

**A Course in Applied Econometrics**  
**Lecture 11: Difference-in-Differences Estimation**

Jeff Wooldridge  
 IRP Lectures, UW Madison, August 2008

1. The Basic Methodology
2. How Should We View Uncertainty in DD Settings?
3. Multiple Groups and Time Periods
4. Individual-Level Panel Data
5. Semiparametric and Nonparametric Approaches

1

- $d2$  captures aggregate factors that would cause changes in  $y$  over time even in the absence of a policy change. The coefficient of interest is  $\delta_1$ .
- The difference-in-differences (DD) estimate is

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

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

3

**1. The Basic Methodology**

- 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. Structure can apply to repeated cross sections or panel data.
- With repeated cross sections, let  $A$  be the control group and  $B$  the treatment group. Write

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

where  $y$  is the outcome of interest.  $dB$  captures possible differences between the treatment and control groups prior to the policy change.

2

- Can refine 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 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 absence of intervention. Instead, use the over-65 population from another state as an additional control. Let  $dE$  be a dummy equal to one for someone over 65:

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

4

- The OLS estimate  $\hat{\delta}_3$  is

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

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 suggested 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, in the “new” view, the additional uncertainty is often assumed to swamp the sampling error in estimating group/time period means.

- 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, for example, the DL approach does not allow inference in such cases.
- 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 before-after setting.

- 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, nor does it change the point estimates.) 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. Multiple Groups and Time Periods

• With many time periods and groups, in BDM (2004) and Hansen (2007b) is useful. At the individual level,

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

$$i = 1, \dots, M_{gt},$$

where  $i$  indexes individual,  $g$  indexes group, and  $t$  indexes time. Full set of time effects,  $\lambda_t$ , full set of group effects,  $\alpha_g$ , group/time period covariates (policy variabls),  $\mathbf{x}_{gt}$ , individual-specific covariates,  $\mathbf{z}_{igt}$ , unobserved group/time effects,  $v_{gt}$ , and individual-specific errors,  $u_{igt}$ . Interested in  $\boldsymbol{\beta}$ .

• As in cluster sample cases, can write

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

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

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

Think of (7) as a 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  $\{v_{gt} : t = 1, 2, \dots, T\}$  but assume independence across  $g$ .

• If we view (7) as ultimately of interest, there are simple ways to proceed. We observe  $\mathbf{x}_{gt}$ ,  $\lambda_t$  is handled with year dummies, and  $\alpha_g$  just represents group dummies. The problem, then, is that we do not observe  $\delta_{gt}$ . Use OLS on the individual-level data to estimate the  $\delta_{gt}$ , 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 minimum distance 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  $\delta_{gt}$ . 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  $t = 1, \dots, T, g = 1, \dots, G$ ,

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

- 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  $\{v_{gt} : t = 1, 2, \dots, T\}$  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  $\Omega = Var(\mathbf{v}_g)$ , where  $\mathbf{v}_g$  is the  $T \times 1$  error vector for each  $g$ , is difficult when group effects have been removed. Bias in estimates based on the FE residuals,  $\hat{v}_{gt}$ , disappears as  $T \rightarrow \infty$ , but can be substantial even for moderate  $T$ . In AR(1) case,  $\hat{\rho}$  comes from

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

- 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 estimator when  $G \rightarrow \infty$  and  $T$  is fixed.

- Even when  $G$  and  $T$  are both large, so that the unadjusted AR coefficients also deliver asymptotic efficiency, the bias-adjusted estimates deliver higher-order improvements in the asymptotic distribution.
- One limitation of Hansen’s results: they assume  $\{x_{gt} : t = 1, \dots, T\}$  are strictly exogenous. If we just use OLS, that is, the usual fixed effects estimate – strict exogeneity is not required for consistency as  $T \rightarrow \infty$ . Of course, GLS approaches to serial correlation generally rely on strict exogeneity. In intervention analysis, might be concerned if the policies can switch on and off over time.

- With large  $G$  and small  $T$ , can estimate an unstricted variance matrix  $\Omega$  ( $T \times T$ ) and proceed with GLS, as studied 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  $\hat{\delta}_{gt}$ . Instead, aggregate over individuals:

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \quad (10)$$

$$t = 1, \dots, T, g = 1, \dots, G.$$

Can estimate this by FE and use fully robust inference (to account for time series dependence) 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

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

where  $d2_t = 1$  if  $t = 2$  and zero otherwise,  $c_i$  is an observed effect  $\tau$  is the treatment effect. Remove  $c_i$  by first differencing:

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

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

If  $E(\Delta w_i \Delta u_i) = 0$ , OLS applied to (13) is consistent.

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

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

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

- It is *not* more general to regress  $y_{i2}$  on  $1, w_{i2}, y_{i1}$ ,  $i = 1, \dots, N$ , even though this appears to free up the coefficient on  $y_{i1}$ . Why? Under (11) with  $w_{i1} = 0$  we can write

$$y_{i2} = \eta + \tau w_{i2} + y_{i1} + (u_{i2} - u_{i1}). \quad (15)$$

Now, if  $E(u_{i2} | w_{i2}, c_i, u_{i1}) = 0$  then  $u_{i2}$  is uncorrelated with  $y_{i1}$ , and  $y_{i1}$  and  $u_{i1}$  are correlated. So  $y_{i1}$  is correlated with  $u_{i2} - u_{i1} = \Delta u_i$ .

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

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it}, \quad t = 1, \dots, T, \quad (16)$$

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 between  $w_{it}$  and  $u_{it}$  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:

$$y_{it} = c_i + g_i t + \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + u_{it} \quad (17)$$

where  $g_i$  is the trend for unit  $i$ . A general analysis allows arbitrary correlation between  $(c_i, g_i)$  and  $w_{it}$ , which requires at least  $T \geq 3$ . If we first difference, we get, for  $t = 2, \dots, T$ ,

$$\Delta y_{it} = g_i + \eta_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\gamma} + \Delta u_{it}. \quad (18)$$

Can difference again or estimate (18) by FE.

- Can derive panel data approaches using the counterfactual 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

$$E[y_{it}(0)|\mathbf{w}_i, \mathbf{c}_i] = E[y_{it}(0)|\mathbf{c}_i] \quad (19)$$

$$E[y_{it}(1)|\mathbf{w}_i, \mathbf{c}_i] = E[y_{it}(1)|\mathbf{c}_i], \quad (20)$$

where  $\mathbf{w}_i = (w_{i1}, \dots, w_{iT})$  is the time sequence of all treatments.

Suppose the gain from treatment only depends on  $t$ ,

$$E[y_{it}(1)|\mathbf{c}_i] = E[y_{it}(0)|\mathbf{c}_i] + \tau_t. \quad (21)$$

Then

$$E(y_{it}|\mathbf{w}_i, \mathbf{c}_i) = E[y_{it}(0)|\mathbf{c}_i] + \tau_t w_{it} \quad (22)$$

where  $y_{i1} = (1 - w_{i1})y_{i1}(0) + w_{i1}y_{i1}(1)$ . If we assume

$$E[y_{it}(0)|\mathbf{c}_i] = \alpha_{i0} + c_{i0}, \quad (23)$$

then

$$E(y_{it}|\mathbf{w}_i, \mathbf{c}_i) = \alpha_{i0} + c_{i0} + \tau_t w_{it}, \quad (24)$$

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

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

$$E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, \mathbf{c}_i) = \alpha_{i0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 + w_{it}(\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i w_{it}, \quad (25)$$

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  $\tau_i = \tau + a_i$  and then use

$$\hat{\tau} = N^{-1} \sum_{i=1}^N \hat{\tau}_i. \quad (26)$$

Can bootstrap the standard error.

25

- Can also get to a random trend model, where  $g_{it}$  is added to (25). Then, can difference followed by a second difference or fixed effects estimation on the first differences. With  $\tau_t = \tau$ ,

$$\Delta y_{it} = \psi_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it} \boldsymbol{\gamma}_0 + [\Delta w_{it}(\mathbf{x}_{it} - \boldsymbol{\xi}_t)] \boldsymbol{\delta} + a_i \Delta w_{it} + g_i + \Delta u_{it}. \quad (27)$$

- Might ignore  $a_i \Delta w_{it}$  or, again, estimate  $\tau_i = \tau + a_i$  and form (26).

26

## 5. Semiparametric and Nonparametric Approaches

- Consider the setup of Heckman, Ichimura, Smith, and Todd (1997) and Abadie (2005), with two time periods. No units treated 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,

$$\tau_{att} = E[Y_1(1) - Y_1(0)|W = 1]. \quad (28)$$

$W = 1$  means treatment in the second time period.

27

- Along with  $Y_0(1) = Y_0(0)$  (no counterfactual in time period zero), key unconfoundedness assumption:

$$E[Y_1(0) - Y_0(0)|X, W] = E[Y_1(0) - Y_0(0)|X] \quad (29)$$

Also need

$$P(W = 1|X) < 1 \quad (30)$$

is critical. Under (29) and (30),

$$\tau_{att} = E \left\{ \frac{[W - p(X)](Y_1 - Y_0)}{[1 - p(X)]} \right\} / P(W = 1), \quad (31)$$

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.

28

- 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

$$\hat{\tau}_{ate} = \hat{\rho}^{-1} N^{-1} \sum_{i=1}^N \left\{ \frac{[W_i - \hat{p}(X_i)] \Delta Y_i}{[1 - \hat{p}(X_i)]} \right\}. \quad (32)$$

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 add

$$E[Y_1(1) - Y_0(1)|X, W] = E[Y_1(1) - Y_0(1)|X], \quad (33)$$

a similar approach works for  $\tau_{ate}$ .

$$\hat{\tau}_{ate} = N^{-1} \sum_{i=1}^N \left\{ \frac{[W_i - \hat{p}(X_i)] \Delta Y_i}{\hat{p}(X_i)[1 - \hat{p}(X_i)]} \right\} \quad (34)$$

- A regression version:

$$\Delta Y_i \text{ on } 1, W_i, \hat{p}(X_i), [\hat{p}(X_i) - \hat{\rho}] \cdot W_i, i = 1, \dots, N. \quad (35)$$

The coefficient on  $W_i$  is the estimated ATE. Requires some functional form restrictions. Preferred to regression in levels.