What's New in Econometrics?
Lecture 10
Difference-in-Differences Estimation

Jeff Wooldridge
NBER Summer Institute, 2007

1. Review of the Basic Methodology

2. How Should We View Uncertainty in DD Settings?

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

4. Individual-Level Panel Data

5. Semiparametric and Nonparametric Approaches

6. Synthetic Control Methods for Comparative Case Studies

1. Review of the Basic Methodology

- The standard case: 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 (panel data), 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.

- With repeated cross sections, let A be the control group and B the treatment group. Write

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u, \qquad (1)$$

where y is the outcome of interest. The dummy dB captures possible differences between the treatment and control groups prior to the policy change. The dummy $d2$ captures aggregate factors that would cause changes in y even in the absense of a policy change. The coefficient of interest is δ_1.

- The difference-in-differences estimate is

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

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.

- More convincing analysis sometimes available by refining the definition of treatment and control groups. Example: change in state health care policy aimed at elderly. Could use data only on people in the state with the policy change, both before and after the change, with the control group being people 55 to 65 (say) and and the treatment group being people over 65. This DD analysis assumes that the paths of health outcomes for the younger and older groups would not be systematically different in the absense of intervention. Instead, might use the over-65 population from another state as an additional control. Let dE be a dummy equal to one for someone over 65.

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \quad (3)$$

The coefficient of interest is 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:

$$\hat{\delta}_3 = (\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1}) \quad (4)$$

where the *A* subscript means the state not implementing the policy and the *N* subscript represents the non-elderly. This is the *difference-in-difference-in-differences (DDD)* estimate.

- Can add covariates to either the DD or DDD analysis to (hopefully) control for compositional changes.
- Can use multiple time periods and groups.

2. **How Should We View Uncertainty in DD Settings**?

- Standard approach: all uncertainty in inference enters through sampling error in estimating the means of each group/time period combination. Long history in 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. Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Hansen (2007a,b), and Abadie, Diamond, and Hainmueller (2007) argue for additional sources of uncertainty.
- In fact, for the most part, the additional uncertainty is assumed to swamp the sampling error in estimating group/time period means. (See DL

approach in 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.
- Issue: In the standard DD and DDD cases, the policy effect is just identified in the sense that we do not have multiple treatment or control groups assumed to have the same mean responses. So, the DL approach does not allow inference.
- 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 the standard

DD setting: a before and after period, where the policy change was to raise the cap on covered earnings; control group is low earners. Using Kentucky and a total sample size of 5,626, the DD estimate of the policy change is about 19.2% (longer time on workers' compensation) with $t = 2.76$. Using Michigan, with a total sample size of 1,524, the DD estimate is 19.1% with $t = 1.22$. (Adding controls does not help reduce the standard error.) There seems to be plenty of uncertainty in the estimate even with a pretty large sample size. Should we conclude that we really have no usable data for inference?

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 DD case, add a full set of time dummies to the equation. This assumes the policy has the same effect in every year; 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 Meyer (1995) for applications.

- 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

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \qquad (5)$$
$$i = 1, \ldots, M_{gt},$$

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 $\boldsymbol{\beta}$.

- As in cluster sample cases, can write

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \qquad (6)$$

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

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \qquad (7)$$

We can think of (7) as a regression model at the group/time period level.

- As discussed by BDM, a common way to estimate and perform inference in (5) is to ignore v_{gt}, so the individual-level observations are treated as independent. When v_{gt} is present, the resulting inference can be very misleading.
- BDM and Hansen (2007b) allow serial correlation in $\{v_{gt} : t = 1, 2, \ldots, T\}$ but assume independence across g.
- If we view (7) as ultimately of interest, there are simple ways to proceed. 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}. Use OLS on the individual-level data to estimate the δ_{gt}, assuming $E(\mathbf{z}_{igt}' u_{igt}) = \mathbf{0}$ and the group/time period sizes, M_{gt}, are reasonably large.

- 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, proceed as if M_{gt} are large enough to ignore the estimation error in the $\hat{\delta}_{gt}$; instead, the uncertainty comes through v_{gt} in (7). The MD approach from cluster sample notes effectively drops v_{gt} from (7) 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 MD estimator uses only sampling variation in the $\hat{\delta}_{gt}$. Here, we proceed ignoring estimation error, and so act as if (7) is, for $t = 1, \ldots, T, g = 1, \ldots, G$,

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt} \qquad (8)$$

We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (8) 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 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 (8) is inefficient when v_{gt} is serially uncorrelated, proposes feasible GLS. When T is small, estimating

the parameters in $\Omega_g = Var(\mathbf{v}_g)$, where $\mathbf{v}_g$ is the $T \times 1$ error vector for each g, is difficult when group effects have been removed. Estimates based on the FE residuals, $\hat{v}_{gt}$, disappear as $T \to \infty$, but can be substantial. In AR(1) case, $\hat{\rho}$ comes from

$$\hat{v}_{gt} \text{ on } \hat{v}_{g,t-1}, \; t = 2, \ldots, T, g = 1, \ldots, G. \tag{9}$$

- One way to account for bias in $\hat{\rho}$: use fully robust inference. But, as Hansen (2007b) shows, this can be very inefficient relative to his suggestion to bias-adjust the estimator $\hat{\rho}$ and then use the bias-adjusted estimator in feasible GLS. (Hansen covers the general $AR(p)$ model.)
- Hansen shows that 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 infeasible GLS etsimator when $G \to \infty$ and T is fixed.

- 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: they assume $\{\mathbf{x}_{gt} : t = 1, \ldots, T\}$ are strictly exogenous. If we just use OLS, that is, the usual fixed effects estimate – strict exogeneity is not required for consistency as $T \to \infty$. Nothing new that GLS relies on strict exogeneity in serial correlation cases. In intervention analyis, might be concerned if the

policies can switch on and off over time.

- 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). Works pretty well with $G = 50$ and $T = 10$, but get substantial size distortions for $G = 50$ and $T = 20$.
- If the M_{gt} are not large, might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. Can instead aggregate the equations over individuals, giving

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \qquad (10)$$
$$t = 1,..,T, g = 1,\ldots,G.$$

Can estimate this by FE and use fully robust inference because the composite error, $\{r_{gt} \equiv v_{gt} + \bar{u}_{gt}\}$, is weakly dependent.

- The Donald and Lang (2007) approach applies in the current setting by using finite sample analysis applied to the pooled regression (10). 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}.

4. Individual-Level Panel Data

- Let w_{it} be a binary indicator, which is unity if unit i participates in the program at time t. Consider

$$y_{it} = \alpha + \eta d2_t + \tau w_{it} + c_i + u_{it}, \; t = 1, 2, \qquad (11)$$

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 :

$$(y_{i2} - y_{i1}) = \eta + \tau(w_{i2} - w_{i1}) + (u_{i2} - u_{i1}) \qquad (12)$$

or

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i. \qquad (13)$$

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 (13) is consistent.

- If $w_{i1} = 0$ for all i, the OLS estimate is

$$\hat{\tau} = \Delta\bar{y}_{treat} - \Delta\bar{y}_{control}, \qquad (14)$$

which is a DD estimate except that we different the means of the same units over time.

- With many time periods and arbitrary treatment patterns, we can use

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it}, \; t = 1, \ldots, T, \qquad (15)$$

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 (with FE having some advantages for larger T if u_{it} is weakly dependent)

- What if designation is correlated with unit-specific trends? "Correlated random trend" model:

$$y_{it} = c_i + g_i t + \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + u_{it} \qquad (16)$$

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, for $t = 2, \ldots, T$,

$$\Delta y_{it} = g_i + \eta_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\gamma} + \Delta u_{it}. \tag{17}$$

Can difference again or estimate (17) by FE.

- 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. Unconfoundedness, conditional on unobserved heterogeneity, can be stated as

$$E(y_{it0}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it0}|c_{i0}) \tag{18}$$

$$E(y_{it1}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it1}|c_{i1}), \tag{19}$$

where $\mathbf{w}_i = (w_{i1}, \ldots, w_{iT})$ is the time sequence of all treatments. If the gain from treatment only depends on t,

$$E(y_{it1}|c_{i1}) = E(y_{it0}|c_{i0}) + \tau_t \tag{20}$$

and then

$$E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it0}|c_{i0}) + \tau_t w_{it}. \tag{21}$$

If we assume

$$E(y_{it0}|c_{i0}) = \alpha_{t0} + c_{i0}, \tag{22}$$

then

$$E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) = \alpha_{t0} + c_{i0} + \tau_t w_{it}, \tag{23}$$

an estimating equation that leads to FE or FD (often with $\tau_t = \tau$).

- If add strictly exogenous covariates, and assume linearity of conditional expectations, and allow the gain from treatment to depend on $\mathbf{x}_{it}$ and an additive unobserved effect a_i, get

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, c_{i0}, a_i) = \alpha_{t0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 \\ + w_{it}(\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i w_{it}, \end{aligned} \tag{24}$$

a correlated random coefficient model because the

coefficient on w_{it} is $(\tau_t + a_i)$. Can eliminate a_i (and c_{i0}). Or, with $\tau_t = \tau$, can "estimate" the τ_i and then use

$$\hat{\tau} = N^{-1} \sum_{i=1}^{N} \hat{\tau}_i. \tag{25}$$

See Wooldridge (2002, Section 11.2) for standard error, or bootstrapping.

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

$$Y_i(0) = h_0(U_i, T_i), \tag{26}$$

where T_i is the time period and

$$h_0(u,t) \text{ strictly increasing in } u \text{ for } t = 0, 1 \tag{27}$$

The random variable U_i represents all unobservable characteristics of individual i. Equation (26) 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 distribution of U_i is allowed to vary across groups, but not over time within groups, so that

$$D(U_i|T_i, G_i) = D(U_i|G_i). \tag{28}$$

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

$$h_0(u,t) = u + \delta \cdot t \tag{29}$$

and

$$U_i = \alpha + \gamma G_i + V_i,\ V_i \perp (G_i, T_i) \tag{30}$$

although, because of the linearity, we can get by with the mean independence assumption $E(V_i|G_i, T_i) = 0$. With constant treatment effect,

$$Y_i = \alpha + \beta T_i + \gamma G_i + \tau G_i T_i + V_i, \tag{31}$$

Because $E(V_i|G_i, T_i) = 0$, the parameters in (31) can be estimated by OLS (usual DD analysis).

- Athey and Imbens call the extension of the usual DD model the *changes-in-changes* (CIC) model. Can recover

$$D(Y_i(0)|G_i = 1, T_i = 1), \tag{32}$$

under their assumptions (with an extra support condition). In fact, if $F^0_{gt}(y)$ 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 $F_{gt}(y)$ is the cdf for the observed outcome Y_i conditional on $G_i = g$ and $T_i = t$, then

$$F_{11}^{(0)}(y) = F_{10}(F_{00}^{-1}(F_{01}(y))), \tag{33}$$

where $F_{00}^{-1}(\cdot)$ is the inverse function of $F_{00}(\cdot)$, which exists under the strict monotonicity assumption. Because $F_{11}^{(1)}(y) = F_{11}(y)$, we can estimate the entire distributions of both counterfactuals conditional on intervention, $G_i = T_i = 1$.

- Can apply to repeated cross sections or panel data. Of course, can also identify the average treatment effect

$$\tau_{CIC} = E(Y_{11}(1)) - E(Y_{11}(0)). \tag{34}$$

In particular,

$$\tau_{CIC} = E(Y_{11}) - E[F_{01}^{-1}(F_{00}(Y_{10}))]. \quad (35)$$

- Other approaches with panel data: Altonji and Matzkin (2005) under exchaneability in $D(U_i|W_{i1}, \ldots, W_{iT})$.
- Heckman, Ichimura, Smith, and Todd (1997) and Abadie (2005). Consider basic setup with two time periods, no treated units in first time period. Without an i subscript, $Y_t(w)$ is the counterfactual outcome for treatment level w, $w = 0, 1$, at time t. Parameter: the average treatment effect on the treated,

$$\tau_{ATT} = E[Y_1(1) - Y_1(0)|W = 1]. \quad (36)$$

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

- Key unconfoundedness assumption:

$$E[Y_1(0) - Y_0(0)|X, W] = E[Y_1(0) - Y_0(0)|X] \qquad (37)$$

Also need

$$P(W = 1|X) < 1 \qquad (38)$$

is critical. Under (37) and (38),

$$\tau_{ATT} = E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{[1 - p(X)]} \right\} / P(W = 1), \qquad (39)$$

where Y_t, $t = 0, 1$ are the observed outcomes (for the same unit) and $p(X) = P(W = 1|X)$ is the propensity score. Dehejia and Wahba (1999) derived (39) for the cross-sectional case. All quantities 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

$$\hat{\tau}_{ATT} = \hat{\rho}^{-1} N^{-1} \sum_{i=1}^{N} \left\{ \frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{[1 - \hat{p}(X_i)]} \right\}. \qquad (40)$$

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.

- Similar approach works for τ_{ATE}.
- Regression version:

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

The coefficient on W_i is the estimated ATE.

Requires some functional form restrictions.

Certainly preferred to 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 $1, W_i$

- Regression adjustment can also be used, as in HIST (1997).

6. Synthetic Control Methods for Comparative Case Studies

- Abadie, Diamond, and Hainmueller (2007) argue that in policy analysis at the aggregate leve, there is little or no estimation uncertainty: the goal is to determine the effect of a policy on an entire population, and the aggregate is measured without error (or very little error). Application: California's tobacco control program on state-wide smoking rates.
- ADH focus on the uncertainty with choosing a

suitable control for California among other states (that did not implement comparable policies over the same period).

- ADH suggest using many potential control groups (38 states or so) to create a single synthetic control group.
- 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 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

$$y_{12} - \sum_{j=2}^{J+1} w_j y_{j2}, \tag{41}$$

where w_j are nonnegative weights that add up to one. How to choose the weights to best estimate the intervention effect?

- ADH propose choosing the weights so as to minimize the distance between $(y_{11}, \mathbf{x}_1)$ and $\sum_{j=2}^{J+1} w_j \cdot (y_{j1}, \mathbf{x}_j)$, say. That is, functions of the pre-treatment outcomes and the predictors of post-treatment outcomes.
- ADH propose permutation methods for inference, which require estimating a placebo treatment effect for each region, using the same synthetic control method as for the region that underwent the intervention.