t} - \widetilde{X}_{\ell_j t}'\hat{\beta})}{\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)^2} < w\right).$$

[6] Stationarity refers to the joint distribution of observations indexed in a Euclidean space being invariant to translation in their indexes. Observations have identical marginal distributions, and sets of observations with indexes that differ only by a translation have identical distributions.

[7] See Conley and Taber (2005) for an alternative model in this framework that allows for heteroskedasticity arising from variation in group populations along with arbitrary serial dependence but with spatial independence.

[8] Of course, the residuals could also be used to estimate any parametric model of their distribution. This may be a preferable practical strategy in applications with moderately large N_0.

We state a consistency result for $\widehat{\Gamma}(w)$ as proposition 2.

Proposition 2. *Under assumptions 1 and 2 and assuming* β *is interior to a compact parameter space, as* $N_0 \to \infty$, $\widehat{\Gamma}(w)$ *converges in probability to* $\Gamma(w)$ *uniformly on any compact subset of the support of* W.

Proof. See the appendix.

Given the consistent estimator $\widehat{\Gamma}(w)$, it is straightforward to conduct hypothesis tests regarding α using $\widehat{\alpha}$ as a test statistic. Under the null hypothesis that the true value of $\alpha = \alpha_0$, the large sample (N_0 large) approximation following from proposition 1 is that $\hat{\alpha}$ is distributed as $\alpha_0 + W$ conditional on $\{d_{jt}, j = 1, \ldots, N_1, t = 1, \ldots, T\}$. Therefore, we consistently estimate the distribution function $\Pr(\hat{\alpha} < c)$ via $\widehat{\Gamma}(c - \alpha_0)$ and use its appropriate quantiles to define an asymptotically valid acceptance region for this null hypothesis.[9] For example, a 90% acceptance region could be estimates as $[\widehat{\alpha}_{lower}, \widehat{\alpha}_{upper}]$ with these end points being the 5th and 95th percentiles of this distribution: $\widehat{\Gamma}(\widehat{\alpha}_{lower} - \alpha_0) \approx .05$ and $\widehat{\Gamma}(\widehat{\alpha}_{upper} - \alpha_0) \approx .95$.[10] A 90% confidence interval for the true value of α can then be constructed as the set of all values of α_0 where one fails to reject the null hypothesis that α_0 is the true value of α.

This might look complicated, but it is actually easy to implement. To see this, consider the example in which we have only one treatment ($N_1 = 1$) and want to test the null hypothesis that $\alpha = 0$. We use the following procedure:

1. Run the regression of $\tilde{Y}$ on $\tilde{X}$.
2. Take the residuals of the regression for the controls from group j and call them $\tilde{\eta}_{jt}$.
3. Use these to form the empirical distribution of
$$\frac{\sum_{t=1}^{T}(d_{1t} - \bar{d}_1)\tilde{\eta}_{jt}}{\sum_{t=1}^{T}(d_{1t} - \bar{d}_1)^2}.$$
4. If $\hat{\alpha}$ is in the tails of this empirical distribution, reject the null hypothesis.

With more than one treatment group or a different null hypothesis, it is only marginally more difficult; step 3 is conducted with a different linear combination of residuals.

An alternative, asymptotically equivalent estimator is heuristically motivated by the literature on permutation or randomization inference (see Rosenbaum, 2002). In randomization inference, random assignment of the treatment is the basis for inference, and the exact, small sample distributions statistics are computable. The applications we have in mind are not situations with random assignment of treatment; at best, they could be described as having treatment randomly assigned conditional on X. In this scenario, even if recentering by subtracting $X'\beta$ were sufficient to accomplish conditioning on X, this would still not be enough to implement exact inference because β must be estimated. However, we anticipate that if $\hat{\beta}$ is a good estimate of β, then plugging $\hat{\beta}$ into a permutation estimator in place of β should provide good approximations of the small sample distribution of W. Such an estimator requires forming residuals under the null hypothesis for the treatment groups ($\widetilde{Y}_{\ell_j t} - \alpha_0 \widetilde{d}_{\ell_j t} - \widetilde{X}'_{\ell_j t}\hat{\beta}$), using them along with residuals from controls and using the distribution of N_1 draws without replacement from $N_1 + N_0$ residuals as the underlying reference distribution in place of the empirical distribution of control residuals. This gives us an estimator:

$$\widehat{\Gamma}^*(w) \equiv \frac{1}{(N_0 + N_1)(N_0 + N_1 - 1)\ldots(N_0)} \times \left[\sum_{\ell_1 \in [1:N_1+N_0]} \sum_{\substack{\ell_2 \in [1:N_1+N_0] \\ \ell_2 \neq \ell_1}} \cdots \sum_{\substack{\ell_{N_1} \in [1:N_1+N_0] \\ \ell_{N_1} \notin \{\ell_1, \ldots, \ell_{N_1-1}\}}} 1\left(\frac{\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt} - \bar{d}_j)(\widetilde{Y}_{\ell_j t} - \alpha_0 \widetilde{d}_{\ell_j t} - \widetilde{X}'_{\ell_j t}\hat{\beta})}{\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt} - \bar{d}_j)^2} < w \right) \right].$$

The summations are over all possible assignments of treatment status to N_1 of the $N_1 + N_0$ total groups. While $\widehat{\Gamma}^*(w)$ is motivated by (infeasible) estimators with known exact distributions, we note that it is not an exact estimate of the distribution of W. The rigorous justification of $\widehat{\Gamma}^*(w)$ is that it is asymptotically equivalent (as $N_0 \to \infty, N_1, T$ fixed) to $\widehat{\Gamma}(w)$.[11]

III. Extensions

This section presents extensions of our base model to accommodate treatment parameter heterogeneity and individual-level data. Extensions of our model to accommodate spatial dependence are presented in the Web appendix.

A. *Treatment Parameter Heterogeneity*

It is straightforward to modify equation (2) to allow heterogeneity in treatment parameters across groups.

[9] We note that no test in this framework can be consistent as $N_1 \to \infty$ since a finite number of observations are informative regarding α. We also make no claim that this test is optimal.

[10] We cannot obtain exact equality in these expressions because $\widehat{\Gamma}$ is a step function, but we can choose the closest point, and asymptotically the coverage probability will converge to 90%.

[11] We expect $\widehat{\Gamma}^*(w)$ to outperform $\widehat{\Gamma}(w)$ in situations for which $\hat{\beta}$ is well estimated but N_1 is still small enough for the empirical distribution in $\widehat{\Gamma}(w)$ to perform poorly. There are certainly applications where this is likely to be the case. For example, suppose that data are collected at the state level and that demographic regressors like income or population have substantial variation. With such large-variance regresssors, $\hat{\beta}$ may be well estimated with, say, $N_1 = 20$ states, while with only twenty observations, the empirical distribution will do a mediocre job at best of estimating conventional critical values. This situation will also arise when the model is extended to individual-level data in section III. With only individual-level regressors, coefficients analogous to β will be estimated extremely well regardless of N_1 if there are many individuals within each group. This situation is routine with repeated cross-section data and arises in our empirical example to merit aid programs discussed in section IV.

Consider the extension to allow group-specific treatment parameters:

$$Y_{jt} = \alpha_j d_{jt} + X'_{jt}\beta + \theta_j + \gamma_t + \eta_{jt}. \quad (6)$$

Using the notation defined above, we can rewrite this as

$$\widetilde{Y}_{jt} = \alpha_j \widetilde{d}_{jt} + \widetilde{X}'_{jt}\beta + \widetilde{\eta}_{jt}.$$

Note that $\widetilde{d}_{jt}$ is 0 for all of the control groups; thus, we estimate treatment parameters only for $j = 1$ to N_1 and stack these estimable parameters in the vector $A = [\alpha_1, \ldots, \alpha_{N_1}]'$. We define D_{jt} to be the $N_1 \times 1$ vector of interactions between d_{jt} and group indicators. That is, the ℓth element of the vector $D_{jt} = d_{jt}$ if $j = \ell$ and is zero otherwise. We can then write

$$\widetilde{Y}_{jt} = \widetilde{D}'_{jt}A + \widetilde{X}'_{jt}\beta + \widetilde{\eta}_{jt}.$$

We refer to OLS estimates of (A, β) in this regression as $(\widehat{A}, \widehat{\beta})$.

Proposition 3. *If assumption 1 holds, then as* $N_0 \to \infty$, $\widehat{\beta} \xrightarrow{p} \beta$ *and* $\widehat{A}$ *converges in probability to* $A + W$, *where* W *is an* $N_1 \times 1$ *random vector with generic element*

$$W(j) = \frac{\sum_{t=1}^{T}(d_{jt} - \overline{d}_j)(\eta_{jt} - \overline{\eta}_j)}{\sum_{t=1}^{T}(d_{jt} - \overline{d}_j)^2}.$$

Proof. See the Web appendix, section A.1.

Testing and inference can proceed exactly as in section II. A consistent sample analog estimator of the distribution of $\widehat{A}$ under the null hypothesis that A_0 is the true value of A can be constructed with residuals from controls. This allows testing any point null hypothesis about the heterogeneous treatment effects, and inversion of this test provides a joint confidence set for the elements of A. Alternatively, the distribution of any function of the elements of A (e.g., their mean across groups) can also be consistently estimated, allowing analogous hypothesis testing and confidence set construction.

We have restricted the form of the treatment effect heterogeneity to vary only with j for ease in exposition. Our method can be extended to allow α_{jt} to vary across j and t by inverting a corresponding set of point hypotheses tests on the α_{jt} for a set of groups and time periods. Extensions to situations where treatment effects depend on an observable covariates, such as the time since the policy was adopted, are also straightforward.[12]

B. *Individual-Level Data*

Our approach can easily be applied with repeated cross-sections or panels of individual data, the relevant data type for many situations. We restrict ourselves to repeated cross-sections for ease of exposition. Let i index an individual who is observed in group $j(i)$ at a single time period $t(i)$. Allowing individual-specific regressors Z_i (for example, demographic characteristics) and noise ε_i, we arrive at a model:

$$Y_i = \lambda_{j(i)t(i)} + Z'_i\delta + \varepsilon_i \quad (7)$$

$$\lambda_{jt} = \alpha d_{jt} + X'_{jt}\beta + \theta_j + \gamma_t + \eta_{jt}. \quad (8)$$

In equation (8), i subscripts have been dropped because its components vary only at the group $\times$ time level: $\lambda_{j(i)t(i)} = \lambda_{jt}$ for all individuals i in group j at time t. The difference between Z_i and X_{jt} is that we assume that Z_i varies within a group $\times$ time cell, while X_{jt} does not.

There are at least three ways to approach estimation of the above model. A one-step approach would plug equation (8) into equation (7), and the resulting model could be estimated by least squares under the assumption that the error terms ε, η were orthogonal to the regressors. The Web appendix, section A.2.4, contains a rigorous demonstration that our methods extend to this approach, and we use this in our empirical example below. Another option would be to first aggregate the data within the group-time cell and proceed to estimate our base model as in section II.

Here, we focus on the third approach: the well-known two-step approach to estimation.[13] We obtain estimates for α by first estimating λ_{jt} in equation (7) for all groups and time periods using a regression of Y_i on a full set of indicators for group $\times$ time and Z_i. In the second step, the estimated λ_{jt} are then used as the outcome variable in equation (8), and the inference procedures described in section II can be applied directly to this second-step regression. The main difference between the three approaches is in the estimation of δ. Estimating δ in the one-step approach uses all variation, averaging first uses only between variation, and the two-step estimator we suggest uses only within variation. Our preference for this two-step approach is driven by its ease of exposition and that it is more flexible than the one-step estimator because it does not require orthogonality between Z and η.

A variety of assumptions could be made about the behavior of the number of individuals per group. Let $M(j, t)$ be the set of individuals observed in group j at time t and $|M(j, t)|$ denote the number of individuals in this set. We focus on the case in which $|M(j, t)|$ is growing with N_0 and continue to assume T is fixed. However, in the Web appendix (section A.2.3) we provide a rigorous demonstration that our test procedures remain asymptotically valid when the number of individuals per group $\times$ time is fixed but common across group $\times$ time cells.[14]

Let I_i be a set of fully interacted indicators for all group $\times$ time cells. Now consider a regression of Y_i on Z_i and I_i. Let

[12] A common example would be an event study analysis such as in Jacobson, LaLonde, and Sullivan (1993). In this approach, one would let the effect of the treatment be time varying relative to when it was introduced—that is, the effect of the policy one year after it was passed may be different from the effect five years later.

[13] See, e.g., Hanushek (1974) or Amemiya (1978), who discuss aspects of this approach.

[14] In Conley and Taber (2005) we present a complementary strategy with fixed sample sizes that vary across group $\times$ time cells. This is considerably more difficult because of the need to solve a deconvolution problem.

$\widehat{\lambda}_{jt}$ be the regression coefficient on the dummy variable for group j at time t. It is straightforward to show that

$$\widehat{\lambda}_{jt} = \lambda_{jt} + \left[\frac{1}{|M(j,t)|} \sum_{i \in M(j,t)} Z_i'(\delta - \widehat{\delta}) + \frac{1}{|M(j,t)|} \sum_{i \in M(j,t)} \varepsilon_i \right], \tag{9}$$

where $\widehat{\delta}$ is the regression coefficient obtained in the first step. As $|M(j,t)|$ grows large, the term in brackets vanishes. The second step is then simply to plug in $\widehat{\lambda}_{jt}$ for λ_{jt} in equation (8) and run a fixed-effect OLS. We recycle notation and use $\widehat{\beta}$ and $\widehat{\alpha}$ in this section to refer to the second-step OLS estimators of equation (8). The results of section II apply to these estimators under a straightforward set of conditions. Aside from the usual orthogonality and rank conditions, we need to specify the rate at which $|M(j,t)|$ grows; these are stated as assumption 3:

Assumption 3. *ε_i is i.i.d., independent of $[Z_i \quad I_i]$ and has a finite second moment. $[Z_i \quad I_i]$ is full rank. For all j, $|M(j,t)|$ grows uniformly at the same rate as N_0.*

Proposition 4. *Under assumptions 1, 2, and 3 and assuming β is interior to a compact parameter space, as $N_0 \to \infty$, the conclusions of propositions 1 and 2 apply to the Amemiya (1978) second-step OLS estimators $\widehat{\beta}$ and $\widehat{\alpha}$ of equation (8): $\widehat{\beta} \xrightarrow{p} \beta$ and $\widehat{\alpha} \xrightarrow{p} \alpha + W$, where W has exactly the same form given by equation (4). Using the notation $\widetilde{Z}$ to refer to the residual from a linear projection of a variable Z on a full set of time and group indicators, define $\widehat{\Gamma}$ as*

$$\widehat{\Gamma}(w) \equiv \left(\frac{1}{N_0}\right)^{N_1} \sum_{\ell_1 = N_1+1}^{N_1+N_0} \cdots \sum_{\ell_{N_1} = N_1+1}^{N_1+N_0} 1\left(\frac{\sum_{j=1}^{N_1} \sum_{t=1}^{T} (d_{jt} - \overline{d}_j)\left(\widetilde{\widehat{\lambda}}_{\ell_j t} - \widetilde{X}_{\ell_j t}'\hat{\beta}\right)}{\sum_{j=1}^{N_1} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)^2} < w \right).$$

$\widehat{\Gamma}(w)$ converges in probability to $\Gamma(w)$ uniformly on any compact subset of the support of W.

Proof. The proof is in the Web appendix, section A.2.2.

With access to data containing a large number of individuals within group × time cells, it is straightforward to extend our approach to models with a nonlinear first step. For example, consider the following latent variable model for a binary outcome Y_i,

$$Y_i = 1(\lambda_{j(i)t(i)} + Z_i'\delta + \varepsilon_i \geq 0) \tag{10}$$

$$\lambda_{jt} = \alpha d_{jt} + X_{jt}'\beta + \theta_j + \gamma_t + \eta_{jt} \tag{11}$$

Equation (11) is, of course, the same as equation (8), with i subscripts dropped because its components vary only at the group × time level. The parameters in equation (10) can easily be consistently estimated in a standard way such as, probit, logit, or even semiparametrically, depending on the assumption one is willing to make on ε_i. The resulting λ_{jt} estimates, $\widehat{\lambda}_{jt}$, are simply the estimated group × time cell intercepts from the first step. Inference regarding α can then be conducted exactly as above with a linear first step. The $\widehat{\lambda}_{jt}$ can used as outcome variables in equation (11), which can again be estimated using OLS and our test procedure applied to the resulting α estimates. We use a logistic first-step procedure in our empirical application in the following section.

IV. Empirical Example: The Effect of Merit Aid Programs on Schooling Decisions

In the past fifteen years a number of states have adopted merit-based college aid programs that provide subsidies for tuition and fees to students who meet certain merit-based criteria. The largest and probably the best-known program is the Georgia HOPE (Helping Outstanding Pupils Educationally) scholarship, which started in 1993. This program provides full tuition as well as some fees to eligible students who attend in-state public colleges.[15] Eligibility for the program requires maintaining a 3.0 grade point average during high school. A number of previous papers have examined the effect of HOPE and other merit-based aid programs.[16] Given the large amount of previous work on this subject, we leave full discussion of the details of these programs to these other papers and focus on our methodological contribution.

Our work most closely relates to Dynarski (2004) by focusing on the effects of HOPE and other merit aid programs on college enrollment of 18 and 19 year olds using the October CPS from 1989 to 2000. Our specifications are motivated by some of hers, but we do not replicate her entire analysis. Our goal is to illustrate the use of our method, and our analysis falls well short of a complete empirical analysis of merit scholarship effects.

During the 1989–2000 time period, 10 states initiated merit aid programs. We use two specifications, with the first focusing on the HOPE program alone. In this case, we ignore data from the other 9 treatment states and use 41 controls (40 states plus the district of Columbia). In the second case, we study the effect of merit-based programs together and use all 51 units.[17] The outcome variable in all cases is an indicator

[15] A subsidy for private colleges is also part of the program.

[16] Examples include Dynarski (2000, 2004); Cornwell, Mustard, and Sridhar (2006); Bugler, Henry, and Rubenstein (1999); and Henry and Rubenstein (2002).

[17] Note that these merit programs are quite heterogeneous. This exercise does not necessarily mean that we are assuming that the impact of all of these programs is the same. One could interpret this as estimation of a weighted average of the treatment effects. Alternatively, we can think of this as a test of the joint null hypothesis that all of the effects are 0. We could estimate more general confidence intervals allowing for heterogeneous treatment effects, but we focus on the simplest case here.

TABLE 1.—ESTIMATES FOR THE EFFECT OF GEORGIA HOPE PROGRAM ON COLLEGE ATTENDANCE

	A	B	C
	Linear Probability	Logit	Population-Weighted Linear Probability
Hope Scholarship	0.078	0.359	0.072
Male	−0.076	−0.323	−0.077
Black	−0.155	−0.673	−0.155
Asian	0.172	0.726	0.173
State dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
95% confidence intervals for hope effect			
Standard cluster by State × Year	(0.025,0.130)	(0.119,0.600) [0.030,0.149]	(0.025, 0.119)
Standard cluster by state	(0.058,0.097)	(0.274,0.444) [0.068,0.111]	(0.050,0.094)
Conley-Taber	(−0.010,0.207)	(−0.039,0.909) [−0.010,0.225]	(−0.015,0.212)
Sample size			
Number of states	42	42	42
Number of individuals	34,902	34,902	34,902

Confidence intervals for parameters are presented in parentheses. We use the $\widehat{\Gamma}^*$ formula to construct the Conley-Taber standard errors. Brackets contain a confidence interval for the program impact on a person whose college attendance probability in the absence of the program would be 45%.

variable representing whether the individual is currently enrolled in college.

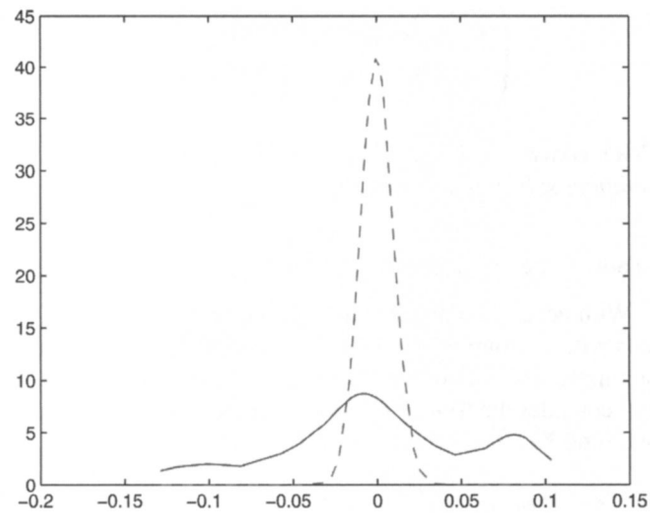
In constructing the confidence intervals, two issues arise due to the fact that we have only 41 control states. The first issue is whether 41 is large enough for the asymptotics to be valid. With that in mind, we use the $\widehat{\Gamma}^*$ estimator described in section II, motivated by its anticipated good finite sample properties. The second issue can be seen in figure 1. The estimated CDF is of course a step function, and with a single treatment state and 41 controls, its probability increments are limited to 1/41. To approximate intervals with conventional, say 95%, coverage probabilities, we use a conservative interval so that the limiting coverage probability is at least 95%. As a practical matter, this is usually relevant only for the case of a single treatment group. With two or more treatment groups, the empirical CDF will have a number of steps on the order of the number of ways to choose the N_1 treatment groups out of the total number of groups ($\frac{(N_1+N_0)!}{N_1!N_0!}$). Thus, the number of steps in the CDF is typically large for two or more treatments with corresponding small probability increments.

In table 1 we present results for the HOPE program with Georgia as the only treatment state. We compare three estimators: column A corresponds to the approach described in section IIIB, equations (7) and (8), and columns B and C present two natural alternatives. The estimates in both columns A and B are obtained from Amemiya's (1978) two-step approach. The estimates reported in column A use a first-step linear probability model (OLS), and in column B, the first step is a logit; regressors in both case include demographics and state × year indicators. The second step in both A and B estimates equation (8) with OLS using the estimated state × year coefficients as the dependent variable. Column C presents results from a one-step estimator, which is simply a linear probability model estimated using OLS using the entire sample. Thus, the column C treatment effect estimates will be population weighted across states, while in column A, states are equally weighted. The top panel of table 1 presents point estimates for all three estimators, and the bottom panel presents interval estimates for the treatment parameter, both using our methods with $\widehat{\Gamma}^*$ and the typical approaches clustering by state and state-by-time.

Although results differ depending on the clustering used, interval estimates in column A using typical methods indicate significant treatment effects. An interval of 2.5% to 13.0% obtains with clustering by state and year, which allows the error terms of individuals within the same state and year to be arbitrarily correlated with each other. This interval shrinks to 5.8% to 9.7% when clustering is done by state, which allows serial correlation in η_{jt}. Clearly one should be worried about the assumption that the number of states changing treatment status is large, which underlies these routine confidence interval estimates since only one state, Georgia, contributes to the estimate of the treatment effect.

The estimated confidence interval using our method reported in the last row of column A is −1% to 21%. This confidence interval is formed by inverting the test statistic $(\hat{\alpha} - \alpha_0)$ using our $\widehat{\Gamma}^*$ estimator. It is centered at a larger value and much wider than the intervals obtained with conventional inference—wide enough to include 0 despite its shift in centering. To better understand these discrepancies, Figure 2 displays a kernel smooth estimate (solid line) of the distribution of $(\widehat{\alpha} - \alpha)$ under the null hypothesis that the true value of α is 0. This distribution is estimated from the control states. For comparison, the dashed line plots an estimate implied by the usual asymptotic approximation with clustering by state. This curve is a gaussian density centered at 0 with a standard deviation equal to 0.0098: the standard error on $\widehat{\alpha}$ from a fixed-effect regression that clusters by state. The pronounced differences between the spread and symmetry (lack thereof) of these distributions are what drive our interval

FIGURE 2.—ESTIMATED DENSITY OF $\hat{\alpha}$ UNDER $H_0 : \alpha_0 = 0$



Solid line: Kernel-smoothed density estimate for Conley-Taber approach. Dashed line: Density estimate using standard asymptotics.

estimates of α to differ from those resulting from conventional methods.

In column B, we present a logit version of the model as in equations (10) and (11) with ε_i logistic. The estimates in this column were obtained in exactly the same manner as for column A, except that in the first step, we use a logit model of the college attendance indicator so the predicted parameter has the interpretation of a logit index coefficient. The pattern is very similar to column A. Intervals from our method are again centered higher than conventional ones, but enough wider that the HOPE treatment effect becomes marginally insignificant. This contrasts with effects that are highly significant using standard inference methods. To display the magnitude of the program impact, we calculate a 95% confidence interval for changes in college attendance probability for a particular individual. We consider an individual (without the treatment) whose logit index puts his probability of college attendance at the sample unconditional average attendance probability of 45% (i.e., an individual with a logit index of $-.20$). The bracketed intervals reported in column 2 are 95% confidence intervals for the change in attendance probability for our reference individual (intervals in parentheses are 95% confidence intervals for α).[18]

In column C we present results from a linear probability that estimates equations (7) to (8) using OLS with all 34,902 observations. The details for constructing the confidence intervals are formally presented in the Web appendix (section A4.4). These results are close to those presented in column A. The difference between these two estimates is that in column A, the states are equally weighted, while in column C, they are population weighted.

In table 2 we present results estimating the effect of merit aid using all ten states that added programs during this time period. The format of the table is identical to table 1. There are a few notable features of this table. First, the weighting matters substantially, as the effect is much smaller when we weight all the states equally as opposed to the population-weighted estimates in column C. Second, in contrast to table 1, the confidence intervals are quite similar when we cluster by state compared to clustering by state $\times$ year. Most important, our approach changes the confidence intervals substantially, but less dramatically than in table 1. In particular, the treatment effect with equal weighting across states is still statistically significant at conventional levels.

TABLE 2.—ESTIMATES FOR MERIT AID PROGRAMS ON COLLEGE ATTENDANCE

	A	B	C
	Linear Probability	Logit	Population-Weighted Linear Probability
Merit scholarship	0.051	0.229	0.034
Male	−0.078	−0.331	−0.079
Black	−0.150	−0.655	−0.150
Asian	0.168	0.707	0.169
State dummies	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
95% confidence intervals for merit aid program effect			
Standard cluster by State × Year	(0.024,0.078)	(0.111,0.346) [0.028,0.086]	(0.006,0.062)
Standard cluster by state	(0.028,0.074)	(0.127,0.330) [0.032,0.082]	(0.008,0.059)
Conley-Taber	(0.012,0.093)	(0.056,0.407) [0.014,0.101]	(−0.003,0.093)
Sample size			
Number of states	51	51	51
Number of individuals	42,161	42,161	42,161

Confidence intervals for parameters are presented in parentheses. We use the $\widehat{\Gamma}^*$ formula to construct the Conley-Taber standard errors. Brackets contain a confidence interval for the program impact on a person whose college attendance probability in the absence of the program would be 45%.

V. Monte Carlo

In this section we discuss the results of a small Monte Carlo study evaluating the performance of our method and comparing it to typical approaches. The specification that we examine is

$$Y_{jt} = \alpha d_{jt} + \beta X_{jt} + \theta_j + \gamma_t + \eta_{jt},$$

in which we focus on the model of section II with group-level data since that is our base case. Note that we focus here on a single regressor. We assign a binary treatment, d_{jt}, that is 0 for controls and at some point in the data turns permanently from 0 to 1 for each treatment group. We assume that the error term within group has a first-order autoregressive structure:

$$\eta_{jt} = \rho\eta_{jt-1} + u_{jt},$$
$$u_{jt} \sim N(0, 1).$$

Finally, we want controlling for X_{jt} to be important (as it often is in real data); therefore, we build in a correlation between X and the treatment:

$$X_{jt} = a_x d_{jt} + \nu_{jt},$$
$$\nu_{jt} \sim N(0, 1).$$

In our base case model, we let the total number of groups $(N_1 + N_0)$ be 100, $T = 10$ and let five groups change treatment status during the time period. The turn-on time periods for the base case are periods 2, 4, 6, 8, and 10. We set the remaining parameters to have the values $\alpha = 1, \rho = 0.5$, $a_x = 0.5, \beta = 1$.

In table 3, we present the results of testing the true null hypothesis ($\alpha = 1$) and a false one ($\alpha = 0$) at the 5% level using 10,000 trials and present the percentage of times the hypothesis is rejected. Thus, if the test works well, we should reject the hypothesis $\alpha = 1$ around 5% of the trials and reject $\alpha = 0$ much more frequently. We present four different approaches: a standard t-test adjusted for degrees of freedom (as suggested by Donald & Lang, 2007), a cluster-by-group approach (as suggested by Bertrand, Duflo, & Mullainathan, 2004), and then our approach using both the $\widehat{\Gamma}$ and $\widehat{\Gamma}^*$

[18] These confidence intervals for changes in attendance probabilities are calculated directly from the 95% CI for α. Specifically, when the CI for α is $[c_1, c_2]$, letting Λ denote the logistic CDF, we report an interval for the change in predicted probability for our reference individual of $(\Lambda(-.2 + c_1) - 45\%)$ to $(\Lambda(-.2 + c_2) - 45\%)$.

TABLE 3.—MONTE CARLO RESULTS: SIZE AND POWER OF TEST OF AT MOST 5% LEVEL
BASIC MODEL
$Y_{jt} = \alpha d_{jt} + \beta X_{jt} + \theta_j + \gamma_t + \eta_{jt}$
$\eta_{jt} = \rho\eta_{jt-1} + \varepsilon_{jt}, \alpha = 1, X_{jt} = a_x d_{jt} + \nu_{jt}$

Percentage of Times Hypothesis Is Rejected out of 10,000 Simulations								
	Size of Test ($H_0 : \alpha = 1$)				Power of Test ($H_0 : \alpha = 0$)			
	Classic Model	Cluster	Conley Taber ($\widehat{\Gamma}^*$)	Conley Taber ($\widehat{\Gamma}$)	Classic Model	Cluster	Conley Taber ($\widehat{\Gamma}^*$)	Conley Taber ($\widehat{\Gamma}$)
Base model[a]	14.23	16.27	4.88	5.52	73.23	66.10	54.08	55.90
Total groups = 1000	14.89	17.79	4.80	4.95	73.97	67.19	55.29	55.38
Total groups = 50	14.41	15.55	5.28	6.65	71.99	64.48	52.21	56.00
Time periods = 2	5.32	14.12	5.37	6.46	49.17	58.54	49.13	52.37
Number treatments = 1[b]	18.79	84.28	4.13	5.17	40.86	91.15	13.91	15.68
Number treatments = 2[b]	16.74	35.74	4.99	5.57	52.67	62.15	29.98	31.64
Number treatments = 10[b]	14.12	9.52	4.88	5.90	93.00	84.60	82.99	84.21
Uniform error[c]	14.91	17.14	5.30	5.86	73.22	65.87	53.99	55.32
Mixture error[d]	14.20	15.99	4.50	5.25	55.72	51.88	36.01	37.49
$\rho = 0$	4.86	15.30	5.03	5.57	82.50	86.42	82.45	83.79
$\rho = 1$	30.18	16.94	4.80	5.87	54.72	34.89	19.36	20.71
$a_x = 0$	14.30	16.26	4.88	5.55	73.38	66.37	54.08	55.93
$a_x = 2$	14.18	16.11	4.82	5.49	73.00	65.91	54.33	55.76
$a_x = 10$	10.36	9.86	11.00	11.90	51.37	47.78	53.29	54.59

In the results for the Conley-Taber ($\widehat{\Gamma}^*$) with smaller sample sizes, we cannot get a size of exactly 5% due to the discreteness of the empirical distribution. When this happens, we choose the size to be the largest value possible that is under 5%.

[a]For the base model, the total number of groups is 100, with five treatments, and ten periods. Parameter values: $\rho = 0.5, a_x = 0.5, \beta = 1, \varepsilon_{jt} \sim N(0,1), \nu_{jt} \sim N(0,1)$.

[b]With T treatments and five periods, the changes occur during periods 2, 4, 6, 8, and 10. For one treatment, it is in period 6; for two treatments, it is in periods 3 and 7; and for ten treatments, it is periods 2, 2, 3, 4, 5, 6, 7, 8, 9, and 10.

[c]The range of the uniform is $[-\sqrt{3}, \sqrt{3}]$ so that it has unit variance.

[d]The mixture model we consider is a mixtures of a $N(0,1)$ and a $N(2,1)$ in which the standard normal is drawn 80% of the time.

formulas. The results for the base case are presented in the first row. One can see that our approach performs much better than either of the alternatives, both of which miss the size by a factor of about three.[19]

We then consider other cases by altering some of the parameters in the data-generating process (DGP), one at a time. The labels in the left column indicate the parameters that differ from the base case setting. For example, the fifth row decreases the number of treatment groups from five to two, holding all other parameters at the base setting. This decrease in information results in a large drop in power for both the $\widehat{\Gamma}$ and $\widehat{\Gamma}^*$ estimators with little size distortion. With treatments reduced to two, the classic estimator suffers a large drop in power and a small increase in size distortion, whereas the cluster estimator suffers a large increase in size distortion along with a small drop in power. In both the $T = 2$ and $\rho = 0$ lines, we see alternate specifications where our Monte Carlo DGP collapses to the classical linear model. However, $\widehat{\Gamma}^*$ appears to perform on par with the classical model here, and $\widehat{\Gamma}$ does reasonably well too. Thus, our methods have comparable size and power characteristics to the classical test in some scenarios where it is ideal.

As anticipated, $\widehat{\Gamma}^*$ does seem to work a little better than $\widehat{\Gamma}$ with smaller samples, as seen in size in the Groups = 50 row. However, across all scenarios, the similarities between the performance of $\widehat{\Gamma}$ and $\widehat{\Gamma}^*$ are more salient than the slight size advantage of $\widehat{\Gamma}^*$.

[19]Their power is higher here, but this is likely in large part because the size is too large; that is, the confidence intervals are tighter than they should be.

We do not expect our approaches to work well when there is a great deal of estimation error in $\widehat{\beta}$. This can be seen in our simulations as the parameter a_x increases. We get a substantial size distortion for both $\widehat{\Gamma}$ and $\widehat{\Gamma}^*$ with $a_x = 10$. This means that the distribution of X_{jt} is $N(0,1)$ without the treatment, but then jumps to $N(10,1)$ after the treatment is implemented. The classical and cluster methods also struggle here, so our method is not dominated by these alternatives even in this case.

Perhaps the starkest result is how poorly the cluster approach works with a small number of treatment changers. The size in the base case is triple what it should be. Performance here is very sensitive to the number of treatment groups. When this is decreased to one or two, the performance is terrible. However, it does better when one gets up to ten treatments and, in results not shown, it works well at forty treatments. However, even with ten treatments, although the size of the test is down to 9.52%, the power is not much better than for our approach. These results show that cluster standard errors can be very misleading when the number of groups changing status is small.

VI. Conclusion

This paper presents an inference method for difference-in-differences fixed-effect models when the number of policy changes observed in the data is small. This method is an alternative to typical asymptotic inference based on a large number of policy changes and classical small sample inference. Our approach will be most valuable in applications where the classical model does not apply—for example, due

to nongaussian or serially correlated errors. We provide an estimator $\hat{\Gamma}^*$ that is large-N_0 asymptotically valid and appears to have good finite sample properties with serially dependent, cross-sectionally i.i.d. data. Our approach can also be applied with much weaker conditions on the data. Many forms of cross-sectional dependence and heteroskedasticity, for example, can be readily accommodated. We provide an example application studying the effect of merit scholarship programs on college attendance for which our approach seems appropriate. It results in very different inference from conventional methods. We also perform a Monte Carlo analysis, which indicates that our approach fares far better than the standard alternatives when the number of treatment groups is small and performs well even in cases that are tailored to ensure good performance of these alternatives.

REFERENCES

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association* 105:490 (2010), 493–505.

Amemiya, Takeshi, "A Note on a Random Coefficients Model," *International Economic Review* 19:3 (1978), 793–796.

Anderson, Patricia, and Bruce Meyer, "The Effects of the Unemployment Insurance Payroll Tax on Wages, Employment, Claims, and Denials," *Journal of Public Economics* 78:1 (2000), 81–106.

Andrews, Donald, "End-of-Sample Tests," *Econometrica* 71:6 (2003), 1661–1694.

Angrist, Joshua, and Alan Krueger, "Empirical Strategies in Labor Economics" (pp. 1277–1366), in Orley Ashenfelter and David Card (Eds.), *Handbook of Labor Economics* (New York: Elsevier, 1999).

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" NBER working paper 8841 (2002).

——— "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 19:1 (2004), 249–275.

Bugler, Daniel, Gary Henry, and Ross Rubenstein, "An Evaluation of Georgia's HOPE Scholarship Program: Effects of HOPE on Grade Inflation, Academic Performance and College Enrollment," (Atlanta: Georgia State University, 1999).

Card, David, and Alan Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 90:5 (1994), 1397–1420.

Conley, Timothy, "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics* 92 (1999), 1–45.

Conley, Timothy, and Christopher Taber, "Inference with 'Difference in Differences' with a Small Number of Policy Changes," NBER working paper no. 0312 (2005).

Cornwell, Christopher, David Mustard, and Deepa Sridhar, "The Enrollment Effects of Merit-Based Financial Aid: Evidence from Georgia's HOPE Scholarship," *Journal of Labor Economics* 24:4 (2006), 761–786.

Donald, Stephen, and Kevin Lang, "Inference with Difference in Differences and Other Panel Data," this REVIEW 89:2 (2007), 221–233.

Dufor, Jean-Marie, Eric Ghysels, and Alastair Hall, "Generalized Predictive Tests and Structural Change Analysis in Econometrics," *International Economics Review* 35:1 (1994), 199–229.

Dynarski, Susan, "Hope for Whom? Financial Aid for the Middle Class and Its Impact on College Attendance," *National Tax Journal* 53:3, pt. 2 (2000), 629–662.

——— "The New Merit Aid" (pp. 63–97), in Caroline Hoxby (Ed.), *College Choices: The Economics of Which College, When College, and How to Pay for It* (Chicago: University of Chicago Press, 2004).

Gruber, Jonathon, Phillip Levine, and Douglas Staiger, "Abortion Legalization and Child Living Circumstances: 'Who Is the Marginal Child?'" *Quarterly Journal of Economics* 114:1 (1999), 263–291.

Hanushek, Eric, "Efficient Estimators for Regressing Regression Coefficients," *American Statistician* 298:1 (1974), 66–67.

Henry, Gary, and Ross Rubinstein, "Paying for Grades: Impact of Merit-Based Financial Aid on Educational Quality," *Journal of Policy Analysis and Management* 21:1 (2002), 93–109.

Hotz, V. J., C. Mullin, and S. Sanders, "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effect of Teenage Childbearing," *Review of Economic Studies* 64 (1997), 576–603.

Jacobson, L., Lalonde, R., and D. Sullivan, "Earnings Losses of Displaced Workers," *American Economic Review* 83:4 (1993), 685–709.

Jenish, N., and I. Prucha "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields," *Journal of Econometrics* 150 (2009), 86–98.

Meyer, Bruce, "Natural and Quasi-Natural Experiments in Economics," *Journal of Business and Economic Statistics* 12 (1995), 151–162.

Moulton, Brent, "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables in Micro Units," this REVIEW 72:2 (1990), 334–338.

Newey, Whitney, and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing" (pp. 2113–2245), in Engle and McFadden (Eds.), *Handbook of Econometrics*, Vol. 4 (New York: Elsevier, 1994).

Rosenbaum, Paul, "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science* 17:3 (2002), 286–327.

Wooldridge, Jeffrey, "Cluster-Sample Methods in Applied Econometrics," *American Economic Review* 93:2 (2003), 133–138.

APPENDIX

A1 Proof of Proposition 1. First, a standard application of the partitioned inverse theorem makes it straightforward to show that

$$\begin{aligned}\widehat{\beta} = \beta + &\left(\frac{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{X}_{jt}\widetilde{X}_{jt}'}{N_0+N_1} - \frac{\left[\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}_{jt}\right]\left[\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}_{jt}'\right]}{(N_0+N_1)\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^{\,2}}\right)^{-1} \\ &\times\left(\frac{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{X}_{jt}\widetilde{\eta}_{jt}}{N_0+N_1} - \frac{\left[\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}_{jt}\right]\left[\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{\eta}_{jt}\right]}{(N_0+N_1)\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^{\,2}}\right).\end{aligned} \tag{A1}$$

Now consider each piece in turn.

First, assumption 1 states that

$$\frac{1}{N_0+N_1}\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{X}_{jt}\widetilde{X}_{jt}' \xrightarrow{p} \Sigma_x < \infty.$$

The mixing components of assumption 1 imply that a strong law of large numbers (LLN) applies here (see, e.g., Jenish & Prucha, 2009). This LLN and the zero-conditional expectation component of assumption 1 imply that

$$\frac{1}{N_0+N_1}\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{X}_{jt}\widetilde{\eta}_{jt} \xrightarrow{p} E\left[\sum_{t=1}^{T}\widetilde{X}_{jt}\widetilde{\eta}_{jt}\right] = 0.$$

For control groups $j > N_1$, the treatment does not vary over time, so $d_{jt} = \overline{d}_j$. Therefore,

$$\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^{\,2} = \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j-\overline{d}_t+\overline{d})^2 + \sum_{j=N_1+1}^{N_1+N_0}\sum_{t=1}^{T}(\overline{d}-\overline{d}_t)^2$$

where

$$\sum_{j=N_1+1}^{N_1+N_0}\sum_{t=1}^{T}(\overline{d}-\overline{d}_t)^2 = N_0\sum_{t=1}^{T}\left(\frac{1}{N_1+N_0}\sum_{\ell=1}^{N_1+N_0}\left[\left(\frac{1}{T}\sum_{\tau=1}^{T}d_{\ell\tau}\right)-d_{\ell t}\right]\right)^2 \xrightarrow{p} 0.$$

Now consider the other term,

$$\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j-\overline{d}_t+\overline{d})^2 \xrightarrow{p} \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)^2,$$

since $(\overline{d}_t-\overline{d})$ converges in probability to 0 due to the finite number of groups with intertemporal variation in treatments. Thus,

$$\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^2 \xrightarrow{p} \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)^2 > 0,$$

since $N_1 \geq 1$.

Now consider

$$\begin{aligned}\frac{1}{\sqrt{N_0+N_1}}\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}_{jt} &= \frac{1}{\sqrt{N_0+N_1}}\sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)\widetilde{X}_{jt} \\ &\quad + \sum_{t=1}^{T}(\overline{d}-\overline{d}_t)\frac{1}{\sqrt{N_0+N_1}}\sum_{j=1}^{N_1+N_0}\widetilde{X}_{jt} \\ &\xrightarrow{p} 0 \text{ as } N_0 \to \infty.\end{aligned}$$

This result follows because the first term involves a sum of a finite number of $O_p(1)$ random variables normalized by an $O(N_0)$ term and the second term is identically 0 due to differencing.

Likewise,

$$\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{\eta}_{jt} = \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)(\eta_{jt}-\overline{\eta}_j-\overline{\eta}_t+\overline{\eta}),$$

which is $O_p(1)$; thus,

$$\frac{1}{\sqrt{N_0+N_1}}\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{\eta}_{jt} \xrightarrow{p} 0.$$

Consistency for $\widehat{\beta}$ follows on plugging the pieces back into equation (A1) and applying Slutsky's theorem.

From the normal equation for $\widehat{\alpha}$, it is straightforward to show that

$$\widehat{\alpha} = \alpha + \frac{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{\eta}_{jt}}{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^2} + \left[\frac{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}'_{jt}}{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^2}\right](\beta-\widehat{\beta}). \quad \text{(A2)}$$

From above, we know that

$$\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^2 \xrightarrow{p} \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)^2$$

$$\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}_{jt} = \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j-\overline{d}_t+\overline{d})\widetilde{X}_{jt}$$

$$(\beta-\widehat{\beta}) \xrightarrow{p} 0.$$

Thus,

$$\left[\frac{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{X}'_{jt}}{\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}^2}\right](\beta-\widehat{\beta}) \xrightarrow{p} 0.$$

We showed above that

$$\sum_{j=1}^{N_0+N_1}\sum_{t=1}^{T}\widetilde{d}_{jt}\widetilde{\eta}_{jt} = \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)(\eta_{jt}-\overline{\eta}_j-\overline{\eta}_t+\overline{\eta}).$$

The variables $\overline{\eta}_t$ and $\overline{\eta}$ both converge to 0 in probability as $N_0 \to \infty$; therefore,

$$\sum_{j=N_1+1}^{N_1+N_0}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)\widetilde{\eta}_{jt} \xrightarrow{p} \sum_{j=1}^{N_1}\sum_{t=1}^{T}(d_{jt}-\overline{d}_j)(\eta_{jt}-\overline{\eta}_j).$$

Plugging these pieces into equation (A2) gives the result.

A2 Proof of Proposition 2. Since Γ is defined conditional on d_{jt} for $j=1,\ldots,N_1, t=1,\ldots,T$, every probability in this proof conditions on this set. To simplify the notation, we omit this explicit conditioning. Thus, every probability statement and distribution function in this proof should be interpreted as conditioning on d_{jt} for $j=1,\ldots,N_1, t=1,\ldots,T$.

It is convenient to define

$$\rho_{jt} = \frac{(d_{jt}-\overline{d}_j)}{\sum_{\ell=1}^{N_1}\sum_{t=1}^{T}(d_{\ell t}-\overline{d}_\ell)^2}.$$

For each $j=1,\ldots,N_1$, define the random variable

$$W_j \equiv \sum_{t=1}^{T}\rho_{jt}\eta_{jt}.$$

Let F_j be the distribution of W_j for $j=1,\ldots,N_1$.

Then note that

$$\begin{aligned}\Gamma(w) &= \Pr\left(\sum_{j=1}^{N_1}\sum_{t=1}^{T}\rho_{jt}\eta_{jt} < w\right) \\ &= \int\cdots\int 1\left(\sum_{j=1}^{N_1}W_j < w\right)dF_1(W_1)\ldots dF_{N_1}(W_{N_1}).\end{aligned}$$

We can also write

$$\widehat{\Gamma}(w) = \int\cdots\int 1\left(\sum_{j=1}^{N_1}W_j < w\right)d\widehat{F}_1(W_1;\widehat{\beta})\ldots d\widehat{F}_{N_1}(W_{N_1};\widehat{\beta}),$$

where $\widehat{F}_j(\cdot;\widehat{\beta})$ is the empirical CDF one gets from the residuals using the control groups only. That is, more generally,

$$\widehat{F}_j(w_j;b) \equiv \frac{1}{N_0}\sum_{m=1}^{N_0}1\left(\sum_{t=1}^{T}\rho_{jt}(\widetilde{Y}_{mt}-\widetilde{X}'_{mt}b) < w_j\right).$$

To avoid repeating the expression, we define

$$\phi_j(w_j,b) \equiv \Pr\left(\sum_{t=1}^{T}\rho_{jt}(\eta_{mt}-X'_{mt}(\beta-b)) < w_j\right).$$

Note that $\phi_j(w_j,\beta) = F_j(w_j)$. The proof strategy is first to demonstrate that $\widehat{F}_j(w_j;\widehat{\beta})$ converges to $\phi_j(w_j,\beta)$ uniformly over w_j. We will then show that $\widehat{\Gamma}(a)$ is a consistent estimate of $\Gamma(a)$.

Define

$$\widehat{\omega}_j = \sum_{t=1}^{T}\rho_{jt}(\overline{\eta}_t-\overline{X}_t(\beta-\hat{\beta})).$$

Let Ω be a compact parameter space for w and Θ a compact subset of the parameter space for $(\widehat{\beta},\widehat{\omega}_j)$ in which $(\beta,0)$ is an interior point.

For each $j = 1, \ldots, N_1$, consider the difference between $\widehat{F}_j(w_j; \widehat{\beta})$ and $\phi_j(w_j, \beta)$:

$$\begin{aligned}
&\sup_{w_j \in \Omega} |\widehat{F}_j(w_j; \widehat{\beta}) - \phi_j(w_j, \beta)| \\
&\quad = \sup_{w_j \in \Omega} \left| \frac{1}{N_0} \sum_{m=N_1+1}^{N_0} 1 \right. \\
&\qquad \left. \times \left(\sum_{t=1}^{T} \rho_{jt}(\eta_{mt} - \overline{\eta}_t - (X_{mt} - \overline{X}_t)'(\beta - \hat{\beta})) < w_j \right) - \phi_j(w_j, \beta) \right| \\
&\quad \leq \sup_{\substack{w_j \in \Omega \\ (b, \omega_j) \in \Theta}} \left| \frac{1}{N_0} \sum_{m=N_1+1}^{N_0} 1 \right. \\
&\qquad \left. \times \left(\sum_{t=1}^{T} \rho_{jt}(\eta_{mt} - X'_{mt}(\beta - b)) < w_j + \omega_j \right) - \phi_j(w_j + \omega_j, b) \right| \\
&\qquad + \Pr((\hat{\beta}, \widehat{\omega}_j) \notin \Theta) + \sup_{w_j \in \Omega} |\phi_j(w_j + \widehat{\omega}_j, \widehat{\beta}) - \phi_j(w_j, \beta)|. \qquad \text{(A3)}
\end{aligned}$$

First, consider $\sup_w |\phi_j(w_j, \widehat{\beta}) - \phi_j(w_j, \beta)|$. Using a standard mean-value expansion of ϕ, for some $(\widetilde{\omega}_j, \widetilde{\beta})$,

$$\begin{aligned}
&\sup_{w_j \in \Omega} |\phi_j(w_j + \widehat{\omega}_j, \widehat{\beta}) - \phi_j(w, \beta)| \\
&\quad = \sup_{w} \left| \frac{\partial \phi_j(w_j + \widetilde{\omega}_j, \widetilde{\beta})}{\partial \beta} (\widehat{\beta} - \beta) + \frac{\partial \phi_j(w_j, \widetilde{\beta})}{\partial w_j} (\widehat{\omega}_j) \right|.
\end{aligned}$$

To see that the derivative $\frac{\partial \phi_j(w_j, b)}{\partial b}$ is bounded, first note that

$$\frac{\partial \phi_j(w_j, b)}{\partial b} = E\left(f_j \sum_{t=1}^{T} \rho_{jt} \widetilde{X}'_{jt} \right),$$

where f_j is the density associated with F_j. Since f_j is bounded and X_{jt} has first moments, this term is bounded. Clearly $\frac{\partial \phi_j(w_j, b)}{\partial w_j}$ is also bounded for the same reason. Thus, $\sup_{w_j \in \Omega} |\phi_j(w + \widehat{\omega}_j, \widehat{\beta}) - \phi_j(w_j, \beta)|$ converges to 0 since $\widehat{\beta}$ is consistent.

Since $(\widehat{\beta}, \widehat{\omega}_j)$ converges in probability to $(\beta, 0)$, an interior point of Θ, $\Pr((\hat{\beta}, \widehat{\omega}_j) \notin \Theta)$ converges to 0.

Next consider the first term on the right side of equation (A3). Note that the function

$$1\left(\sum_{t=1}^{T} \rho_{jt}(\widetilde{Y}_{mt} - \widetilde{X}'_{mt} b) < w_j + \omega_j \right)$$

is continuous at each (b, w, ω) with probability 1, and its absolute value is bounded by 1, so applying lemma 2.4 of Newey and McFadden (1994), $\widehat{F}_j(w_j; b)$ converges uniformly to $\phi(w_j, b)$. Thus, putting the three pieces of equation (A3) together gives

$$\sup_{w_j \in \Omega} |\widehat{F}(w_j; \widehat{\beta}) - \phi(w_j, \beta)| \xrightarrow{p} 0.$$

Now to see that $\widehat{\Gamma}(w)$ converges to $\Gamma(w)$, we can write

$$\begin{aligned}
&|\widehat{\Gamma}(w) - \Gamma(w)| \\
&\quad = \left| \left\{ \int \left[\widehat{F}_1 \left(\left[w - \sum_{j=2}^{N_1} W_j \right]; \widehat{\beta} \right) - F_1 \left(w - \sum_{j=2}^{N_1} W_j \right) \right] \right. \right. \\
&\qquad \left. \times d\widehat{F}_2(W_2; \widehat{\beta}) \ldots d\widehat{F}_{N_1}(W_{N_1}; \widehat{\beta}) \right\} \\
&\qquad + \left\{ \int \left[\widehat{F}_2 \left(\left[w - \sum_{\substack{j=1 \\ j \neq 2}}^{N_1} W_j \right]; \widehat{\beta} \right) - F_2 \left(w - \sum_{\substack{j=1 \\ j \neq 2}}^{N_1} W_j \right) \right] \right. \\
&\qquad \left. \times dF_1(W_1) d\widehat{F}_3(W_3; \widehat{\beta}) \ldots d\widehat{F}_{N_1}(W_{N_1}; \widehat{\beta}) \right\} \\
&\qquad + \ldots \\
&\qquad + \left\{ \int \left[\widehat{F}_{N_1} \left(\left[w - \sum_{j=1}^{N_1 - 1} W_j \right]; \widehat{\beta} \right) - F_{N_1} \left(w - \sum_{j=1}^{N_1 - 1} W_j \right) \right] \right. \\
&\qquad \left. \left. \times dF_1(W_1) \ldots dF_{N_1 - 1}(W_{N_1}) \right\} \right|.
\end{aligned}$$

Since each $\widehat{F}_j(w; \widehat{\beta})$ converges uniformly to $F_j(w)$, the right-hand side of this expression must converge to 0, so $\widehat{\Gamma}(a)$ converges to $\Gamma(a)$.