DIFFERENCE-IN-DIFFERENCES ESTIMATION

Jeff Wooldridge
Michigan State University

LABOUR Lectures, EIEF
October 18-19, 2011

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

1. The Basic Methodology

- In the basic setting, 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

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u, \tag{1}$$

where y is the outcome of interest.

- dB captures possible differences between the treatment and control groups prior to the policy change. $d2$ captures aggregate factors that would cause changes in y over time even in the absense of a policy change. The coefficient of interest is δ_1.
- The difference-in-differences (DD) estimate is

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

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.

- 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 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, use the same two groups from another (“untreated”) state as an additional control. Let *dE* be a dummy equal to one for someone over 65 and *dB* be the dummy for living in the “treatment” state:

$$\begin{aligned} y = {} & \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 \\ & + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \end{aligned} \tag{3}$$

where δ_3 is the average treatment effect.

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

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

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. Even if the intervention is independent of observed covariates, adding those covariates may improve precision of the DD or DDD estimate.

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, QJE), Donald and Lang (2007, REStat), Hansen (2007a,b, JE), and Abadie, Diamond, and Hainmueller (2010, JASA) 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.
- ADH show how to construct a synthetic control group (for California) using pre-treatment characteristics of other states (that were not subject to cigarette smoking restrictions) to choose the “best” weighted average of states in constructing the control.

- 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 Donald and Lang 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.

```
. use injury

. reg ldurat afchnge highearn afhigh if ky, robust

Linear regression                                      Number of obs =    5626
                                                       F(  3,  5622) =   38.97
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0207
                                                       Root MSE      =  1.2692

------------------------------------------------------------------------------
             |               Robust
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     afchnge |   .0076573   .0440344     0.17   0.862     -.078667    .0939817
    highearn |   .2564785   .0473887     5.41   0.000     .1635785    .3493786
      afhigh |   .1906012    .068982     2.76   0.006     .0553699    .3258325
       _cons |   1.125615   .0296226    38.00   0.000     1.067544    1.183687
------------------------------------------------------------------------------
```

```
. reg ldurat afchnge highearn afhigh if mi, robust

Linear regression                                      Number of obs =    1524
                                                       F(  3,  1520) =    5.65
                                                       Prob > F      =  0.0008
                                                       R-squared     =  0.0118
                                                       Root MSE      =  1.3765

------------------------------------------------------------------------------
             |               Robust
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     afchnge |   .0973808   .0832583     1.17   0.242    -.0659325    .2606941
    highearn |   .1691388   .1070975     1.58   0.114    -.0409358    .3792133
      afhigh |   .1919906   .1579768     1.22   0.224     -.117885    .5018662
       _cons |   1.412737   .0556012    25.41   0.000     1.303674      1.5218
------------------------------------------------------------------------------
```

3. Multiple Groups and Time Periods

- With many time periods and groups, setup in Bertrand, Duflo, and Mullainathan (2004) (BDM) and Hansen (2007a) is useful. At the individual level,

$$\begin{aligned} y_{igt} &= \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \\ i &= 1,\ldots,M_{gt}, \end{aligned} \tag{5}$$

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

- We can write a model at the individual level as

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

where intercepts and slopes are allowed to differ across all (g,t) pairs. Then, think of δ_{gt} as

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

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 the individual-level equation

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

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.
- We cannot replace $\lambda_t + \alpha_g$ a full set of group/time interactions because that would eliminate $\mathbf{x}_{gt}$.

- If we view $\boldsymbol{\beta}$ in $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}$ as ultimately of interest – which is usually the case because $\mathbf{x}_{gt}$ contains the aggregate policy variables – 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}.
- But we can use OLS on the individual-level data to estimate the δ_{gt} in

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

assuming $E(\mathbf{z}_{igt}'u_{igt}) = \mathbf{0}$ and the group/time period sample 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 across groups and/or time can be used to impose the restrictions.
- However we obtain the $\hat{\delta}_{gt}$, 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 $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}$.
- A minimum distance (MD) approach (later) effectively drops v_{gt} 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, proceed ignoring estimation error, and act *as if*

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

- 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 $\mathbf{x}_{gt}$ – then the OLS estimator has satisfying large-sample properties as G and T both increase, provided $\{v_{gt} : t = 1, 2, \ldots, T\}$ is a weakly dependent time series for all g.
- Simulations in BDM and Hansen (2007a) indicate cluster-robust inference works 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

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

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

- Even when G and T are both large, so that the unadjusted AR coefficients also deliver asymptotic efficiency, the bias-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 $T \to \infty$.
- Of course, GLS approaches to serial correlation generally rely on strict exogeneity. In intervention analyis, 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 Ω ($T \times T$) 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:

$$\begin{aligned} \bar{y}_{gt} &= \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \\ t &= 1,..,T, g = 1,\ldots,G. \end{aligned} \tag{10}$$

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.

4. Estimation with a Small Number of Groups

- Suppose we have only a small number of groups, G, but where the number of units per group is fairly large. This setup – first made popular by Moulton (1990) in economics – has been recently studied by Donald and Lang (2007) (DL).

- DL treat the problem as a small number of random draws from a large number of groups (because they assume independence). This may not be the most realistic way to view the data.
- Simplest case: A single regressor that varies only by group:

$$\begin{aligned} y_{gm} &= \alpha + \beta x_g + c_g + u_{gm} \\ &= \delta_g + \beta x_g + u_{gm}. \end{aligned}$$

In second equation, common slope, β, but intercept, δ_g, that varies across g.

- DL focus on first equation, where c_g is assumed to be independent of x_g with zero mean.

- Note: Because c_g is assumed independent of x_g, the DL criticism of standard methods for standard DD analysis is not one of endogeneity. It is one of inference.
- DL highlight the problems of applying standard inference leaving c_g as part of the error term, $v_{gm} = c_g + u_{gm}$.

- Pooled OLS inference applied to

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm}$$

can be badly biased because it ignores the cluster correlation. Hansen's results do not apply. (And we cannot use fixed effects estimation here.)

- DL propose studying the regression in averages:

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G.$$

- Add some strong assumptions: $M_g = M$ for all g, $c_g | x_g \sim Normal(0, \sigma_c^2)$ and $u_{gm} | x_g, c_g \sim Normal(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim Normal(0, \sigma_c^2 + \sigma_u^2/M)$. Then the model in averages satisfies the classical linear model assumptions (we assume independent sampling across g).

- So, we can just use the "between" regression

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G.$$

- The estimates of α and β are identical to pooled OLS across g and m

when $M_g = M$ for all g.

- Conditional on the x_g, $\hat{\beta}$ inherits its distribution from $\{\bar{v}_g : g = 1,\ldots,G\}$, the within-group averages of the composite errors.

- We can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$.

- If G is small, the requirements for a significant t statistic using the t_{G-2} distribution are much more stringent then if we use the $t_{M_1+M_2+\ldots+M_G-2}$ distribution – which is what we would be doing if we use the usual pooled OLS statistics.

- Using the averages in an OLS regression is *not* the same as using cluster-robust standard errors for pooled OLS. Those are not justified and, anyway, we would use the wrong df in the t distribution.
- We can apply the DL method without normality of the u_{gm} if the group sizes are large because $Var(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$ so that $\bar{u}_g$ is a negligible part of $\bar{v}_g$. But we still need to assume c_g is normally distributed.
- If $\mathbf{z}_{gm}$ appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \bar{\mathbf{z}}_g\boldsymbol{\gamma} + \bar{v}_g, g = 1,\ldots,G,$$

provided $G > K + L + 1$.

- Inference can be carried out using the $t_{G-K-L-1}$ distribution.

- Regressions on averages are reasonably common, at least as a check on results using disaggregated data, but usually with larger G then just a handful.

- If $G = 2$ in the DL setting, we cannot do inference (there are zero degrees of freedom).

- Suppose x_g is binary, indicating treatment and control ($g = 2$ is the treatment, $g = 1$ is the control). The DL estimate of β is the usual one: $\hat{\beta} = \bar{y}_2 - \bar{y}_1$. But we cannot compute a standard error for $\hat{\beta}$.

- So according the the DL framework the traditional comparison-of-means approach to policy analysis cannot be used. Should we just give up when $G = 2$?
- In a sense the problem is an artifact of saying there are three group-level parameters. If we write

$$y_{gm} = \delta_g + \beta x_g + u_{gm}$$

where $x_1 = 0$ and $x_2 = 1$, then $E(y_{1m}) = \delta_1$ and $E(y_{2m}) = \delta_2 + \beta$. There are only two means but three parameters.

- The usual approach simply defines $\mu_1 = E(y_{1m})$, $\mu_2 = E(y_{2m})$, and then uses random samples from each group to estimate the means. Any "cluster effect" is contained in the means.
- Same is true for the DD framework with $G = 4$ (control and treatment, before and after).
- Remember, in the DL framework, the cluster effect is independent of x_g, so the DL criticism is not about systematic bias.

- Applies to simple difference-in-differences settings. Let $y_{gm} = w_{gm2} - w_{gm1}$ be the change in a variable w from period one to two for . So, we have a before period and an after period, and suppose a treated group ($x_2 = 1$) and a control group ($x_1 = 0$). So $G = 2$.
- The estimator of β is the DD estimator:

$$\hat{\beta} = \overline{\Delta w}_2 - \overline{\Delta w}_1$$

where $\overline{\Delta w}_2$ is the average of changes for the treament group and $\overline{\Delta w}_1$ is the average change for the control.

- Card and Krueger (1994) minimum wage example: $G = 2$ so, according to DL, cannot put a confidence interval around the estimated change in employment.
- If we go back to

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm}$$

when $x_1 = 0, x_2 = 1$, one can argue that c_g should just be part of the estimated mean for group g. It is assumed assignment is exogenous.

- In the traditional view, we are estimating $\mu_1 = \alpha + c_1$ and $\mu_2 = \alpha + \beta + c_2$ and so the estimated policy effect is $\beta + (c_2 - c_1)$.

- The same DL criticism arises in the standard difference-in-differences setting with two groups and two time periods. From the traditional perspective, we have four means to estimate: μ_{A1}, μ_{A2}, μ_{B1}, μ_{B2}. From the DL perspective, we instead have

$$y_{gtm} = \mu_{gt} + c_{gt} + u_{gtm},\, m = 1,\ldots,M_{gt}; t = 1,2; g = A,B;$$

the presence of c_{gt} makes it impossible to do inference.

- Even when DL approach applies, should we use it? Suppose $G = 4$ with two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$), and we impose the same means within control and treatment. DL involves the OLS regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$; inference is based on the t_2 distribution. Can show

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2,$$

which shows $\hat{\beta}$ is approximately normal (for most underlying population distributions) even with moderate group sizes M_g.

- In effect, the DL approach rejects usual inference based on means from large samples because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. Why not allow heterogeneous means?
- Could just define the treatment effect as, say,

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2,$$

and then plug in the unbiased, consistent, asymptotically normal estimators of the μ_g under random sampling within each g.

● The expression $\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2$ hints at a different way to view the small G, large M_g setup. We estimated two parameters, α and β, given four moments that we can estimate with the data.

● The OLS estimates of α and β can be interpreted as minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. In the general MD notation, $\boldsymbol{\pi} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and

$$\mathbf{h}(\boldsymbol{\theta}) = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta \\ \alpha + \beta \end{pmatrix}.$$

- Can show that if we use the 4×4 identity matrix as the weight matrix, we get the DL estimates, $\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2$ and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$.

- In the general setting, with large group sizes M_g, and whether or not G is especially large, we can put the problem into an MD framework, as done by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group.
- Idea is to think of a set of G linear models at the invididual (m) level with group-specific intercepts (and possibly slopes).

- For each group g, write

$$\begin{aligned} y_{gm} &= \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm} \\ E(u_{gm}) &= 0, E(\mathbf{z}_{gm}' u_{gm}) = \mathbf{0}. \end{aligned}$$

Within-group OLS estimators of δ_g and $\boldsymbol{\gamma}_g$ are $\sqrt{M_g}$ -asymptotically normal under random sampling within group.

- The presence of aggregate features $\mathbf{x}_g$ can be viewed as putting restrictions on the intercepts:

$$\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta}, g = 1,\ldots,G.$$

- With K attributes ($\mathbf{x}_g$ is $1 \times K$) we must have $G \geq K + 1$ to determine α and $\boldsymbol{\beta}$.
- In the first stage, obtain $\hat{\delta}_g$, either by group-specific regressions or pooling to impose some common slope elements in $\boldsymbol{\gamma}_g$.
- If we impose some restrictions on the $\boldsymbol{\gamma}_g$, such as $\boldsymbol{\gamma}_g = \boldsymbol{\gamma}$ for all g, the $\hat{\delta}_g$ are (asymptotically) correlated.

- Let $\hat{\mathbf{V}}$ be the $G \times G$ estimated (asymptotic) variance of the $G \times 1$ vector $\hat{\boldsymbol{\delta}}$. Let $\mathbf{X}$ be the $G \times (K+1)$ matrix with rows $(1, \mathbf{x}_g)$. The MD estimator is

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\delta}}$$

The asymptotics are as each group size gets large, and $\hat{\boldsymbol{\theta}}$ has an asymptotic normal distribution; its estimated asymptotic variance is $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$.

- Estimator looks like “GLS,” but inference is with G (number of rows in $\hat{\boldsymbol{\delta}}$ and $\mathbf{X}$) fixed and M_g growing.

- When separate group regressions are used for each g, the $\hat{\delta}_g$ are independent and $\hat{\mathbf{V}}$ is diagonal, and $\hat{\boldsymbol{\theta}}$ looks like a weighted least squares estimator. That is, treat the $\{(\hat{\delta}_g, \mathbf{x}_g) : g = 1, \ldots, G\}$ as the data and use WLS of $\hat{\delta}_g$ on $1, \mathbf{x}_g$ using weights $1/[se(\hat{\delta}_g)]^2$.
- Can test the $G - (K + 1)$ overidentification restrictions using the SSR from the "weighted least squares" as approximately χ^2_{G-K-1}.

- What happens if the overidentifying restrictions reject?

(1) Can search for more features to include in $\mathbf{x}_g$. If $G = K + 1$, no restrictions to test.

(2) Think about whether a rejection is important. In the program evaluation applications, rejection generally occurs if group means within the control groups or within the treatment groups differ. For example, in the $G = 4$ case with $x_1 = x_2 = 0$ and $x_3 = x_4 = 1$, the test will reject if $\mu_1 \neq \mu_2$ or $\mu_3 \neq \mu_4$. But why should we care? We might want to allow heterogeneous policy effects and define the parameter of interest as

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2.$$

(3) Apply the DL approach on the group-specific intercepts. That is, write

$$\delta_g = \alpha + \mathbf{x}_g \boldsymbol{\beta} + c_g, g = 1, \ldots, G$$

and assume that this equation satisfies the classical linear model assumptions.

- With large group sizes, we can act *as if*

$$\hat{\delta}_g = \alpha + \mathbf{x}_g \boldsymbol{\beta} + c_g, g = 1, \ldots, G$$

because $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$ and we can ignore the $O_p(M_g^{-1/2})$ part. But we must assume c_g is homoskedastic, normally distributed, and independent of $\mathbf{x}_g$.

- Note how we only need $G > K + 1$ because the $\mathbf{z}_{gm}$ have been accounted for in the first stage in obtaining the $\hat{\delta}_g$. But we are ignoring the estimation error in the $\hat{\delta}_g$.

5. 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

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

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

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

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i. \tag{13}$$

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

$$\hat{\tau}_{FD} = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}, \tag{14}$$

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

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

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

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 $u_{i2} - u_{i1} = \Delta u_i$.

- In fact, if we add the standard no serial correlation assumption, $E(u_{i1}u_{i2}|w_{i2},c_i) = 0$, and write the linear projection $w_{i2} = \pi_0 + \pi_1 y_{i1} + r_{i2}$, then can show that

$$plim(\hat{\tau}_{LDV}) = \tau + \pi_1(\sigma^2_{u_1}/\sigma^2_{r_2})$$

where

$$\pi_1 = Cov(c_i, w_{i2})/(\sigma^2_c + \sigma^2_{u_1}).$$

- For example, if w_{i2} indicates a job training program and less productive workers are more likely to participate ($\pi_1 < 0$), then the regression y_{i2} (or Δy_{i2}) on 1, w_{i2}, y_{i1} underestimates the effect.

- If more productive workers participate, regressing y_{i2} (or Δy_{i2}) on 1, w_{i2}, y_{i1} overestimates the effect of job training.
- Following Angrist and Pischke (2009, MHE), suppose we use the FD estimator when, in fact, unconfoundedness of treatment holds conditional on y_{i1} (and the treatment effect is constant). Then we can write

$$\begin{aligned} y_{i2} &= \gamma + \tau w_{i2} + \psi y_{i1} + e_{i2} \\ E(e_{i2}) &= 0,\ Cov(w_{i2}, e_{i2}) = Cov(y_{i1}, e_{i2}) = 0. \end{aligned}$$

- Write the equation as

$$\begin{aligned} \Delta y_{i2} &= \gamma + \tau w_{i2} + (\psi - 1) y_{i1} + e_{i2} \\ &\equiv \gamma + \tau w_{i2} + \lambda y_{i1} + e_{i2} \end{aligned}$$

Then, of course, the FD estimator generally suffers from omitted variable bias if $\psi \neq 1$. We have

$$plim(\hat{\tau}_{FD}) = \tau + \lambda \frac{Cov(w_{i2}, y_{i1})}{Var(w_{i2})}$$

- If $\lambda < 0$ ($\psi < 1$) and $Cov(w_{i2}, y_{i1}) < 0$ – workers observed with low first-period earnings are more likely to participate – the $plim(\hat{\tau}_{FD}) > \tau$, and so FD overestimates the effect.

- We might expect ψ to be close to unity for processes such as earnings, which tend to be persistent. (ψ measures persistence without conditioning on unobserved heterogeneity.)
- As an algebraic fact, if $\hat{\lambda} < 0$ (as it usually will be even if $\psi = 1$) and w_{i2} and y_{i1} are negatively correlated in the sample, $\hat{\tau}_{FD} > \hat{\tau}_{LDV}$. But this does not tell us which estimator is consistent.
- If either $\hat{\lambda}$ is close to zero or w_{i2} and y_{i1} are weakly correlated, adding y_{i1} can have a small effect on the estimate of τ.

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

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it},\ t = 1, \ldots, T, \tag{16}$$

which accounts for aggregate time effects and allows for controls, $\mathbf{x}_{it}$.

- Estimation by fixed effects or first differencing 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 advantages for larger T if u_{it} is weakly dependent).

- What if designation is correlated with unit-specific trends?

“Correlated Random Trend” model:

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

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 $t = 2, \ldots, T$,

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

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

$$E[y_{it}(0)|\mathbf{w}_i,\mathbf{c}_i] = E[y_{it}(0)|\mathbf{c}_i] \tag{19}$$
$$E[y_{it}(1)|\mathbf{w}_i,\mathbf{c}_i] = E[y_{it}(1)|\mathbf{c}_i], \tag{20}$$

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

Suppose the gain from treatment only depends on t,

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

Then

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

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

$$E[y_{it}(0)|\mathbf{c}_i] = \alpha_{t0} + c_{i0}, \tag{23}$$

then

$$E(y_{it}|\mathbf{w}_i, \mathbf{c}_i) = \alpha_{t0} + c_{i0} + \tau_t w_{it}, \tag{24}$$

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

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, \mathbf{c}_i) &= \alpha_{t0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 \\ &\quad + w_{it} \cdot (\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i \cdot w_{it}, \end{aligned} \tag{25}$$

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

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

- With $T \geq 3$, can also get to a random trend model, where $g_i t$ is added to (25). Then, can difference followed by a second difference or fixed effects estimation on the first differences. With $\tau_t = \tau$,

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

- Might ignore $a_i \Delta w_{it}$, using the results on the robustness of the FE estimator in the presence of certain kinds of random coefficients, or, again, estimate $\tau_i = \tau + a_i$ for each i and form (26).

- As in the simple $T = 2$ case, using unconfoundedness conditional on unobserved heterogeneity and strictly exogenous covariates leads to different strategies than assuming unconfoundedness conditional on past responses and outcomes of other covariates.
- In the latter case, we might estimate propensity scores, for each t, as $P(w_{it} = 1|y_{i,t-1},\ldots,y_{i1},w_{i,t-1},\ldots,w_{i1},\mathbf{x}_{it})$.

6. 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. $Y_t(w)$ is the counterfactual outcome for treatment level w, $w = 0, 1$, at time t. Main parameter: the average treatment effect on the treated,

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

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

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

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

Also the (partial) overlap assumption is critical for τ_{att}

$$P(W = 1|X) < 1 \tag{30}$$

or the full overlap assumption for $\tau_{ate} = E[Y_1(1) - Y_1(0)]$,
$0 < P(W = 1|X) < 1$.

Panel Data

Let Y_0 and Y_1 be the observed outcomes in the two periods for a unit from the population. Then, under (29) and (30),

$$\tau_{att} = E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{\rho[1 - p(X)]} \right\} \tag{31}$$

where Y_t, $t = 0, 1$ are the observed outcomes (for the same unit), $\rho = P(W = 1)$ is the unconditional probability of treatment, and $p(X) = P(W = 1|X)$ is the propensity score.

- All quantities are observed or, in the case of $p(X)$ and ρ, 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

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

is consistent and $\sqrt{N}$-asymptotically normal. In other words, just apply propensity score weighting to $(\Delta Y_i, W_i, X_i)$.

- If we add

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

a similar approach works for τ_{ate}.

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

- Regression on the propensity score:

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

The coefficient on W_i is the estimated τ_{ate}. Not ideal, but preferred to a pooled OLS method using levels Y_{it}.

- Matching can be used, too, but now we compute averages based on ΔY_i.
- In fact, any ATE or ATE estimator can be applied to $(\Delta Y_i, W_i, X_i)$.

Pooled Cross Sections

● Heckman, Ichimura, and Todd (1997) show that, under the previous unconfoundedness assumption (29),

$$\begin{aligned}\{E(Y_1|X,W=1)-E(Y_1|X,W=0)\}\ &-\{E(Y_0|X,W=1)-E(Y_0|X,W=0)\}\\ &=E[Y_1(1)-Y_1(0)|X,W=1]. \qquad (\end{aligned}$$

Each of the four expected values on the left hand side of (36) is estimable given random samples from the two time periods. For example, we can use flexible parametric models, or even nonparametric estimation, to estimate $E(Y_1|X,W=1)$ using the data on those receiving treatment at $t=1$.

- Under the stronger form of unconfoundedness, (29) plus (33), it can be shown that

$$\begin{aligned}\{E(Y_1|X,W=1)-E(Y_1|X,W=0)\} &- \{E(Y_0|X,W=1)-E(Y_0|X,W=0)\} \\ &= E[Y_1(1)-Y_1(0)|X]. \qquad (\end{aligned}$$

- Now use iterated expectations to obtain τ_{ate}.

- A regression adjustment estimator would look like

$$\hat{\tau}_{ate,reg} = N_1^{-1} \sum_{i=1}^{N_1} [\hat{\mu}_{11}(X_i) - \hat{\mu}_{10}(X_i)] - N_0^{-1} \sum_{i=1}^{N_0} [\hat{\mu}_{01}(X_i) - \hat{\mu}_{00}(X_i)], \qquad (38)$$

where $\hat{\mu}_{tw}(x)$ is the estimated regression function for time period t and treatment status w, N_1 is the total number of observations for $t = 1$, and N_0 is the total number of observations for time period zero.

- Strictly speaking, (38) consistently estimates τ_{ate} only when the distribution of the covariates does not change over time. The usual DD approach avoids the issue by assuming the treatment effect does not depend on the covariates.
- Equation (38) reduces to the standard DD estimator with controls when the mean functions are linear in X_i with constant coeffcients.

- Abadie (2005) obtained the propensity score weighting versions, also under a stationarity requirement:

$$\hat{\tau}_{att,ps} = N_1^{-1} \sum_{i=1}^{N_1} \left\{ \frac{[W_i - \hat{p}(X_i)]Y_{i1}}{\hat{\rho}[1 - \hat{p}(X_i)]} \right\} - N_0^{-1} \sum_{i=1}^{N_0} \left\{ \frac{[W_i - \hat{p}(X_i)]Y_{i0}}{\hat{\rho}[1 - \hat{p}(X_i)]} \right\}, \quad (39)$$

where $\{Y_{i1} : i = 1, \ldots, N_1\}$ are the data for $t = 1$ and $\{Y_{i0} : i = 1, \ldots, N_0\}$ are the data for $t = 0$.

- Equation (39) has a straightforward interpretation. The first average would be the standard propensity score weighted estimator if we used only $t = 1$ and assumed unconfoundedness in levels. The second average is the same estimate but using the $t = 0$ data. Equation (39) differences across the two time periods – hence the DD interpretation.