# ESTIMATING AVERAGE TREATMENT EFFECTS: DIFFERENCE-IN-DIFFERENCES

Jeff Wooldridge
Michigan State University

BGSE/IZA Course in Microeconometrics
July 2009

1. The Basic Methodology
2. How Should We View Uncertainty in DD Settings?
3. The Donald and Lang Approach
4. Multiple Groups and Time Periods
5. Semiparametric and Nonparametric Methods
6. Unit-Level Panel Data

## 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

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

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 $\delta_1$.
- The difference-in-differences (DD) estimate is

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

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:

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

- 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), 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.
- ADH show how to construct a synthetic control group (for California) using pre-training characteristics of other states (that were not subject to cigarette smoking restrictions) to choose the “best” weighted average of states in constructing the control.

- 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.

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

Linear regression                                      Number of obs =    5626
                                                       F(  3,  5622) =   38.
                                                       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.
                                                       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
------------------------------------------------------------------------------
```

```
. reg ldurat afchnge highearn afhigh male married age head neck upextr trunk
   lowextr occdis manuf construc if ky, robust

Linear regression                                      Number of obs =    5347
                                                       F( 14,  5332) =   18.
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0452
                                                       Root MSE      =  1.2476

------------------------------------------------------------------------------
             |               Robust
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
     afchnge |   .0130565   .0444454     0.29   0.769    -.0740747    .1001876
    highearn |   .1530299   .0506912     3.02   0.003     .0536543    .2524054
      afhigh |   .2244972   .0696846     3.22   0.001     .0878869    .3611075
        male |  -.0560689   .0446726    -1.26   0.209    -.1436455    .0315077
     married |   .0775528   .0390977     1.98   0.047     .0009054    .1542003
         age |   .0066663   .0014459     4.61   0.000     .0038318    .0095008
        head |   -.503178   .1027703    -4.90   0.000    -.7046498   -.3017062
        neck |   .2962081   .1435099     2.06   0.039       .01487    .5775461
      upextr |  -.1655011   .0458495    -3.61   0.000    -.2553849   -.0756172
       trunk |   .1294822   .0596328     2.17   0.030     .0125775     .246387
     lowextr |  -.1097762   .0477096    -2.30   0.021    -.2033066   -.0162458
      occdis |   .2620801   .2197785     1.19   0.233    -.1687757    .6929359
       manuf |    -.16232    .040204    -4.04   0.000    -.2411364   -.0835036
    construc |   .1107367    .049864     2.22   0.026     .0129829    .2084906
       _cons |    1.01803   .0718698    14.16   0.000     .8771354    1.158924
------------------------------------------------------------------------------
```

## 3. The Donald and Lang Approach and an MD Approach

### Background: Inference with "Cluster" Samples

- For each group or cluster $g$, let $\{(y_{gm}, \mathbf{x}_g, \mathbf{z}_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where $M_g$ is the number of units in cluster $g$, $y_{gm}$ is a scalar response, $\mathbf{x}_g$ is a $1 \times K$ vector containing explanatory variables that vary only at the group level, and $\mathbf{z}_{gm}$ is a $1 \times L$ vector of covariates that vary within (as well as across) groups.

- The linear model with an additive cluster effect and unit-specific unobservables is

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + c_g + u_{gm}$$

for $m = 1,\ldots,M_g$, $g = 1,\ldots,G$.

- If we can random sample a large number of groups or clusters, $G$, from a large population of relatively small clusters (with sizes $M_g$), inference is straightforward, even for $\boldsymbol{\beta}$, provided we assume

$$E(v_{gm}|\mathbf{x}_g, \mathbf{z}_{gm}) = 0$$

where $v_{gm} = c_g + u_{gm}$. Just use pooled OLS: then pooled OLS estimator of $y_{gm}$ on $1, \mathbf{x}_g, \mathbf{z}_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G$. Consistent for $\boldsymbol{\lambda} \equiv (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}')'$ (as $G \to \infty$ with $M_g$ fixed) and $\sqrt{G}$-asymptotically normal.

- Robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in $Var(v_{gm}|\mathbf{x}_g, \mathbf{z}_{gm})$, or both. Write $\mathbf{W}_g$ as the $M_g \times (1 + K + L)$ matrix of all regressors for group $g$. Then the $(1 + K + L) \times (1 + K + L)$ variance matrix estimator is

$$\left(\sum_{g=1}^{G} \mathbf{W}_g' \mathbf{W}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{W}_g' \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \mathbf{W}_g\right) \left(\sum_{g=1}^{G} \mathbf{W}_g' \mathbf{W}_g\right)^{-1}$$

where $\hat{\mathbf{v}}_g$ is the $M_g \times 1$ vector of pooled OLS residuals for group $g$. This “sandwich” estimator is now computed routinely using “cluster” options.

- Can use the random effects estimator, too, to exploit the presence of $c_g$, which must cause within-cluster correlation. But one should still use fully robust inference via a “sandwich” estimator. (Might have heteroskedasticity in variances; might have other sources of cluster correlation due to neglected random slopes.)
- Even if use fixed effects to estimate just $\boldsymbol{\gamma}$, still use fully robust inference (just like with panel data to account for neglected serial correlation).

- Recent work by Hansen (2007, Journal of Econometrics): Can use the usual “cluster-robust” inference even if the group sizes, $M_g$, are comparable in magnitude to the number of groups, $G$, provided each is not “too small.” (About $G \approx M_g \approx 30$ seems to do it.) So, the so-called “Moulton problem” where, say, we have $G = 50$ U.S. states and not too many individuals per state has a solution.

- However, when $M_g$ is the large dimension, the usual cluster-robust inference does not work. (For example, $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest varies only at the hospital level, tempting to use pooled OLS with cluster-robust inference. But we have no theoretical justification for doing so, and reasons to expect it will not work well.

- If the explanatory variables of interest vary within group, FE is attractive. First, allows $c_g$ to be arbitrarily correlated with the $\mathbf{z}_{gm}$. Second, with large $M_g$, can treat the $c_g$ as parameters to estimate – because we can estimate them precisely – and then assume that the observations are independent across $m$ (as well as $g$). This means that the usual inference is valid, perhaps with adjustment for heteroskedasticity.

- But what if our interest is in coefficients on the group-level covariates, $\mathbf{x}_g$, and $G$ is small with large $M_g$?
- When $G$ is small and each $M_g$ is large, we often have a different sampling scheme: large random samples are drawn from different segments of a population. Except for the relative dimensions of $G$ and $M_g$, the resulting data set is essentially indistinguishable from a data set obtained by sampling entire clusters.

- Enter Donald and Lang (2007). DL treat the parameters associated with the different groups as outcomes of random draws.
- Simplest case: a single regressor that varies only by group:

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

- Think of the very simple case where $x_g$ is a treatment indicator at the group level.
- DL focus on the first equation, where $c_g$ is assumed to be independent of $x_g$ with zero mean.

- In other words, the DL criticism of the standard difference-of-differences approach has nothing to do with whether the DD quasi-experiment is a good one or not. It is entirely about inference on $\beta$ with small $G$, "large" $M_g$.
- The problem, as set up by DL, is the $c_g$ in the error term.
- Cannot use pooled OLS standard errors which ignore $c_g$ and cannot use clustering because the asymptotics do not work. And cannot use group fixed effects.

- DL propose studying the regression in averages:

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

If we add some strong assumptions, we can perform inference on using standard methods. In particular, assume that $M_g = M$ for all $g$, $c_g|x_g$ ~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 $\bar{v}_g \sim Normal(0,\sigma_c^2 + \sigma_u^2/M)$. Because we assume independence across $g$, the equation in averages satisfies the classical linear model assumptions.

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

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

identical to pooled OLS across $g$ and $m$ with same group sizes.

- 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 $\beta$, 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 $t_{M_1+M_2+...+M_G-2}$ distribution.
- Using OLS on the averages is *not* the same as using cluster-robust standard errors for pooled OLS. Those are not justified *and* 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 $G > K + L + 1$.

- If $c_g$ is independent of $(\mathbf{x}_g, \bar{\mathbf{z}}_g)$ with a homoskedastic normal distribution, and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution. Regressions on aggregate averages are reasonably common, at least as a check on results using disaggregated data, but usually with larger $G$ then just a handful.

- Now the conundrum: If $G = 2$, should we give up? Suppose $x_g$ is binary, indicating treatment and control ($g = 2$ is the treatment, $g = 1$ is the control). The DL estimate of $\beta$ is the usual one: $\hat{\beta} = \bar{y}_2 - \bar{y}_1$. But in the DL setting, we cannot do inference (there are zero df). So, the DL setting rules out the standard comparison of means.

- Can we still obtain inference on estimated policy effects using randomized or quasi-randomized interventions when the policy effects are just identified? Not according the DL approach.
- If $y_{gm} = \Delta w_{gm}$ – the change of some variable over time – then the simplest model

$$\Delta w_{gm} = \alpha + \beta x_g + c_g + u_{gm},$$

using the DL approach, where $x_g$ is a binary treatment estimator, leads to a difference in mean changes, $\hat{\beta} = \overline{\Delta w}_2 - \overline{\Delta w}_1$. This approach has been a workhorse in the quasi-experimental literature [Card and Krueger (1994), for example.]

- According to DL, the comparison of of mean changes using the usual formulas from statistics (possibly allowing for heteroskedasticity) produces the wrong inference, and there is no available inference. The estimate is the same as the usual DD estimator, but there is no way to estimate its sampling variance in the DL scheme.
- This is always true when the treatment effect are just identified.

- Even when DL approach applies, should we? Suppose $G = 4$ with two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). 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 $\mu_3 = \mu_4$.

- Could just define the treatment effect as

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

- 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, $\alpha$ and $\beta$, given four moments that we can estimate with the data. The OLS estimates can be interpreted as minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. If we use the $4 \times 4$ identity matrix as the weight matrix, we get $\hat{\beta}$ and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$.

- With large group sizes, 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.

For each group $g$, write

$$y_{gm} = \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm}.$$

Again, random sampling within group and independence across groups. OLS estimates withing group are $\sqrt{M_g}$ -asymptotically normal.

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

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

where we now think of $x_g$ as fixed, observed attributes of heterogeneous groups. With $K$ attributes we must have $G \geq K + 1$ to determine $\alpha$ 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 $\boldsymbol{\gamma}_g$.

Let $\hat{\mathbf{V}}$ be the $G \times G$ estimated (asymptotic) variance of $\hat{\boldsymbol{\delta}}$. 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 $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. When separate group regressions are used, the $\hat{\delta}_g$ are independent and $\hat{\mathbf{V}}$ is diagonal.

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

- Can test the overidentification restrictions. If reject, can go back to the DL approach or find more elements to put in $\mathbf{x}_g$. With large group sizes, can analyze

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

as a classical linear model because $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, provided $c_g$ is homoskedastic, normally distributed, and independent of $\mathbf{x}_g$.

- In the case of policy analysis, we can just define policy effects in terms of the $\delta_g$, which have been estimated using large random samples, and use the usual kind of inference. The policy effects are just linear combinations of the $\delta_g$.
- The case of small $G$, small $M_g$ is very difficult, and one is forced to use a small-sample analysis on the averages, as in DL. But it can be very sensitive to nonnormality and heteroskedasticity (say, if $y$ is binary).

## 4. Multiple Groups and Time Periods

- With many time periods and groups, setup in BDM (2004) and Hansen (2007a) is useful. With random samples at the individual level for each $(g,t)$ pair,

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

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 variabels), $\mathbf{x}_{gt}$, individual-specific covariates, $\mathbf{z}_{igt}$, unobserved group/time effects, $v_{gt}$, and individual-specific errors, $u_{igt}$. Interested in $\boldsymbol{\beta}$.

- Can write

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

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

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

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 (2007a) 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 $\mathbf{x}_{gt}$, $\lambda_t$ is handled with year dummies,and $\alpha_g$ just represents group dummies. The problem, then, is that we do not observe $\delta_{gt}$.
- But we can use OLS on the individual-level data to estimate the $\delta_{gt}$ 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 $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}$.
- The minimum distance (MD) approach (see cluster sample notes) 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 $\delta_{gt}$. Inference using the efficient MD estimator uses only sampling variation in the $\hat{\delta}_{gt}$.

- Here, proceed ignoring estimation error, and act *as if*

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

- We can apply the BDM findings and Hansen (2007) results directly to this equation. Namely, if we estimate this equation 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 (2007) indicate cluster-robust inference works reasonably well when $\{v_{gt}\}$ follows a stable AR(1) model and $G$ is moderately large.
- If the $M_{gt}$ are not large, might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. Instead, aggregate over individuals:

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

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 previous pooled regression. 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 $v_{gt}$.

## 5. Semiparametric and Nonparametric Approaches

- As in Heckman, Ichimura, and Todd and Abadie (2005), first consider estimating

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

where $Y_t(w)$ the denotes counterfactual outcome with treatment level $w$ in time period $t$. Because no units are treated prior to the initial time period, $W = 1$ means an intervention prior to the second time period.

- For estimating $\tau_{att}$, the key unconfoundedness assumpton is

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

so that, conditional on $X$, treatment status is not related to the gain over time in the absense of treatment. For $\tau_{att}$, need the partial overlap assumption

$$P(W = 1|X = x) < 1, \text{ all } x.$$

- As in HIT, can use regression to first estimate $E[Y_1(1) - Y_1(0)|X, W = 1]$. This expectation is identified under the previous unconfoundedness and overlap assumptions. Let $Y_1 = (1 - W) \cdot Y_1(0) + W \cdot Y_1(1)$ be the observed response for $t = 1$, and let $Y_0 = Y_0(0) = Y_0(1)$ be the response at $t = 0$. Then can show (see lecture notes at provided links)

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

- Each of the four expected values 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 $t = 1$. So, use the data for $t = 0$ to estimate $E(Y_0|X, W = 1) - E(Y_0|X, W = 0)$ – just as we would in the usual regression adjustment – and use the $t = 1$ data to estimate $E(Y_1|X, W = 1) - E(Y_1|X, W = 0)$.

- Analysis for

$$\tau_{ate} = E[Y_1(1) - Y_1(0)]$$

is similar under the stonger overlap assumption and we add to the original unconfoundedness assumption

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

which means that treatment status is unconfounded with respect to the gain under treatment, too.

- Then

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

and so now the ATE conditional on $X$ can be estimated using the estimates of the conditional means for the four time period/treatment status groups.

- The regression-adjustment estimate of $\tau_{ate}$ has the general form

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

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, the previous formula leads to $\tau_{ate}$ (after averaging out the distribution of $X$) only when the distribution of the covariates does not change over time. Of course, one reason to include covariates is to allow for compositional changes in the relevant populations over time. The usual DD approach, based on linear regression, avoids the issue by assuming the treatment effect does not depend on the covariates.
- The HIT approach allows for treatment effects to differ by $X$, but the two averages in practice are necessarily for different time periods.

- Abadie (2005) shows how propensity score weighting can recover $\tau_{att}$ with repeated cross sections and, not surprisingly, also requires a stationarity condition. For $\tau_{att}$,

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

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 $t = 0$.

- Straightforward interpretation: The first average is the standard propensity score weighted estimator if we used only $t = 1$ and assumed unconfoundedness in levels while the second is the same but for $t = 0$. This is why it, like the HIT estimator, is a DD estimator.
- As in the HIT case, we really are replacing $X_i$ with $X_{i1}$ in the first sum and $X_i$ with $X_{i0}$ in the second sum.

- Athey and Imbens (2006) generalize the standard DD model. 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 take the view that the time period, $T_i$, is drawn randomly, too. The key representation is

$$Y_i(0) = h_0(U_i, T_i)$$

where $U_i$ is unobserved. Key assumption is

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

- $Y_i(0) = h_0(U_i, T_i)$ 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. Strict monotonicity assumption rules out discrete responses (but can get bounds under weak monotonicity; with additional assumptions, can recover point identification).

- The distribution of $U_i$ is allowed to vary across groups, but not over time within groups:

$$D(U_i|T_i, G_i) = D(U_i|G_i).$$

- Standard DD model takes

$$h_0(u,t) = u + \delta \bullet t$$

and

$$U_i = \alpha + \gamma G_i + V_i, \; V_i \perp (G_i, T_i)$$

- 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

$$D(Y_i(0)|G_i = 1, T_i = 1).$$

- Uses nonparametric estimation of cumulative distribution functions for pairs $(g, t)$ pair.

- For example, the average treatment effect is estimated as

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

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.

## 6. Unit-Level Panel Data

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

$$y_{it} = \alpha + \eta d2_t + \tau w_{it} + c_i + u_{it}, t = 0, 1,$$

where $d1_t = 1$ if $t = 1$ and zero otherwise, $c_i$ is an observed effect $\tau$ is the treatment effect. Remove $c_i$ by first differencing:

$$(y_{i1} - y_{i0}) = \eta + \tau(w_{i1} - w_{i0}) + (u_{i1} - u_{i0})$$

- Apply OLS on the first differenced equation

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i$$

under $E(\Delta w_i \Delta u_i) = 0$.

- If $w_{i0} = 0$ for all $i$ – no intervention prior to the initial time period – , the OLS estimate is

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

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

- It is *not* more general to regress $y_{i1}$ on $1, w_{i1}, y_{i0}$, $i = 1, \ldots, N$, even though this appears to free up the coefficient on $y_{i0}$. Why? With $w_{i0} = 0$ we can write

$$y_{i1} = \eta + \tau w_{i1} + y_{i0} + (u_{i1} - u_{i0}).$$

Now, if $E(u_{i1}|w_{i1}, c_i, u_{i0}) = 0$ then $u_{i1}$ is uncorrelated with $y_{i0}$, and $y_{i0}$ and $u_{i0}$ are correlated. So $y_{i0}$ is correlated with $u_{i1} - u_{i0} = \Delta u_i$.

- In fact, if we add the standard no serial correlation assumption, $E(u_{i0}u_{i1}|w_{i1}, c_i) = 0$, and write the linear projection $w_{i1} = \pi_0 + \pi_1 y_{i0} + r_{i1}$, then can show that

$$plim(\hat{\tau}_{LDV}) = \tau + \pi_1(\sigma_{u0}^2/\sigma_{r_1}^2)$$

where

$$\pi_1 = Cov(c_i, w_{i1})/(\sigma_c^2 + \sigma_{u0}^2).$$

- For example, if $w_{i1}$ indicates a job training program and less productive workers are more likely to participate ($\pi_1 < 0$), then the regression $y_{i1}$ (or $\Delta y_{i1}$) on 1, $w_{i1}, y_{i0}$ underestimates the effect.

- If more productive workers participate, regressing $\Delta y_{i1}$ on 1, $w_{i1}, y_{i0}$ overestimates the effect of job training.
- Now consider the other way around. Following Angrist and Pischke (2009), 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

$$\begin{aligned} y_{i1} &= \gamma + \tau w_{i1} + \psi y_{i0} + e_{i1} \\ E(e_{i1}) &= 0,\ Cov(w_{i1}, e_{i1}) = Cov(y_{i0}, e_{i1}) = 0. \end{aligned}$$

- Write the equation as

$$\begin{aligned}\Delta y_{i1} &= \gamma + \tau w_{i1} + (\psi - 1) y_{i0} + e_{i1} \\ &\equiv \gamma + \tau w_{i1} + \lambda y_{i0} + e_{i1}\end{aligned}$$

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

$$plim(\hat{\tau}_{FD}) = \tau + \lambda \frac{Cov(w_{i1}, y_{i0})}{Var(w_{i1})}$$

- If $\lambda < 0$ ($\psi < 1$) and $Cov(w_{i1}, y_{i0}) < 0$ – workers observed with low first-period earnings are more likely to participate – the $plim(\hat{\tau}_{FD}) > \tau$, and so FD overestimates the effect.

- Generally, it is possible to derive the standard unobserved effects models – leading to the basic estimation methods of fixed effects and extensions – in a counterfactual setting. And this is with general patterns of treatment. For example, 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

$$\begin{aligned} E[y_{it}(0)|\mathbf{w}_i,\mathbf{c}_i] &= E[y_{it}(0)|\mathbf{c}_i] \\ E[y_{it}(1)|\mathbf{w}_i,\mathbf{c}_i] &= E[y_{it}(1)|\mathbf{c}_i], \end{aligned}$$

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

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

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

Then

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

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

- If we further assume

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

then

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

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

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

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

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

- And so on. Can get random trend models, with $g_i t$, say. Then, can difference followed by a second difference or fixed effects estimation on the first differences. With $\tau_t = \tau$,

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

- 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 the average.

- Altonji and Matzkin (2005), Wooldridge (2005) can be used without specifying functional forms. If we assume unconfoundedness contional on $\mathbf{c}_i$,

$$E(Y_{it}(g)|\mathbf{W}_i, \mathbf{c}_i) = h_{tg}(\mathbf{c}_i)$$

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

$$\tau_t = E[h_{t1}(\mathbf{c}_i) - h_{t0}(\mathbf{c}_i)].$$

- Suppose

$$D(\mathbf{c}_i|W_{i1}, \ldots, W_{iT}) = D(\mathbf{c}_i|\bar{W}_i)$$

which means that only the intensity of treatment is correlated with heterogeneity. (Or, can break the average into more than one time period.)

- Then can show the following class of estimators is consistent for $\tau_t$ provided we consistently estimate the mean responses given

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

where $\mu_t^Y(1, \bar{W}_i) = E(Y_{it} | W_{it} = 1, \bar{W}_i)$ and similarly for $\mu_t^Y(0, \bar{W}_i)$.

- With two periods and no treatment in the first period, can use the Abadie (2005) with unit-level panel data. For example,

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

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

- These are just the usual propensity score weighted estimators but applied to the changes in the responses over time.
- So matching based on the covariates or PS is available, too, as is regression adjustment, using the time change in the response.
- Much more convincing than regressions such as

$$Y_{i1} \text{ on } 1, W_i, \hat{p}(X_i)$$

which is worse than just the usual DD estimator.

- Abadie’s approach does not extend immediately to more than two time periods with complicated treatment patterns. The usual kind of panel data models assume unconfoundedness of the entire history of treatments given unobserved heterogeneity. Does this describe how treatments are determined?

- Lechner (1999), Gill and Robins (2001), and Lechner and Miquel (2005) use unit-level panel data and assume sequential unconfoundedness, also with more than two treatment states. Dynamic regression adjustment, inverse propensity score weighting, matching are all available solutions, as well as combined methods.

- In the binary treatment case, the assumption is that $[Y_{it}(0), Y_{it}(1)]$ is independent of $W_{it}$ (treatment assignment) conditional on $\{Y_{i,t-1}, \ldots, Y_{i1}, W_{i,t-1}, \ldots, W_{i1}, \mathbf{X}_{it}\}$ where $\mathbf{X}_{it}$ is all observed covariates up through time $t$. The propensity score is

$$p_t(\mathbf{R}_{it}) = P(W_{it} = 1 | Y_{i,t-1}, \ldots, Y_{i1}, W_{i,t-1}, \ldots, W_{i1}, \mathbf{X}_{it})$$

and then an estimate of $\tau_{t,ate}$ is

$$\hat{\tau}_{t,ate} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{[W_{it} - \hat{p}_t(\mathbf{R}_{it})]Y_{it}}{\hat{p}_t(\mathbf{R}_{it})[1 - \hat{p}_t(\mathbf{R}_{it})]} \right\}$$

- With more than two treatment possibilities, say $W_{it} \in \{0, 1, \ldots, G\}$, the observed response can be written as

$$Y_{it} = 1[W_{it} = 0]Y_{it}(0) + 1[W_{it} = 1]Y_{it}(1) + \ldots + 1[W_{it} = 1]Y_{it}(1)$$

and a sufficient unconfoundedness assumption is

$$E[Y_{it}(g)|W_{it}, \mathbf{R}_{it}] = E[Y_{it}(g)|\mathbf{R}_{it}], g = 1, \ldots, G$$

and all $t$. Then, the means $\mu_{tg} = E[Y_{it}(g)]$ are identified from, for example,

$$\mu_{tg} = E\left[\frac{1[W_{it} = g]Y_{it}}{p_{tg}(\mathbf{R}_{it})}\right],$$

where

$$p_{tg}(\mathbf{R}_{it}) = P(W_{it} = g|\mathbf{R}_{it})$$

IPW estimators take the form

$$\hat{\mu}_{tg} = N^{-1}\sum_{i=1}^{N}\left\{\frac{1[W_{it} = g]Y_{it}}{\hat{p}_{tg}(\mathbf{R}_{it})}\right\}$$

and these estimates can be used to construct contrasts, such as $\hat{\mu}_{tg} - \hat{\mu}_{t,g-1}$.