Georgetown University
Spring 2015
Advanced Applied Econometrics
(PPOL 754-20)

Andrew H. McCallum, Ph.D.

April 14, 2016

1. The key to causal is control for observed confounding variables.
2. If important confounders are unobserved, causal effects can be estimated using IV
3. Good instruments are rare so want to consider other tools to deal with unobserved confounders.
4. Use data with a time or cohort dimension to control for unobserved-but-fixed omitted variables.
5. Do not allow for comparisons in levels AND require counterfactual trends to be the same.

Union membership and wages. Are collective bargaining wages higher than non-bargained wages due to union power or would these workers earn more anyway? (experienced, skill etc.).

Y_{it} equal the (log) earnings of worker i at time t and let D_{it} denote her union status. The observed Y_{it} is either Y_{0it} or Y_{1it}, depending on union status.

Suppose further that

$$E\left(Y_{0it} \mid A_i, X_{it}, t, D_{it}\right) = E\left(Y_{0it} \mid A_i, X_{it}, t\right),$$

where Y_{it} is (log) earnings of worker i at time t, D_{it} is union status and as before observed Y_{it} is either Y_{0it} or Y_{1it}, depending on union status.

D_{it} randomly assigned conditional on unobserved ability, A_i, and observed covariates X_{ti}.

Assume unobserved A_i appears without a time subscript . . . which is the key

$$E\left(Y_{0it} \mid A_i, X_{it}, t\right) = \alpha + \lambda_t + A_i'\gamma + X_{it}\delta,$$

$$E\left(Y_{1it} \mid A_i, X_{it}, t\right) = E\left(Y_{0it} \mid A_i, X_{it}, t\right) + \rho.$$

$$E\left(Y_{it} \mid A_i, X_{it}, t, D_{it}\right) = \alpha + \lambda_t + \rho D_{it} = A_i'\gamma + X_{it}\delta,$$

ρ is the causal effect of union membership.

These assumptions are more restrictive than usual: linear additive functional allows including *unobserved* confounders in panels without IV.

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}\delta + \varepsilon_{it}.$$

where

$$\alpha_i \equiv \alpha + A_i'\gamma.$$

This is a *fixed-effects model.*
Panel data has repeated two subscripts (often observations on individuals over time)
The causal effect of union status on wages can be estimated by treating α_i, the fixed effect, as a parameter to be estimated.
The *year effect*, λ_t, is also treated as a parameter to be estimated.

$$E\left(Y_{1it} - Y_{0it} \mid A_i, X_{it}, t\right) = \rho_i.$$

Many parameters to be estimated in the fixed effects model . . . but instead of estimating we can remove them.

first we calculate the individual averages

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \rho\overline{D}_i + \overline{X}_i\delta + \overline{\varepsilon}_i.$$

and then form

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(D_{it} - \overline{D}_i\right) + \left(X_{it} - \overline{X}_i\right)\delta + \left(\varepsilon_{it} - \overline{\varepsilon}_i\right),$$

deviations from means removes unobserved individual effects
We could also usefirst differences

$$\Delta Y_{it} = \Delta\lambda_t + \rho\Delta D_{it} + \Delta X_{it}\delta + \Delta\varepsilon_{it},$$

where $\Delta Y_{ti} = Y_{ti} - Y_{ti-1}$ is the change over time.

- Both FD and WG work and are the same when only have two periods.
- Why is removing the FE give the same estimate as estimating each as a parameter?
- Regression anatomy implies any multivariate regression coefficients can be estimated in two steps.
- The residuals from a regression on a full set of person-dummies in a person-year panel are deviations from person means.
- FE are no consistent in a panel where the number of periods T is fixed while $N \to \infty$ (incidental parameters problem) since the number of parameters grows with the sample size
- Nevertheless, other parameters in the fixed effects model — the ones we care about — are consistently estimated.
- Freeman (1984) estimated union wage effects allowing joining a union to be related to unobserved-but-fixed individual characteristics.

Problems with fixed effects

- FE estimates susceptible to attenuation bias from measurement error.
- Economic variables are often persistent (union membership this year is a good predictor of union membership next year)
- measurement error often changes from year-to-year
- observed year-to-year change may be mostly noise. likely more measurement error in diffferenced data than in levels of the regressors.
- This measurement error leads to lower FE estimates
- Another view is that FD or WG remove both good and bad variation.
- These transformations may toss out some of the OMV bathwater But they also remove much of the useful variation in the variables, the "baby"

Pre and Post, Treatment and Control

- On April 1, 1992, New Jersey raised the state minimum from \$4.25 to \$5.05. The minimum wage in Pennsylvania stayed at \$4.25 until the federal min was bumped in 1996
- Card and Krueger (1994) compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum. A perfect DD setup.
- DD is a version of fixed-effects estimation using aggregate data. Let

$$\begin{aligned} Y_{1ist} &= \text{fast food employment at restaurant } i \text{ and period } t \text{ if there is a high state minimum wage} \\ Y_{0ist} &= \text{fast food employment at restaurant } i \text{ and period } t \text{ if there is a low state minimum wage} \end{aligned}$$

The DD Setup

- The heart of the DD setup is an additive model for potential outcomes in the no-treatment state:

$$E\left[Y_{0ist} \mid s,t\right] = \gamma_s + \lambda_t$$

  where s denotes state (New Jersey or Pennsylvania), t denotes period (February, before the minimum wage increase or November, after the increase), and γ_s and λ_t are state and year effects
- Let D_{st} be a dummy for high-minimum-wage states. Assuming that $E\left[Y_{1ist} - Y_{0ist} \mid s,t\right]$ is a constant, denoted δ, we have:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}$$

  where $E\left(\varepsilon_{ist} \mid s,t\right) = 0$

The DD Setup (cont.)

- From here, we get

$$E\left[Y_{ist} \mid s = NJ, t = Nov\right] - E\left[Y_{ist} \mid s = NJ, t = Feb\right] = \lambda_{Nov} - \lambda_{Feb} + \delta$$

$$E\left[Y_{ist} \mid s = PA, t = Nov\right] - E\left[Y_{ist} \mid s = PA, t = Feb\right] = \lambda_{Nov} - \lambda_{Feb}$$

- The population difference-in-differences, is

$$E\left[Y_{ist} \mid s = NJ, t = Nov\right] - E\left[Y_{ist} \mid s = NJ, t = Feb\right] -$$

$$E\left[Y_{ist} \mid s = PA, t = Nov\right] - E\left[Y_{ist} \mid s = PA, t = Feb\right] = \delta$$

  given parallel trends, this is the causal effect of interest
- Table 5.2.1 (based on Table 3 in Card and Krueger, 1994) reports average employment at NJ and PA fast food restaurants before and after the change in the NJ minimum wage. The table shows a surprising positive DD

Card and Kruger (1994)

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	−2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	−2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	−2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a]Stores in New Jersey were classified by whether starting wage in wave 1 equals $4.25 per hour ($N = 101$), is between $4.26 and $4.99 per hour ($N = 140$), or is $5.00 per hour or higher ($N = 73$).

[b]Difference in employment between low-wage ($4.25 per hour) and high-wage (≥ $5.00 per hour) stores; and difference in employment between midrange ($4.26–$4.99 per hour) and high-wage stores.

[c]Subset of stores with available employment data in wave 1 and wave 2.

[d]In this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.

Common Trends

- The key DD assumption here is parallel employment *trends* in NJ and PA in the absence of treatment
- Common trends can be applied to transformed data, e.g.,

$$E\left[\log Y_{0ist} \mid s, t\right] = \gamma_s + \lambda_t$$

  Common in logs does not imply (indeed, contradicts) common in levels. DD identification is fickle!
- Deviations from this common trend are mistaken for treatment effects, as illustrated in Figure 5.2.1
- The common trends assumption can be investigated using multiple periods: CK (2000) update in Figure 5.2.2
  - Vertical lines indicate original CK surveys and the 10/96 increase in the federal min to \$4.75, which affected PA but not NJ
  - These data reveal time series variation that differs substantially in the two states. PA may not be a good control for NJ
- We'll soon explore strategies to manage this

Regression DD: All this and SEs too

- Let NJ_s be a dummy for NJ restaurants and let d_t be a dummy for post-min obs. Then,

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta \left(NJ_s \cdot d_t \right) + \varepsilon_{ist}$$

  is the same as (2) where $NJ_s \cdot d_t = D_{st}$. This is a saturated model

- Parameters in (3) link to the DD model for the CEF (2) as follows:

$$\begin{aligned}
\alpha &= E\left[Y_{ist} \mid s = PA, t = Feb\right] = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E\left[Y_{ist} \mid s = NJ, t = Feb\right] - E\left[Y_{ist} \mid s = PA, t = Feb\right] \\
&= \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E\left[Y_{ist} \mid s = PA, t = Nov\right] - E\left[Y_{ist} \mid s = PA, t = Feb\right] \\
&= \lambda_{Nov} - \lambda_{Feb} \\
&= \left\{E\left[Y_{ist} \mid = NJ, t = Nov\right] - E\left[Y_{ist} \mid s = NJ, t = Feb\right]\right\} \\
&\quad - \left\{E\left[Y_{ist} \mid = PA, t = Nov\right] - E\left[Y_{ist} \mid s = PA, t = Feb\right]\right\}
\end{aligned}$$

  Regression DD generalizes easily to additional states and periods

- Beware clustering and serial correlation

Card (1992): Many States, Variably Treated

Card (1992) cleverly makes the federal min into a DD experiment using an equation like

$$Y_{ist} = \gamma_s + \lambda_t + \delta\left(FA_s \cdot d_t\right) + \varepsilon_{ist},$$

where FA_s is *fraction affected* in each state (pre-increase proportion of teen labor force earning $<$ \$3.80) and d_t is a dummy for observations in 1990, after the federal min went from \$3.35 to \$3.80

- Card (1992) used two periods, before and after (1989 and 1990) and 51 states: $N = 102$
- $FA_s \cdot dt$ is an interaction term, like $NJ_s \cdot d_t$ in (3), though here the interaction term is not a dummy
- Two periods: levels w/fixed (state) effects $=$ first differences:

$$\Delta\overline{Y}_s = \lambda^* + \delta FA_s + \Delta\overline{\varepsilon}_s,$$

where $\Delta\overline{Y}_s$ is the change in teen employment in state s and $\Delta\overline{\varepsilon}_s$ is the differenced error

Card (1992): Variably Treated States (cont.)

- Table 5.2.2, based on Card (1992), shows that wages increased more in states where the minimum wage increase is likely to have had more bite (see the estimate of $.15$ in column 1)
- Employment, on the other hand, seems largely unrelated to *fraction affected*, as can be seen in column 3
- It's easy to add covariates in this framework, though OVB must be at the state-year level since that's where $FA_s \cdot d_t$ lives. (Card controls for adult employment; arguably, a bad idea)
- Equivalently, we can pool micro data from the CPS to estimate

$$Y_{ist} = \gamma_s + \lambda_t + \delta\left(FA_s \cdot d_t\right) + X_{ist}'\beta + \varepsilon_{ist},$$

  where X_{ist} includes individual characteristics such as race. Without micro covs, this generates the same estimates as sample-size weighted estimation using averages (Recall the regression CEF theorem)
- Micro covs won't matter unless the sample composition *changes dierentially* by state (why?)

DD Spec Checks: Granger Causality

Granger tests ask whether contemporaneous and lagged values of a time-varying policy, D_{st}, predict Y_{ist}, while future D_{st} does not.

- We check this by estimating equations like:

$$Y_{ist} = \gamma_s + \lambda_t + \sum_{\tau=0}^{m} \delta_{-\tau} D_{s,t-\tau} + \sum_{\tau=1}^{q} \delta_{+\tau} D_{s,t+\tau} + X'_{ist}\beta + \varepsilon_{ist},$$

  allowing for m lags $(\delta_{-1}, \delta_{-2}, \ldots, \delta_{-m})$ or post-treatment effects and q leads $(\delta_{+1}, \delta_{+2}, \ldots, \delta_{+q})$ or anticipatory effects
- The pattern of lagged effects is usually of substantive interest as well: Perhaps causal effects grow or fade
- Autor (2003) investigates the effect of EPL on employers' use of temporary help. He relates temp employment to state court rulings that allow exceptions to employment-at-will
- Autor's reg'n-DD includes leads and lags, running from 2 years ahead to 4 years behind, plotted in Figure 5.2.4 ($3+$ prior is ref.)
  - No effects before; sharply increasing after

DD Spec Checks: State-Specific Trends

A closely related check adds state-specific time trends:

$$Y_{ist} = \gamma_{0s} + \gamma_{1s}t + \lambda_t + \delta D_{st} + X'_{ist}\beta + \varepsilon_{ist},$$

where γ_{0s} is a state-specific intercept as before and γ_{1s} is a state-specific trend coefficient multiplying the linear trend variable, t.

- We need at least 3 periods to estimate a model with state-specific trends (the more the better)
- Besley and Burgess (2004) include state trends in a study of the effect of labor regulation on businesses in Indian states. Table 5.2.3 reproduces the key results
- Without state-specific trends, labor regulation leads to lower output per capita . . . but alas, this anguished though familiar cry did echo in the narrow alleys of Westminster: "State trends kill it!"
- On the other hand, our MLDA death rate regressions survive

Higher-Order DD

A modiffication of the two-way DD set-up uses higher-order contrasts for causal inference.

- In the 1980s, some states extended Medicaid coverage to young children in families ineligible for AFDC
- Yelowitz (1995) uses this to study Medicaid LFP effects by estimating:

$$Y_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \delta D_{ast} + X'_{iast}\beta + \varepsilon_{iast},$$

  where s is state, t time, and a age of youngest child. This controls for state-speciffic time effects common across age groups (γ_{st}), time-varying age effects (λ_{at}), and state-speciffic age effects (θ_{as}). Here, D_{ast}, indicates children in affected age groups in state-years where coverage is provided
- Triple-DD may be more convincing than a pure state-by-year analysis (allows for flexible state trends), but typically has less power

DD vs LDV (lagged dependent variables) in Short Panels

DD and panel program evaluation are closely linked

- DD can be motivated by "FE ignorability":

$$E\left(Y_{0it} \mid \alpha_i, X_{it}, D_{it}\right) = E\left(Y_{0it} \mid \alpha_i, X_{it}\right),$$

  where α_i is an unobserved *fixed effect*, assumed to be the only source of selection bias
  - If $E\left(Y_{0it} \mid \alpha_i, X_{it}\right)$ is additive, regression-DD identifies causal effects: $Y_{it} = \alpha_i + X_{it}'\beta + \delta D_{it} + \varepsilon_{it}$
- The distinctive earnings histories of trainees motivates DW99's selection-observables inditification using LDV controls:

$$E\left(Y_{0it} \mid Y_{i\tau-h}, X_{it}, D_{it}\right) = E\left(Y_{0it} \mid Y_{i\tau-h}, X_{it}\right),$$

  where $t > \tau$, the treatment (training) year ($Y_{i\tau-h}$ might be a vector)
  - With linear $E\left(Y_{0it} \mid Y_{i\tau-h}, X_{it}\right)$, causal effects are identified by the LDV regression:

$$Y_{it} = \alpha + \theta Y_{i\tau-h} + \delta D_{it} + X_{it}'\beta + \varepsilon_{it}$$

LDV and DD Are Not Nested Part I: Mistaken DD

- Suppose treatment is determined by low Y_{it-1}, so LDV regression gives the right answer:

$$Y_{it} = \alpha + \theta Y_{it-1} + \delta D_{it} + \varepsilon_{it}$$

- You mistakenly do DD. Here $D_{it-1} = 0$ for everyone, so DD is

$$\frac{\text{Cov}\left(Y_{it} - Y_{it-1}, D_{it} - D_{it-1}\right)}{V\left(D_{it} - D_{it-1}\right)} = \frac{\text{Cov}\left(Y_{it} - Y_{it-1}, D_{it}\right)}{V\left(D_{it}\right)}$$

- Subtracting Y_{it-1} from both sides of (11),

$$Y_{it} - Y_{it-1} = \alpha + (\theta - 1) Y_{it-1} + \delta D_{it} + \varepsilon_{it}$$

  Substituting this into (12), bad DD yields

$$\frac{\text{Cov}\left(Y_{it} - Y_{it-1}, D_{it}\right)}{V\left(D_{it}\right)} = \delta + (\theta - 1)\left[\frac{\text{Cov}\left(Y_{it-1}, D_{it}\right)}{V\left(D_{it}\right)}\right]$$

- θ lies strictly in (0,1), else Y_{it} is non-stationary. Since $\text{Cov}\left(Y_{it} - Y_{it-1}, D_{it}\right) < 0$, this DD estimate is too big

LDV and DD Are Not Nested Part II: Mistaken LDV

- Suppose treatment is determined by a fixed effect, α_i
- You mistakenly estimate LDV model (11), ignoring fixed effects. By regression anatomy, this generates $\frac{\text{Cov}\left(Y_{it},\tilde{D}_{it}\right)}{V\left(\tilde{D}_{it}\right)}$, where $\tilde{D}_{it} = D_{it} - \gamma Y_{it-1}$
- Assuming ε_{it} is serially uncorrelated, we can show (see MHE 5.4)

$$\frac{\text{Cov}\left(Y_{it},\tilde{D}_{it}\right)}{V\left(\tilde{D}_{it}\right)} = \delta + \frac{\gamma\sigma_{\varepsilon}^{2}}{V\left(\tilde{D}_{it}\right)}.$$

  - Use regression to write

$$Y_{it} = \alpha_i + \delta D_{it} + \varepsilon_{it}$$

    where ε_{it} is uncorrelated with α_i and D_{it}; note that $Y_{it-1} = \alpha_i + \varepsilon_{it-1}$
- The treated have low Y_{it-1} so $\gamma < 0$ and this estimate is too small

LDV and DD Summary

- FE and LDV conditional independence assumptions are distinct: models w/LDV controls are not "robust" to bias from FE selection
- Why not rock FEs and LDVs? An AC/DC model strives to estimate causal effects like this:

$$Y_{it} = \alpha_i + \theta Y_{it-1} + \delta D_{it} + \varepsilon_{it}$$

  - OLS in first diffs (a DD-style model) fails for (14):

$$\Delta Y_{it} = \theta \Delta Y_{it-1} + \delta D_{it} + \Delta \varepsilon_{it}$$

    has residuals correlated with $\Delta Y_{i\tau-1}$ (See Nickell, 1981)
  - With many periods, FE estimates (ANCOVA) are consistent
- We might try to instrument ΔY_{it-1} with higher-order lags, as in Arellano and Bond (1991)
  - Identification then turns on serial correlation assumptions that are hard to interpret or defend

Sharp RD

- RD arises from the paradoxical idea that *rules* — which at first appear to reduce or even eliminate the scope for randomness — create natural experiments
- Sharp RD is used when treatment status is a deterministic and discontinuous function of a covariate, x_i, sometimes called the *running variable*:

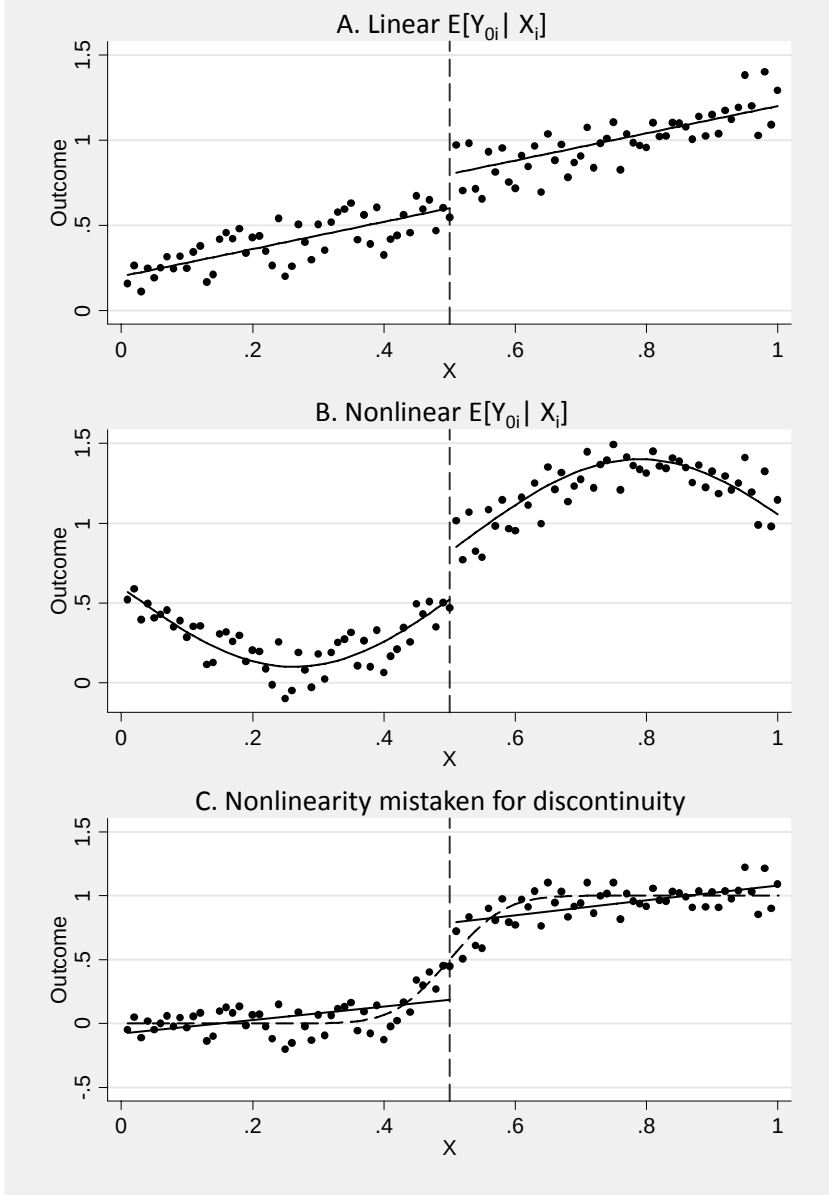
$$D_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0 \end{cases}$$

  where x_0 is a known threshold or cutoff value
  - D_i is a deterministic function of x_i: once we know x_i, we know D_i
  - D_i is discontinuous because no matter how close x_i gets to x_0, treatment is unchanged until $x_i = x_0$

The RD Thing

- There is *no* value of x_i at which we get to observe both treatment and control observations
- Unlike regression and matching strategies, which are based on treatment-control comparisons *conditional* on covariate values, the validity of RD turns on our willingness to extrapolate across covariate values
- Thistlethwaite and Campbell (1960) contributed the index example: applicants for National Merit Scholarships; the outcome is occupation; the running variable is applicant PSAT
- Figure 6.1.1 illustrates a hypothetical RD scenario where those with $x_i > 0.5$ are treated. In Panel A, the trend relationship between Y_i and x_i is linear, while in Panel B, it's nonlinear. In both cases, the relationship between $E[Y_i \mid x_i]$ and x_i is discontinuous at x_0, suggesting a treatment effect

Example of RD



Doing Sharp RD

- Suppose potential outcomes can be described by a linear, constant-effects model

$$\begin{aligned} E\left[Y_{0i} \mid x_i\right] &= \alpha + \beta x_i \\ Y_{1i} &= Y_{0i} + \rho \end{aligned}$$

  leading to the regression,

$$Y_i = \alpha + \beta x_i + \rho D_i + \eta_i$$

  where ρ is the causal effect of interest
- D_i is not only correlated with x_i, it's *determined* by x_i.
- RD distinguishes the nonlinear and discontinuous function, $1\left(x_i \geq x_0\right)$, from the smooth and (in this case) linear function, x_i
- Given (1) and (2), there isn't — cannot be! — any OVB in OLS estimates of (4)

Nonlinear Trend Functions

- We can allow for nonlinear $E[Y_{0i} \mid x_i]$
- Figure 6.1.1 (Panel B) suggests causal effects are identified even if $E[Y_{0i} \mid x_i] = f(x_i)$ for some smooth but nonlinear function, $f(x_i)$
  - Panel C of the figure suggests we must get this right
- Assuming we do, RD works by fitting:

$$Y_i = f(x_i) + \rho D_i + \eta_i$$

  where again, $D_i = 1(x_i \geq x_0)$ is discontinuous at x_0
- Typically, $f(x_i)$ is taken to be polynomial,

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \rho D_i + \eta_i$$

- We hope our results are robust to the choice of degree, at least for $p \geq 3$

Changing Trends

- The behavior of $E\left[Y_{0i} \mid x_i\right]$ and $E\left[Y_{1i} \mid x_i\right]$ may differ:

$$\begin{aligned} E\left[Y_{0i} \mid x_i\right] &= f_0\left(x_i\right) = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \ldots + \beta_{0p}\tilde{x}_i^p \\ E\left[Y_{1i} \mid x_i\right] &= f_1\left(x_i\right) = \alpha + \rho + \beta_{11}\tilde{x}_i + \beta_{12}\tilde{x}_i^2 + \ldots + \beta_{1p}\tilde{x}_i^p, \end{aligned}$$

  where $\tilde{x}_i \equiv x_i - x_0$ centers these polynomials at x_0

- Substituting these in $E\left[Y_i \mid x_i\right] = E\left[Y_{0i} \mid x_i\right] + E\left[Y_{1i} - Y_{0i} \mid x_i\right] D_i$, gives:

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \ldots + \beta_{0p}\tilde{x}_i^p \\ &\quad + \rho D_i + \beta_1^* D_i\tilde{x}_i + \beta_2^* D_i\tilde{x}_i^2 + \ldots + \beta_p^* D_i\tilde{x}_i^p + \eta_i, \end{aligned}$$

  where the error term, η_i, is the CEF residual

- The interaction terms are defined so that,

$$\beta_1^* = \beta_{11} - \beta_{01},\ \beta_2^* = \beta_{12} - \beta_{02}, \ldots, \beta_p^* = \beta_{1p} - \beta_{0p}$$

Interpreting RD Interactions

- Conditional treatment effects:

$$E\left[Y_{1i} - Y_{0i} \mid x_i\right] = \rho + \beta_1^* \tilde{x}_i + \beta_2^* \tilde{x}_i^2 + \ldots + \beta_p^* \tilde{x}_i^p$$

- In other words,

$$\begin{aligned} &\rho \quad \text{at } x_i = x_0 \\ &\rho \quad + \beta_1^* c + \beta_2^* c^2 + \ldots + \beta_p^* c^p \text{ at } x_i - x_0 = c \end{aligned}$$

- Equation (6) is a special case of (7) where $\beta_1^* = \beta_2^* = \ldots = \beta_p^* = 0$
- RD estimates of ρ from models without interactions often come out similar to those from models with
- Increasing nonlinearity in the additive specification, (6), makes it more like the interacted, equation (7)

Lee (2008)

- To capture causal effects of party incumbency, Lee exploits the fact that election winners are determined by $D_i = 1(x_i \geq 0)$, where x_i is the *vote share margin of victory* (Democrat-Republican vote shares)
- Because D_i is a deterministic function of x_i, there are no confounding variables other than x_i
- Figure 6.1.2 (A) plots the probability of Democrat wins against the difference between Democratic and Republican vote shares
  - Dots in the figure are local averages (the average win rate in non-overlapping windows of share margins that are .005 wide);
  - Fitted values are from a Logit model for the probability of winning as a function of D_i , a 4^{th}-order polynomial in x_i , and interactions between the polynomial terms and D_i
- Results here aren't subtle: Incumbency appears to raise party re-election probabilities by about 40 percentage points.

Lee (2008)

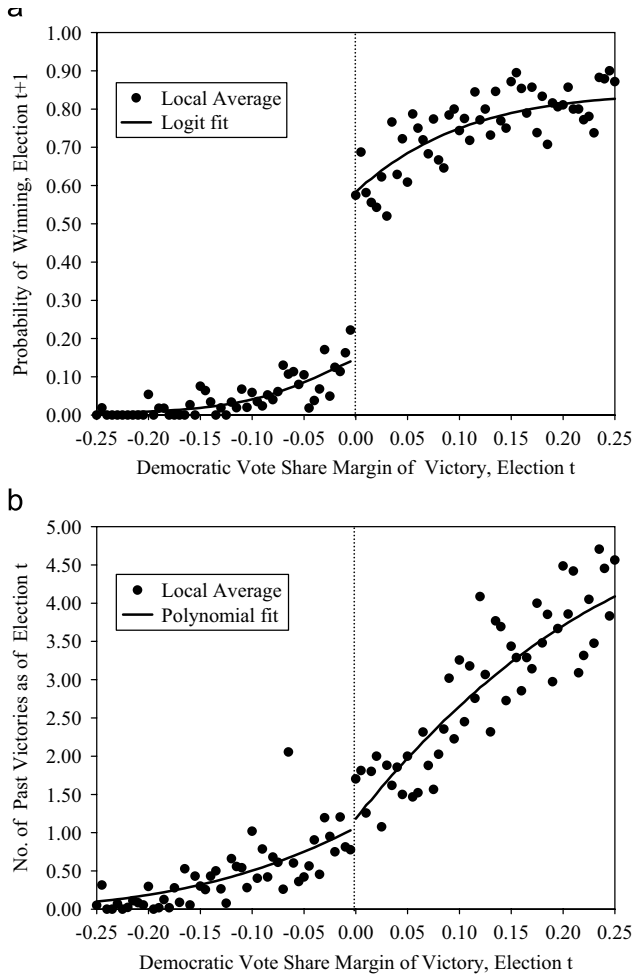


Figure 6.1.2: Probability of winning an election by past and future vote share (from Lee, 2008). (a) Candidate's probability of winning election $t+1$, by margin of victory in election t: local averages and parametric fit. (b) Candidate's accumulated number of past election victories, by margin of victory in election t: local

Spec Checks and Worries

- Purposeful sorting around the cutoff messes with RD: when agents strive to avoid or cross the threshhold, $E\left[Y_{0i} \mid x_i\right]$ is unlikely to be smooth near x_0
- Figure 6.1.2 (B) checks sharp RD identification by looking at Democratic victories *before* the last election. Democratic win rates in older elections should be smooth through the cutoff in the last election, a spec check that works out well
- We can also check the density of x_i for evidence of sorting, looking for bunching near x_0. McCrary (2008) proposes a formal test for this
  - Urquiola and Verhoogen (2009) show this fails for class size caps in Chile's privatized school system (Chilean enrollment is "endogenous")
- *Heaps of trouble*: Almond, et al. (2010) estimate effects of the extra medical care VLBW babies get; This is RD at a 1500 gram cutoff
- Barecca, et al. (2011) show that heaping biases these estimates because mothers heaped at the cutoff are different; Almond, et al. (2011) reply

Negotiating Sharp Turns: Nonparametric RD

- Figure 6.1.1 (Panel C) shows how a sharp turn in $E\left[Y_{0i} \mid x_i\right]$ might be mistaken for a jump from one CEF to another.
- To reduce the likelihood of such mistakes, we can look only at data in a neighborhood of the discontinuity:

$$\begin{aligned} E\left[Y_i \mid x_0 - \delta < x_i < x_0\right] &\simeq E\left[Y_{0i} \mid x_i = x_0\right] \\ E\left[Y_i \mid x_0 < x_i < x_0 + \delta\right] &\simeq E\left[Y_{1i} \mid x_i = x_0\right], \end{aligned}$$

so that

$$\begin{aligned} & \lim_{\delta \to 0} E\left[Y_i \mid x_0 < x_i < x_0 + \delta\right] - E\left[Y_i \mid x_0 - \delta < x_i < x_0\right] \\ = \;& E\left[Y_{1i} \mid x_i = x_0\right] - E\left[Y_{0i} \mid x_i = x_0\right] \\ = \;& E\left[Y_{1i} - Y_{0i} \mid x_i = x_0\right] \end{aligned}$$

- This limiting argument obviates the need to model $E\left[Y_{0i} \mid x_i\right]$
- But it raises the question of how best to estimate mean Y_i in neighborhoods to the right and left of x_0

Fuzzy Logic

- Let D_i denote the treatment as before, though here D_i is no longer deterministically related to the threshold-crossing rule, $x_i \geq x_0$. Rather, the probability of treatment jumps at x_0:

$$P\left[D_i = 1 \mid x_i\right] = \begin{cases} g_0\left(x_i\right) & \text{if } x_i < x_0 \\ g_1\left(x_i\right) & \text{if } x_i \geq x_0 \end{cases} \quad \text{where } g_1\left(x_i\right) \neq g_0\left(x_i\right)$$

- We'll assume $g_1\left(x_0\right) > g_0\left(x_0\right)$, so $x_i \geq x_0$ makes treatment more likely, and that the $g_j\left(x_i\right)$ are smooth near x_0
- We can write the conditional probability of treatment given x_i as

$$E\left[D_i \mid x_i\right] = P\left[D_i = 1 \mid x_i\right] = g_0\left(x_i\right) + \left[g_1\left(x_i\right) - g_0\left(x_i\right)\right] T_i,$$

where

$$T_i = 1\left(x_i \geq x_0\right)$$

Fuzzy RD is IV

- Using polynomials to model $g_1(x_i)$ and $g_0(x_i)$, we have

$$\begin{aligned} E[D_i \mid x_i] &= \gamma_{00} + \gamma_{01} x_i + \gamma_{02} x_i^2 + \ldots + \gamma_{0p} x_i^p \\ &\quad + \gamma_0^* T_i + \gamma_1^* x_i T_i + \gamma_2^* x_i^2 T_i + \ldots + \gamma_p^* x_i^p T_i \end{aligned}$$

- T_i, as well as the interaction terms $\{x_i T_i, x_i^2 T_i, \ldots, x_i^p T_i\}$, are instruments for D_i in,

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \rho D_i + \eta_i$$

- Simple fuzzy RD uses only T_i as an instrument in an additive model:

$$\begin{aligned} D_i &= \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \ldots + \gamma_p x_i^p + \pi T_i + \xi_{1i} \\ Y_i &= \mu + \kappa_1 x_i + \kappa x_i^2 + \ldots + \kappa_p x_i^p + \rho\pi T_i + \xi_{2i}, \end{aligned}$$

  where $\kappa_j = \beta_j + \rho\gamma_j$ for $j = 1, \ldots, p$. This also works after centering, i.e., swapping $\tilde{x}_i$ for x_i

Nonparametric and Fuzzy

- As with sharp RD, identification in the fuzzy case turns on the discontinuity in $T_i = 1(x_i \geq x_0)$: The fuzzy RF is sharp
- The nonparametric reduced-form near x_0 is

$$E[Y_i \mid x_0 < x_i < x_0 + \delta] - E[Y_i \mid x_0 - \delta < x_i < x_0] \simeq \rho\gamma_0^*$$

- Similarly, for the nonparametric first stage:

$$E[D_i \mid x_0 < x_i < x_0 + \delta] - E[D_i \mid x_0 - \delta < x_i < x_0] \simeq \gamma_0^*$$

Therefore

$$\lim_{\delta \to 0} = \frac{E[Y_i \mid x_0 < x_i < x_0 + \delta] - E[Y_i \mid x_0 - \delta < x_i < x_0]}{E[D_i \mid x_0 < x_i < x_0 + \delta] - E[D_i \mid x_0 - \delta < x_i < x_0]} = \rho$$

- The sample analog of (16) is a Wald estimator using T_i as an instrument for D_i in a δ — neighborhood of x_0
- IK (2011) derive an optimal bandwidth for this ratio; In practice, this is close to the optimal bw for the RF

Just Do it . . . in Stata

- Nonparametric and fuzzy LLR/RD is easily done with conventional 2SLS software
- Estimate these first and second stages using kernel weights and the bw of your choice:

$$\begin{aligned} D_i &= \gamma_{00} + \gamma_{01}\tilde{x}_i + \gamma_0^* T_i + \gamma_1^* \tilde{x}_i T_i + \xi_{1i} \\ Y_i &= \alpha + \rho D_i + \beta_{01}\tilde{x}_i + \beta_1^* T_i \tilde{x}_i + \eta_i \end{aligned}$$

- Note that this 2SLS estimator uses a linear version of the Fancy Fuzzyfirst Stage, (12)
- Something funny about this version of fuzzy: the T_i main effect is an instrument, yet the interaction $\tilde{x}_i T_i$ is a control
  - Dong (2012) shows that $\tilde{x}_i T_i$ is also available as an instrument; without T_i in the first stage, this is the "regression kink design" (Card, Lee, and Zhu, 2009)