Matching, with special Emphasis on Propensity Score Matching
Overview of Next Three Days

- Tom is going to outline (a) the design under consideration for the next three days, (b) broad strategies for ameliorating it, and (c) the roles of intact group and individual case matching
- Peter is going to do a technical exposition of PSM
- Will will walk us through one data set you will use that tests a contingency theory of the validity of matching and contrasts it with OLS analyses
- I am going to walk you through some empirical studies of design elements that improve matching
“Work Horse” Design: The Most Common NECGD

- _O _X_O_
  O   O

Two design elements here that could help causally: pretest and control group not formed at random

- The change in the non-equivalent comparison group is meant to indicate how much change there would have been in the treated group had it not been treated

- Said differently, the posttest means are assumed to be similar conditional on adjusting for pretest difference

- How plausible is this?
Chief Internal Validity Threats to rule out with this Design

Selection–Maturation – differences in growth rates
Selection-History -- Local History
Selection–Instrumentation – measures change differentially by group
Selection-Statistical Regression – Aschenfelter dip
Selection-Attrition – creates even more complex group selection differences

So why not make the groups equivalent to start with? The intuition here is that the problem goes away.
Four Generic Strategies to Equate Non-Equivalent Groups

- DiD with one pretest time point only
- OLS – e.g., ANCOVA
- Matching of some kind
- Refuse to accept the design limitation we have imposed and add design elements—e.g., more pretest time points for CITS. Thursday and Fri on this.
- Because we are discussing this simple design for so long does not mean we like it. The time spent reflects its use in the social sciences, not its value for identification in practice. We generally see it as a last resort to be avoided wherever possible.
Strategy One: Difference in Differences

- The key assumption here is what? -- at posttest the unobserved potential outcome in the treatment group is different from the observed posttest comparison mean by the same group difference observed at pretest.

- When is this plausible and implausible in general and in Education in particular?

- How can you diagnose that the assumption is true?

- Is this why we like to add many pretest time points to DiD in order to convert it into CITS and observe baseline time trend differences.
Strategy 2: OLS, incl ANCOVA

- How is its model of change different from DiD?
- Consider time trend differences in achievement within just the comparison group—fan spread hypothesis—describe fan spread in original metric and in standardized form
- Then in the treatment group
- The analysis assumes that the difference between groups would have changed DIFFERENTIALLY over time in the same way the pattern of change averaged within groups
ANCOVA continued

- IS this a better counterfactual for some situations?
- Diagnostic tests of similar within-group patterns are?
- No diagnostic test of whether between-group differences mirror within-group ones.
- Can the pattern of within-group differences fail to reflect the pattern of between-group differences?
- How often does this happen?
- We’re going to get a sense of this when we compare how well OLS does relative to PSM.
Very Generic Matching and the Intuition

- Assume the two groups have “somehow” been matched to make initially equivalent on pretest.
- In DiD perspective, the null hypothesis is no difference at posttest, given no difference at pretest.
- In ANCOVA perspective, the assumption is that within-group trends more likely to be similar and the average of them is more likely like what the between-group difference would have been.
- Is it a better null?
- What about the possibility of groups with similar immediate pretest means but different trends?
Escaping from the Design Limitation imposed – Friday in particular

- Add more pretest time points
- Add more comparison groups
- Add non-equivalent dependent variables
- Add even atypical and small RCTs
Two often Forgotten Facets of Matching

- Matching that causes bias rather than reduces or eliminates it
- Matching populations prior to matching individual cases if needed
Bad Matching for Comparability

Simple Regression illustrated with one group

Frequency of such regression in our society

It is a function of all imperfect correlations

Amount of Regression gross $f(\text{unreliability/population difference})$

Simple one-on-one case matching from low degree of overlap in the next slide
T: Decreases
C: Increases

Both forces makes T look ineffective
The Net Effect is…

- If either treatment decreases or controls increase due to regression, then bias results
- If both change in opposite directions, then the bias is exacerbated
- Matching individual units from extremes is not recommended
In this Predicament You Might…

- The Cicirelli Head Start Evaluation had this problem, concluding Head Start was harmful.
- LISREL reanalyses by Magidson using multiple measures at pretest led to different conclusion, likely by reducing unreliability.
- Reliability is higher using aggregate scores like schools--but beware here as with effective schools literature or effective teacher (value-added) literature.
Escaping the Predicament through the Study Sampling Design: Matching Intact Comparison Groups from least different Population that is feasible
Select intact comparison group with more overlap than below

T: Decreases
C: Increases

Both forces makes T look ineffective
Better is to get out of the Pickle

- Don’t match from extremes! Use intact groups instead, selecting for comparability/overlap on pretest
- Comer Detroit study as an education example
- Sample schools in same district – why?
- Match by multiple years of prior achievement and by race composition of school body--why?
- Choose multiple matches per intervention school, bracketing so that one close match above and the other below intervention schools
The Value of Intact Group Matching?

- We explore the ability of NECGDs to achieve unbiased causal conclusions by means of within-study comparisons (WSCs).

- And so we need to explain what a WSC is, given that its purpose is to test how closely the results of an (often statistically adjusted) NECGD approximate those of an RCT with the same or equivalent treatment group.
What is a Within-Study Comparison?

Overall Population
sampled/selected into

Randomized Experiment
randomly assigned to

Control Group

Treatment Group

ITT of RCT
What is a Within-Study Comparison?

Overall Population sampled/selected into Randomized Experiment randomly assigned to

Randomized Experiment randomly assigned to

Control Group  Treatment Group  Comparison Group

ITT of RCT  ITT of OS
What is a Within-Study Comparison?

Overall Population sampled/selected into

Randomized Experiment randomly assigned to

Control Group  Treatment Group  Comparison Group

ITT of RCT  ?  ITT of OS
Logic of WSC

Function of the WSC design is to compare ES from RCT and NECGD sharing the same treatment group

- Since treatment group is a constant, it can be ignored. Thus, the test is of how the control group is formed -- at random or not.

- WSCs contribute to learning when NECGDs more closely approximate RCT results, assuming that RCT offers best estimate of an unbiased effect.

- We will deal with them quite a bit this week as part of the evidence for what is a better QED.

- On many QED topics: Now, consequences of selecting an intact local group
Criteria for better WSC Studies

- Clear variation in mode of forming control group—random or not
- RCT merits being considered a “gold standard” when it demonstrably meets assumptions, but probabilistic
- Experiment and non-experiment difference is not confounded with 3rd variables like measurement
- The quasi-experiment should be a good example of its type—otherwise one compares a good experiment to a poor quasi-experiment
Criteria continued

- The experiment and quasi-experiment should estimate the same causal quantity— not LATE vs ATE or ITT vs TOT
- Criteria for inferring correspondence of results should be clear
- The non-experimental analyses should be done blind to the experimental results
- Historical change in probability of meeting these criteria
Examples of bias-reducing Effects of Intact Local Group Matching

- Bloom, Michalopoulos et al.
- Aiken, West et al.
- Diaz & Handa
Bloom, Michalopoulos et al

Issue in Bloom et al. is: Will the randomly and non-randomly formed control groups differ after standard adjustments for differences in mean, slope or other covariates across 8 pretest observation points.

Our Issue is: Is it possible to sample non-equivalent comparison groups so as to eliminate, or at least minimize, the initial group differences one would otherwise have to adjust away.
The Context

- RCT is on job training at 11 sites
- Bloom et al restrict their analysis to 5 within-state comparisons, 4 of them within-city. This last essentially matches on city characteristics, including many (most?) job market ones.
- The non-random comparison cases are all those from job training centers in same city as treatment sites. Thus, use everyone at control site switching design.
- The outcome is measured the same way at all sites. So these factors not confounded with treatment
Results: 3 within-city Samples

Mean Quarterly Earnings: Oklahoma City

Mean Quarterly Earnings: Riverside

Mean Quarterly Earnings: Detroit
What you see in Graphs

- Hardly differ at all --- one advantage of TS is that we can see group differences
- Statistical tests confirm no differences in intercept or slope
- In these cases, equivalence of randomly and non-randomly formed comparison groups is achieved thru sampling design alone
- Thus, no need for statistical tests to render them “equivalent”
- What is the generalizability of selecting comparison cases this way?
Two other Sites

Portland--sample size smallest and least stable

Detroit vs Grand Rapids--a within-state but not within-city comparison. Hence, this is not a very local comparison
Bloom et al. Results (2)

Mean Quarterly Earnings: Grand Rapids and Detroit

Mean Quarterly Earnings: Portland

Quarters from Random Assignment

Mean Earnings (1995 $)

- Grand Rapids (control) — Detroit (comparison)

- Portland West Office (control) — Portland East and North Offices (comparison)
Here you see

- TS are not equivalent overall
- TS especially not equivalent around the crucial intervention point
- Thus use of a random or non-random control group would produce different results
- 10 types of statistical analyses were used to make the series equivalent:
Results of these Analyses

- OLS, propensity scores, Heckman selection models, random growth models—all failed to give the same results as the experiment under these conditions.
- But the more the pretest time points, the less the bias.
- Only the random growth model took advantage of the TS nature of the data.
- Why did it fail too?
Selecting Intact Groups locally matched on pretest outcomes

- Without intending it, Bloom et al’s choice of within-city non-equivalent controls achieved comparability with the randomly formed experimental controls. That is, there was

- No bias across 3 of the 4 within-city samples; nor for the weighted average of all 4 sites

- So, overlap on observables was achieved through the sampling design alone, precluding need for statistical adjustments, incl PSM of individual cases

- Remember: There was bias in across-state comparisons used in the paper that could not be adjusted away with the data and models used
Selecting Intact Groups with Maximal Overlap: 2nd Example

- Aiken et al. ASU--effects of remedial writing
- Sample selection in their Quasi-Experiment was from the same range of ACTs and SATs as in their experiment
- Differed by failure of researchers to contact them over summer and later registration
- What will the role of unobserved variables correlated with these two features differentiating randomly and non-randomly formed control units?
- Note that the measurement framework was identical in the experiment and quasi-experiment, as were the intervention and control group experiences other than remedial writing
Results

The randomly and non-randomly formed comparison groups did not differ on SAT/CAT or on 2 pretest writing measures

- So close correspondence of groups on observables without any need for subsequent adjustment; and

- So in Q-E, OLS controls for pretest add power but do not to reduce bias since there was little to control away.

- Results for multiple choice writing test in SD units = \( RCT = 59 \) and \( NECGD = .57 \)--both sig.

- Results for essay \( RCT = .06 \) and \( NECGD = .16 \) - both non-sig
3rd Example: Diaz & Handa (2006)

- Progresa: Design = Matched Villages with and without the program, and eligibles assigned to T and C status within these villages

- Village eligibility depended on being in bottom quintile on test of material resources. But for a variety of reasons some eligible villages were not in RCT. These are the non-equivalent control villages.

- Within them were otherwise eligible families, and these became the non-equivalent controls. How similar were they to controls in RCT?
Were the Results unbiased in the Non-Equivalent Village Comparison

The eligible families in the non-equivalent villages were generally not different from the eligible comparison families in the treated villages on pretests. (They were also not initially different from treated families in treated villages).

There were a few small differences, though, and these were added as controls in the final outcome analysis.

The outcome results remained the same in the RCT as in the design with eligible non-equivalent families from non-equivalent but otherwise eligible villages.
Implications of all Three Studies

Aiken et al and Bloom et al. created non-equivalent control groups that were not different on observables from the treatment group and produced unbiased results.

These observables included a pretest on the same scale as the outcome.

Diaz and Handa created considerable overlap but there were some differences. They did not affect outcome.

Moral is: How you sample intact groups can eliminate all observed bias without anything else being needed – like individual case matching.
But what about Unobservables?

- We can never know without an RCT as the causal benchmark comparison.
- But chances of little bias are greater with intact groups that do not differ on the pretest and
- Do not differ on the known strongest correlates of the outcome, especially those weakly correlated with pretest measures of the outcome.
- The lesson here is: How you sample non-equivalent comparison cases is crucial for bias reduction. It is not just a case of how you adjust for observed non-equivalence.
What is a Local, Focal, Non-Equivalent Intact Control Group 1

- Local because…
- Focal because…
- Non-Equivalent because…
- Intact because…
What is a Local, Focal, Non- Equivalent Intact Control Group 2

- Identical Twins
- Fraternal Twins
- Siblings
- Successive Grade Cohorts within the same School
- Same Cohort within different Schools in same District
- Same Cohort within different Schools in different districts in same state
- Same Cohort within different schools in different states, etc.
The Trade Offs here are…

- **Identity vs. Comparability.** We cannot assume that siblings are identical, for example. They have some elements of non-shared genes and environments.

- **Comparability vs. Contamination.** Closer they are in terms of space and presumed receptivity to the intervention, the greater the risk of contamination.

- To reduce an inferential threat is not to prevent it entirely.