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

Timothy Conley
Christopher Taber

Inference with "Difference in Differences" with a Small Number of Policy Changes
Timothy Conley and Christopher Taber
NBER Technical Working Paper No. 312
July 2005
JEL No.

ABSTRACT

Difference in differences methods have become very popular in applied work. This paper provides a new method for inference in these models when there are a small number of policy changes. This situation occurs in many implementations of these estimators. Identification of the key parameter typically arises when a group "changes" some particular policy. The asymptotic approximations that are typically employed assume that the number of cross sectional groups, N, times the number of time periods, T, is large. However, even when N or T is large, the number of actual policy changes observed in the data is often very small. In this case, we argue that point estimators of treatment effects should not be thought of as being consistent and that the standard methods that researchers use to perform inference in these models are not appropriate. We develop an alternative approach to inference under the assumption that there are a finite number of policy changes in the data, using asymptotic approximations as the number of non-changing groups gets large. In this situation we cannot obtain a consistent point estimator for the key treatment effect parameter. However, we can consistently estimate the finite-sample distribution of the treatment effect estimator, up to the unknown parameter itself. This allows us to perform hypothesis tests and construct confidence intervals. For expositional and motivational purposes, we focus on the difference in differences case, but our approach should be appropriate more generally in treatment effect models which employ a large number of controls, but a small number of treatments. We demonstrate the use of the approach by analyzing the effect of college merit aide programs on college attendance. We show that in some cases the standard approach can give misleading results.

Timothy Conley
University of Chicago GSB
1101 E 58th Street
Chicago, IL 60637
tconley1@gsb.uchicago.edu

Christopher Taber
Department of Economics
Nortwestern University
2003 Sheridan Road
Evanston, IL 60208
and NBER
ctaber@northwestern.edu

1   Introduction

Difference in differences methods have become very popular in applied work. These models are typically quite easy to implement and to interpret. However, performing inference with these models is often difficult. The goal of this paper is to address one particular aspect that is likely to be very important in many implementations of these estimators. Identification of the key parameter often arises when a group "changes" some particular policy. We use the notation N_0 to refer to the number of "treatment" groups that change their policy in the data and N_1 to refer to the number of "control" groups who do not change their policy. The asymptotic approximations that are typically employed assume that the number of both groups, N_0 and N_1, are large. However, even when the total number of groups is large, the number of actual policy changes observed in the data is often very small. In this case, we argue that point estimators of treatment effects should not be thought of as being consistent and that the standard methods that researchers use to perform inference in these models are not appropriate. We develop an alternative approach to inference under the assumption that N_0 is finite, using asymptotic approximations that let N_1 grow large. While our point estimator of the treatment effect parameter is not consistent, we can consistently estimate its finite-sample distribution up to the true value of the parameter itself. This allows us to test the hypothesis that this parameter takes on any given value and to construct a confidence interval for it by 'inverting' a test statistic. For expositional and motivational purposes, we focus on the difference in differences case, but our approach is appropriate more generally in treatment effect models in which there are a large number of controls, but a small number of treatments.

Our approach is related to a large body of existing work on difference and difference models and inference in more general group effect models.[1] It is complementary to typical approaches focusing on situations where the number of treatment and control groups, N_0 and N_1, are both large (e.g. Moulton, 1990) or both small (e.g. Donald and Lang, 2002).

Our approach is in the spirit of comparisons of changes in treatment groups to control groups often done by careful applied researchers. Anderson and Meyer (2000) provide a nice example of the type of question for which our methodology is particularly well suited. They examine the effect of changes in unemployment insurance payroll in Washington state

---

[1]See for example Angrist and Krueger (1999) and Meyer (1995) for overviews of difference in difference methods. Wooldrige (2003) provides a concise survey of group effect models.

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 in time in order determine whether it is an outlier consistent with a policy effect. This application of exact inference is very much in the spirit of our approach. Our approach can also be thought of as a generalization/formalization of other exact inference type procedures like the 'placebo laws' experiments that Bertrand, Duflo, and Mullainathan (2004) use to obtain critical values for hypotheses testing under a particular null hypothesis about the distribution of the treatment indicator.[2]

There are so many examples of difference-in-differences-style empirical work that we do not attempt to survey them. Bertrand, Duflo, and Mullainathan (2004) provide a nice overview. However, we will mention a few examples for which our approach seems appropriate. As mentioned above, Anderson and Meyer (2000) look at changes in Washington state using other states as controls. Another example is the effects of merit aid programs on college attendance. For example, in some of her specifications Dynarski (2004) identifies the effect using a policy change from a single state (Georgia). Finally, Gruber, Levine, and Staiger (1999) use comparisons between the five treatment states that legalized abortion prior to Roe v. Wade versus the remaining states.

One can also find many studies which use a small number of both 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. 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 includes only one control group (eastern Pennsylvania), but they could have collected data from many "control states" to contrast with the available treatment state. Another famous example is Card (1990) who examines the effect of the Mariel Boatlift on

---

[2]Bertrand, Duflo, and Mullainathan (2004) concern themselves primarily with serial correlation and mostly use a standard asymptotic approach, but at one point also discuss an exact test using a 'placebo laws' experiment. The placebo laws experiment of Bertrand et. al. recovers the exact distribution of a treatment effect parameter (conditional on state and time fixed effects) for group-time aggregates under a particular null hypothesis. Our thought experiment is somewhat different as we use the control groups to obtain a consistent estimate of the distribution of a treatment effect parameter, which is then used to conduct small sample inference for the treatment group. Our setup allows for a richer set of models in terms of regressors and unobservable structure; special cases of our setup will result in inference analogous to that obtained via the Bertrand et. al. simulation.

the Miami labor market. He uses four comparison cities as controls, but could have used many additional cites.

The closest analog to our approach to inference in econometrics is work on testing for structural breaks. In particular, work on testing for end-of-sample stability/structural breaks such as that by, e.g., Dufor, Ghysels, and Hall (1994) and Andrews (2003) is quite related to our basic approach. These authors consider the problem of testing for a structural break over a fixed and perhaps very short interval at the end of a sample, analogous to our N_0 observations on policy changers. They develop tests that are asymptotically valid as the number of observations before the potential break point grows, holding fixed the number of points after the break point. This is analogous to our taking large N_1 limits with fixed N_0. Asymptotically valid critical values for these tests rely on using the time span before the potential break to get consistent estimates of the distribution of a test statistic formed from data during the fixed end-of-sample interval. Andrews accomplishes this via a procedure akin to subsampling and Dufor, Ghysels, and Hall (1994) use semi-nonparametric density estimators. Again, our method for constructing interval estimates is roughly analogous in that we use consistent model estimates obtained from the N_1 non-changers to characterize the small-sample distribution of the treatment parameter.

Basic Model and Problem

We consider a case in which we have repeated cross section data[3] from different groups (e.g. U.S. states) and time periods. To give the main intuition for the result consider a simple version of the model with an individual i, with outcome Y_i who is in group $j(i)$, and observed at time $t(i)$. We model his outcome as

$$Y_i = \alpha d_{j(i)t(i)} + \theta_{j(i)} + \gamma_{t(i)} + \eta_{j(i)t(i)} + \varepsilon_i \tag{1}$$

where d_{jt} is the policy variable of interest.[4] The parameter θ_j is a fixed effect for group $j = 1, ..., N_0 + N_1$ that will be common to group j across time, γ_t is a time effect that is common across all groups but varies across time $t = 1, ..., T$, η_{jt} is a group×time random effect that varies across groups and time, and ε_i is an individual specific error term. We assume that ε_i is i.i.d. with $E(\varepsilon_i) = 0$ and that it is independent of all other terms in the

---

[3]Extension of these results to panel data is straight forward. We assume throughout that we are using cross sectional data to economize already complicated notation.

[4]We focus on linear models, but extensions to nonlinear models seem feasible combining the approach here with Athey and Imbens (2002).

model. 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 assume throughout this paper that T is fixed. The primary goal is to estimate the treatment parameter α.

Initial work using this model ignored η_{jt} which leads to the classic difference in differences estimator. In this case one can obtain a consistent estimate of α using only two groups and two time periods. In particular assume that

$$\eta_{jt} \equiv 0 \text{ for all } j,t \tag{2}$$

and denote the two groups $j = \{0,1\}$ and two time periods $t = \{0,1\}$. Suppose further that the policy variable is binary, and for group 0, there is no change in the treatment $(d_{00} = d_{01} = 0)$, but for group 1 the treatment is enacted between the periods zero and one $(d_{10} = 0, d_{11} = 1)$. We define the notation $\overline{Y}_{jt}$ and $\overline{\varepsilon}_{jt}$ to denote the averages of Y_i and ε_i across all the individuals in group j at time t, (i.e. $\overline{Y}_{jt} = \frac{1}{|M(j,t)|} \sum_{i \in M(j,t)} Y_i$). The classic difference in differences estimator is:

$$
\begin{aligned}
\widehat{\alpha}_{DD} &\equiv \left(\overline{Y}_{11} - \overline{Y}_{10}\right) - \left(\overline{Y}_{01} - \overline{Y}_{00}\right) \\
&= (\alpha + \theta_1 + \gamma_1 - \theta_1 - \gamma_0) - (\theta_0 + \gamma_1 - \theta_0 - \gamma_0) + (\overline{\varepsilon}_{11} - \overline{\varepsilon}_{10}) - (\overline{\varepsilon}_{01} - \overline{\varepsilon}_{00}) \\
&= (\alpha + \gamma_1 - \gamma_0) - (\gamma_1 - \gamma_0) + (\overline{\varepsilon}_{11} - \overline{\varepsilon}_{10}) - (\overline{\varepsilon}_{01} - \overline{\varepsilon}_{00}) \\
&= \alpha + (\overline{\varepsilon}_{11} - \overline{\varepsilon}_{10}) - (\overline{\varepsilon}_{01} - \overline{\varepsilon}_{00}) \\
&\xrightarrow{p} \alpha.
\end{aligned}
$$

The group and time effects of course drop out due to the differencing, with large samples within each group/time the ε terms vanish, and if (2) holds $\widehat{\alpha}_{DD}$ is a consistent estimator of α as $|M(j,t)|$ gets large for each group/period.

In the past decade or so, researchers have recognized that (2) is an extremely strong assumption and they have tried to account for η effects in estimation (see e.g. Moulton, 1990). It is easy to show that two group/two time period differences in difference is not consistent without assuming (2). In that case

$$
\begin{aligned}
\widehat{\alpha}_{DD} &= \alpha + (\eta_{11} - \eta_{10}) - (\eta_{01} - \eta_{00}) + \\
&\quad (\overline{\varepsilon}_{11} - \overline{\varepsilon}_{10}) - (\overline{\varepsilon}_{01} - \overline{\varepsilon}_{00}) \\
&\xrightarrow{p} \alpha + (\eta_{11} - \eta_{10}) - (\eta_{01} - \eta_{00}).
\end{aligned}
$$

The term involving $(\eta_{11} - \eta_{10}) - (\eta_{01} - \eta_{00})$ does not vanish as the number of observed individuals at each group/time period increases. Our focus is on analogs of this situation

where a fixed number of groups with policy changes imply that the randomness due to η cannot be eliminated by cross-group averaging.[5]

Many empirical economists recognize this problem and augment their 'natural experiment' by collecting data from additional groups that do not experience treatment changes and/or additional time periods. For simplicity, assume that only the first group experiences a treatment change after period t^*, so the binary treatment indicator for group one can be written as:

$$d_{1t} = 1(t > t^*)$$

(where $1(\cdot)$ is the indicator function) and for all other groups $d_{jt} = d_{j\tau}$ for all t and τ. Also to keep the exposition simple, assume that all cell sizes are the same ($|M(j,t)| = m$). Note that d_{jt} for control group j could be all zeros or all ones. Consider estimating the model (1) by using fixed effects regression, controlling for group and time effects through dummy variables. Let $\widehat{\alpha}_{FE}$ be the regression estimate of α. It is straight forward to show that this can be written as a difference of differences

$$\widehat{\alpha}_{FE} = \alpha + \left[\frac{1}{T - t^*} \sum_{t=t^*+1}^{T} (\eta_{1t} + \overline{\varepsilon}_{1t}) - \frac{1}{t^*} \sum_{t=1}^{t^*} (\eta_{1t} + \overline{\varepsilon}_{1t}) \right] \qquad (3)$$
$$- \left(\frac{1}{(N-1)} \sum_{j=2}^{N} \frac{1}{(T - t^*)} \sum_{t=t^*+1}^{T} \left(\eta_{jt} + \overline{\varepsilon}_{jt} \right) - \frac{1}{(N-1)} \sum_{j=2}^{N} \frac{1}{t^*} \sum_{t=1}^{t^*} \left(\eta_{jt} + \overline{\varepsilon}_{jt} \right) \right)$$

The terms involving $\overline{\varepsilon}_{jt}$ will all vanish as within-group sample sizes grow (i.e. $m \to \infty$). If $E(\eta_{jt} \mid d_{jt}) = 0$ then this yields an unbiased estimate of α. However, $\widehat{\alpha}_{FE}$ is not consistent as the number of groups grows since the term in brackets approaches $\left(\frac{1}{T-t^*} \sum_{t=t^*+1}^{T} \eta_{1t} - \frac{1}{t^*} \sum_{t=1}^{t^*} \eta_{1t} \right)$ as either m or N get large.

This problem is rarely acknowledged in empirical work and researchers often ignore it when calculating standard errors. In practice, if the error terms are truly normally distributed, standard methods will yield the correct standard errors (if degree of freedom corrections are used, see Donald and Lang, 2001). However, if the distribution of η_{jt} is sufficiently different from normal, the standard errors may be very misleading.

The example presented in equation (3) considered the case of a single treatment group. Clearly the same problem holds when the number of treatment groups is small.[6] The goal of

---

[5] Of course with access to many groups that experience a policy change, averaging across groups can yield a consistent estimator of α under suitable assumptions about η.

[6] Clearly the precise sample size that constitutes "small" is an empirical question that is beyond the scope of this paper.

this paper is to show that even though one can not obtain consistent estimates of α in these cases, it is still possible to perform inference. We assume that there are a finite number of policy changes in the data N_0, but approximate the distribution of our estimator of α taking limits as the number of control groups (N_1) gets large.

The remainder of this paper is organized into four sections. In Section 2, we present regression models for both group and individual-level data. In each case we show how to perform inference about the parameter α. Extensions to limited dependent variables are discussed in Section 3. Section 4 of the paper provides an illustrative example application estimating the effect of merit aid programs upon college attendance. Finally, Section 5 offers brief conclusions.

2 Models

This section presents two models. In the first, we assume that we have one observation per group×time cell (e.g. data that is collected at the state×year level). In the second, we allow multiple observations per group×time. For the second model we focus on approximations in which the number of individuals in a group×time cell remains fixed, suitable for applications where at least some of the groups are small.

2.1 Model 1

We start by discussing the analog of equation (1) defined at the group×time level and allowing for regressors. We assume that

$$Y_{jt} = \alpha d_{jt} + X'_{jt}\beta + \theta_j + \gamma_t + \eta_{jt}. \tag{4}$$

Note that we no longer restrict d_{jt} to be binary.

The crucial assumption for difference in differences is that changes in η_{jt} are unrelated to imposition of the treatment. In order to perform inference in our case, we also assume that $(\eta_{j1}, ..., \eta_{jT})$ is independent and identically distributed across groups. Within a group, we allow arbitrary correlation over time.

Assumption 1.1 $\left(\left(X_{j1}, \eta_{j1} \right), ..., \left(X_{jT}, \eta_{jT} \right) \right)$ *is independent and identically distributed across units;* $\left(\eta_{j1}, ..., \eta_{jT} \right)$ *is independent of* $(d_{j1}, ..., d_{jT})$ *and* $(X_{j1}, ..., X_{jT})$ *and has a bounded density and bounded support; and all random variables have finite second moments.*

6

The key problem motivating our approach is that for many groups there is little variation in d_{jt}. Following the notation in the introduction, define N_0 as the number of groups for which d_{jt} changes during the sample period and let N_1 represent the number of remaining groups. We will refer to the N_0 changers as treatment groups and the remaining non-changing groups as controls. Without loss of generality, define the index j so that the $j = 1, ..., N_0$ represents the observations for which d_{jt} changes at some time t and $j = N_0 + 1, ..., N_0 + N_1$ represents the observations for which d_{jt} is unchanged for the whole sample. Thus if $j > N_0$ then for any $t = 1, ..., T, d_{jt} = d_{j1}$. We treat N_0 and T as fixed, taking limits as N_1 grows large. We are assuming throughout that at least one group changes its policy so that $N_0 \geq 1$.

For any random variable Z_{jt}, define

$$\overline{Z}_j = \frac{1}{T} \sum_{t=1}^{T} Z_{jt}$$

$$\overline{Z}_t = \frac{1}{N_1 + N_0} \sum_{j=1}^{N_1 + N_0} Z_{jt}$$

$$\overline{Z} = \frac{1}{T} \frac{1}{N_1 + N_0} \sum_{t=1}^{T} \sum_{j=1}^{N_1 + N_0} Z_{jt}$$

$$\widetilde{Z_{jt}} = Z_{jt} - \overline{Z}_j - \overline{Z}_t + \overline{Z}$$

The essence of 'difference in differences' is that we can rewrite regression model (4) as

$$\widetilde{Y}_{jt} = \alpha \widetilde{d}_{jt} + \widetilde{X}'_{jt} \beta + \widetilde{\eta}_{jt}. \tag{5}$$

One 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 (5).

We need an assumption to guarantee that after taking out time and fixed effects, $\widetilde{X}_{jt}$ is not collinear.

Assumption 1.2

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

where Σ_x is finite and of full rank.

In Proposition 1.1 we show that OLS yields a consistent estimator of β and we derive the limiting distribution of $\widehat{\alpha}$.

Proposition 1.1 *Under Assumptions 1.1-1.2,*

$$\widehat{\beta} \xrightarrow{p} \beta$$

$$(\widehat{\alpha} - \alpha) \xrightarrow{p} \frac{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right) \left(\eta_{jt} - \overline{\eta}_j\right)}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)^2}$$

as $N_1 \to \infty$.

In the expression above, $\left(\eta_{jt} - \overline{\eta}_j\right)$ appears rather than the original residual $\widetilde{\eta}_{jt}$. This results because both $\overline{\eta}_t$ and $\overline{\eta}$ converge in probability to zero as N_1 gets large.

The fact that $\widehat{\alpha}$ is not consistent does not prevent us from conducting inference about the true value of α. The difference between $\widehat{\alpha}$ and α depends on two variables: d_{jt} and $\left(\eta_{jt} - \overline{\eta}_j\right)$. The d_{jt} are observable and the distribution of $\left(\eta_{jt} - \overline{\eta}_j\right)$ can be estimated from the control groups, $j > N_0$. Therefore, we can estimate the asymptotic $(N_1 \to \infty)$ conditional distribution of $(\widehat{\alpha} - \alpha)$ given d_{jt} for the treatment groups. We state this as Proposition 1.2 below. Estimation of the distribution of $\widehat{\alpha}$ allows hypothesis testing on α and construction of confidence intervals for $(\widehat{\alpha} - \alpha)$.

To see how the distribution of $\left(\eta_{jt} - \overline{\eta}_j\right)$ can be estimated, consider estimation of the residual for a member of the control group (i.e. $j > N_0$),

$$\widetilde{Y}_{jt} - \widetilde{X}'_{jt}\widehat{\beta} = \widetilde{X}'_{jt}(\widehat{\beta} - \beta) + \left(\eta_{jt} - \overline{\eta}_j - \overline{\eta}_t + \overline{\eta}\right)$$
$$\xrightarrow{p} \left(\eta_{jt} - \overline{\eta}_j\right)$$

hence the distribution of $\left(\eta_{jt} - \overline{\eta}_j\right)$ is trivially identified using residuals for groups $j > N_0$.

From this it is straight forward to show how to estimate the asymptotic distribution of $\widehat{\alpha}$ up to α. Let

$$\Gamma(a) \equiv \plim_{N_1 \to \infty} \Pr((\widehat{\alpha} - \alpha) < a \mid \{d_{jt}, j = 1, .., N_0, t = 1, ..., T\}).$$

We will estimate $\Gamma(a)$ with the analogous empirical distribution of residuals from the control groups. For the $N_0=1$ case we can estimate $\Gamma(a)$ using

$$\widehat{\Gamma}(a) \equiv \frac{1}{N_1} \sum_{\ell=N_0+1}^{N_0+N_1} 1\left(\frac{\sum_{t=1}^{T}\left(d_{1t} - \overline{d}_1\right)\left(\widetilde{Y}_{\ell t} - \widetilde{X}'_{\ell t}\widehat{\beta}\right)}{\sum_{t=1}^{T}\left(d_{1t} - \overline{d}_1\right)^2} < a\right).$$

More generally

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

Proposition 1.2 *Under Assumptions 1.1 and 1.2, $\widehat{\Gamma}(a)$ converges uniformly to $\Gamma(a)$.*

To see the usefulness of this result, first consider testing the null hypothesis

$$H_0 : \alpha = \alpha_0$$

conditioning on the observed sequence d_{jt}, $j = 1, .., N_0$, $t = 1, ..., T$. We could define an approximate 95% acceptance region by $\left(\widehat{A}_1, \widehat{A}_2\right)$ as the maximum value of A_{lower} and minimum value of A_{upper} such that

$$
\begin{aligned}
\widehat{\Gamma}\left(A_{\text{upper}} - \alpha_0\right) &\geq 0.975 \\
\widehat{\Gamma}\left(A_{\text{lower}} - \alpha_0\right) &\leq 0.025.
\end{aligned}
$$

Then we reject if $\widehat{\alpha}$ is outside $\left[\widehat{A}_1, \widehat{A}_2\right]$. Under the null hypothesis, the rejection probability will converge to 5% as $N_1 \to \infty$. We define an approximate confidence interval of α as the set of α_0 for which we do not reject the null hypothesis. As $N_1 \to \infty$, the coverage probability of this interval will converge to 95%.

2.2   Model 2

Now we augment the model to allow for individual data. Since difference-in-differences methods are most commonly used with repeated cross-section data, we let i index an individual who is observed within a single group at a single time period. As in the introduction, we use the notation $j(i)$ to represent the group to which individual i belongs, and $t(i)$ to represent the time period in which we observe individual i. We also continue to assume that the data come from repeated cross sections so that we only observe individual i during one time period. We see no reason why extension to panel data would be problematic. Our model is analogous to (1) with the addition of regressors:

$$Y_i = \alpha d_{j(i)t(i)} + X_i'\beta + \theta_{j(i)} + \gamma_{t(i)} + \eta_{j(i)t(i)} + \varepsilon_i. \tag{6}$$

Given that the model is defined somewhat differently than in the previous section, we need to modify the assumptions slightly:

Assumption 2.1 $\left\{\eta_{jt}, \{X_i : i \in M(j,t)\}\right\}_{t=1}^T$ *is i.i.d. across groups X_i is i.i.d. within group for all j and t, and all second moments exist. Furthermore the distribution of $\left(\eta_{j1}, ..., \eta_{jT}\right)$ is independent of $(d_{j1}, ..., d_{jT})$ and $\{X_i : i \in M(j,t)\}_{t=1}^T$ and has a bounded density and bounded support.*

We add the additional assumption that

Assumption 2.2 ε_i *is i.i.d. across individuals and is independent of* (d_{jt}, X_i, η_{jt}) *and* $E(\varepsilon_i) = 0$.

We use notation analogous to the above for Model 1. First, we modify the notation for averages across time within a group. For a generic variable Z_i define

$$\overline{Z}_j = \frac{\sum_{t=1}^{T} \sum_{i \in M(j,t)} Z_i}{\sum_{t=1}^{T} |M(j,t)|}.$$

Since in general, the number of individuals varies across (j,t) cells, derivation of the difference in differences operator requires additional notation. We need to formally define the full set of indicators for groups $\{g_{\ell i}\}_{\ell=1}^{N_0+N_1}$ and time periods, $\{p_{\tau i}\}_{\tau=1}^{T-1}$ so that

$$g_{\ell i} \equiv 1(\ell = j(i)) \tag{7}$$

$$p_{\tau i} \equiv 1(\tau = t(i)). \tag{8}$$

Further define G_i and P_i as the vectors of these dummy variables,

$$G_i \equiv \begin{bmatrix} g_{1i} & g_{2i} & \cdots & g_{N_o+N_1,i} \end{bmatrix}' \tag{9}$$

$$P_i \equiv \begin{bmatrix} p_{1i} & p_{2i} & \cdots & p_{T-1,i} \end{bmatrix}'. \tag{10}$$

Then for any individual-specific random variable Z_i, let $\widetilde{Z}_i$ be the residual from a linear regression of Z_i on $\{g_{\ell i}\}_{\ell=1}^{N_0+N_1}$ and $\{p_{\tau i}\}_{\tau=1}^{T-1}$. That is

$$\widetilde{Z}_i \equiv Z_i - \begin{bmatrix} G_i \\ P_i \end{bmatrix}' \left(\sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{h \in M(j,t)} \begin{bmatrix} G_h \\ P_h \end{bmatrix} \begin{bmatrix} G_h \\ P_h \end{bmatrix}' \right)^{-1} \left(\sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{h \in M(j,t)} \begin{bmatrix} G_h \\ P_h \end{bmatrix} Z_h \right).$$

We need a regularity condition to guarantee enough degrees of freedom that regressions upon time and group indicators can be run.

Assumption 2.3

$$\frac{\sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i \in M(j,t)} P_i P_i'}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|} -$$

$$\frac{\sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i \in M(j,t)} P_i G_i' \left(\sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i \in M(j,t)} G_i G_i' \right)^{-1} \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i \in M(j,t)} G_i P_i'}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|}$$

$$\xrightarrow{p} \Omega$$

where Ω *is of full rank.*

Under this condition, we can rewrite the model as:

$$\widetilde{Y}_i = \alpha \widetilde{d}_{j(i)t(i)} + \widetilde{X}'_i \beta + \widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i. \tag{11}$$

We estimate α and β in equation (11) by OLS, letting $\widehat{\alpha}$ and $\widehat{\beta}$ denote the corresponding estimators. This requires the usual OLS rank condition stated as

Assumption 2.4

$$\frac{\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} \sum_{i \in M(j,t)} \widetilde{X}_i \widetilde{X}'_i}{\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} |M(j,t)|} \xrightarrow{p} \Sigma_x$$

where Σ_x is finite and of full rank.

When each (j,t) cell has a large sample, inference in model (11) can be conducted in essentially the same manner as for Model 1 since averaging within time×group cells effectively eliminates $\widetilde{\varepsilon}_i$. For the sake of completeness, in the Appendix, we present a consistency result for $\widehat{\beta}$ and the distribution of $\widehat{\alpha}$, when $|M(j,t)|$ and N_1 grow.

However, we focus on the fixed-$|M(j,t)|$ case because we anticipate that it will be more appropriate for a majority of applications. This is because large $|M(j,t)|$ approximations must work in *all* group/time period cells-not just on average–in order for the resulting approximation for the distribution of $(\widehat{\alpha} - \alpha)$ to perform well. There will routinely be substantial heterogeneity in $|M(j,t)|$ across groups, e.g. states, with the smallest $|M(j,t)|$ perhaps best considered a small rather than large sample. For example, in our illustrative example application using states as groups, $|M(j,t)|$ ranges from 383 to 15. We characterize the fixed $|M(j,t)|$ case in the following manner:

Assumption 2.5 *For each $j = 1, .., N_0 + N_1$, $|M(j,t)|$ for $t = 1, ..., T$ is fixed and finite. In addition, $(|M(j,t)|, t = 1, ..., T)$ is independent and identically distributed across j for $j > N_0$ and jointly independent of η and ε.*

Note that we have assumed that $|M(j,t)|$ is i.i.d. for $j > N_0$, but we allow the distribution of $|M(j,t)|$ for $j \leq N_0$ to differ from the distribution of $|M(j,t)|$ for $j > N_0$. For example, if larger states were likely to implement policy changes earlier, the distribution of $|M(j,t)|$ for $j \leq N_0$ would stochastically dominate the distribution of $|M(j,t)|$ for $j > N_0$.

Proposition 2.1 provides a statement of consistency for $\widehat{\beta}$ as N_1 grows large and the asymptotic distribution of $(\widehat{\alpha} - \alpha)$.

Proposition 2.1 *Under Assumptions 2.1-2.5,*

$$\widehat{\beta} \xrightarrow{p} \beta$$

$$(\widehat{\alpha} - \alpha) \xrightarrow{p} \frac{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(\sum_{i \in M(j,t)} \left(d_{jt} - \overline{d}_j \right) \left(\eta_{jt} - \overline{\eta}_j + \varepsilon_i - \overline{\varepsilon}_j \right) \right)}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left(d_{jt} - \overline{d}_j \right)^2}$$

as $N_1 \to \infty.$

Analogous to model 1, the expression for $(\widehat{\alpha} - \alpha)$ involves $(\eta_{jt} - \overline{\eta}_j + \varepsilon_{jt} - \overline{\varepsilon}_j)$ rather than $\left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i \right)$. To see why, consider the regression of $\eta_{j(i)t(i)} + \varepsilon_i$ on group and time indicators. The coefficient on each group indicator converges to $\left(\overline{\eta}_j + \overline{\varepsilon}_j \right)$ while the coefficients on the time indicators converge to zero since these random variables both have expectation zero.

A number of different options are available for estimating the distribution of $(\widehat{\alpha} - \alpha)$. In principle, with enough groups, one could simply estimate the distribution of residuals conditional on the values of $|M(j,t)|$ for the treatment states. We suspect that this procedure would not work well in most applications since the number of control groups is likely not large enough for this to be a useful approximation. Instead we take advantage of our model's structure to estimate the distribution of components of $(\eta_{jt} - \overline{\eta}_j + \varepsilon_{jt} - \overline{\varepsilon}_j)$.

More specifically define

$$
\begin{aligned}
v_i &\equiv Y_i - X_i'\beta \\
&= \left[\alpha d_{j(i)t(i)} + \gamma_{t(i)} + \theta_{j(i)} + \eta_{j(i)t(i)} \right] + \varepsilon_i \\
&\equiv \eta^*_{j(i)t(i)} + \varepsilon_i.
\end{aligned}
\tag{12}
$$

Note that since we are using control groups only, the term in brackets is constant across individuals within the same time and group and is independent of ε_i. Our goal is to simulate the distribution of $(\eta_{jt} - \overline{\eta}_j)$ and $(\varepsilon_i - \overline{\varepsilon}_j)$. Note that since $\left(\alpha d_{j(i)t(i)} + \theta_{j(i)} \right)$ does not vary across time within a group and γ_t does not vary across groups, knowledge of the joint distribution of η^*_{jt} is sufficient for knowledge of $(\eta_{jt} - \overline{\eta}_j)$. If we have a consistent estimate of the distribution of ε_i, we can consistently estimate the distribution of $(\varepsilon_i - \overline{\varepsilon}_j)$.

Thus our goal is to obtain consistent estimates of the distribution of η^*_{jt} and the distribution of ε_i. This is a standard deconvolution problem. We will first show that these distributions are identified making use of a well known result. We report Theorem 2.1.1 in Prakasa Rao (1992) which he attributes to Kotlarski (1967) as Theorem 2.2.

Theorem 2.2 (Kotlarski, Prakasa Rao) *Suppose that $X_1, X_2,$and X_3 are independent real valued random variables. Define*

$$
\begin{aligned}
Z_1 &= X_1 - X_3 \\
Z_2 &= X_2 - X_3
\end{aligned}
$$

if the characteristic function of (Z_1, Z_2) does not vanish then the joint distribution of (Z_1, Z_2) determines the distributions of (X_1, X_2, X_3) up to a change of the location.

To apply the theorem we need one additional assumption.

Assumption 2.6 *The characteristic functions of ε_i and η_{jt}^* do not vanish.*

Given that, we can show identification of the distribution of $\widehat{\alpha}$.

Proposition 2.3 *Under Assumptions 2.1-2.6, the distribution of $(\widehat{\alpha} - \alpha)$ is identified from knowledge of d_{jt} and $|M(j,t)|$ from the treatment groups and the joint distribution of v_i for the control groups.*

Many options are available to estimate the distributions of ε_i and η_{jt}^*. In this section we present one possible estimator which is perhaps the most common way to estimate this type of mixture model in economics. We derive a sieve estimator assuming that $(\eta_{j1}^*, ..., \eta_{jT}^*)$ has finite support. This approach is most commonly associated with Heckman and Singer (1984). We propose to estimate the model in two steps. First we run the fixed effects model (11). We can construct the residual for each individual in the control set

$$
\begin{aligned}
\widehat{v}_i &\equiv Y_i - X_i'\widehat{\beta} \tag{13} \\
&= X_i'\left(\beta - \widehat{\beta}\right) + \eta_{j(i)t(i)}^* + \varepsilon_i.
\end{aligned}
$$

Our goal is to separately estimate the distribution of ε_i from η_{jt}^*. We parameterize η_{jt}^* to take on K_1 values with each value taking the value $\eta^{(\kappa_1)}$ with probability $P_1^{(\kappa_1)}$for $\kappa_1 = 1, .., K_1$. We let ε be a mixture of normals that take on K_2 values with mean and standard deviation $(\mu^{(\kappa_2)}, \sigma)$ with probability $P_2^{(\kappa_2)}$ for $\kappa_2 = 1, ..., K_2$. The objective function is

$$
\sum_{j=N_0+1}^{N_0+N_1} \log \left(\sum_{\kappa_1=1}^{K_1} \prod_{t=1}^{T} \prod_{i \in M(j,t)} \sum_{\kappa_2=1}^{K_2} \phi \left(\frac{\widehat{v}_i - \eta_t^{(\kappa_1)} - \mu^{(\kappa_2)}}{\sigma} \right) P_1^{(\kappa_1)} P_2^{(\kappa_2)} \right) \tag{14}
$$

13

where σ is prespecified. Asymptotically we allow K_1 and K_2 to grow with the sample size which is why we interpret this model as a sieve model. Showing consistency of this estimator is a straightforward application of sieve methodology, but involves introducing much new notation. Since this is only one of numerous estimation options and to avoid introducing this notation in the text, we leave the details of the estimation to the Appendix Section A.7 where we show this provides a consistent estimator of the two distribution functions. With consistent estimates of distributions of ε and the η_{jt}^* in hand, we can simulate the distribution of $(\widehat{\alpha} - \alpha)$ for any hypothesized value of α.

3 Empirical Example: The Effect of Merit-Aid Programs on Schooling Decisions

3.1 Merit-Aid Programs

In the last fifteen years a number of states have adopted merit-based aid programs. These programs are run at the state level and 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.[7] 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.[8] 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 1989-2000. However, our analysis differs from hers in several ways. Perhaps most importantly, we use all states as controls while she just uses those from the South. Of course her paper is a more complete empirical analysis while our primary goal is to demonstrate the use of our method.

During the 1989-2000 time period, ten different states initiated merit-aid programs. We

---

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

[8] Examples include Dynarski (2000, 2004), Cornwell, Mustard, and Sridhar (2003), Cornwell, Lee, and Mustard, (2003), Cornwell, Leidner, and Mustard (2003), Bugler, Henry, and Rubenstein (1999), Berker(2001), Bugler and Henry (1997,1998), Henry and Rubenstein (2002).

use two specifications with the first focusing on the HOPE program alone. In this case, we ignore data from the other nine "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.[9] The dependent variable in our model is a dummy variable representing whether the individual is currently enrolled in college. Given that we obtain multiple observations of individuals in the same state at the same time, Model 2 is appropriate. However, since our dependent variable is binary we modify our approach somewhat to deal with binary dependent variables. We discuss this approach in section 3.2. For estimation we assume that the number of individuals in a state×year is large and present these results in section 3.3. In section 3.4 we treat group size as fixed. We control for race and gender throughout.

3.2 Limited Dependent Variable Models

Since our college attendance dependent variable is discrete, the analysis above can not be applied directly. In this Subsection, we discuss an extension of Model 2 to handle limited dependent variables.

We redefine the model letting the regression equation define a latent variable Y_i^* and where the researcher observes only an indicator of its sign: Y_i.

$$Y_i^* = \alpha d_{j(i)t(i)} + X_i'\beta + \theta_{j(i)} + \gamma_{t(i)} + \eta_{j(i)t(i)} + \varepsilon_i \tag{15}$$

$$Y_i = 1(Y_i^* > 0). \tag{16}$$

For computational simplicity, we assume that the distribution of ε_i is known with logistic distribution Λ. We first discuss the natural extension to the case in which $|M(j,t)| \to \infty$. We then turn to the discussion of the more difficult case where $|M(j,t)|$ is finite.

Consider the case in which $|M(j,t)| \to \infty$. As in section 2.2, define

$$\eta_{jt}^* = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt},$$

so that it incorporates all of the group×time variation. Then we can write

$$\Pr(Y_i = 1 \mid X_i, \; j(i), t(i)) = \Lambda(X_i'\beta + \eta_{jt}^*).$$

---

[9]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 zero. Our methods could be extended to incorporate heterogeneous effects in which case one could look at complicated joint tests of the effects of the programs.

Since $|M(j,t)| \to \infty$, this is a standard discrete choice model and we can obtain consistent estimates of β and η^*_{jt} for each j and t by maximum likelihood where η^*_{jt} can be estimated as the coefficient on group×time dummy variables (a strategy analogous to that in Amemiya, 1978). Alternatively, we could relax the assumption that ε_i is logistic and use a semiparametric estimator. Having obtained consistent estimates of η^*_{jt} we are essentially in the conditions of Model 1 and can apply the methodology in that Section using η^*_{jt} as the dependent variable.

When $|M(j,t)|$ is assumed fixed, we can no longer obtain consistent estimates of η^*_{jt} in this model and thus can not use the Model 1 methodology. To complicate things further, the fixed effects θ_j cannot be differenced out in this nonlinear model. Typical solutions to the presence of fixed effects like Chamberlain's (1980) conditional logit model or the fixed effects maximum score estimator (Manski, 1987) could be used to estimate β, but this is not enough to perform hypothesis tests on α which essentially require estimation of the joint distribution of $\left(\eta_{j1}, ..., \eta_{jT} \right)$.

Thus, in order to obtain estimates of the distribution of the error term we use somewhat stronger assumptions. We have defined η^*_{jt} so that

$$Y^*_i = X'_i\beta + \eta^*_{j(i)t(i)} + \varepsilon_i, \tag{17}$$

and we assume that for the control groups, $\eta^*_{j(i)t(i)}$ is independent of X_i.[10] As long as the support of $X'_i\beta$ is sufficiently large we can identify the joint distribution of $(\eta^*_{j1}+\varepsilon, ..., \eta^*_{jT}+\varepsilon)$ up to scale. Given that this joint distribution is identified for various values of $|M(j,t)|$, one can use an argument analogous to that in the proof of proposition 2.5 to show how to identify the marginal distribution of ε_i and the joint distribution of $(\eta^*_{j1}, ..., \eta^*_{jT})$.[11]

Given knowledge of X_i, and the distribution of η^*_j and ε_i, for any α, we can simulate the conditional distribution of Y_i given X_i and $d_{j(i)t(i)}$. This allows us to identify the distribution of any test statistic that is a function of observed variables, up to the parameter α. Thus, we can obtain interval estimates by 'inverting' a test statistic. First, we must choose a test statistic that depends on $(Y_i, X_i, d_{j(i)t(i)})$. Since we have estimated a model that gives us the

---

[10]Note that αd_{jt} is part of η^*_{jt} so that it seems as if we are assuming that d_{jt} is independent of X_i. In our example this is not the case because $d_{jt} = 0$ for all of the control states (in all time periods). In other cases, one may want to modify this assumption to allow for dependence.

[11]Cameron and Taber (1998) discuss identification of panel data logit models with unobserved heterogeneity. This model is more complicated in that η is a vector, but this does not substantially complicate the analysis.

distribution of Y_i conditional on X_i, d_{jt}, and α; we can simulate the distribution of the test statistic under any null hypothesis $\alpha = \alpha_0$.

The question then becomes which test statistic we should use. A natural choice would be the difference-in-difference parameter from a linear probability model. That is we can estimate the linear regression model

$$Y_i = a d_{j(i)t(i)} + X_i'b + G_{j(i)}'c + P_{t(i)}'f + e_i \tag{18}$$

which has group effects and time effects. Here we use different notation than in the models above because the "true" structural model is (17) while (18) represents a "reduced form" regression equation for which the parameters are defined by the linear projection. We can then use the estimated value of a (call it $\widehat{a}$) as the test statistic itself. Given our estimated model and a null hypothesis on α, we can simulate the distribution of $\widehat{a}$. While the estimator is not a standard fixed effect estimator, it still embodies the central idea behind difference in differences; we would reject the null hypothesis that $\alpha = 0$ when the difference between the pretreatment and posttreatment outcomes is substantially different than what one might predict based on variation from the control sample.

A number of different options exist for estimation of (18). For our application the most convenient was to first run the regression model using only the control states to produce consistent estimates of b and f (call these estimates $\widehat{b}$ and $\widehat{f}$). We then estimate α by running a (state) fixed effect regression of $\left(Y_i - X_i'\widehat{b} - P_{t(i)}'\widehat{f} \right)$ on $d_{j(i)t(i)}$. The advantage of this approach is that when we simulate the distribution of the test statistic we only need to simulate the error distribution for the treatments which is all that we need in the second stage of this procedure.

3.3 Confidence Interval Estimation under Standard Approach and Large Group Sizes

We compare three estimation approaches in this subsection: linear probability estimators with both population weighting across groups and equal weighting across groups, and a logit estimator. For each estimator, we compare interval estimates for the treatment parameter using our methods to those obtained under the typical approaches allowing clustering by group and group-by-time.

To obtain population-weighted estimates, we estimate equation (5) via OLS using all 34,902 observations. These results are presented in the first column of Table 1. The de-

pendent variable is a dummy variable for college enrollment and the sample only includes individuals aged 18 and 19. The point estimates suggests that the HOPE scholarship increased schooling enrollment of students who live in Georgia by about seven percentage points. Interval estimates of the HOPE effect are presented in the second panel of the table. The first clusters by state and year, allowing the error terms of individuals within the same state and year to be arbitrarily correlated with each other. One can see that the coefficient is highly significant. We next cluster by state which allows for serial correlation in η_{jt}. Bertrand et. al (2004) discuss a case in which accounting for serial correlation can lead to standard errors to increase, but in our case we find the opposite. The standard errors fall substantially when one clusters by state. Clearly one should be worried about the asymptotic assumptions underlying these routine confidence interval estimates. The key assumption justifying them is that the number of states that change status is large, but only one state (Georgia) contributes to the estimate of the treatment effect.

The estimated confidence intervals using our method are presented in the last row of Column 1. These confidence intervals are formed by inverting the test statistic $(\hat{\alpha} - \alpha_0)$ using our large-sample approximation for its distribution. (For details see Appendix Section A.4). These confidence intervals are substantially different from those obtained with typical methods. The confidence interval increases by a factor of about 3 and the coefficient is not significant. To see why, in Figure 1 we display the estimated distribution of $(\hat{\alpha} - \alpha)$ under the null hypothesis that the true value of α is zero (after using a kernel smoother). This distribution is estimated from the other 41 states. It appears very different from normal so it is not surprising that the asymptotic approximation is very different.

In the second column we present linear probability estimates resulting from a commonly used two-step approach (Amemiya 1978). In the first stage we regress schooling on the individual X's and on the full interacted state×year dummies. In a second stage we regress the predicted state×year dummies on the HOPE indicator controlling for state dummies and year dummies (separately). These results are presented in the second column and are remarkably close to the first. The difference between these estimates and those in the first column is that the states are equally weighted while in the first column they are population-weighted.

Finally we present a logit version of the model. The estimates in the third column were obtained in exactly the same manner as in the second column, except that in the first stage

we run a logit model of the school dummy on our $X's$ and state×year dummy variables. In the second stage we once again regress the state×year dummies on the hope indicator controlling for state dummies and year dummies (separately). Thus the predicted parameter has the interpretation of a logit index. The pattern is very similar. In all three cases the HOPE variable becomes marginally insignificant when we use our approach even though the variable is highly significant using standard 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 three are 95% confidence intervals for the change in attendance probability for our reference individual.[12]

In Table 2 we present results estimating the effect of merit aid using all ten states who added programs during this time period. The format of the table is identical to Table 1. There are a few notable features of the 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. Second, in contrast to Table 1, the confidence intervals are quite similar when we cluster by state compared to clustering by state×year. Most importantly our approach changes the confidence intervals substantially, but less dramatically than in Table 1.

3.4 Confidence Interval Estimation assuming Small Group Sizes

We next turn to the case in which $|M(j,t)|$ is fixed. Given that we have 34,902 observations one may wonder why we are worried about the number of individuals in the sample not being substantially high. The problem is for the asymptotic approximation in Model 2 to work well we need that the asymptotic approximation works well in *all* states×time periods not just on average. The largest is California in 1991 with 383 people while the smallest is New Hampshire in 1992 with 15 people. One very well might expect that individual components contribute a substantial amount to the variance of the state component for the smaller states. This would lead the variance of the effect to be substantially larger for the smaller states

---

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

than the larger ones invalidating the previous exercise.

The deconvolution we discuss in Section 2 required that ε be independent of η. This is not possible in a linear probability model since the dependent variable must be one or zero. We instead use logit model (15)-(16).

We perform inference in this model in three stages. First we obtain consistent estimates of β using Chamberlain's (1986) fixed effect logit model using state×year fixed effects. Second we estimate the joint distribution of $\widetilde{\eta}_j$ up to a location normalization. Finally, after choosing a test statistic, we simulate the distribution of the test statistic from the estimated model.

The first stage is straightforward, so we now describe the second. We use a Heckman and Singer (1994) style nonparametric maximum likelihood method analogous to that in (14). The Log-likelihood takes the form

$$\sum_{j=N_0+1}^{N_0+N_1} \log \left(\sum_{\ell=1}^{L} \prod_{t=1}^{T} \prod_{i \in I(j,t)} \Lambda(X_i'\widehat{\beta} + \eta_t^\ell)^{Y_i} \left(1 - \Lambda(X_i'\widehat{\beta} + \eta_t^\ell) \right)^{1-Y_i} \mu_\ell \right).$$

We maximize this likelihood in terms of the η_t^ℓ and μ_ℓ parameters. In practice we use $L=13$ and we have 12 years of data.[13] That yields 168 parameters.[14] Naturally, local optima are a problem in these cases so we randomly selected many different starting values to search for a global optima.[15] Given the number of parameters and their limited interpretation we do not report these numbers.

The next goal is to obtain a confidence interval for α. We argue in section 3.2 that a natural choice for a test statistic is the coefficient in the difference in difference model. Following the discussion there, we can write the test statistic as

$$\tau = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N_0} \sum_{k \in M(j,t)} \widetilde{d}_{j(k)t(k)} \left(Y_i - X_i'\widehat{b} - P_{t(i)}'\widehat{f} \right)}{\sum_{t=1}^{T} \sum_{j=1}^{N_0} \sum_{k \in M(j,t)} \widetilde{d}_{j(k)t(k)}^2}.$$

We first estimate τ using the actual data.

Once we have estimated the data generation model, we can use it simulate the distribution of τ under the null hypothesis $\alpha = \alpha_0$. Note that τ will vary in these simulations both because of heterogeneity in η and because $|M(j,t)|$ is finite. We reject the null hypothesis if τ is less than the 0.025 quantile or greater than the 0.975 quantile of this simulated distribution. The confidence intervals is the set of parameters for which the null hypothesis is not rejected.

---

[13] We experimented with alternative values, and the results are not sensitive to the choice.

[14] That is 13×12 η_t^ℓ parameters, and 12 μ_ℓ parameters (since probabilities must add to one).

[15] Many in this case was 5000. We found that this procedure ran surprisingly fast taking only about two days to complete all 5000 optimizations on a linux machine.

In Table 3 we present confidence intervals constructed using this approach. The results are similar, but not identical to those in Tables 1 and 2. The confidence interval for the HOPE program is slightly bigger than those in the third column of Table 1. The interval for all merit programs is similar in size but skewed slightly to the left of that in Table 2. For this treatment effect, a one sided test probably is perhaps most interesting. At the 5% level a one-sided test rejects the null hypothesis of no effect.

One may worry that the model we have estimated is too stylized or too flexible to approximate the data well. To examine this, we tried the following experiment somewhat like the placebo law used in Bertrand, Duflo, and Mullainathan. We use all 41 of our control states and construct the test statistic that we used for Georgia for testing the null hypothesis that $\alpha = 0$. That is, for each of the 41 control states in turn, we act as if the HOPE program were operating in the state after 1993 and used the remaining 40 states as controls. For each alternate pretend treatment state we calculate the p-value for the test that $\alpha = 0$ using our method. Since this null hypothesis is true by construction, these p-values should have a uniform [0,1] distribution. We plot the distribution of p-values in Figure 2. We present a histogram of the values and along the horizontal axis plot the actual p-values. The fit of the model looks surprisingly strong in the sense that the p-values are spread throughout the distribution. This logit approach with this test statistic is not the only way to obtain confidence intervals for α, and is almost certainly not the most efficient, but it appears to work well.

4   Conclusions

The main goal of this paper is to construct a method to perform inference for difference-in-differences models when the number of policy changes observed in the data is small. We argue that point estimates of treatment effects should not be thought of as being consistent and that the standard methods that researchers use to perform inference in these models are not appropriate. The main contribution of our work is to show how to perform inference under the assumption that there are a finite number of policy changes in the data, using asymptotic approximations as the number of control groups gets large. In this case, we cannot obtain a consistent point estimator for the key parameter but are able to consistently estimate its distribution, up to the unknown parameter itself. This allows us to perform inference on the key parameter and construct confidence intervals.

We develop this methodology in a number of different cases. Model 1 considers a regression model in which one observes group×time level data. Model 2 extend the idea to cases in which we observe individual level data. Within Model 2 we focus on the case in which the number of observations in a group/time cell is fixed.

We demonstrate the methodology by applying it the study of the effects of merit-aid programs on schooling. We think this application is a good example of a situation with a few treatment groups changing policy and many controls with unchanged policies. To accommodate our particular example, we extend the methodology to a logit model. Our empirical results suggest that conventional methods understate the magnitude of the standard errors considerably. However, we still find evidence of a positive effect of merit aid programs.

We think our combination of large and small sample inference will be appropriate in many other situations as well. For example, in applications studying the effect of a law change in a small number of states using other states as controls. While we have focused on difference in differences estimators, our approach is more general and is straightforward to extend to any type of regression model in which there are a large number of control observations, but only a small number of treatments.

5 References

Amemiya, Takeshi "A Note on a Random Coefficients Model International Economic Review Vol. 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 W.K. "End-of-SampleTests" *Econometrica* 71(6) 1661-1694, 2003.

Angrist, Joshua, and Alan Krueger, "Empirical Strategies in Labor Economics, " Handbook of Labor Economics, 1999, Elsevier, New York, Ashenfelder and Card eds.

Athey, Susan, and Guido Imbens, "Identification and Inference in Nonlinear Difference in Difference Models," unpublished manuscript, 2002.

Berker, Ali Murat, "The Impact of Merit-Based Aid on College Enrollment: Evidence from HOPE-Like Scholarship Programs," unpublished manuscript, Michigan State University, 2001.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Truce Differences-in-Differences Estimates?," *Quarterly Journal of Economics* 19:1, 2004. 249-275.

Bugler, Daniel and Gary Henry, "Evaluating the Georgia HOPE Scholarship Program: Impact on Students Attending Public Colleges and Universities," unpublished manuscript, Coucil for School Performance, Georgia State University, 1997.

Bugler, Daniel and Gary Henry, "An Evaluation of Georgia's HOPE Scholarship Program: Impact of College Attendance and Performace," unpublished manuscript, Council for School Performance, Georgia State University, 1998.

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," Councis for School Performace, Georgia State Univerisity, Atlanta, GA, 1999.

Cameron, Stephen and Christopher Taber, "Evaluation and Identification of Semiparametric Maximum Likelihood Models of Dynamic Discrete Choice," unpublished manuscript, Northwestern University, 1998.

Card, David, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review,* January, 1990.

Card, David, and A. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review,* September, 1994.

Chamberlain, Gary, "Analysis of Covariance with Qualitative Data," Review of Economic Studies, 47 (1), 1980, pp 225-238.

Cornwell, Christopher, Kyung Hee Lee, and David Mustard, "The Effects of Merit-Based Financial Aid on Academic Choices in College,", unpublished manuscript, University of Georgia, 2003.

Cornwell, Christopher, Mark Leidner, and David Mustard, "Rules, Incentives and Policy Implications of Large-Scale Merit-Aid Programs," unpublished manuscript, University of Georgia, 2003.

Cornwell, Christopher, David Mustard, and Deepa Sridhar, "The Enrollment Effects of Merit-Based Financial Aid: Evidence from Georgia's HOPE Scholarship, unpublished manuscript, University of Georgia, 2003.

Donald, Stephen, and Kevin Lang, "Inference with Difference in Differences and Other Panel Data," unpublished manuscript, Boston University, 2001.

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

Dynarski, Susan, "Hope for Whom? Financial Aid for the Middle Class and its Impact on College Attendance," National Tax Journal, Vol 53 no. 3 Part 2, 2000, pp 629-662.

Dynarski, Susan, "The New Merit Aid," in *College Choices: The Economics of Which College, When College, and How to pay for it.* Caroline Hoxby, ed. University of Chicago Press, 2004.

Gallant, A.R. and D.W. Nychka, "Seminonparametric Maximum Likelihood Estimation," Econometrica, 55, 1987, 363-390.

Gruber, Jonathon, Pillip Levine, and Douglas Staiger, "Abortion Legalization and Child Living Circumstances: "Who is the Marginal Child?", *The Quarterly Journal of Economics*, February, 1999.

Hansen, Christian, "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects" MIT Working Paper 2004.

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, pp 93-109.

Kotlarski, Ignacy, "On Characterizing the Gamma and Normal Distributions, Pacific Journal of Mathematics, vol 20, 69-76, 1967

Manski, Charles, "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," Econometrica, 1987, 357-362.

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

Moulton, Brent R. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units" Review of Economics and Statistics, 1990, 334-338.

Newey, Whitney and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing." *Handbook of Econometrics, Vol. 4*, Engle and McFadden eds. Elsevier, 1994,2113-2245.

Matzkin, Rosa, "Restrictions of Economic Theory in Nonparametric Methods" *Handbook of Econometrics, Vol. 4*, Engle and McFadden eds. Elsevier, 1994, 2524-2558.

Prakasa Rao, B.L.S., *Identifiability in Stochastic Models*, Academic Press, 1992.

Wooldridge, Jeffrey, "Cluster-Sample Methods in Applied Econometrics," unpublished manuscript, Michigan State University.

Technical Appendix

A.1 Proof of Proposition 1.1

First a standard application of the partitioned inverse theorem makes it straight forward to show that

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

$$
\times \left(\frac{\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} \widetilde{X}_{jt}\widetilde{\eta}_{jt}}{N_1+N_0} - \frac{\left[\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} \widetilde{d}_{jt}\widetilde{X}_{jt}\right]\left[\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} \widetilde{d}_{jt}\eta_{jt}\right]}{(N_1+N_0)\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} \widetilde{d}_{jt}^{\,2}} \right).
$$

Now consider each piece in turn.

First Assumption 1.2 states that

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

The i.i.d. sampling and conditional independence components of Assumption 1.1 imply that:

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

For control groups $j > N_0$, the treatment does not vary over time so $d_{jt} = \overline{d}_j$. Therefore,

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

where

$$
\sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \left(\overline{d} - \overline{d}_t \right)^2 = N_1 \sum_{t=1}^{T} \left(\overline{d} - \overline{d}_t \right)^2
$$

$$
= N_1 \sum_{t=1}^{T} \left(\frac{1}{N_0+N_1} \sum_{\ell=1}^{N_0+N_1} \left[\left(\frac{1}{T} \sum_{\tau=1}^{T} d_{\ell\tau} \right) - d_{\ell t} \right] \right)^2
$$

$$
= \frac{N_1}{(N_0+N_1)^2} \sum_{t=1}^{T} \left(\sum_{\ell=1}^{N_0} \left[\left(\frac{1}{T} \sum_{\tau=1}^{T} d_{\ell\tau} \right) - d_{\ell t} \right] \right)^2
$$

$$
\xrightarrow{p} 0.
$$

Now consider the other term

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

25

since $\overline{d}_t$ and $\overline{d}$ both have the same limit due to the finite number of groups with intertemporal variation in treatments. Thus

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

since $N_0 \geq 1$.

Now consider

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

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

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

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

$$\xrightarrow{p} 0 \text{ as } N_1 \to \infty.$$

This result follows because the first term involves a sum of a finite number of $O_p(1)$ random variables normalized by an $O(N_1^{-1/2})$ term and the second term is identically zero due to differencing:

$$\sum_{j=1}^{N_1+N_0} \widetilde{X}_{jt} = \sum_{j=1}^{N_0+N_1} \left(X_{jt} - \overline{X}_j - \overline{X}_t + \overline{X}\right)$$

$$= (N_0 + N_1) \left(\overline{X}_t - \overline{X} - \overline{X}_t + \overline{X}\right)$$

$$= 0.$$

Likewise

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

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

which is $O_p(1)$, thus

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

Consistency for $\widehat{\beta}$ follows upon plugging the pieces back into (A-1) and applying Slutsky's theorem.

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

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

Now from above we know that

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

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

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

Thus

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

We showed above that

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

The variables $\overline{\eta}_t$ and $\overline{\eta}$ both converge to zero in probability as $N_1 \to \infty$, therefore

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

This gives the result. ∎

A.2   Proof of Proposition 1.2

Since Γ is defined conditional on d_{jt} for $j = 1, ... N_0, t = 1, ..., T$, 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 $j = 1, ...N_0$, $t = 1, ..., T$.

For each $j = 1, ..., N_0$ define the random variable

$$W_j \equiv \frac{\sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)\left(\eta_{jt} - \overline{\eta}_j\right)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} \left(d_{\ell t} - \overline{d}_\ell\right)^2}$$

and let F_j be the distribution of W_j for $j = 1, .., N_0$.

Then note that

$$\Gamma(a) = \Pr\left(\frac{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)\left(\eta_{jt} - \overline{\eta}_j\right)}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)^2} < a\right)$$

$$= \int \cdots \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1)...dF_{N_0}(W_{N_0}).$$

We can also write

$$\widehat{\Gamma}(a) = \int \cdots \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) d\widehat{F}_1(W_1; \widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0}; \widehat{\beta}),$$

where $\widehat{F}_j(\cdot; \widehat{\beta})$ is the empirical c.d.f. one gets from the residuals using the control states only. That is more generally

$$\widehat{F}_j(w; b) \equiv \frac{1}{N_1} \sum_{m=1}^{N_1} 1\left(\frac{\sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)\left(\widetilde{Y}_{mt} - \widetilde{X}'_{mt}b\right)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} \left(d_{\ell t} - \overline{d}_\ell\right)^2} < w\right).$$

To avoid repeating the expression we define

$$\phi_j(w, b) \equiv \Pr\left(\frac{\sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right)\left(\widetilde{Y}_{mt} - \widetilde{X}'_{mt}b\right)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} \left(d_{\ell t} - \overline{d}_\ell\right)^2} < w\right).$$

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

First, for each $j = 1, ..., N_0$ consider the difference between $\widehat{F}_j(w; \widehat{\beta})$ and $\phi_j(w, \beta)$

$$\sup_w |\widehat{F}_j(w;\widehat{\beta}) - \phi_j(w,\beta)| \tag{A-2}$$

$$\leq \sup_w \left| \frac{1}{N_1} \sum_{m=N_0+1}^{N_1} 1\left(\frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)(\widetilde{Y}_{mt} - \widetilde{X}'_{mt}\widehat{\beta})}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} < w \right) - \phi_j\left(w,\widehat{\beta}\right) \right|$$

$$+ \sup_w \left| \phi_j\left(w,\widehat{\beta}\right) - \phi_j(w,\beta) \right|$$

$$\leq \sup_{b,w} \left| \frac{1}{N_1} \sum_{m=N_0+1}^{N_1} 1\left(\frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)(\widetilde{Y}_{mt} - \widetilde{X}'_{mt}b)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} < w \right) - \phi_j(w,b) \right|$$

$$+ \sup_w \left| \phi_j\left(w,\widehat{\beta}\right) - \phi_j(w,\beta) \right|.$$

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

$$\sup_w \left| \phi_j\left(w,\widehat{\beta}\right) - \phi_j(w,\beta) \right| = \sup_w \left| \frac{\partial \phi_j\left(w,\widetilde{\beta}\right)}{\partial \widetilde{\beta}} \left(\widehat{\beta} - \beta\right) \right|.$$

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

$$\phi_j(w,b) = \Pr\left(\frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)\left(\widetilde{\eta}_{jt} + \widetilde{X}'_{jt}(\beta - b)\right)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} < w \right)$$

$$= \Pr\left(W_j < w - \frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)\widetilde{X}'_{jt}(\beta - b)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} \right)$$

So

$$\frac{\partial \phi_j(w,b)}{\partial b} = E\left(f_j\left(w - \frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)\widetilde{X}'_{jt}(\beta - b)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} \right) \frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)\widetilde{X}'_{jt}}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} \right)$$

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. Thus $\sup_w \left| \phi_j\left(w,\widehat{\beta}\right) - \phi_j(w,\beta) \right|$ converges to zero since $\widehat{\beta}$ is consistent.

Next consider the first term on the right side of (A-2). Note that the function

$$1\left(\frac{\sum_{t=1}^{T} (d_{jt} - \overline{d}_j)(\widetilde{Y}_{mt} - \widetilde{X}'_{mt}b)}{\sum_{\ell=1}^{N_0} \sum_{t=1}^{T} (d_{\ell t} - \overline{d}_\ell)^2} < w \right)$$

is continuous at each b,w with probability one and its absolute value is bounded by 1, so applying Lemma 2.4 of Newey and McFadden, 1994, $\widehat{F}_j(w;b)$ converges uniformly to $\phi(w,b)$. Thus putting the two pieces of (A-2) together,

$$\sup_w |\widehat{F}(w;\widehat{\beta}) - \phi(w,\beta)| \overset{p}{\to} 0.$$

29

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

$$\left|\widehat{\Gamma}(a) - \Gamma(a)\right|$$

$$= \left| \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) d\widehat{F}_1(W_1;\widehat{\beta}) d\widehat{F}_2(W_2;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right.$$

$$\left. - \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1) dF_2(W_2)...dF_{N_0}(W_{N_0}) \right|$$

$$= \left| \left\{ \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) d\widehat{F}_1(W_1;\widehat{\beta}) d\widehat{F}_2(W_2;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right.\right.$$

$$\left. - \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1) d\widehat{F}_2(W_2;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right\}$$

$$\left\{ + \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1) d\widehat{F}_2(W_2;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right.$$

$$\left. - \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1) dF_2(W_2) d\widehat{F}_3(W_3;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right\}$$

$$+...$$

$$\left\{ + \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1)...dF_{N_0-1}(W_{N_0-1}) d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right.$$

$$\left.\left. - \int 1\left(\sum_{j=1}^{N_0} W_j < a\right) dF_1(W_1)...dF_{N_0-1}(W_{N_0-1}) dF_{N_0}(W_{N_0}) \right\} \right|$$

$$= \left| \left\{ \int \left[\widehat{F}_1\left(\left[a - \sum_{j=2}^{N_0} W_j\right];\widehat{\beta}\right) - F_1\left(a - \sum_{j=2}^{N_0} W_j\right) \right] d\widehat{F}_2(W_2;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right\} + \right.$$

$$\left\{ \int \left[\widehat{F}_2\left(\left[a - \sum_{\substack{j=1 \\ j\neq 2}}^{N_0} W_j\right];\widehat{\beta}\right) - F_2\left(a - \sum_{\substack{j=1 \\ j\neq 2}}^{N_0} W_j\right) \right] dF_1(W_1) d\widehat{F}_3(W_3;\widehat{\beta})...d\widehat{F}_{N_0}(W_{N_0};\widehat{\beta}) \right\}$$

$$+...$$

$$\left. + \left\{ \int \left[\widehat{F}_{N_0}\left(\left[a - \sum_{j=1}^{N_0-1} W_j\right];\widehat{\beta}\right) - F_{N_0}\left(a - \sum_{j=1}^{N_0-1} W_j\right) \right] dF_1(W_1)...dF_{N_0-1}(W_{N_0}) \right\} \right|$$

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

A.3 Projection Lemma

We use the following lemma in sections A.4 and A.5:

Lemma A.1 *Consider a regression of $d_{j(i)t(i)}$ on group dummies (G_i) and time dummy variables (P_i) as defined in equations (7)-(10). Let $\widehat{a}_t$ be coefficient on the time variable for time period $t = 1, .., T - 1$ and $\widehat{a}_T \equiv 0$. Under Assumption 2.3, and either Assumption 2.5 or Assumption A.1,*

$$\widetilde{d}_{j(i)t(i)} = d_{j(i)t(i)} - \overline{d}_{j(i)} - \left(\widehat{a}_{t(i)} - \frac{\sum_{\tau=1}^{T-1} |M(j(i), \tau| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j(i), \tau|} \right)$$

and $\widehat{a}_\tau = O_p(\frac{1}{N_1}), \tau = 1, ..., T - 1$.

Proof. To streamline the notation, let $\sum_i$ denote $\sum_{j=1}^{N_1+N_0} \sum_{t=1}^{T} \sum_{i \in M(j,t)}$ and let

$$m_0 \equiv \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|$$

$$m_1 \equiv \sum_{j=N_0+1}^{N_1+N_0} \sum_{t=1}^{T} |M(j,t)|$$

$$m \equiv m_0 + m_1$$

Note that m_0 is fixed but m_1 and m get large as $N_1 \to \infty$. We will use this notation in a number of proofs.

Now consider a regression of $d_{j(i)t(i)}$ on group dummies and time dummies. We will write this regression equation as

$$d_{j(i)t(i)} = P_i'\widehat{a} + G_i'\widehat{b} + \widetilde{d}_{j(i)t(i)}$$

where P_i and G_i are as defined equations (7)-(10).

The first part of our lemma is a standard regression result with dummy variables. Note that we can rewrite this regression equation as

$$d_{j(i)t(i)} - \widehat{a}_{t(i)} = G_i'\widehat{b} + \widetilde{d}_{j(i)t(i)}.$$

Since $\widetilde{d}_{j(i)t(i)}$ is orthogonal to G_i we could construct residuals by regressing $d_{j(i)t(i)} - \widehat{a}_{t(i)}$ on a full set of group dummies and taking residuals. However, it is well known that this will lead to taking deviations of the left hand side variable from group means so that

$$\begin{aligned} \widetilde{d}_{j(i)t(i)} &= \left(d_{j(i)t(i)} - \widehat{a}_{t(i)} \right) - \frac{\sum_{\tau=1}^{T} \sum_{\ell \in M(j(i),\tau)} \left(d_{j(\ell)\tau} - \widehat{a}_{t(\ell)} \right)}{\sum_{\tau=1}^{T} |M(j(i), \tau)|} \\ &= \left(d_{j(i)t(i)} - \overline{d}_{j(i)} \right) - \left(\widehat{a}_{t(i)} - \frac{\sum_{\tau=1}^{T-1} |M(j(i), \tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j(i), \tau)|} \right). \end{aligned}$$

Next consider the derivation of $\widehat{a}$. Using the partitioned inverse theorem,

$$\begin{aligned} \widehat{a} &= \frac{1}{m} \left(\frac{1}{m} \sum_i P_i P_i' - \frac{1}{m} \sum_i P_i G_i' \left(\sum_i G_i G_i' \right)^{-1} \sum_i G_i P_i' \right)^{-1} \times \\ &\qquad \left[\sum_i P_i d_{j(i)t(i)} - \sum_i P_i G_i' \left(\sum_i G_i G_i' \right)^{-1} \sum_i G_i d_{j(i)t(i)} \right]. \end{aligned}$$

Assumption 2.3 implies that we can rewrite this as

$$\hat{a} = \frac{1}{m} \left(\Omega + o_p(1) \right)^{-1} \left[\sum_i P_i d_{j(i)t(i)} - \sum_i P_i G_i' \left(\sum_i G_i G_i' \right)^{-1} \sum_i G_i d_{j(i)t(i)} \right].$$

Now consider the last term, $\sum P_i G_i' \left(\sum G_i G_i' \right)^{-1} \sum G_i d_{j(i)t(i)}$. It is straightforward to show that this is a $(T-1) \times 1$ vector with generic element t

$$\sum_{j=1}^{N_0+N_1} \frac{|M(j,t)| \sum_{\tau=1}^{T} |M(j,\tau)| \, d_{j\tau}}{\sum_{\tau=1}^{T} |M(j,\tau)|} = \sum_{j=1}^{N_0+N_1} |M(j,t)| \, \overline{d}_j.$$

Thus the $(T-1) \times 1$ vector $\left[\sum P_i d_{j(i)t(i)} - \sum P_i G_i' \left(\sum G_i G_i' \right)^{-1} \sum G_i d_{j(i)t(i)} \right]$ has generic t element

$$\sum_{j=1}^{N_0+N_1} |M(j,t)| \, d_{jt} - \sum_{j=1}^{N_0+N_1} |M(j,t)| \, \overline{d}_j \;=\; \sum_{j=1}^{N_0+N_1} |M(j,t)| \left(d_{jt} - \overline{d}_j \right)$$

$$= \sum_{j=1}^{N_0} |M(j,t)| \left(d_{jt} - \overline{d}_j \right).$$

Under Assumption 2.5 this is just a random variable which is $O_p(1)$ so since

$$\hat{a} = \frac{1}{m} \left(\Omega + o_p(1) \right)^{-1} \left[\sum_i P_i d_{j(i)t(i)} - \sum_i P_i G_i' \left(\sum_i G_i G_i' \right)^{-1} \sum_i G_i d_{j(i)t(i)} \right],$$

$\hat{a}$ is $O_p(\frac{1}{N_1})$.

Under Assumption A.1 we can write

$$\hat{a} = \frac{1}{N_0 + N_1} \left(\Omega + o_p(1) \right)^{-1} \times$$

$$\left[\frac{N_0 + N_1}{m} \sum_i P_i d_{j(i)t(i)} - \frac{N_0 + N_1}{m} \sum_i P_i G_i' \left(\sum_i G_i G_i' \right)^{-1} \sum_i G_i d_{j(i)t(i)} \right].$$

As above the last term in brackets is a $(T-1) \times 1$ vector with a generic element t that can be written as

$$\frac{N_0 + N_1}{m} \sum_{j=1}^{N_0} |M(j,t)| \left(d_{jt} - \overline{d}_j \right) = \frac{\sum_{j=1}^{N_0} |M(j,t)| \left(d_{jt} - \overline{d}_j \right)}{\frac{1}{N_0+N_1} \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} |M(j,t)|}$$

$$= \frac{\sum_{j=1}^{N_0} \frac{|M(j,t)|}{\sum_{\ell=1}^{N_0} \sum_{\tau=1}^{T} |M(\ell,\tau)|} \left(d_{jt} - \overline{d}_j \right)}{\frac{1}{N_0+N_1} \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \left(\frac{|M(j,t)|}{\sum_{\ell=1}^{N_0} \sum_{\tau=1}^{T} |M(\ell,\tau)|} \right)}$$

$$\xrightarrow{p} \frac{\sum_{j=1}^{N_0} \phi_{jt} \left(d_{jt} - \overline{d}_j \right)}{\sum_{t=1}^{T} \phi_t}$$

which is $O_p(1)$. ∎

A.4 Consistency Result for Large |M(j,t)|

In this Appendix we present a consistency result analogous to Proposition 2.1, but for the case in which $|M(j,t)|$ can grow with the sample size. We assume that group sizes grow at the same rate so that no group dominates in the limit. Formally we state this as

Assumption A.1 *For each* $j = 1, ..., N_0 + N_1$, $|M(j,t)|$ *grows at the same rate as* N_1. *For all j and t, defining*

$$\phi_{jt} \equiv \lim_{N_1 \to \infty} \frac{|M(j,t)|}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|},$$

we assume that where $\phi_{jt} > 0$ and bounded from above. For all t, defining

$$\phi_t \equiv \lim_{N_1 \to \infty} \frac{1}{N_0 + N_1} \sum_{j=1}^{N_0+N_1} \frac{|M(j,t)|}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|},$$

we assume that $0 < \phi_t < \infty$.

For this case, Proposition A.2 states that $\widehat{\beta}$ is consistent and derives the asymptotic distribution of $\widehat{\alpha}$.

Proposition A.2 *Under Assumptions 2.1-2.4, and A.1*

$$\widehat{\beta} \xrightarrow{p} \beta$$

$$\widehat{\alpha} \xrightarrow{p} \alpha + \frac{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \phi_{jt} \left(d_{jt} - \overline{d}_j \right) \left(\eta_{jt} - \overline{\eta}_j \right)}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \phi_{jt} \left(d_{jt} - \overline{d}_j^2 \right)}$$

as $N_1 \to \infty$.

Proof:

In this proof we make use of the notation defined in the proof of the Lemma A.1 in Section A.3.

First a standard application of the partitioned inverse theorem makes it straightforward to show that

$$
\widehat{\beta} = \beta + \left(\frac{1}{m} \sum_i \widetilde{X}_i \widetilde{X}_i' - \frac{m_0}{m} \frac{\left[\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}_i \right] \left[\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}_i' \right]}{\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)}^2} \right)^{-1}
$$
$$
\times \left(\frac{1}{m} \sum_i \widetilde{X}_i \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i \right) - \frac{m_0}{m} \frac{\left[\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}_i \right] \left[\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)} \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i \right) \right]}{\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)}^2} \right).
$$

Now consider each piece in turn.

Assumption 2.4 states that

$$\frac{1}{m} \sum_i \widetilde{X}_i \widetilde{X}_i' \xrightarrow{p} \Sigma_x.$$

Using Assumptions 2.1-2.2 and invoking the law of large numbers,

$$\frac{1}{m} \sum_i \widetilde{X}_i \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i \right) \xrightarrow{p} 0.$$

Define $\widehat{a}_t$ as in the statement of Lemma $A.1$ and then define

$$\widehat{\widetilde{a}}_{jt} \equiv \left(\widehat{a}_t - \frac{\sum_{\tau=1}^{T-1} |M(j,\tau| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau|} \right).$$

Lemma A.1 states that $\widetilde{d}_{j(i)t(i)} = d_{j(i)t(i)} - \overline{d}_{j(i)} - \widehat{\widetilde{a}}_{j(i)t(i)}$. Note also that for $j > N_0$, $d_{jt} - \overline{d}_j = 0$.
Thus

$$\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)}^2 = \frac{1}{m_0} \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \, \widetilde{d}_{jt}^2 + \frac{1}{m_0} \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} |M(j,t)| \, \widetilde{d}_{jt}^2$$

$$= \frac{1}{m_0} \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left[\left(d_{jt} - \overline{d}_j \right)^2 - 2\widehat{\widetilde{a}}_{jt} \left(d_{jt} - \overline{d}_j \right) + \widehat{\widetilde{a}}_{jt}^2 \right] + \frac{1}{m_0} \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \widehat{\widetilde{a}}_{jt}^2 \, |M(j,t)|$$

$$\xrightarrow{p} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \phi_{jt} \left(d_{jt} - \overline{d}_j \right)^2.$$

This result follows because

$$\frac{1}{m_0} \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \widehat{\widetilde{a}}_{jt}^2 \, |M(j,t)|$$

$$= \frac{1}{m_0} \sum_{j=N_0+1}^{N_0+N_1} \left[\sum_{t=1}^{T} \left(\widehat{a}_t - \frac{\sum_{\tau=1}^{T-1} |M(j,\tau| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau|} \right)^2 |M(j,t)| \right]$$

$$= \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T-1} \widehat{a}_t^2 \frac{|M(j,t)|}{m_o} - 2 \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T-1} \widehat{a}_t \frac{\sum_{\tau=1}^{T-1} |M(j,\tau| \widehat{a}_\tau}{m_0 \sum_{\tau=1}^{T} |M(j,\tau|}$$

$$+ \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \left(\frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|} \right)^2 \frac{|M(j,t)|}{m_0}$$

$$= \sum_{t=1}^{T-1} \widehat{a}_t^2 \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,t)|}{m_0} - 2 \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} \widehat{a}_t \widehat{a}_\tau \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)|}{m_0 \sum_{s=1}^{T} |M(j,s)|}$$

$$+ \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} \widehat{a}_\tau \widehat{a}_t \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)| \, |M(j,t)|}{m_0 \sum_{s=1}^{T} |M(j,s)|}$$

$$= \sum_{t=1}^{T-1} O_p \left(\frac{1}{N_1^2} \right) \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,t)|}{m_0} - \frac{2}{m_0} \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} O_p \left(\frac{1}{N_1^2} \right) \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)|}{\sum_{s=1}^{T} |M(j,s)|}$$

$$+ \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} O_p \left(\frac{1}{N_1^2} \right) \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)| \, |M(j,t)|}{m_0 \sum_{s=1}^{T} |M(j,s)|}$$

$$\xrightarrow{p} 0.$$

Next consider the object

$$
\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}_i = \frac{1}{m_0} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i\in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i + \frac{1}{m_0} \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i\in m(j,t)} \widehat{\widetilde{a}}_{jt} \widetilde{X}_i
$$

$$
= \frac{1}{m_0} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i\in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i - \frac{1}{m_0} \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T-1} \sum_{i\in m(j,t)} \left(\widehat{a}_t - \frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|}\right) \widetilde{X}_i
$$

$$
+ \frac{1}{m_0} \sum_{j=1}^{N_0+N_1} \sum_{i\in m(j,t)} \frac{\sum_{\tau=1}^{T-1} |M(j,T)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,T)|} \widetilde{X}_i
$$

$$
= \frac{1}{m_0} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i\in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i - \frac{1}{m_0} \sum_{t=1}^{T-1} \widehat{a}_t \sum_{j=1}^{N_0+N_1} \sum_{i\in m(j,t)} \widetilde{X}_i
$$

$$
+ \frac{1}{m_0} \sum_{j=1}^{N_0+N_1} \frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|} \sum_{t=1}^{T} \sum_{i\in m(j,t)} \widetilde{X}_i
$$

$$
= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right) \left[\frac{|M(j,t)|}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|}\right] \left[\frac{1}{|M(j,t)|} \sum_{i\in m(j,t)} \widetilde{X}_i\right]
$$

$$
\xrightarrow{p} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \phi_{jt} \left(d_{jt} - \overline{d}_j\right) E(\widetilde{X}_i \mid i \in M(j,t))
$$

$$
= O_p(1).
$$

We used the fact that $\widetilde{X}_i$ is the residual from a regression on time and state dummies so $\sum_{j=1}^{N_0+N_1} \sum_{i\in m(j,t)} \widetilde{X}_i = 0$ and $\sum_{t=1}^{T} \sum_{i\in m(j,t)} \widetilde{X}_i = 0$.

An analogous argument gives

$$
\frac{1}{m_0} \sum_i \widetilde{d}_{j(i)t(i)} \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right) = \frac{1}{m_0} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i\in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\widetilde{\eta}_{jt} + \widetilde{\varepsilon}_i\right)
$$

$$
= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \left(d_{jt} - \overline{d}_j\right) \left[\frac{|M(j,t)|}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)|}\right] \left[\frac{1}{|M(j,t)|} \sum_{i\in m(j,t)} \left(\widetilde{\eta}_{jt} + \widetilde{\varepsilon}_i\right)\right]
$$

$$
\xrightarrow{p} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \phi_{jt} \left(d_{jt} - \overline{d}_j\right) E\left(\widetilde{\eta}_{jt} + \widetilde{\varepsilon}_i \mid i \in M(j,t)\right)
$$

$$
= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \phi_{jt} \left(d_{jt} - \overline{d}_j\right) \left(\eta_{jt} - \overline{\eta}_j\right)
$$

$$
= O_p(1).
$$

The last term follows because for any $\tau = 1,..,T$ $E\left(\eta_{j\tau} - \overline{\eta}_j\right) = E(\varepsilon_i - \overline{\varepsilon}_j \mid t(i) = \tau) = 0$. So for a regression of either $\left(\eta_{j\tau} - \overline{\eta}_j\right)$ or $(\varepsilon_i - \overline{\varepsilon}_j)$ on time dummies, the coefficient on the dummy variables will converge to zero so $\widetilde{\eta}_{jt} \xrightarrow{p} \left(\eta_{j\tau} - \overline{\eta}_j\right)$ and $\widetilde{\varepsilon}_i \xrightarrow{p} (\varepsilon_i - \overline{\varepsilon}_j)$.

Putting all the objects into the expression for $\widehat{\beta}$, one can see that $\widehat{\beta}$ is consistent. Now consider $\widehat{\alpha}$. It is straight forward to show that

$$(\widehat{\alpha} - \alpha) = \frac{\frac{1}{m_0}\sum_i \widetilde{d}_{j(i)t(i)}\left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right)}{\frac{1}{m_0}\sum_i \widetilde{d}^2_{j(i)t(i)}} + \frac{\frac{1}{m_0}\sum_i \widetilde{d}_{j(i)t(i)}\widetilde{X}'_{j(i)t(i)}\left(\beta - \widehat{\beta}\right)}{\frac{1}{m_0}\sum_i \widetilde{d}^2_{j(i)t(i)}}$$

We have shown that

$$\frac{1}{m_0}\sum_i \widetilde{d}^2_{jt} \xrightarrow{p} \sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right)^2$$

$$\frac{1}{m_0}\sum_i \widetilde{d}_{j(i)t(i)}\widetilde{X}'_{j(i)t(i)} \xrightarrow{p} \sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right) E(\widetilde{X}_i \mid i \in M(j,t))$$

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

$$\frac{1}{m_0}\sum_i \widetilde{d}_{j(i)t(i)}\left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right) \xrightarrow{p} \sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right)\left(\eta_{jt} - \overline{\eta}_j\right).$$

Thus we are left with:

$$
\begin{aligned}
(\widehat{\alpha} - \alpha) &= \frac{\sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right)\left(\eta_{jt} - \overline{\eta}_j\right) + o_p(1)}{\sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right)^2 + o_p(1)} + o_p(1) \\
&\xrightarrow{p} \frac{\sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right)\left(\eta_{jt} - \overline{\eta}_j\right)}{\sum_{j=1}^{N_0}\sum_{t=1}^{T} \phi_{jt}\left(d_{jt} - \overline{d}_j\right)^2}.
\end{aligned}
$$

This gives the result.

∎

A.5 Proof of Proposition 2.1

We use the notation defined at the beginning of the proof of Lemma A.1 above. This proof is almost identical to that of Proposition A.2.

First a standard application of the partitioned inverse theorem makes it straightforward to show that

$$
\begin{aligned}
\widehat{\beta} = {}& \beta + \left(\frac{1}{m}\sum_i \widetilde{X}_i\widetilde{X}'_i - \frac{\frac{1}{\sqrt{m}}\left[\sum_i \widetilde{d}_{j(i)t(i)}\widetilde{X}_i\right]\frac{1}{\sqrt{m}}\left[\sum_i \widetilde{d}_{j(i)t(i)}\widetilde{X}'_i\right]}{\sum_i \widetilde{d}^2_{j(i)t(i)}}\right)^{-1} \\
&\times \left(\frac{1}{m}\sum_i \widetilde{X}_i\left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right) - \frac{1}{m}\frac{\left[\sum_i \widetilde{d}_{j(i)t(i)}\widetilde{X}_i\right]\left[\sum_i \widetilde{d}_{j(i)t(i)}\left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right)\right]}{\sum_i \widetilde{d}^2_{j(i)t(i)}}\right).
\end{aligned}
$$

Now consider each piece in turn.

Assumption 2.4 states that

$$\frac{1}{m}\sum_i \widetilde{X}_i\widetilde{X}'_i \xrightarrow{p} \Sigma_x.$$

Using Assumptions 2.1-2.2 and invoking the law of large numbers,

$$\frac{1}{m} \sum_i \widetilde{X}_i \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i \right) \xrightarrow{p} 0.$$

Define $\widehat{a}_t$ as in the statement of Lemma $A.1$ and then define

$$\widehat{\widetilde{a}}_{jt} \equiv \left(\widehat{a}_t - \frac{\sum_{\tau=1}^{T-1} |M(j,\tau| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau|} \right).$$

In Lemma A.1 it is shown that $\widetilde{d}_{j(i)t(i)} = d_{j(i)t(i)} - \overline{d}_{j(i)} - \widehat{\widetilde{a}}_{j(i)t(i)}$. Note also that for $j > N_0$, $d_{jt} - \overline{d}_j = 0$. Thus

$$\sum_i \widetilde{d}^2_{j(i)t(i)} = \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \, \widetilde{d}^2_{jt} + \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} |M(j,t)| \, \widetilde{d}^2_{jt}$$

$$= \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left[\left(d_{jt} - \overline{d}_j \right)^2 - 2\widehat{\widetilde{a}}_{jt} \left(d_{jt} - \overline{d}_j \right) + \widehat{\widetilde{a}}^2_{jt} \right] + \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \widehat{\widetilde{a}}^2_{jt} |M(j,t)|$$

$$\xrightarrow{p} \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left(d_{jt} - \overline{d}_j \right)^2.$$

This result follows because

$$\sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \widehat{\widetilde{a}}_{jt}^2 \, |M(j,t)|$$

$$= \sum_{j=N_0+1}^{N_0+N_1} \left[\sum_{t=1}^{T} \left(\widehat{a}_t - \frac{\sum_{\tau=1}^{T-1} |M(j,\tau| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau|} \right)^2 |M(j,t)| \right]$$

$$= \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T-1} \widehat{a}_t^2 \, |M(j,t)| - 2 \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T-1} \widehat{a}_t \frac{\sum_{\tau=1}^{T-1} |M(j,\tau| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau|}$$

$$+ \sum_{j=N_0+1}^{N_0+N_1} \sum_{t=1}^{T} \left(\frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|} \right)^2 |M(j,t)|$$

$$= \sum_{t=1}^{T-1} \widehat{a}_t^2 \sum_{j=N_0+1}^{N_0+N_1} |M(j,t)| - 2 \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} \widehat{a}_t \widehat{a}_\tau \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)|}{\sum_{s=1}^{T} |M(j,s)|}$$

$$+ \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} \widehat{a}_\tau \widehat{a}_t \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)| \, |M(j,t)|}{\sum_{s=1}^{T} |M(j,s)|}$$

$$= \sum_{t=1}^{T-1} O_p \left(\frac{1}{N_1^2} \right) \sum_{j=N_0+1}^{N_0+N_1} |M(j,t)| - 2 \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} O_p \left(\frac{1}{N_1^2} \right) \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)|}{\sum_{s=1}^{T} |M(j,s)|}$$

$$+ \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} O_p \left(\frac{1}{N_1^2} \right) \sum_{j=N_0+1}^{N_0+N_1} \frac{|M(j,\tau)| \, |M(j,t)|}{\sum_{s=1}^{T} |M(j,s)|}$$

$$\overset{p}{\rightarrow} 0.$$

Next consider the object

$$\sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}_i = \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i + \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \widehat{\widetilde{a}}_{jt} \widetilde{X}_i$$

$$= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i - \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T-1} \sum_{i \in m(j,t)} \left(\widehat{a}_t - \frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|} \right) \widetilde{X}_i$$

$$+ \sum_{j=1}^{N_0+N_1} \sum_{i \in m(j,t)} \frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|} \widetilde{X}_i$$

$$= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i - \sum_{t=1}^{T-1} \widehat{a}_t \sum_{j=1}^{N_0+N_1} \sum_{i \in m(j,t)} \widetilde{X}_i$$

$$+ \sum_{j=1}^{N_0+N_1} \frac{\sum_{\tau=1}^{T-1} |M(j,\tau)| \widehat{a}_\tau}{\sum_{\tau=1}^{T} |M(j,\tau)|} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \widetilde{X}_i$$

$$= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \widetilde{X}_i$$

$$= O_p(1).$$

We used the fact that $\widetilde{X}_i$ is the residual from a regression on time and state dummies so $\sum_{j=1}^{N_0+N_1} \sum_{i \in m(j,t)} \widetilde{X}_i = 0$ and $\sum_{t=1}^{T} \sum_{i \in m(j,t)} \widetilde{X}_i = 0$.

An analogous argument gives

$$\sum_i \widetilde{d}_{j(i)t(i)} \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right) = \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\widetilde{\eta}_{jt} + \widetilde{\varepsilon}_i\right) + \sum_{j=1}^{N_0+N_1} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \widehat{\widetilde{a}}_{jt} \left(\widetilde{\eta}_{jt} + \widetilde{\varepsilon}_i\right)$$

$$= \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\widetilde{\eta}_{jt} + \widetilde{\varepsilon}_i\right)$$

$$\overset{p}{\to} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\eta_{jt} - \overline{\eta}_j + \varepsilon_i - \overline{\varepsilon}_j\right)$$

$$= O_p(1).$$

The last term follows because for any $\tau = 1,..,T$ $E\left(\eta_{j\tau} - \overline{\eta}_j\right) = E(\varepsilon_i - \overline{\varepsilon}_j \mid t(i) = \tau) = 0$. So for a regression of either $\left(\eta_{j\tau} - \overline{\eta}_j\right)$ or $(\varepsilon_i - \overline{\varepsilon}_j)$ on time dummies, the coefficient on the dummy variables will converge to zero so $\widetilde{\eta}_{jt} \overset{p}{\to} \left(\eta_{j\tau} - \overline{\eta}_j\right)$ and $\widetilde{\varepsilon}_i \overset{p}{\to} (\varepsilon_i - \overline{\varepsilon}_j)$.

Putting all the objects into the expression for $\widehat{\beta}$, one can see that $\widehat{\beta}$ is consistent.

Now consider $\widehat{\alpha}$. It is straight forward to show that

$$(\widehat{\alpha} - \alpha) = \frac{\sum_i \widetilde{d}_{j(i)t(i)} \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right)}{\sum_i \widetilde{d}_{j(i)t(i)}^2} + \frac{\sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}'_{j(i)t(i)} \left(\beta - \widehat{\beta}\right)}{\sum_i \widetilde{d}_{j(i)t(i)}^2}.$$

We showed above that

$$\sum_i \widetilde{d}^2_{j(i)t(i)} \xrightarrow{p} \sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left(d_{jt} - \overline{d}_j\right)^2$$

$$\sum_i \widetilde{d}_{j(i)t(i)} \widetilde{X}'_{j(i)t(i)} = O_p(1)$$

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

$$\sum_i \widetilde{d}_{j(i)t(i)} \left(\widetilde{\eta}_{j(i)t(i)} + \widetilde{\varepsilon}_i\right) \xrightarrow{p} \sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in m(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\eta_{jt} - \overline{\eta}_j + \varepsilon_i - \overline{\varepsilon}_j\right).$$

Thus we are left with:

$$(\widehat{\alpha} - \alpha) = \frac{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in M(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\eta_{jt} - \overline{\eta}_j + \varepsilon_i - \overline{\varepsilon}_j\right) + o_p(1)}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left(d_{jt} - \overline{d}_j\right)^2 + o_p(1)} + o_p(1)$$

$$\xrightarrow{p} \frac{\sum_{j=1}^{N_0} \sum_{t=1}^{T} \sum_{i \in M(j,t)} \left(d_{jt} - \overline{d}_j\right) \left(\eta_{jt} - \overline{\eta}_j + \varepsilon_i - \overline{\varepsilon}_j\right)}{\sum_{j=1}^{N_0} \sum_{t=1}^{T} |M(j,t)| \left(d_{jt} - \overline{d}_j\right)^2}$$

$$= O_p(1).$$

This gives the result.
∎

A.6 Proof of Proposition 2.3

As in the text recall that

$$\begin{aligned} v_i &\equiv \eta^*_{j(i)t(i)} + \varepsilon_i \\ \eta^*_{jt} &\equiv \alpha d_{jt} + \gamma_t + \theta_j + \eta_{jt}. \end{aligned}$$

Since v_i is the error term from the regression (6) after taking out time effects and observables, for each i, this is identified. η^*_{jt} is the component of this error term that is group and time-specific while ε_i is idiosyncratic.

Define $\iota_1(j,t)$ and $\iota_2(j,t)$ as any two different individuals from group j at time t. We can identify the joint distribution of

$$\left(v_{\iota_1(j,t)}, v_{\iota_2(j,t)}\right) = \left(\eta^*_{jt} + \varepsilon_{\iota_1(j,t)}, \eta^*_{jt} + \varepsilon_{\iota_2(j,t)}\right).$$

Since η^*_{jt} is independent of ε, applying Theorem 2.2, from this joint distribution we can identify the marginal distributions of ε and η^*_{jt}.

We next need to show that one can identify the joint distribution of $\eta^*_j \equiv \left(\eta^*_{j1}, ..., \eta^*_{jT}\right)$. Since there is a unique mapping between characteristic functions and distributions, we know that the characteristic function of ε is identified. Define this to be $\phi_\varepsilon(\cdot)$.

Using a similar argument to above, take $\iota(j,t)$ to be any individual from group j at time t, we can identify the joint distribution of

$$\Psi_j \equiv \left(v_{\iota(j,1)}, ..., v_{\iota(j,T)}\right) = \left(\eta^*_{j1} + \varepsilon_{\iota(j,1)}, ..., \eta^*_{jT} + \varepsilon_{\iota(j,T)}\right).$$

Let $\Gamma = (\gamma_1, ..., \gamma_T)'$ be a $T \times 1$ vector. Then since Ψ_j is identified directly from the residuals of the regression for the controls, we can identify

$$\frac{E\left(\exp\left(i\Gamma'\Psi_j\right)\right)}{\prod_{t=1}^{T}\phi_\varepsilon\left(\gamma_t\right)} = \frac{E\left(\exp\left(i\Gamma'\eta_j^* + i\sum_{t=1}^{T}\gamma_t\varepsilon_{\iota(j,t)}\right)\right)}{\prod_{t=1}^{T}\phi_\varepsilon\left(\gamma_t\right)}$$

$$= \frac{E\left(\exp\left(i\Gamma'\eta_j^*\right)\right)\prod_{t=1}^{T}E\left(i\gamma_t\varepsilon_{\iota(j,t)}\right)}{\prod_{t=1}^{T}\phi_\varepsilon\left(\gamma_t\right)}$$

$$= E\left(\exp\left(i\Gamma'\eta_j^*\right)\right)$$

which is the characteristic function of $\widetilde{\eta}_j^*$. Thus the distribution of $\widetilde{\eta}_j^*$ is identified.

From the distribution of $\widetilde{\eta}_j^*$ and ε_i and with knowledge of d_{jt} and $|M(j,t)|$ for the control states we can identify the distribution of

$$\frac{\sum_{j=1}^{N_0}\sum_{t=1}^{T}\left(\sum_{i\in M(j,t)}\left(d_{jt}-\overline{d}_j\right)\left(\eta_{jt}^* - \overline{\eta_j^*} - E\left(\eta_{jt}^* - \overline{\eta_j^*}\right) + \varepsilon_i - \overline{\varepsilon}_j\right)\right)}{\sum_{j=1}^{N_0}\sum_{t=1}^{T}|M(j,t)|\left(d_{jt}-\overline{d}_j\right)^2}$$

$$= \frac{\sum_{j=1}^{N_0}\sum_{t=1}^{T}\left(\sum_{i\in M(j,t)}\left(d_{jt}-\overline{d}_j\right)\left(\gamma_t + \eta_{jt} - \overline{\gamma}_j - \overline{\eta_{jt}} - \left(\gamma_t - \overline{\gamma}_j\right) + \varepsilon_i - \overline{\varepsilon}_j\right)\right)}{\sum_{j=1}^{N_0}\sum_{t=1}^{T}|M(j,t)|\left(d_{jt}-\overline{d}_j\right)^2}$$

$$= \frac{\sum_{j=1}^{N_0}\sum_{t=1}^{T}\left(\sum_{i\in M(j,t)}\left(d_{jt}-\overline{d}_j\right)\left(\eta_{jt} - \overline{\eta_{jt}} + \varepsilon_i - \overline{\varepsilon}_j\right)\right)}{\sum_{j=1}^{N_0}\sum_{t=1}^{T}|M(j,t)|\left(d_{jt}-\overline{d}_j\right)^2}$$

which is the distribution of $\left(\widehat{\alpha} - \alpha\right)$. ∎

A.7   Consistent Estimation of distribution of $\left[\eta_{j1}^*, ..., \eta_{jT}^*\right]$ and ε

Our goal is to show consistency of the Sieve estimator (14). Since the likelihood function is a continuously differentiable function of β and γ, we ignore the fact that they are estimated which can be addressed in the standard way. Our goal is to estimate the joint distribution of $\left[\eta_{j1}^*, ..., \eta_{jT}^*\right]$ and also the density of ε from the joint distribution of v_i. Call the first distribution F. We assume that we can write ε as the convolution between a random variable with distribution G and a normal random variable with mean 0 and standard deviation σ.

In order to keep our underlying sets compact we assume that the support of G and F are compact. Formally

Assumption A.2 *$G \in \mathcal{G}$ where $\mathcal{G}$ is the set of distribution functions with support Ξ which is a compact subset of $\Re$.*

Assumption A.3 *$F \in \mathcal{F}_T$ where $\mathcal{F}_T$ is the set of distribution functions with support Θ which is a compact subset of $\Re^T$.*

Thus our space of interest is $\mathcal{F}_T \times \mathcal{G}$. We use an L_2 norm:

$$d\left((F,G),\left(\widetilde{F},\widetilde{G}\right)\right) = \int_\Theta \left(F(x) - \widetilde{F}(x)\right)^2 dx + \int_\Xi \left(G(z) - \widetilde{G}(z)\right)^2 dz.$$

Our model is a Sieve estimator in that we do not maximize the likelihood function with respect to $G \in \mathcal{G}$ and $F \in \mathcal{F}_T$, but rather maximizes relative to a subset of these distributions $\mathcal{G}^N$ and $\mathcal{F}_T^N$ which restrict the distributions to be step functions. The number of mass points expand asymptotically in N so that $\mathcal{G}^N \times \mathcal{F}_T^N$ becomes dense in $\mathcal{G}$ and $\mathcal{F}_T$. We denote

$$F^K(x) \equiv \sum_{j_1=1}^{K_1} 1\left(\eta^{(\kappa_1)} \le x\right) P_1^{(j_1)}$$

$$G^K(z) \equiv \sum_{j_2=1}^{K_2} 1\left(\mu^{(\kappa_1)} \le z\right) P_2^{(j_2)}.$$

Under these conditions our model is consistent. That is

Proposition A.3 *Let the objective function be*

$$\mathcal{L}(F^{K_1}, G^{K_2}) = \frac{1}{N_1} \sum_{j=N_0+1}^{N_0+N_1} \log\left(\sum_{j_1=1}^{K_1} \prod_{t=1}^{T} \prod_{i \in M(j,t)} \sum_{j_2=1}^{K_2} \phi\left(\frac{v_i - \eta_t^{(j_1)} - \mu^{(j_2)}}{\sigma}\right) P_1^{(j_1)} P_2^{(j_2)}\right)$$

where we have parameterized $\left[\eta_{j(i)t(i)}^*\right]$ to take on K_1 values with each value taking the value $\eta^{(j_1)} = \left(\eta_1^{(j_1)}, ..., \eta_T^{(j_1)}\right)$ with probability $P_1^{(j_1)}$ for $j_1 = 1, .., K_1$. We let ε be a mixture of normals that take on K_2 values with mean and standard deviation $(\mu^{(j_2)}, \sigma)$ with probability $P_2^{(j_2)}$ for $j_2 = 1, ..., K_2$. Let $\left(\widehat{F}_T^{K_1}, \widehat{G}^{K_2}\right)$ denote the maximum of the objective function. Under Assumptions 2.1-2.6, and $A.1 - A.3$, $\left(\widehat{F}_T^{K_1}, \widehat{G}^{K_2}\right)$ converges in probability to the true values of (F_T, G) as long as K_1 and $K_2 \to \infty$ as $N_1 \to \infty$.

Proof: We will verify the condition of the theorem in Matzkin (1994) section 3.2 which is a restatement of Theorem 0 in Gallant and Nychka (1987).

The asymptotic limit of the likelihood function is

$$L(F,G) = E\left(\log\left(\int \prod_{t=1}^{T} \prod_{i \in M(j,t)} \int \phi\left(\frac{v_i - \eta_{jt} - \varepsilon}{\sigma}\right) dG(\varepsilon) dF(\eta_j)\right)\right).$$

We prove consistency by verifying each of the four conditions that Matzkin (1994) requires.
Condition (iii) the set $\mathcal{F}_T \times \mathcal{G}$ is compact relative to the metric d.

Helly's Selection Theorem guarantees that any sequence of distribution functions will have a convergent subsequence which is a valid distribution function except that it may not converge to zero as $x \to -\infty$ and may not converge to 1 as $x \to \infty$. The fact that Ξ and Θ are compact guarantees that the limit of a subsequence in $\mathcal{F}_T \times \mathcal{G}$ will be an element of $\mathcal{F}_T \times \mathcal{G}$, therefore the set is compact.

Condition (i) The function $L_N(F,G)$ converges uniformly over M to a nonrandom continuous function $L : M \to \Re$

For this we apply Lemma 2.4 in Newey and McFadden (1994). The likelihood function is clearly continuous and for any z the log likelihood will be bounded since the support is compact.

Condition (iv) There exists a sequence of function $\{g_n\} \subset M$ such that $g_N \in M_N$ for all $N = 1, 2, ... and$ $d(g_N, m^*) \to 0$

We can always find a sequence of step functions that converges to the actual CDF. One obvious way is to do this would be to take the number of support points $M = P^T$ where P is an integer that depends on N. We then divide the support of η into M cubes, take $P_1^{(\kappa_1)}$ to be the probability of lying in each cube, and take $P_1^{(\kappa_1)}$ to be the median point. This will converge to F_T as M gets large.

Condition (ii) the function m^* uniquely maximizes L over the set M

We proved that the model is identified in Proposition 2.3. The fact that m^* uniquely maximizes the likelihood comes from the standard result that log likelihood function is maximized at the true distribution (e.g. Lemma 2.2 of Newey and McFadden, 1994). ∎

Table 1

Estimated Parameters for

Effect of Georgia Hope Program on College Attendance

	Population Weighted Linear Probability	State Weighted Linear Probability	Logit Parameters
Hope Scholarship	0.072	0.077	0.359
Male	-0.077	-0.076	-0.323
Black	-0.155	-0.155	-0.673
Asian	0.173	0.172	0.726
State Dummies	yes	yes	yes
Year Dummies	yes	yes	yes
95% Confidence intervals for Hope Effect			
Standard Cluster by State×Year	(0.025, 0.119)	(0.025,0.130)	(0.119,0.600) [0.030,0.149]
Standard Cluster by State	(0.050,0.094)	(0.058,0.097)	(0.274,0.444) [0.068,0.111]
Conley-Taber	(-0.006,0.212)	(-0.008,0.207)	(-0.030,0.905) [-0.007,0.219]
Sample Size			
Number States	41	41	41
Number of Individuals	34902	34902	34902

Note: Confidence intervals for parameters are presented in parentheses. Brackets contain a confidence interval for the program impact upon a person whose college attendance probability in the absence of the program would be 45%.

Table 2

Estimated Parameters for

Effect of Merit Aide Programs on College Attendance

	Population Weighted Linear Probability	State Weighted Linear Probability	Logit Parameters
Merit Scholarship	0.034	0.051	0.229
Male	-0.079	-0.078	-0.331
Black	-0.150	-0.150	-0.655
Asian	0.169	0.168	0.707
State Dummies	yes	yes	yes
Year Dummies	yes	yes	yes
95% Confidence intervals for Hope Effect			
Standard Cluster by State×Year	(0.006,0.062)	(0.024,0.078)	(0.111,0.346) [0.028,0.086]
Standard Cluster by State	(0.008,0.059)	(0.028,0.074)	(0.127,0.330) [0.032,0.082]
Conley-Taber	(0.001,0.095)	(0.012,0.075)	(0.061,0.405) [0.015,101]
Sample Size			
Number States	51	51	51
Number of Individuals	42161	42161	42161

Note: Confidence intervals for parameters are presented in parentheses. Brackets contain a confidence interval for the program impact upon a person whose college attendance probability in the absence of the program would be 45%.

Table 3

Confidence Interval for

Effect Merit Aide Programs on College Attendance

Using Model 2

	Georgia Hope Program	Merit Aide Programs
95% Confidence Intervals	(-0.170,0.917)	(-0.017,0.367)
	[-0.041,0.222]	[-0.004,0.092]
90% Confidence Intervals	(-0.080,0.796)	(0.033,0.343)
	[-0.020,0.195]	[0.008,0.086]

Note: Confidence intervals for parameters are presented in parentheses. Brackets contain a confidence interval for the program impact upon a person whose college attendance probability in the absence of the program would be 45%.

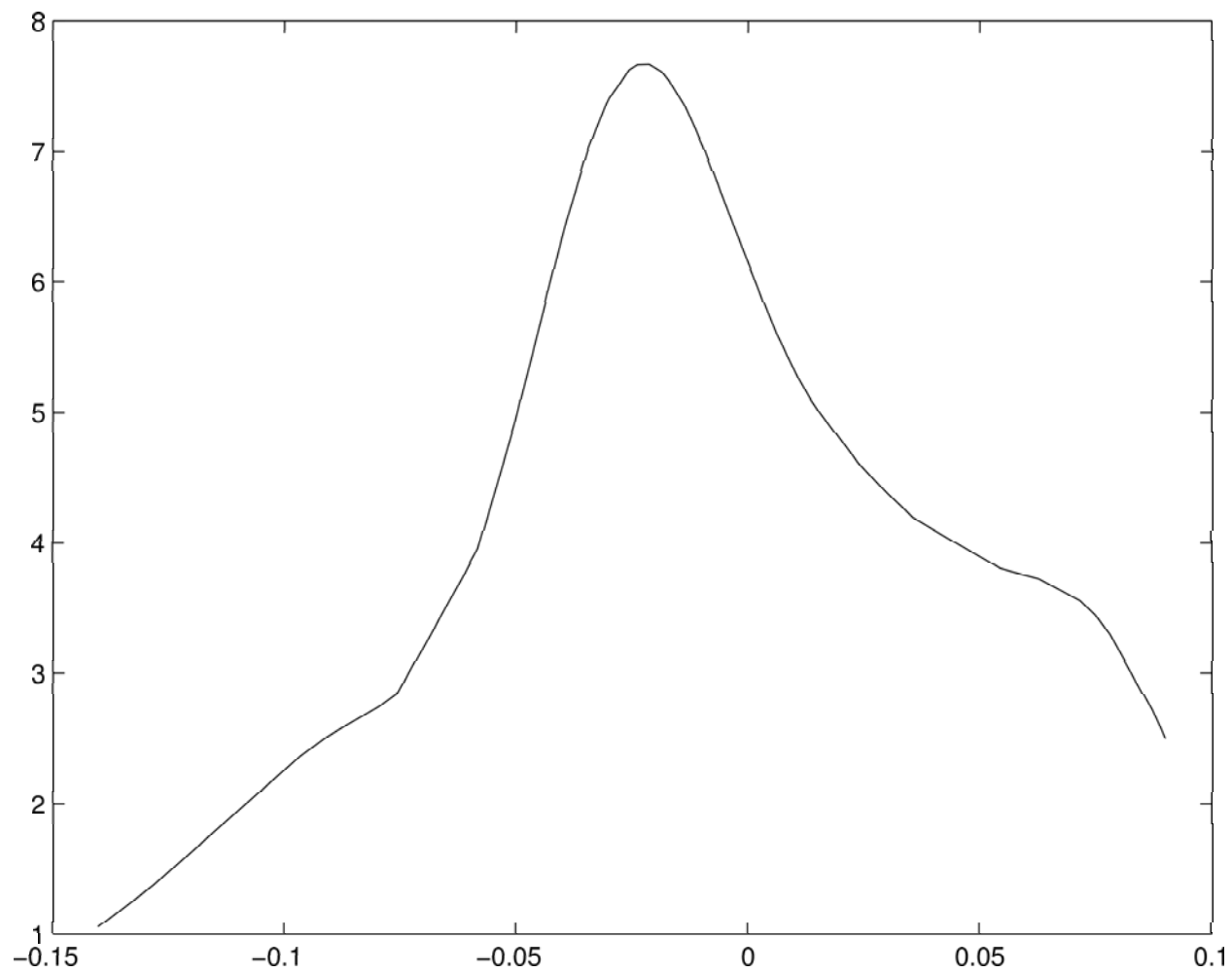Figure 1: Estimated Density of $\hat{\alpha}$ under $H_0 : \alpha_0 = 0$

Figure 2: Distribution of P-values from Control States