Gov 2002: 5. Matching

Matthew Blackwell

October 1, 2015

Where are we? Where are we going?

- Discussed randomized experiments, started talking about observational data.
- Last week: no unmeasured confounders and how it identifies the ATE.
- This week: one way to estimate causal effects under no unmeasured confounders, matching.
- Coming up: other ways of estimating causal effects: weighting, regression.

1/ Identification for Matching

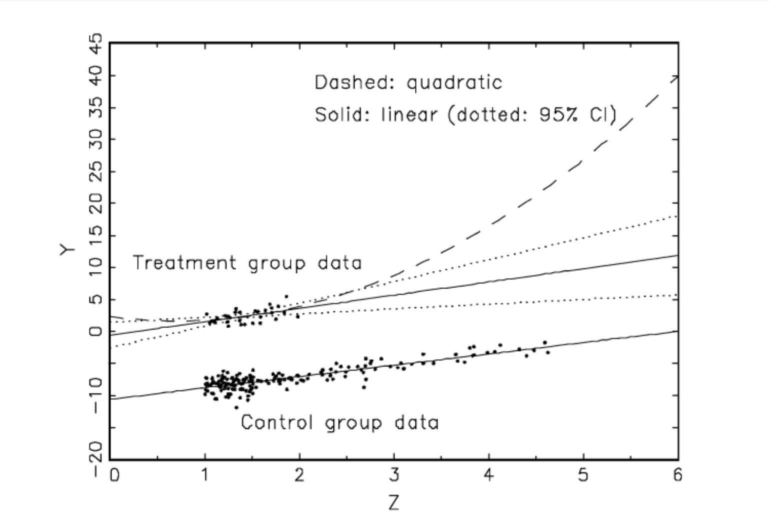
Why match?

- No unmeasured confounding holds, but we need to adjust for X_i
- Common solution: write a parametric model for $\mathbb{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbb{E}[Y_i(d)|X_i] = X_i'\beta$
  - Regression, MLE, Bayes, etc.
- But this model might be wrong $\rightsquigarrow$ wrong causal estimates.
- **Matching** has two benefits:
  1. Can simplify the analysis of causal effects
  2. Reduces dependence of estimates on parametric models.

Model dependence

- Use parametric models $M_1, \ldots, M_J$ to estimate the ATE: $\widehat{\tau}_j$
    - include X_i , X_i^2, $\log(X_i)$, $X_i \times Z_i$, X_i^4, etc
- **Model dependence**: large variation in the estimates, $\widehat{\tau}_j$
- Why does this occur?
    - Parametric models extrapolate to regions with only treated or only control.
    - Modeling assumptions will greatly affect these extrapolations.

Model dependence example



Caution

- No unmeasured confounders identifies the causal effect.
- Matching doesn't make this more plausible
- $\rightsquigarrow$ Matching doesn't justify a causal effect.
- Matching just allows for relatively nonparametric ways of estimating the causal effect.
- Sekhon:

  > *Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive.*

Assumptions

1. No unmeasured confounders:

$$D_i \perp\!\!\!\perp \left(Y_i(0), Y_i(1)\right) | X_i$$

2. Positivity/overlap:

$$0 < \mathbb{P}(D_i = 1 | X_i = x) < 1$$

Exact matching

- Let X_i take on a finite number of values, $x \in \mathscr{X}$.
- Let $\mathbb{I}_t = \{1, 2, \dots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$
  - Randomly select one of these control units to be the match, indicated $j(i)$.
- Let $\mathbb{I}_c = \{j(1), \dots, j(N_t)\}$ be the set of matched controls.
- Last, discard all unmatched control units.
- The distribution of X_i will be **exactly** the same for treated and matched control:

$$\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$$

Identification of the ATT

- Let's show that the ATT is identified if the data is exactly matched:

$$
\begin{aligned}
\tau_{\text{ATT}} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \\
&= \underbrace{E[Y_i|D_i = 1]}_{\text{consistency}} - E[Y_i(0)|D_i = 1] \\
&= E[Y_i|D_i = 1] - \underbrace{\sum_{x \in \mathscr{X}} E[Y_i(0)|X_i = x, D_i = 1] \Pr(X_i|D_i = 1)}_{\text{iterated expectations}} \\
&= E[Y_i|D_i = 1] - \sum_{x \in \mathscr{X}} E[Y_i(0)|X_i = x, \underbrace{D_i = 0}_{\text{n.u.c.}}] \Pr(X_i|D_i = 1) \\
&= E[Y_i|D_i = 1] - \sum_{x \in \mathscr{X}} \underbrace{\mathbb{E}[Y_i|X_i}_{\text{consis.}} = x, D_i = 0] \Pr(X_i|D_i = 1) \\
&= E[Y_i|D_i = 1] - \sum_{x \in \mathscr{X}} E[Y_i|X_i = x, D_i = 0] \underbrace{\Pr(X_i|D_i = 0, \mathbb{I}_c)}_{\text{exact matches}}
\end{aligned}
$$

Weakening the identification assumptions

- No unmeasured confounders, consistency, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to **conditional mean independence** (CMI):

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0]$$

- Two nice features of CMI:
  1. Only have to make assumptions about $Y_i(0)$ not $Y_i(1)$
  2. Only places restrictions on the means, not other parts of the distribution (variance, skew, kurtosis, etc)

Analyzing exactly matched data

- How do we analyze the exactly matched data?
- Dead simple difference in means:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_i - \frac{1}{N_c} \sum_{j \in \mathbb{I}_c} Y_j$$

- Notice that we matched 1 treated to 1 control exactly, so we have:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - Y_{j(i)})$$

- $\rightsquigarrow$ average of the within matched-pair differences.

Variance with exact matches

- Notice that with 1:1 treated/control matching, similar to a matched-pair experiment.
- Variance estimators are a little different for these.
- Variance estimator:

$$\widehat{\mathbb{V}}(\widehat{\tau}_m) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - Y_{j(i)} - \widehat{\tau}_m \right)^2$$

- In-sample variance of the within-pair differences.

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.
- Let S be a **matching solution**: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.
- Suppose that this procedure produces balance:

$$D_i \perp\!\!\!\perp X_i | S$$

- This implies that no unmeasured confounders holds in that subset:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp D_i | S$$

- Balance is checkable $\rightsquigarrow$ are D_i and X_i related in the matched data?

2/ Matching details

The matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)
4. Check balance
5. Repeat (1)-(4) until balance is acceptable
6. Calculate the effect of the treatment on the outcome in the matched dataset.

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to "impute" missing potential outcome.
- For $i \in \mathbb{I}_t$ define:

$$\widehat{Y}_i(0) = \frac{1}{M} \sum_{j \in J_M(i)} Y_j$$

- New estimator for the effect:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \widehat{Y}_i(0))$$

- Under no unmeasured confounding, $\widehat{Y}_i(0)$ is a good predictor of the true potential outcome under control, Y_i.

Number of matches

- How many control matches should we include?
  - Small $M \rightsquigarrow$ small sample sizes
  - Large $M \rightsquigarrow$ worse matches (each additional match is further away).
- If M varies by treated unit, need to weight observations to ensure balance.

With or without replacement

- **Matching with replacement**: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.
- Drawbacks:
  - Inference is more complicated.
  - $\rightsquigarrow$ need to account for multiple appearances with weights.
  - Potentially higher uncertainty (using the same data multiple times = relying on less data).

3/ Distance metrics

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.
- How do we define distance/similarity on X_i if it is high dimensional?
- We need a **distance metric** which maps two covariates vectors into a single number.
  - Lower values $\rightsquigarrow$ more similar values of X_i.
  - Choice of distance metric will lead to different matches.

Exact distance metric

- **Exact**: only match units to other units that have the same exact values of X_i.

$$D_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$

Propensity scores, redux

- **Propensity scores**: $e(X_i) = \mathbb{P}(D_i = 1 | X_i)$
- Remember that we only need to condition on the **true** PS:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp D_i | e(X_i)$$

- $\rightsquigarrow$ sufficient to balance on the **true** propensity score.
- Rubin et al. have shown that PS matching has good properties if covariates are roughly normal.
    - Though, see King and Nielsen working paper on PS matching.

Propensity score distances

- Intuitive to use the raw absolute differences in the PS:

$$D_{ij} = |e(X_i) - e(X_j)|$$

- Better to use the **linear propensity score**, $\text{logit}(e(X_i)) = X_i\beta$:

$$D_{ij} = |\text{logit}(e(X_i)) - \text{logit}(e(X_j))|$$

- Accounts for non-linearity in the substantive differences in the PS:
  - 0.05 → 0.10 is more important than 0.50 → 0.55.

True vs. estimated propensity scores

- Balancing properties of the PS depend on knowing the true PS function, $e(x)$.
- In observational studies we never know the true PS $\rightsquigarrow$ estimate it $\hat{e}(x)$.
- Is balancing on $\hat{e}(X_i)$ sufficient? **No idea!!**
  - Have to check if X_i is actually balanced.
  - Somewhat deflates the benefits of PS matching/balancing.
- $\rightsquigarrow$ "propensity score tautology"

Euclidean distance

- The **Euclidean distance metric** just uses the sum of the normalized distances for each covariate.
  - "Closeness" is standardized across covariates.
- Suppose that $X_i = (X_{i1}, \dots, X_{iK})'$, so that there are K covariates.
- Then the Euclidean distance metric is:

$$D_{ij} = \sqrt{\sum_{k=1}^{K} \frac{(X_{ik} - X_{jk})^2}{\widehat{\sigma}_k}}$$

- Here, $\widehat{\sigma}_k$ is the standard deviation of the kth variable:

$$\widehat{\sigma}_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ik} - \bar{X}_k)$$

Mahalanobis distance

- **Mahalanobis distance**: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
  - Easy to get close on correlated covariates $\leadsto$ downweight.
  - Harder to get close on uncorrelated covariates $\leadsto$ upweight.
- Metric:

$$D_{ij} = \sqrt{(X_i - X_j)'\widehat{\Sigma}^{-1}(X_i - X_j)}$$

- $\widehat{\Sigma}$ is the estimated variance-covariance matrix of the observations:

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T$$

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.
  - Dropping those matches (treated and control) improves balance.
  - Dropping treated units changes the quantity of interest.
- Implementation: a **caliper**, c, is the maximum distance we would accept:

$$D_{ij} = \begin{cases} \sqrt{(X_i - X_j)'\widehat{\Sigma}^{-1}(X_i - X_j)} & \text{if } |\text{logit}(e(X_i)) - \text{logit}(e(X_j))| \leq c \\ \infty & \text{if } |\text{logit}(e(X_i)) - \text{logit}(e(X_j))| > c \end{cases}$$

4/ Estimands and Matching Methods

Estimands

- Matching easiest to justify for the ATT.
  - Dropping control units doesn't affect this identification.
- Can also identify the ATC by finding matched treated units for the controls.
- Combine the two to get the ATE:

$$\tau = \tau_{ATT}\mathbb{P}(D_i = 1) + \tau_{ATC}\mathbb{P}(D_i = 0)$$

- Estimated:

$$\widehat{\tau} = \widehat{\tau}_{ATT}\left(\frac{N_t}{N}\right) + \widehat{\tau}_{ATC}\left(\frac{N_c}{N}\right)$$

Moving the goalposts

- **Common support**: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1|X_i < 18) = 0$).
  - Empirical: no/extremely few treated units in a sea of controls.
  - Solution: restrict analysis to common support (dropping treated and controls).
- **Moving the goalposts**: dropping treated units.
  - We move away from being able to identify the ATT.
  - Now it's the ATT in the matched subsample (sometimes called the feasible ATT).
  - Good to be clear about this.

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is **nearest neighbor**: "Find the control unit with the smallest distance metric."
- Do the same for all treated units.
- What about ties?
  - Randomly choose between them.

Order effects

- With NN matching, the order matters.
  - Treated: $X_1 = 0.5$ and $X_2 = 0.7$
  - Control: $X_3 = 0.8$ and $X_4 = 0.15$
  - Match 1 first: 1← 3 and then 2 ← 4, $\sum D_{ij} = 0.85$
  - Match 2 first: 2← 3 and then 1 ← 4, $\sum D_{ij} = 0.45$
  - NN is "greedy."
- **Optimal matching**: Finds the matching solution that minimizes overall distance.
  - Find $j(1), \ldots, j(N_t)$ to minimize: $\sum_{i=i}^{N_t} D_{ij(i)}$
  - Tends to find the same set of controls, just matched to different treated groups.
  - Useful for finding matched pairs.

GenMatch

- We could extend Mahalanobis distance to weight covariates by their importance to producing balance.
  - Bad balance after matching $\rightsquigarrow$ tweak these weights and re-match.
  - Can we automate this?
- **GenMatch** is a genetic algorithm that attempts to find the Mahalanobis weights that produce the best balance.
  - Randomly a population of different starting vectors (weight vectors).
  - Evaluate the "fitness" of each vector (the balance it produces).
  - Randomly create new population focused on the vectors with best balance.
  - Mimics natural selection.

CEM

- **Coarsened Exact Matching** is akin to stratification.
  - Stratify/coarsen all continuous covariates into bins: X_i^*
  - X_i^* now has a discrete number of possible values.
  - Exact match on X_i^*: keep data in strata $X_i^* = x^*$ if there are at least 1 treated and 1 control with $X_i^* = x^*$, drop others.
  - Use uncoarsened data, X_i, in the analysis stage.
- Example:
  - Coarsen years of education into: (less than H.S., H.S. degree, some college, B.A./B.S., Advanced degree)
- Benefits:
  - Allows you to control the amount of imbalance up front
  - Coarser $\rightsquigarrow$ more imbalanace, finer $\rightsquigarrow$ less imbalance

Assessing balance

- All matching methods seek to minimize balance:

$$\mathbb{P}(X_i = x|D_i = 1, S) = \mathbb{P}(X_i = x|D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:
  - Differences-in-means/medians, standardized.
  - Quantile-quantile plots/KS statistics for comparing the entire distribution of X_i.
  - L_1: multivariate histogram.

5/ Post-matching Analysis

What to do with matched data?

- You matched and pruned the data of non-matches, now what?
- Exact matching: simple difference in means.
- Inexact matching: there will be **matching discrepancy**:

$$W_i = X_i - X_{j(i)}$$

- If balance is good then W_i should be quite small, but could still be large and produce bias.
- Matching discrepancy will grow with the dimension of X_i

Bias of inexact matching

- Let $\mu_c(x) = \mathbb{E}[Y_i(0)|X_i = x]$ be how the mean of $Y_i(0)$ changes as a function of X_i.
- Take a single matched pair produced by matching:

$$\widehat{\tau}_{mi} = Y_i - Y_{j(i)}$$

- We hope this estimates $\tau(X_i)$, but there is actually bias:

$$\mathbb{E}[\widehat{\tau}_{mi}|D_i = 1, X_i, X_{j(i)}] = \tau(X_i) + \underbrace{(\mu_c(X_i) - \mu_c(X_{j(i)}))}_{\text{unit-level bias}}$$

- If X_i has a big effect on the mean of $Y_i(0)$ then this bias could be big!

Bias-corrected estimators

$$B_i = \mu_c(X_i) - \mu_c(X_{j(i)})$$

- How do we get rid of this bias?
  - Estimate it, $\widehat{B}_i$, and subtract it off, $(Y_i - Y_{j(i)}) - \widehat{B}_i$
- Specify a parametric model for $\mu_c(x) = \alpha_c + x'\beta_c$ and estimate $\widehat{\beta}_c$ from the control data:

$$\widehat{B}_i = \widehat{\mu}_c(X_i) - \widehat{\mu}_c(X_{j(i)}) = (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Specification of $\mu_c(x)$ will matter less after matching.
- Create bias-corrected/adjusted imputations for $Y_i(0)$:

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Plug this into the same estimator:

$$\widehat{\tau}_{m,bc} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - \widehat{Y}_i(0)\right)$$

- Variance estimation for this quantity is easiest without replacement.
- Simply take the variance of the within-match differences:

$$\widehat{\mathbb{V}}[\widehat{\tau}_m] = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - \widehat{Y}_i(0) - \widehat{\tau}_{m,bc}\right)^2$$

Fully pooled model

- What if we simply run our original analysis model on the pooled, matching data:

$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{D}_i + \tilde{X}_i' \beta_p + \nu_i$$

- $\tilde{Y}_i$ is the $2 \times N_t$ matched treated and control units stacked.
- $\widehat{\tau}_p$ from OLS on this model is a bias-corrected estimate where we assume that:

$$\mu_c(x) = \mu_t(x)$$

- Still corrects for some of the residual bias left over from the matching.
- SEs from these models might make additional assumptions (homoskedasticity, etc).

6/ Wrap-up

Conclusion

- Matching is a technique to reduce model dependence and avoid parametric modeling assumptions when no unmeasured confounders holds.
- Lots of different ways to match, each has advantages and disadvantages.
- Pay careful attention to the quantity of interest when you drop units.
- Next week:
  - Weighting methods and posttreatment bias.