What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 10, **Tuesday**, **July 31st**, **4.30-5.30 pm**

Difference-in-Differences Estimation

These notes provide an overview of standard difference-in-differences methods that have been used to study numerous policy questions. We consider some recent advances in Hansen (2007a,b) on issues of inference, focusing on what can be learned with various group/time period dimensions and serial independence in group-level shocks. Both the repeated cross sections and panel data cases are considered. We discuss recent work by Athey and Imbens (2006) on nonparametric approaches to difference-in-differences, and Abadie, Diamond, and Hainmueller (2007) on constructing synthetic control groups.

1. Review of the Basic Methodology

Since the work by Ashenfelter and Card (1985), the use of difference-in-differences methods has become very widespread. The simplest set up is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. In the case where the same units within a group are observed in each time period, the average gain in the second (control) group is substracted from the average gain in the first (treatment) group. This removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between those groups, as well as biases from comparisons over time in the treatment group that could be the result of trends. We will treat the panel data case in Section 4.

With repeated cross sections, we can write the model for a generic member of any of groups as

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u \tag{1.1}$$

where y is the outcome of interest, $d2$ is a dummy variable for the second time period. The dummy variable dB captures possible differences between the treatment and control groups prior to the policy change. The time period dummy, $d2$, captures aggregate factors that would cause changes in y even in the absense of a policy change. The coefficient of interest, δ_1, multiplies the interaction term, $d2 \cdot dB$, which is the same as a dummy variable equal to one for those observations in the treatment group in the second period. The difference-in-differences estimate is

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}). \tag{1.2}$$

Inference based on even moderate sample sizes in each of the four groups is straightforward, and is easily made robust to different group/time period variances in the regression framework.

In some cases a more convincing analysis of a policy change is available by further refining the definition of treatment and control groups. For example, suppose a state implements a change in health care policy aimed at the elderly, say people 65 and older, and the response variable, y, is a health outcome. One possibility is to use data only on people in the state with the policy change, both before and after the change, with the control group being people under 65 and the treatment group being people 65 and older. The potential problem with this DD analysis is that other factors unrelated to the state's new policy might affect the health of the elderly relative to the younger population, for example, changes in health care emphasis at the federal level. A different DD analysis would be to use another state as the control group and use the elderly from the non-policy state as the control group. Here, the problem is that *changes* in the health of the elderly might be systematically different across states due to, say, income and wealth differences, rather than the policy change.

A more robust analysis than either of the DD analyses described above can be obtained by using both a different state and a control group within the treatment state. If we again label the two time periods as one and two, let B represent the state implementing the policy, and let E denote the group of elderly, then an expanded verson of (1.1) is

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \quad (1.3)$$

The coefficient of interest is now δ_3, the coefficient on the triple interaction term, $d2 \cdot dB \cdot dE$. The OLS estimate $\hat{\delta}_3$ can be expressed as follows:

$$\hat{\delta}_3 = (\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1}) \quad (1.4)$$

where the A subscript means the state not implementing the policy and the N subscript represents the non-elderly. For obvious reasons, the estimator in (1.4) is called the *difference-in-difference-in-differences (DDD)* estimate. [The population analog of (1.4) is easily established from (1.3) by finding the expected values of the six groups appearing in (1.4).] If we drop either the middle term or the last term, we obtain one of the DD estimates described in the previous paragraph. The DDD estimate starts with the time change in averages for the elderly in the treatment state and then nets out the change in means for elderly in the control state and the change in means for the non-elderly in the treatment state. The hope is that this controls for two kinds of potentially confounding trends: changes in health status of

elderly across states (that would have nothing to do with the policy) and changes in health status of all people living in the policy-change state (possibly due to other state policies that affect everyone's health, or state-specific changes in the economy that affect everyone's health). When implemented as a regression, a standard error for $\hat{\delta}_3$ is easily obtained, including a heteroskedasticity-robust standard error. As in the DD case, it is straightforward to add additional covariates to (1.3) and inference robust to heteroskedasticity.

2. How Should We View Uncertainty in DD Settings?

The standard approach just described assumes that all uncertainty in inference enters through sampling error in estimating the means of each group/time period combination. This approach has a long history in statistics, as it is equivalent to analysis of variance. Recently, different approaches have been suggest that focus on different kinds of uncertainty – perhaps in addition to sampling error in estimating means. Recent work by Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Hansen (2007a,b), and Abadie, Diamond, and Hainmueller (2007) argues for additional sources of uncertainty. In fact, in most cases the additional uncertainty is assumed to swamp the sampling error in estimating group/time period means. We already discussed the DL approach in the cluster sample notes, although we did not explicitly introduce a time dimension. One way to view the uncertainty introduced in the DL framework – and a perspective explicitly taken by ADH – is that our analysis should better reflect the uncertainty in the quality of the control groups.

Before we turn to a general setting, it is useful to ask whether introducing more than sampling error into DD analyses is necessary, or desirable. As we discussed in the cluster sample notes, the DL approach does not allow inference in the basic comparison-of-mean case for two groups. While the DL estimate is the usual difference in means, the error variance of the cluster effect cannot be estimated, and the *t* distribution is degenerate. It is also the case that the DL approach cannot be applied to the standard DD or DDD cases covered in Section 1. We either have four different means to estimate or six, and the DL regression in these cases produces a perfect fit with no residual variance. Should we conclude nothing can be learned in such settings?

Consider the example from Meyer, Viscusi, and Durbin (1995) on estimating the effects of benefit generosity on length of time a worker spends on workers' compensation. MVD have a before and after period, where the policy change was to raise the cap on covered earnings. The treatment group is high earners, and the control group is low earners – who should not have

been affected by the change in the cap. Using the state of Kentucky and a total sample size of 5,626, MVD find the DD estimate of the policy change is about 19.2% (longer time on workers' compensation). The *t* statistic is about 2.76, and the estimate changes little when some controls are added. MVD also use a data set for Michigan. Using the same DD approach, they estimate an almost identical effect: 19.1%. But, with "only" 1,524 observations, the *t* statistic is 1.22. It seems that, in this example, there is plenty of uncertainty in estimation, and one cannot obtain a tight estimate without a fairly large sample size. It is unclear what we gain by concluding that, because we are just identifying the parameters, we cannot perform inference in such cases. In this example, it is hard to argue that the uncertainty associated with choosing low earners within the same state and time period as the control group somehow swamps the sampling error in the sample means.

3. General Settings for DD Analysis: Multiple Groups and Time Periods

The DD and DDD methodologies can be applied to more than two time periods. In the first case, a full set of time-period dummies is added to (1.1), and a policy dummy replaces $d2 \cdot dB$; the policy dummy is simply defined to be unity for groups and time periods subject to the policy. This imposes the restriction that the policy has the same effect in every year, and assumption that is easily relaxed. In a DDD analysis, a full set of dummies is included for each of the two kinds of groups and all time periods, as well as all pairwise interactions. Then, a policy dummy (or sometimes a continuous policy variable) measures the effect of the policy. See Gruber (1994) for an application to mandated maternity benefits.

With many time periods and groups, a general framework considered by BDM (2004) and Hansen (2007b) is useful. The equation at the individual level is

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{3.1}$$

where i indexes individual, g indexes group, and t indexes time. This model has a full set of time effects, λ_t, a full set of group effects, α_g, group/time period covariates, x_{gt} (these are the policy variables), individual-specific covariates, $\mathbf{z}_{igt}$, unobserved group/time effects, v_{gt}, and individual-specific errors, u_{igt}. We are interested in estimating $\boldsymbol{\beta}$. Equation (3.1) is an example of a *multilevel model*.

One way to write (3.1) that is useful is

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{3.2}$$

which shows a model at the individual level where both the intercepts and slopes are allowed to differ across all (g,t) pairs. Then, we think of δ_{gt} as

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \tag{3.3}$$

Equation (3.3) is very useful, as we can think of it as a regression model at the group/time period level.

As discussed by BDM, a common way to estimate and perform inference in (3.1) is to ignore v_{gt}, in which case the observations at the individual level are treated as independent. When v_{gt} is present, the resulting inference can be very misleading. BDM and Hansen (2007b) allow serial correlation in $\{v_{gt} : t = 1,2,\ldots,T\}$ and assume independence across groups, g.

A simple way to proceed is to view (3.3) as ultimately of interest. We observe $\mathbf{x}_{gt}$, λ_t is handled with year dummies,and α_g just represents group dummies. The problem, then, is that we do not observe δ_{gt}. But we can use the individual-level data to estimate the δ_{gt}, provided the group/time period sizes, M_{gt}, are reasonably large. With random sampling within each (g,t), the natural estimate of δ_{gt} is obtained from OLS on (3.2) for each (g,t) pair, assuming that $E(\mathbf{z}_{igt}'u_{igt}) = \mathbf{0}$. (In most DD applications, this assumption almost holds by definition, as the individual-specific controls are included to improve estimation of δ_{gt}.) If a particular model of heteroskedasticity suggests itself, and $E(u_{it}|\mathbf{z}_{igt}) = 0$ is assumed, then a weighted least squares procedure can be used. Sometimes one wishes to impose some homogeneity in the slopes – say, $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}_g$ or even $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}$ – in which case pooling can be used to impose such restrictions. In any case, we proceed as if the M_{gt} are large enough to ignore the estimation error in the $\hat{\delta}_{gt}$; instead, the uncertainty comes through v_{gt} in (3.3). Hansen (2007b) considers adjustments to inference that accounts for sampling error in the $\hat{\delta}_{gt}$, but the methods are more complicated. The minimum distance approach we discussed in the cluster sampling notes, applied in the current context, effectively drops v_{gt} from (3.3) and views $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta}$ as a set of deterministic restrictions to be imposed on δ_{gt}. Inference using the efficient minimum distance estimator uses only sampling variation in the $\hat{\delta}_{gt}$, which will be independent across all (g,t) if they are separately estimated, or which will be correlated if pooled methods are used.

Because we are ignoring the estimation error in $\hat{\delta}_{gt}$, we proceed simply by analyzing the panel data equation

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}, \quad t = 1,\ldots,T, g = 1,\ldots,G, \tag{3.4}$$

where we keep the error as v_{gt} because we are treating $\hat{\delta}_{gt}$ and δ_{gt} interchangeably. If we assume that We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (3.4) by OLS – which means full year and group effects, along with x_{gt} – then the OLS estimator has satisfying properties as G and T both increase, provided $\{v_{gt} : t = 1, 2, \ldots, T\}$ is a weakly dependent (mixing) time series for all g. The simulations in BDM and Hansen (2007a) indicate that cluster-robust inference, where each cluster is a set of time periods, work reasonably well when $\{v_{gt}\}$ follows a stable AR(1) model and G is moderately large.

Hansen (2007b), noting that the OLS estimator (the fixed effects estimator) applied to (3.4) is inefficient when v_{gt} is serially uncorrelated (and possibly heteroskedastic), proposes feasible GLS. As is well known, if T is not large, estimating parameters for the variance matrix $\Omega_g = Var(\mathbf{v}_g)$, where $\mathbf{v}_g$ is the $T \times 1$ error vector for each g, is difficult when group effects have been removed. In other words, using the FE residuals, $\hat{v}_{gt}$, to estimate Ω_g can result in severe bias for small T. Solon (1984) highlighted this problem for the homoskedastic AR(1) model. Of course, the bias disappears as $T \to \infty$, and regression packages such as Stata, that have a built-in command to do fixed effects with AR(1) errors, use the usual AR(1) coefficient $\hat{\rho}$, obtained from

$$\hat{v}_{gt} \text{ on } \hat{v}_{g,t-1}, \; t = 2, \ldots, T, g = 1, \ldots, G. \tag{3.5}$$

As discussed in Wooldridge (2003) and Hansen (2007b), one way to account for the bias in $\hat{\rho}$ is to still use a fully robust variance matrix estimator. But Hansen's simulations show that this approach is quite inefficient relative to his suggestion, which is to bias-adjust the estimator $\hat{\rho}$ and then use the bias-adjusted estimator in feasible GLS. (In fact, Hansen covers the general $AR(p)$ model.) Hansen derives many attractive theoretical properties of his the estimator.An iterative bias-adjusted procedure has the same asymptotic distribution as $\hat{\rho}$ in the case $\hat{\rho}$ should work well: G and T both tending to infinity. Most importantly for the application to DD problems, the feasible GLS estimator based on the iterative procedure has the same asymptotic distribution as the GLS etsimator when $G \to \infty$ and T is fixed. When G and T are both large, there is no need to iterated to achieve efficiency.

Hansen further shows that, even when G and T are both large, so that the unadjusted AR coefficients also deliver asymptotic efficiency, the bias-adusted estimates deliver higher-order improvements in the asymptotic distribution. One limitation of Hansen's results is that they assume $\{\mathbf{x}_{gt} : t = 1, \ldots, T\}$ are strictly exogenous. We know that if we just use OLS – that is,

the usual fixed effects estimate – strict exogeneity is not required for consistency as $T \to \infty$. GLS, in exploiting correlations across different time periods, tends to exacerbate bias that results from a lack of strict exogeneity. In policy analysis cases, this is a concern if the policies can switch on and off over time, because one must decide whether the decision to implement or remove a program is related to past outcomes on the response.

With large G and small T, one can estimate an unstricted variance matrix Ω_g and proceed with GLS – this is the approach suggested by Kiefer (1980) and studied more recently by Hausman and Kuersteiner (2003). It is equivalent to dropping a time period in the time-demeaned equation and proceeding with full GLS (and this avoids the degeneracy in the variance matrix of the time-demeaned errors). Hausman and Kuersteiner show that the Kiefer approach works pretty well when $G = 50$ and $T = 10$, although substantial size distortions exist for $G = 50$ and $T = 20$.

Especially if the M_{gt} are not especially large, we might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. One simple way to avoid ignoring the estimation error in $\hat{\delta}_{gt}$ is to aggregate equation (3.1) over individuals, giving

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \; t = 1, .., T, g = 1, \ldots, G. \tag{3.6}$$

Of course, this equation can be estimated by fixed effects, too, and fully robust inference is available using Hansen (2007a) because the composite error, $\{r_{gt} \equiv v_{gt} + \bar{u}_{gt}\}$, is weakly dependent. Fixed Effects GLS using an unrestricted variance matrix can be used with large G and small T. The complication with using specific time series model for the error is the presence of $\bar{u}_{gt}$. With different M_{gt}, $Var(\bar{u}_{gt})$ is almost certainly heteroskedastic (and might be with the same M_{gt}, of course). So, even if we specify, say, an AR(1) model $v_{gt} = \rho v_{g,t-1} + e_{gt}$, the variance matrix of $\mathbf{r}_g$ is more complicated. One possibility is to just assume the composite error, r_{gt}, follows a simple model, implement Hansen's methods, but then use fully robust inference.

The Donald and Land (2007) approach applies in the current setting by using finite sample analysis applied to the pooled regression (3.4). However, DL assume that the errors $\{v_{gt}\}$ are uncorrelated across time, and so, even though for small G and T it uses small degrees-of-freedom in a t distribution, it does not account for uncertainty due to serial correlation in $\{v_{gt} : t = 1, \ldots, T\}$.

4. Individual-Level Panel Data

Individual-level panel data is a powerful tool for estimating policy effects. In the simplest

case we have two time periods and a binary program indicator, w_{it}, which is unity if unit i participates in the program at time t. A simple, effective model is

$$y_{it} = \alpha + \eta d2_t + \tau w_{it} + c_i + u_{it}, \; t = 1, 2, \tag{4.1}$$

where $d2_t = 1$ if $t = 2$ and zero otherwise, c_i is an observed effect, and u_{it} are the idiosyncratic errors. The coefficient τ is the treatment effect. A simple estimation procedure is to first difference to remove c_i :

$$(y_{i2} - y_{i1}) = \eta + \tau(w_{i2} - w_{i1}) + (u_{i2} - u_{i1}) \tag{4.2}$$

or

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i. \tag{4.3}$$

If $E(\Delta w_i \Delta u_i) = 0$, that is, the change in treatment status is uncorrelated with changes in the idiosyncratic errors, then OLS applied to (4.3) is consistent. The leading case is when $w_{i1} = 0$ for all i, so that no units we exposed to the program in the initial time period. Then the OLS estimator is

$$\hat{\tau} = \Delta\bar{y}_{treat} - \Delta\bar{y}_{control}, \tag{4.4}$$

which is a difference-in-differences estimate except that we different the means of the same units over time.This same estimate can be derived without introducing heterogeneity by simply writing the equation for y_{it} with a full set of group-time effects. Also, (4.4) is not the same estimate obtained from the regression y_{i2} on $1, y_{i1}, w_{i2}$ – that is, using y_{i1} as a control in a cross section regression. The estimates can be similar, but their consistency is based on different assumptions.

More generally, with many time periods and arbitrary treatment patterns, we can use

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it}, \; t = 1, \ldots, T, \tag{4.5}$$

which accounts for aggregate time effects and allows for controls, $\mathbf{x}_{it}$. Estimation by FE or FD to remove c_i is standard, provided the policy indicator, w_{it}, is strictly exogenous: correlation beween w_{it} and u_{ir} for any t and r causes inconsistency in both estimators, although the FE estimator typically has smaller bias when we can assume conteporaneous exogeneity, $Cov(w_{it}, u_{it}) = 0$. Strict exogeneity can be violated if policy assignment changes in reaction to past outcomes on y_{it}. In cases where $w_{it} = 1$ whenever $w_{ir} = 1$ for $r < t$, strict exogeneity is usually a reasonable assumption.

Equation (4.5) allows policy designation to depend on a level effect, c_i, but w_{it} might be

correlated with unit-specific trends in the response, too. This suggests the "correlated random trend" model

$$y_{it} = c_i + g_i t + \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + u_{it}, \; t = 1, \ldots, T, \tag{4.6}$$

where g_i is the trend for unit i. A general analysis allows arbitrary corrrelation between (c_i, g_i) and w_{it}, which requires at least $T \geq 3$. If we first difference, we get

$$\Delta y_{it} = g_i + \eta_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\gamma} + \Delta u_{it}, \; t = 2, \ldots, T, \tag{4.7}$$

where $\eta_t = \lambda_t - \lambda_{t-1}$ is a new set of time effects. We can estimate (4.7) by differencing again, or by using FE. The choice depends on the serial correlation properties in $\{\Delta u_{it}\}$ (assume strict exogeneity of treatment and covariates). If Δu_{it} is roughly uncorrelated, FE is preferred. If the original errors $\{u_{it}\}$ are essentially uncorrelated, applying FE to (4.6), in the general sense of sweeping out the linear trends from the response, treatment, and covariates, is preferred. Fully robust inference using cluster-robust variance estimators is straightforward. Of course, one might want to allow the effect of the policy to change over time, which is easy by interacting time dummies with the policy indicator.

We can derive standard panel data approaches using the counterfactural framework from the treatment effects literature.For each (i, t), let $y_{it}(1)$ and $y_{it}(0)$ denote the counterfactual outcomes, and assume there are no covariates. One way to state the assumption of unconfoundedness of treatment is that, for time-constant heterogeneity c_{i0} and c_{i1},

$$E(y_{it0}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it0}|c_{i0}) \tag{4.8}$$

$$E(y_{it1}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it1}|c_{i1}), \tag{4.9}$$

where $\mathbf{w}_i = (w_{i1}, \ldots, w_{iT})$ is the time sequence of all treatments. We saw this kind of strict exogeneity assumption conditional on latent variables several times before. It allows treatment to be correlated with time-constant heterogeneity, but does not allow treatment in any time period to be correlated with idiosyncratic changes in the counterfactuals. Next, assume that the expected gain from treatment depends at most on time:

$$E(y_{it1}|c_{i1}) = E(y_{it0}|c_{i0}) + \tau_t, \; t = 1, \ldots, T. \tag{4.10}$$

Writing $y_{it} = (1 - w_{it})y_{it0} + w_{it}y_{i1}$, and using (4.8), (4.9), and (4.10) gives

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) &= E(y_{it0}|c_{i0}) + w_{it}[E(y_{it1}|c_{i1}) - E(y_{it0}|c_{i0})] \\ &= E(y_{it0}|c_{i0}) + \tau_t w_{it}. \end{aligned} \tag{4.11}$$

If we now impose an additive structure on $E(y_{it0}|c_{i0})$, namely,

$$E(y_{it0}|c_{i0}) = \alpha_{t0} + c_{i0}, \tag{4.12}$$

then we arrive at

$$E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) = \alpha_{t0} + c_{i0} + \tau_t w_{it}, \tag{4.13}$$

an estimating equation that leads to well-known procedures. Because $\{w_{it} : t = 1, \ldots, T\}$ is strictly exogenous conditional on c_{i0}, we can use fixed effects or first differencing, with a full set of time period dummies. A standard analysis would use $\tau_t = \tau$, but, of course, we can easily allow the effects of the policy to change over time.

Of course, we can add covariates $\mathbf{x}_{it}$ to the conditioning sets and assume linearity, say $E(y_{it0}|\mathbf{x}_{it}, c_{i0}) = \alpha_{t0} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 + c_{i0}$. If (4.8) becomes

$$E(y_{it0}|\mathbf{w}_i, \mathbf{x}_i, c_{i0}, c_{i1}) = E(y_{it0}|\mathbf{x}_{it}, c_{i0}), \tag{4.14}$$

and similarly for (4.9), then the estimating equation simply adds $\mathbf{x}_{it}\boldsymbol{\gamma}_0$ to (4.13). More interesting models are obtained by allowing the gain from treatment to depend on heterogeneity. Suppose we assume, in addition to the ignorability assumption in (4.14) (and the equivalent condition for y_{it1})

$$E(y_{it1} - y_{it0}|\mathbf{x}_{it}, c_{i0}, c_{i1}) = \tau_t + a_i + (\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} \tag{4.15}$$

where a_i is a function of (c_{i0}, c_{i1}) normalized so that $E(a_i) = 0$ and $\boldsymbol{\xi}_t = E(\mathbf{x}_{it})$. Equation (4.15) allows the gain from treatment to depend on time, unobserved heterogeneity, and observed covariates. Then

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, c_{i0}, a_i) = {} & \alpha_{t0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 \\ & + w_{it}(\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i w_{it}. \end{aligned} \tag{4.16}$$

This is a correlated random coefficient model because the coefficient on w_{it} is $(\tau_t + a_i)$, which has expected value τ_t. Generally, we want to allow w_{it} to be correlated with a_i and c_{i0}. With small T and large N, we do not try to estimate the a_i (nor the c_{i0}). But an extension of the within transformation effectively eliminates $a_i w_{it}$. Suppose we simplify a bit and assume $\tau_t = \tau$ and drop all other covariates. Then, a regression that appears to suffer from an incidental parameters problem turns out to consistently estimate τ: Regress y_{it} on year dummies, dummies for each cross-sectional observation, and latter dummies interacted with w_{it}. In other words, we estimate

$$\hat{y}_{it} = \hat{\alpha}_{t0} + \hat{c}_{i0} + \hat{\tau}_i w_{it}. \tag{4.17}$$

While $\hat{\tau}_i$ is usually a poor estimate of $\tau_i = \tau + a_i$, their average is a good estimator of τ :

$$\hat{\tau} = N^{-1} \sum_{i=1}^{N} \hat{\tau}_i. \tag{4.18}$$

A standard error can be calculated using Wooldridge (2002, Section 11.2) or bootstrapping.

We can apply the results from the linear panel data notes to determine when the usual FE estimator – that is, the one that ignores $a_i w_{it}$ – is consistent for τ. In addition to the unconfoundedness assumption, sufficient is

$$E(\tau_i | \ddot{w}_{it}) = E(\tau_i) = \tau, t = 1, \ldots, T, \tag{4.19}$$

where $\ddot{w}_{it} = w_{it} - \bar{w}_i$. Essentially, the individual-specific treatment effect can be correlated with the average propensity to recieve treatment, $\bar{w}_i$, but not the deviations for any particular time period.

Assumption (4.19) is not completely general, and we might want a simple way to tell whether the treatment effect is heterogeneous across individuals. Here, we an exploit correlation between the τ_i and treatment. Recalling that $\tau_i = \tau + a_i$, a useful assumption (that need not hold for obtaining a test) is

$$E(a_i | w_{i1}, \ldots w_{iT}) = E(a_i | \bar{w}_i) = \rho(\bar{w}_i - \mu_{\bar{w}_i}), \tag{4.20}$$

where other covariates have been suppressed. Then we can estimate the equation (with covariates)

$$\begin{aligned} y_{it} = \alpha_{t0} + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 + w_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)\boldsymbol{\delta} \\ + \rho w_{it}(\bar{w}_i - \bar{w}) + c_{i0} + e_{it} \end{aligned} \tag{4.21}$$

by standard fixed effects. Then, we use a simple t test on $\hat{\rho}$, robust to heteroskedasticity and serial correlation. If we reject, it does not mean the mean usual FE estimator is inconsistent, but it could be.

5. Semiparametric and Nonparametric Approaches

Return to the setting with two groups and two time periods. Athey and Imbens (2006) generalize the standard DD model in several ways. Let the two time periods be $t = 0$ and 1 and label the two groups $g = 0$ and 1. Let $Y_i(0)$ be the counterfactual outcome in the absense of intervention and $Y_i(1)$ the counterfactual outcome with intervention. AI assume that

$$Y_i(0) = h_0(U_i, T_i), \tag{5.1}$$

where T_i is the time period and

$$h_0(u, t) \text{ strictly increasing in } u \text{ for } t = 0, 1 \tag{5.2}$$

The random variable U_i represents all unobservable characteristics of individual i. Equation (5.1) incorporates the idea that the outcome of an individual with $U_i = u$ will be the same in a given time period, irrespective of group membership. The strict monotonicity assumption in (5.2) rules out discrete responses, but Athey and Imbens (2006) provide bounds under weak monotonicity, and show how, with additional assumptions, point identification be be recovered.

The distribution of U_i is allowed to vary across groups, but not over time within groups, so that

$$D(U_i|T_i, G_i) = D(U_i|G_i). \tag{5.3}$$

This assumption implies that, within group, the population distribution is stable over time.

The standard DD model can be expressed in this way, with

$$h_0(u,t) = u + \delta \cdot t \tag{5.4}$$

and

$$U_i = \alpha + \gamma G_i + V_i,\ V_i \perp (G_i, T_i) \tag{5.5}$$

although, because of the linearity, we can get by with the mean independence assumption $E(V_i|G_i, T_i) = 0$. If the treatment effect is constant across individuals, $\tau = Y_i(1) - Y_i(0)$, then we can write

$$Y_i = \alpha + \beta T_i + \gamma G_i + \tau G_i T_i + V_i, \tag{5.6}$$

where $Y_i = (1 - G_i T_i) Y_i(0) + G_i T_i Y_i(1)$ is the realized outcome. Because $E(V_i|G_i, T_i) = 0$, the parameters in (5.6) can be estimated by OLS.

Athey and Imbens call the extension of the usual DD model the *changes-in-changes* (CIC) model. They show not only how to recover the average treatment effect, but also that the distribution of the counterfactual outcome conditional on intervention, that is

$$D(Y_i(0)|G_i = 1, T_i = 1), \tag{5.7}$$

is identified. The distribution of $D(Y_i(1)|G_i = 1, T_i = 1)$ is identified by the data because $Y_i = Y_i(1)$ when $G_i = T_i = 1$. The extra condition AI use is that the support of the distribution of $D(U_i|G_i = 1)$ is contained in the support of $D(U_i|G_i = 0)$, written as

$$\mathbb{U}_1 \subseteq \mathbb{U}_0. \tag{5.8}$$

Let $F^0_{gt}(y)$ the be cumulative distribution function of $D(Y_i(0)|G_i = g, T_i = t)$ for $g = 1, 2$ and $t = 1, 2$, and let $F_{gt}(y)$ be the cdf for the observed outcome Y_i conditional on $G_i = g$ and

$T_i = t$. By definition, $F_{gt}(y)$ is generally identified from the data, assuming random sampling for each (g,t) pair. AI show that, under (5.1), (5.2), (5.3), and (5.8),

$$F_{11}^{(0)}(y) = F_{10}(F_{00}^{-1}(F_{01}(y))), \tag{5.9}$$

where $F_{00}^{-1}(\cdot)$ is the inverse function of F_{00}^{-1}, which exists under the strict monotonicity assumption. Notice that all of the cdfs appearing on the right hand size of (5.9) are estimable from the data; they are simply the cdfs for the observed outcomes conditional on different (g,t) pairs. Because $F_{11}^{(1)}(y) = F_{11}(y)$, we can estimate the entire distributions of both counterfactuals conditional on intervention, $G_i = T_i = 1$.

The average treatment effect in the CIC framework as

$$\begin{aligned} \tau_{CIC} &= E[Y(1)|G = 1, T = 1] - E[Y(0)|G = 1, T = 1]. \\ &= E(Y_{11}(1)) - E(Y_{11}(0)), \end{aligned} \tag{5.10}$$

where we drop the i subscript, $Y_{gt}(1)$ is a random variable having distribution $D(Y(1)|G = g,t)$, and $Y_{gt}(0)$ is a random variable having distribution $D(Y(0)|G = g,t)$. Under the same assumptions listed above,

$$\tau_{CIC} = E(Y_{11}) - E[F_{01}^{-1}(F_{00}(Y_{10}))] \tag{5.11}$$

where Y_{gt} is a random variable with distribution $D(Y|G = g,t)$. Given random samples from each subgroup, a generally consistent estimator of τ_{CIC} is

$$\hat{\tau}_{CIC} = N_{11}^{-1} \sum_{i=1}^{N_{11}} Y_{11,i} - N_{10}^{-1} \sum_{i=1}^{N_{10}} \hat{F}_{01}^{-1}(\hat{F}_{00}(Y_{10,i})), \tag{5.12}$$

for consistent estimators $\hat{F}_{00}$ and $\hat{F}_{01}$ of the cdfs for the control groups in the initial and later time periods, respectively. Now, $Y_{11,i}$ denotes a random draw on the observed outcome for the $g = 1, t = 1$ group and similarly for $Y_{10,i}$. Athey and Imbens establish weak conditions under which $\hat{\tau}_{CIC}$ is $\sqrt{N}$-asymptotically normal (where, naturally, observations must accumulate within each of the four groups). In the case where the distributions of Y_{10} and Y_{00} are the same, a simple difference in means for the treatment group over time.

The previous approach can be applied either with repeated cross sections or panel data. Athey and Imbens discuss how the assumptions can be relaxed with panel data, and how alternative estimation strategies are available. In particular, if U_{i0} and U_{i1} represent unobservables for unit i in the initial and later time periods, respectively, then (5.3) can be modified to

$$D(U_{i0}|G_i) = D(U_{i1}|G_i), \tag{5.13}$$

which allows for unobservd components structures $U_{it} = C_i + V_{it}$ where V_{it} has the same distribution in each time period.

As discussed by AI, with panel data there are other estimation approaches. As discussed earlier, Altonji and Matzkin (2005) use exchangeability assumptions to identify average partial effects. To illustrate how their approach might applie, suppose the counterfactuals satisfy the ignorability assumption

$$E(Y_{it}(g)|W_{i1},\ldots,W_{iT},U_i) = h_{tg}(U_i), t = 1,\ldots,T, g = 0,1. \tag{5.14}$$

The treatment effect for unit i in period t is $h_{t1}(U_i) - h_{t0}(U_i)$, and the average treatment effect is

$$\tau_t = E[h_{t1}(U_i) - h_{t0}(U_i)], t = 1,\ldots,T. \tag{5.15}$$

Suppose we make the assumption

$$D(U_i|W_{i1},\ldots,W_{iT}) = D(U_i|\bar{W}_i), \tag{5.16}$$

which means that only the intensity of treatment is correlated with heterogeneity. Under (5.14) and (5.16), it can be shown that

$$E(Y_{it}|W_i) = E[E(Y_{it}|W_i,U_i)|W_i] = E(Y_{it}|W_{it},\bar{W}_i). \tag{5.17}$$

The key is that $E(Y_{it}|W_i)$ does not depend on $\{W_{i1},\ldots,W_{iT}\}$ in an unrestricted fashion; it is a function only of $(W_{it},\bar{W}_i)$. If W_{it} are continuous, or take on numerous values, we can use local smoothing methods to estimate $E(y_{it}|W_{it},\bar{W}_i)$. In the treatment effect case, estimation is very simple because $(W_{it},\bar{W}_i)$ can take on only $2T$. The average treatment effect can be estimated as

$$\hat{\tau}_t = N^{-1}\sum_{i=1}^{n}[\hat{\mu}_t^Y(1,\bar{W}_i) - \hat{\mu}_t^Y(0,\bar{W}_i)]. \tag{5.18}$$

If we pool across t (as well as i) and use a linear regression, Y_{it} on $1, d2_t,\ldots,dT_t, W_{it}, \bar{W}_i, t = 1,\ldots,T; i = 1,\ldots,N$, we obtain the usual fixed effects estimate $\hat{\tau}_{FE}$ as the coefficient on W_{it}. Wooldridge (2005) describes other scenarios and compares this strategy to other approaches. As we discussed earlier, a condtional MLE logit can estimate parameters by not generally ATEs, and require conditional independence. Chamberlain's correlated random effects probit models the heterogeneity as $U_i|W_i \sim \text{Normal}(\xi_0 + \xi_1 W_{i1} + \ldots + \xi_T W_{iT}, \eta^2)$, which identifies the ATEs without assuming exchangeability but maintaining a distributional assumption (and functional form for the

response probability).

For the leading case of two time periods, where treatment does not occur in the initial time period for any unit, but does for some units in the second time period, Abadie (2005) provides methods for both repeated cross sections and panel data that use unconfoundedness assumptions on changes over time. Here we describe the panel data approach. Omitting the i subscript, for any unit from the population there are counterfactual outcomes, which we write as $Y_t(w)$, where $t = 0, 1$ are the two time periods and $w = 0, 1$ represent control and treatment. In this setup, interest lies in two parameters, the average treatment effect in the second time period,

$$\tau_{ATE} = E[Y_1(1) - Y_1(0)], \tag{5.19}$$

or the average treatment effect on the treated,

$$\tau_{ATT} = E[Y_1(1) - Y_1(0)|W = 1]. \tag{5.20}$$

Remember, in the current setup, no units are treated in the initial time period, so $W = 1$ means treatment in the second time period.

As in Heckman, Ichimura, Smith, and Todd (1997), Abadie uses unconfoundedness assumptions on changes over time to identify τ_{ATT}, and straightforward extensions serve to identify τ_{ATE}. Given covariates X (that, if observed in the second time period, should not be influenced by the treatment), Abadie assumes

$$E[Y_1(0) - Y_0(0)|X, W] = E[Y_1(0) - Y_0(0)|X], \tag{5.21}$$

so that, conditional on X, treatment status is not related to the gain over time in the absense of treatment. In addition, the overlap assumption,

$$0 < P(W = 1|X) < 1 \tag{5.22}$$

is critical. (Actually, for estimating τ_{ATT}, we only need $P(W = 1|X) < 1$.) Under (5.21) and (5.22), it can be shown that

$$\tau_{ATT} = [P(W = 1)]^{-1} E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{[1 - p(X)]} \right\},$$

where Y_1 is the observed outcome in period 1, Y_0, is the outcome in period 0, and $p(X) = P(W = 1|X)$ is the propensity score. Dehejia and Wahba (1999) derived (5.23) for the cross-sectional case; see also Wooldridge (2002, Chapter 18). All quantities in (5.23) are observed or, in the case of the $p(X)$ and $\rho = P(W = 1)$, can be estimated. As in Hirano, Imbens, and Ridder (2003), a flexible logit model can be used for $p(X)$; the fraction of units

treated would be used for $\hat{\rho}$. Then

$$\hat{\tau}_{ATT} = \hat{\rho}^{-1}N^{-1}\sum_{i=1}^{N}\left\{\frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{[1-\hat{p}(X_i)]}\right\} \tag{5.23}$$

is consistent and $\sqrt{N}$-asymptotically normal. HIR discuss variance estimation. Imbens and Wooldridge (2007) provide a simple adjustment available in the case that $\hat{p}(\cdot)$ is treated as a parametric model.

If we also add

$$E[Y_1(1) - Y_0(1)|X, W] = E[Y_1(0) - Y_0(0)|X], \tag{5.24}$$

so that treatment is mean independent of the gain in the treated state, then

$$\tau_{ATE} = E\left\{\frac{[W - p(X)](Y_1 - Y_0)}{p(X)[1-p(X)]}\right\}, \tag{5.25}$$

which dates back to Horvitz and Thompson (1952); see HIR. Now, to estimate the ATE over the specified population, the full overlap assumption in (5.22) is needed, and

$$\hat{\tau}_{ATE} = N^{-1}\sum_{i=1}^{N}\left\{\frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{\hat{p}(X_i)[1-\hat{p}(X_i)]}\right\}. \tag{5.26}$$

Hirano, Imbens, and Ridder (2003) study this estimator in detail where $\hat{p}(x)$ is a series logit estimator. If we treat this estimator parametrically, a simple adjustment makes valid inference on $\hat{\tau}_{ATE}$ simple. Let $\hat{K}_i$ be the summand in (5.26) less $\hat{\tau}_{ATE}$, and let $\hat{D}_i = h(X_i)[W_i - \Lambda(h(X_i)\hat{\gamma})]$ be the gradient (a row vector) from the logit estimation. Compute the residuals, $\hat{R}_i$ from the OLS regression $\hat{K}_i$ on $\hat{D}_i$, $i = 1,\ldots,N$. Then, a consistent estimator of $Avar\sqrt{N}(\hat{\tau}_{ATE} - \tau_{ATE})$ is just the sample variance of the $\hat{R}_i$. This is never greater than if we ignore the estimation of $p(x)$ and just use the sample variance of the $\hat{K}_i$ themselves.

Under the unconfoundedness assumption, other strategies are available for estimating the ATE and ATT. One possibility is to run the regression

$$\Delta Y_i \text{ on } 1, W_i, \hat{p}(X_i),\ i = 1,\ldots,N,$$

which was studied by Rosenbaum and Rubin (1983) in the cross section case. The coefficient on W_i is the estimated ATE, although it requires some functional form restrictions for consistency. This is much preferred to pooling across t and running the regression Y_{it} on 1, $d1_t$, $d1_t \cdot W_i$, $\hat{p}(X_i)$. This latter regression requires unconfoundedness in the levels, and as dominated by the basic DD estimate from ΔY_i on $1, W_i$: putting in any time-constant function

as a control in a pooled regression is always less general than allowing an unobserved effect and differencing it away.

Regression adjustment is also possible under the previous assumptions. As derived by HIST,

$$E[Y_1(1) - Y_0(1)|X, W = 1] = \{[E(Y_1|X, W = 1) - E(Y_1|X, W = 0)] - [E(Y_0|X, W = 1) - E(Y_0|X, W = 0)]\} \tag{5.27}$$

where, remember, Y_t denotes the observed outcome for $t = 0, 1$. Each of the four conditional expectations on the right hand side is estimable using a random sample on the appropriate subgroup. Call each of these $\hat{\mu}_{wt}(x)$ for $w = 0, 1$ and $t = 0, 1$. Then a consistent estimator of τ_{ATT} is

$$N_1^{-1} \sum_{i=1}^{N} W_i \{[\hat{\mu}_{11}(X_i) - \hat{\mu}_{01}(X_i)] - [\hat{\mu}_{10}(X_i) - \hat{\mu}_{00}(X_i)]\}. \tag{5.28}$$

Computationally, this requires more effort than the weighted estimator proposed by Abdadie. Nevertheless, with flexible parametric functional forms that reflect that nature of Y, implementing (5.28) is not difficult. If Y is binary, then the $\hat{\mu}_{wt}$ should be obtained from binary response models; if Y is nonnegative, perhaps a count variable, then $\mu_{wt}(x) = \exp(x\beta_{wt})$ is attractive, with estimates obtained via Poisson regression (quasi-MLE).

6. Synthetic Control Methods for Comparative Case Studies

In Section 3 we discussed difference-in-differences methods that ignore sampling uncertain in the group/time period means (more generally, regression coefficients). Abadie, Diamond, and Haimmueller (2007), building on the work of Abadie and Gardeazabal (2003), argue that in policy analysis at the aggregate leve, there is no estimation uncertainty: the goal is to determine the effect of a policy on an entire population – say, a state – and the aggregate is measured without error (or very little error). The application in ADH is the effects of California's tobacco control program on state-wide smoking rates.

Of course, one source of uncertainty in any study using data with a time series dimension is the change in outcomes over time, even if the outcomes are aggregates measured without error. Event study methodology is one such example: often, time series regressions for a single entity, such as a state, are used to determine the effect of a policy (speed limit change, tobacco control program, and so on) on an aggregate outcome. But such event studies can suffer because they do not use a control group to account for aggregate effects that have nothing to

do with the specific state policy.

In the context of case control studies, where a time series is available for a particular unit – the treatment group – there are often many potential control groups. For example, in the tobacco control example, each state in the U.S. is a potential control for California (provided a state did not undergo a similar policy). ADH study this setup and emphasize the uncertainty associated with choosing suitable control groups. They point out that, even in the absense of sampling error, surely someone analyzing a state-level policy must nevertheless seal with uncertainty.

The approach of ADH is to allow one to select a synthetic control group out of a collection of possible controls. For example, in the California tobacco control case, ADH identify 38 states that did not implement such programs during the time period in question. Rather than just use a standard fixed effects analysis – which effectively treats each state as being of equal quality as a control group – ADH propose choosing a weighted average of the potential controls. Of course, choosing a suitable control group or groups is often done informally, including matching on pre-treatment predictors. ADH formalize the procedure by optimally choosing weights, and they propose methods of inference.

Consider a simple example, with only two time periods: one before the policy and one after. Let y_{it} be the outcome for unit i in time t, with $i = 1$ the (eventually) treated unit. Suppose there are J possible controls, and index these as $\{2,\ldots,J+1\}$. Let $\mathbf{x}_i$ be observed covariates for unit i that are not (or would not be) affected by the policy; $\mathbf{x}_i$ may contain period $t = 2$ covariates provided they are not affected by the policy. Generally, we can estimate the effect of the policy as

$$y_{12} - \sum_{j=2}^{J+1} w_j y_{j2},$$

where w_j are nonnegative weights that add up to one. The question is: how can we choose the weights – that is, the synthetic control – to obtain the best estimate of the intervention effect? ADH propose choosing the weights so as to minimize the distance between, in this simple case, $(y_{11}, \mathbf{x}_1)$ and $\sum_{j=2}^{J+1} w_j \cdot (y_{j1}, \mathbf{x}_j)$, or some linear combinations of elements of $(y_{11}, \mathbf{x}_1)$ and $(y_{j1}, \mathbf{x}_j)$. The optimal weights – which differ depending on how we define distance – produce the synthetic control whose pre-intervention outcome and predictors of post-intervention outcome are "closest." With more than two time periods, one can use averages of

pre-intervention outcomes, say, or weighted averages that give more weight to more recent pre-intervention outcomes.

ADH propose permutation methods for inference, which require estimating a placebo treatment effect for each region (potential control), using the same synthetic control method as for the region that underwent the intervention. In this way, one can compare the estimated intervention effect using the synthetic control method is substantially larger than the effect estimated from a region chosen at random. The inference is exact even in the case the aggregate outcomes are estimated with error using individual-level data.

References

(To be added.)