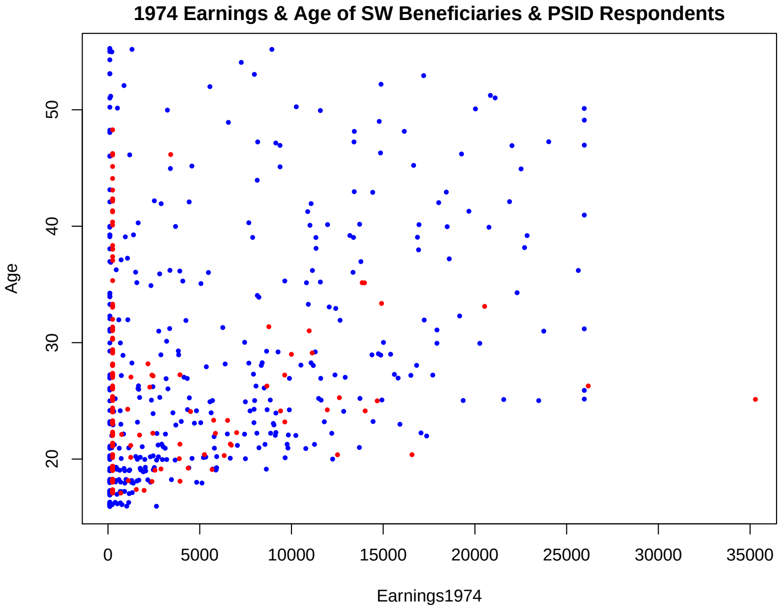
Lecture 6: Matching and Weighting Methods to Estimate Causal Effects Under CIA

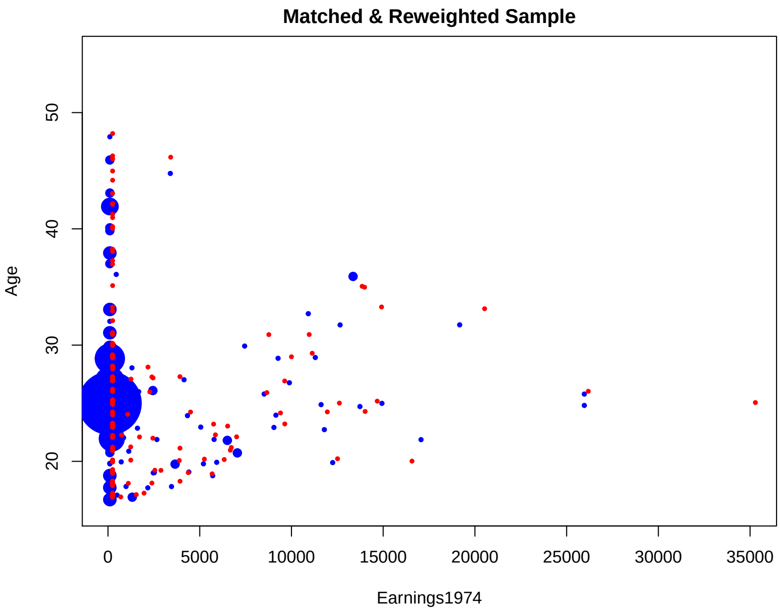
POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

February 8, 2022

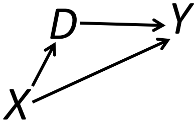
Overview

- Looked at conditioning strategies that allow us to choose identifying sets of covariates.
- Recognize problems with classical regression modeling — specification challenges and odd reweighting.
- Today we look at matching and weighting.
- Keeping to the low dimensional case (just a few Xs).
- High dimensional stuff will be later in semester.


1974 Earnings & Age of SW Beneficiaries & PSID Respondents
Age
50
40
30
20
0
5000
10000
15000
20000
25000
30000
35000
Earnings1974

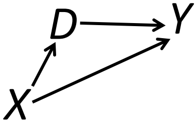

Matched & Reweighted Sample
Age
Earnings1974
50
40
30
20
0
5000
10000
15000
20000
25000
30000
35000

Matching and Causal Effects



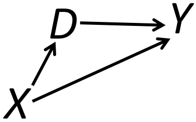
- Given "conditional independence" with X_i, what's the best way to "condition" on X_i?

Matching and Causal Effects



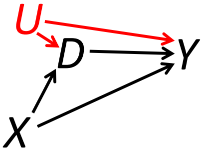
- Given "conditional independence" with X_i, what's the best way to "condition" on X_i?
- One way is to use X_i in a multiple regression, but we have seen some problems arising from classical regression.

Matching and Causal Effects



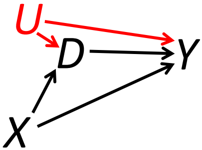
- Given "conditional independence" with X_i, what's the best way to "condition" on X_i?
- One way is to use X_i in a multiple regression, but we have seen some problems arising from classical regression.
- Matching and weighting relieve us of certain modeling assumptions in exploiting conditional independence.

Matching and Causal Effects



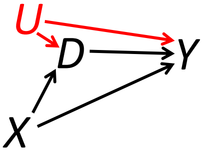
- NB: matching and weighting do not make conditional independence any more plausible than was the case for regression!

Matching and Causal Effects



- NB: matching and weighting do not make conditional independence any more plausible than was the case for regression!
- Causal identification from the data (e.g., via CIA) must be established first using your substantive understanding of the situation and good choices in selecting covariates.

Matching and Causal Effects



- NB: matching and weighting do not make conditional independence any more plausible than was the case for regression!
- Causal identification from the data (e.g., via CIA) must be established first using your substantive understanding of the situation and good choices in selecting covariates.
- Thus, matching and weighting are not, per se, ways to "identify" a causal effect.

Identification of ATT Under CMI

We can actually proceed with an identifying assumption that is somewhat less restrictive than CIA:

Recall ATT:

$$\begin{aligned} \mathrm{E}[\rho_i | D_i = 1] &= \mathrm{E}[Y_{1i} - Y_{0i} | D_i = 1] \\ &= \underbrace{\mathrm{E}[Y_{1i} | D_i = 1]}_{\text{observable}} - \underbrace{\mathrm{E}[Y_{0i} | D_i = 1]}_{\text{counterfactual}} \end{aligned}$$

Identification of ATT Under CMI

We can actually proceed with an identifying assumption that is somewhat less restrictive than CIA:

Recall ATT:

$$\begin{aligned} \mathrm{E}[\rho_i|D_i = 1] &= \mathrm{E}[Y_{1i} - Y_{0i}|D_i = 1] \\ &= \underbrace{\mathrm{E}[Y_{1i}|D_i = 1]}_{\text{observable}} - \underbrace{\mathrm{E}[Y_{0i}|D_i = 1]}_{\text{counterfactual}} \end{aligned}$$

Suppose

$$\mathrm{E}[Y_{0i}|X_i, D_i = 1] = \mathrm{E}[Y_{0i}|X_i, D_i = 0] \text{ and } \Pr[D_i = 1|X_i] < 1.$$

Thus, we assume *mean independence for baseline outcomes* conditional on X_i. Call this “conditional mean independence” (CMI) .

Identification of ATT Under CMI

We can actually proceed with an identifying assumption that is somewhat less restrictive than CIA:

Recall ATT:

$$\begin{aligned} \mathrm{E}\left[\rho_i | D_i = 1\right] &= \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\ &= \underbrace{\mathrm{E}\left[Y_{1i} | D_i = 1\right]}_{\text{observable}} - \underbrace{\mathrm{E}\left[Y_{0i} | D_i = 1\right]}_{\text{counterfactual}} \end{aligned}$$

Suppose

$$\mathrm{E}\left[Y_{0i} | X_i, D_i = 1\right] = \mathrm{E}\left[Y_{0i} | X_i, D_i = 0\right] \text{ and } \Pr[D_i = 1 | X_i] < 1.$$

Thus, we assume *mean independence for baseline outcomes* conditional on X_i. Call this "conditional mean independence" (CMI) .
Of course, CIA $\Rightarrow$ CMI.

Identification of ATT Under CMI

Under CMI (or CIA), we can approximate the counterfactual mean for the ATT:

$$\begin{aligned}
\mathrm{E}[Y_{1i}|D_i=1]-\mathrm{E}[Y_{0i}|D_i=1] &= \mathrm{E}[Y_{1i}|D_i=1]-\mathrm{E}_{X|D_i=1}\{\mathrm{E}[Y_{0i}|X_i,D_i=1]\} \\
&= \mathrm{E}[Y_{1i}|D_i=1]-\mathrm{E}_{X|D_i=1}\{\mathrm{E}[Y_{0i}|X_i,D_i=0]\},
\end{aligned}$$

where $\mathrm{E}_{X|D_i=1}\{\}$ means to take the expectation with respect to $F(x|D_i=1)$. This is identified in the data, and so CMI identifies the ATT.

Identification of ATT Under CMI

Let's study this expression:

$$\mathrm{E}[Y_{1i}|D_i = 1] - \mathrm{E}_{X|D_i=1}\{\mathrm{E}[Y_{0i}|X_i, D_i = 0]\} \tag{1}$$

Identification of ATT Under CMI

Let's study this expression:

$$\mathrm{E}[Y_{1i}|D_i = 1] - \mathrm{E}_{X|D_i=1}\{\mathrm{E}[Y_{0i}|X_i, D_i = 0]\} \tag{1}$$

- Sample mean of Y_i for the treated is unbiased and consistent for $\mathrm{E}[Y_{1i}|D_i = 1]$. So the first term is simple.

Identification of ATT Under CMI

Let's study this expression:

$$\mathrm{E}\left[Y_{1i}|D_i=1\right]-\mathrm{E}_{X|D_i=1}\left\{\mathrm{E}\left[Y_{0i}|X_i,D_i=0\right]\right\} \tag{1}$$

- Sample mean of Y_i for the treated is unbiased and consistent for $\mathrm{E}\left[Y_{1i}|D_i=1\right]$. So the first term is simple.
- The second term is a bit trickier. Now by definition,

$$\mathrm{E}_{X|D_i=1}\left\{\mathrm{E}\left[Y_{0i}|X_i,D_i=0\right]\right\}=\int_X \mathrm{E}\left[Y_{0i}|X_i=x,D_i=0\right]dF(x|D_i=1).$$

- Let's see how this could be estimated non-parametrically...

Exact Matching

- Suppose X_i is discrete.

Exact Matching

- Suppose X_i is discrete.
- Let,
$$\widehat{p}(x|D_i = 1) \equiv \frac{\sum_{i=1}^{N} 1(X_i = x) D_i}{\sum_{i=1}^{N} D_i}$$
be the empirical mass in S at x for those with $D_i = 1$.

Exact Matching

- Suppose X_i is discrete.
- Let,
$$\widehat{p}(x|D_i = 1) \equiv \frac{\sum_{i=1}^{N} 1(X_i = x)D_i}{\sum_{i=1}^{N} D_i}$$
be the empirical mass in S at x for those with $D_i = 1$.
- By WLLN this is consistent for $F(x|D_i = 1)$ (cf. Glivenko-Cantelli thm.).
- Then, the sample analogue to $\mathrm{E}_{X|D_i=1}\{\mathrm{E}[Y_{0i}|X_i, D_i = 0]\}$ is given by,
$$\sum_{x \in \{X_i : D_i = 1\}} \frac{\sum_{i=1}^{N} 1(X_i = x)(1 - D_i)Y_i}{\sum_{i=1}^{N} 1(X_i = x)(1 - D_i)} \widehat{p}(x|D_1 = 1).$$
- If we can exact match on $X_i = x$, then we can compute this directly.

Exact Matching

- Most matching algorithms don't actually compute $\widehat{p}(x|D_i = 1)$, rather the marginalization with respect to $F(x|D_i = 1)$ arises as a byproduct of the matching solution.

Exact Matching

- Most matching algorithms don't actually compute $\widehat{p}(x|D_i = 1)$, rather the marginalization with respect to $F(x|D_i = 1)$ arises as a byproduct of the matching solution.
- Assume exact matching is possible for all x values for members of the treatment group.
- Define M_i as an indicator for whether a unit from the control group is included by some matching algorithm. Then, $M_i = 1$ if the unit is kept by the matching algorithm, and $M_i = 0$ if it is discarded.

Exact Matching

- Most matching algorithms don't actually compute $\widehat{p}(x|D_i = 1)$, rather the marginalization with respect to $F(x|D_i = 1)$ arises as a byproduct of the matching solution.
- Assume exact matching is possible for all x values for members of the treatment group.
- Define M_i as an indicator for whether a unit from the control group is included by some matching algorithm. Then, $M_i = 1$ if the unit is kept by the matching algorithm, and $M_i = 0$ if it is discarded.
- Then, we can set as an objective for the matching algorithm to set the M_i values such that for x values for the treatment group,

$$\underbrace{\widehat{p}(x|D_i = 1)}_{\text{actual treated cov. dist'n}} = \underbrace{\frac{\sum_{i=1}^{N} 1(X_i = x)(1 - D_i)M_i}{\sum_{i=1}^{N}(1 - D_i)M_i} \equiv \widehat{p}(x|D_i = 0, M_i = 1)}_{\text{desired control cov. dist'n}} \tag{2}$$

How do we ensure (2)? Well, we can match treated and control units with respect to X_i and either

1. select one control unit for every treated unit at each value in $\{X_i | D_i = 1\}$, or
2. take all control units at each value in $\{X_i | D_i = 1\}$ for which $\widehat{p}(x | D_i = 1) > 0$ and then weight them to ensure (2).

The result would be a stratified set of treated and control observations where the (weighted) distribution of controls would match the distribution of the treated over the strata.

How do we ensure (2)? Well, we can match treated and control units with respect to X_i and either

1. select one control unit for every treated unit at each value in $\{X_i | D_i = 1\}$, or
2. take all control units at each value in $\{X_i | D_i = 1\}$ for which $\widehat{p}(x | D_i = 1) > 0$ and then weight them to ensure (2).

The result would be a stratified set of treated and control observations where the (weighted) distribution of controls would match the distribution of the treated over the strata.

Under CMI, the (weighted) mean of Y_i's for the matched controls proxies for $\mathrm{E}[Y_{0i} | D_i = 1]$.

Exact Matching

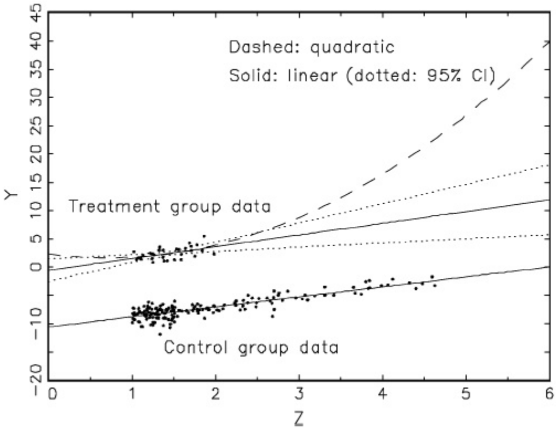


Fig. 4 An illustration of how the degree of extrapolation bias is more severe (and model dependent) than interpolation bias.

Exact Matching

Condition (2) is exact "covariate balance." If (2) or something very close to it holds, then we should see,

- Covariate means and variances are equal across treated and (weighted) matched controls.
- Quantile-quantile plots should be linear.
- Covariate histograms should mirror each other (tested by e.g., Kolgomorov-Smirnov distance, Kullback-Leibler divergence, distance correlation, etc.).

Exact Matching

Condition (2) is exact "covariate balance." If (2) or something very close to it holds, then we should see,

- Covariate means and variances are equal across treated and (weighted) matched controls.
- Quantile-quantile plots should be linear.
- Covariate histograms should mirror each other (tested by e.g., Kolgomorov-Smirnov distance, Kullback-Leibler divergence, distance correlation, etc.).

If not, then we reject (2) and it is not clear that our matching solution will yield an unbiased estimate of $\mathrm{E}_{X|D_i=1}\left\{\mathrm{E}\left[Y_{0i}|X_i, D_i = 0\right]\right\}$.

Continuous Covariates

- With continuous covariates, exact matching not feasible.

Continuous Covariates

- With continuous covariates, exact matching not feasible.
- We can approximate it by selecting "nearby neighbors" in X_i.

Continuous Covariates

- With continuous covariates, exact matching not feasible.
- We can approximate it by selecting “nearby neighbors” in X_i.
- Convergence to the truth at a good rate requires smoothness of potential outcome in the covariates (Abadie & Imbens 2006).

Continuous Covariates

- With continuous covariates, exact matching not feasible.
- We can approximate it by selecting "nearby neighbors" in X_i.
- Convergence to the truth at a good rate requires smoothness of potential outcome in the covariates (Abadie & Imbens 2006).
- In finite samples you may have poor matches, which suggests the desire to impose constraints (e.g., "caliper" constraints).

Continuous Covariates

- With continuous covariates, exact matching not feasible.
- We can approximate it by selecting "nearby neighbors" in X_i.
- Convergence to the truth at a good rate requires smoothness of potential outcome in the covariates (Abadie & Imbens 2006).
- In finite samples you may have poor matches, which suggests the desire to impose constraints (e.g., "caliper" constraints).
- Selecting nearby neighbors is not straightforward with high dimensional X_i.
- As dimension of X_i grows, holding N fixed, matching can become unreliable (Samii et al. 2016).

Continuous Covariates

- With continuous covariates, exact matching not feasible.
- We can approximate it by selecting "nearby neighbors" in X_i.
- Convergence to the truth at a good rate requires smoothness of potential outcome in the covariates (Abadie & Imbens 2006).
- In finite samples you may have poor matches, which suggests the desire to impose constraints (e.g., "caliper" constraints).
- Selecting nearby neighbors is not straightforward with high dimensional X_i.
- As dimension of X_i grows, holding N fixed, matching can become unreliable (Samii et al. 2016).
- One way to evaluate the quality of a proposed matching solution in terms of the balance that results.

Exact Matching

- There is no one “right” way to check* balance. General advice is that you check as many ways as possible, including covariate-by-covariate and using multivariate tests.

Exact Matching

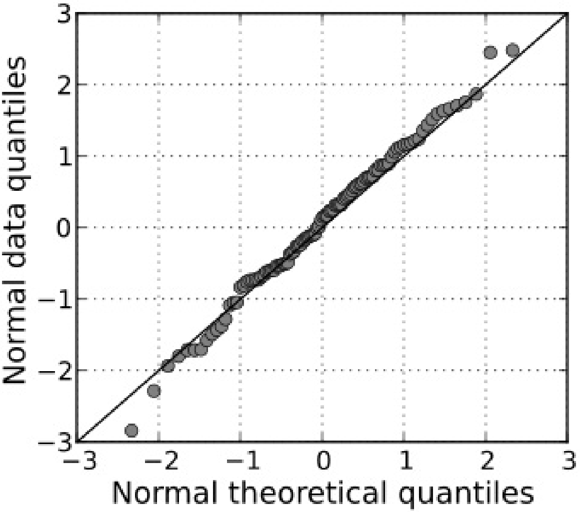
- There is no one "right" way to check* balance. General advice is that you check as many ways as possible, including covariate-by-covariate and using multivariate tests.
- Standard practice is to tweak the matching algorithm until you achieve satisfactory balance by all of these measures. This does not *imply* that identifying assumptions are valid.

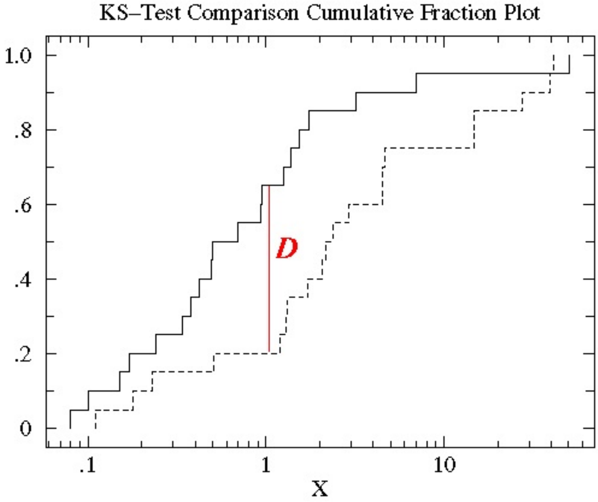
Exact Matching

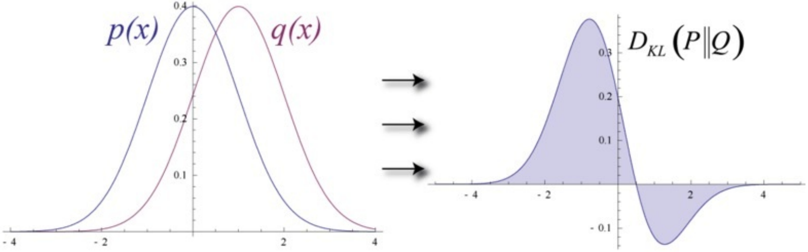
- There is no one "right" way to check* balance. General advice is that you check as many ways as possible, including covariate-by-covariate and using multivariate tests.
- Standard practice is to tweak the matching algorithm until you achieve satisfactory balance by all of these measures. This does not *imply* that identifying assumptions are valid.
- *Beware the "balance test fallacy" (Imai et al. 2008): don't confuse a *lack of statistical power*, and thus "failure to reject zero difference", with balance.

Exact Matching

- There is no one “right” way to check* balance. General advice is that you check as many ways as possible, including covariate-by-covariate and using multivariate tests.
- Standard practice is to tweak the matching algorithm until you achieve satisfactory balance by all of these measures. This does not *imply* that identifying assumptions are valid.
- *Beware the “balance test fallacy” (Imai et al. 2008): don’t confuse a *lack of statistical power*, and thus “failure to reject zero difference”, with balance.
- To *test* for balance, you would want to use an “equivalence test” to see if you could reject some degree of *imbalance*, e.g.,
  $H_0 : |E[X|D = 1] - E[X|D = 0]| \geq c$ vs.
  $H_A : |E[X|D = 1] - E[X|D = 0]| < c$
  (cf. Hartman and Hidalgo 2017).


Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


KS–Test Comparison Cumulative Fraction Plot
1.0
.8
.6
.4
.2
0
D
.1
1
10
X



Original Gaussian PDF's

KL Area to be Integrated

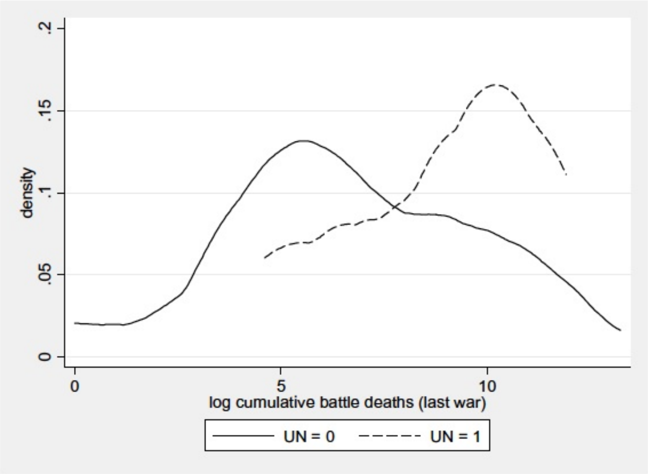


Figure 1. Kernel densities of log cumulative battle deaths in cases where the UN intervened (UN = 1) and where it did not (UN = 0).

(Gilligan and Sergenti, 2008)

Table 1. Balance statistics, post-conflict sample

Variable		Mean treated	Mean control	t–test p–value	K–S test p–value	Var. ratio (Tr/Co)	Mean Std. eCDF diff
Log (Cumulative battle deaths)	Before matching	8.982	6.647	0.001	0.004	0.582	0.226
	After matching	8.982	8.373	0.218	0.27	1.213	0.111
Duration of last war	Before matching	80.526	50.279	0.141	0.042	0.908	0.192
	After matching	80.526	66.842	0.254	0.242	0.790	0.109
Ethnic fractionalization	Before matching	49.213	56.504	0.299	0.344	1.154	0.097
	After matching	49.213	40.726	0.214	0.226	0.894	0.082
Log population size	Before matching	8.754	9.509	0.004	0.008	0.304	0.177
	After matching	8.754	8.894	0.384	0.216	0.971	0.094
Log mountainous	Before matching	2.797	2.221	0.105	0.09	0.883	0.121
	After matching	2.797	2.381	0.337	0.436	0.932	0.086
Log military personnel	Before matching	3.254	3.874	0.081	0.066	0.495	0.137
	After matching	3.254	3.462	0.231	0.678	1.224	0.077
Log GDP per Capita	Before matching	6.588	6.56	0.921	0.858	1.160	0.051
	After matching	6.588	6.518	0.810	0.74	0.779	0.075
Polity	Before matching	−2.579	−0.838	0.194	0.38	0.661	0.088
	After matching	−2.579	−2.263	0.775	0.868	1.059	0.045

(Gilligan and Sergenti, 2008)

Creating Pseudo-Experiments via Balance Directly

A theoretical aside: Many matching algorithms actually pursue *balance* directly, without necessarily considering the need to create exact matches. This may be justified as follows,

- Let $\mathscr{M}$ be the matching solution. Then, by Dawid (1979, Lemma 4.3),

$$\underbrace{D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}) | X_i, \mathscr{M}}_{\text{restricted form of CIA; assumed}} \quad \text{and} \quad \underbrace{D_i \perp\!\!\!\perp X_i | \mathscr{M}}_{\text{balance; testable}}$$

  implies

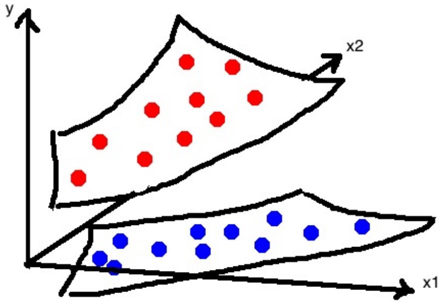
$$D_i \perp\!\!\!\perp ((Y_{1i}, Y_{0i}), X_i) | \mathscr{M}.$$

- Thus, *given* this particular form of CIA, balance provides a sufficient condition for having an "as if" randomized experiment on the notional/synthetic subpopulation defined by $\mathscr{M}$.

Creating Pseudo-Experiments via Balance

- Once we appreciate that overall balance over treatment values *in itself* is a goal, a whole new vista opens up.
- Rather than thinking in terms of matching per se, we can think in terms of reweighting all control units to best reproduce the covariate distribution of the treated.
- See Diamond & Sekhon (2012) on genetic matching to maximize balance.
- See Hainmueller (2011), Imai & Ratkovic (2014), Athey, Imbens & Wager (2018), Arbour & Dimmery (2019) on balancing weights — estimation addressed later with IPW.

Continuous Covariates



Three main approaches to continuous, high dimension X_i:

1. Minimum multivariate distance matching.
2. Propensity score matching.
3. Coarsened exact matching.

Current methods combine these in different ways.
Other choices include "many to one" matching and "calipers."

Minimizing Multivariate Distance

- Most common metric is Mahalanobis distance (MD).
- Similar to Euclidean distance (ED), but MD is scale free and incorporates correlations.

Minimizing Multivariate Distance

- Most common metric is Mahalanobis distance (MD).
- Similar to Euclidean distance (ED), but MD is scale free and incorporates correlations.
- MD between two covariate vectors, X_i and X_j, given by,

$$d_M(X_i, X_j) = \sqrt{(X_i - X_j)' \mathbf{S}^{-1} (X_i - X_j)},$$

  where $\mathbf{S}$ is a covariance matrix.
- The inverse weighting by § means that you put more emphasis on "novel" variation in the $\mathbf{X}$ space (that is, variation that is non-redundant relative to other variables in $\mathbf{X}$).

Minimizing Multivariate Distance

- Most common metric is Mahalanobis distance (MD).
- Similar to Euclidean distance (ED), but MD is scale free and incorporates correlations.
- MD between two covariate vectors, X_i and X_j, given by,

$$d_M(X_i, X_j) = \sqrt{(X_i - X_j)' \mathbf{S}^{-1} (X_i - X_j)},$$

  where $\mathbf{S}$ is a covariance matrix.
- The inverse weighting by § means that you put more emphasis on "novel" variation in the $\mathbf{X}$ space (that is, variation that is non-redundant relative to other variables in $\mathbf{X}$).
- Mahalanobis distance *matching* selects nearby neighbors, where proximity is based on $d_M(X_i, X_j)$.
- A popular package known as GenMatch starts with Mahalanobis distance, but then tweaks the $\mathbf{S}$ matrix to maximize balance (Sekhon, 2009 and papers cited therein; Diamond & Sekhon 2012).

Minimizing Multivariate Distance

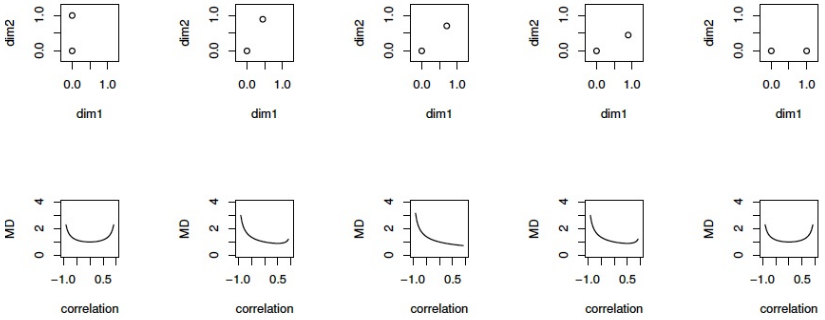


Figure 1: MD by corr for different cases where ED is always 1

Propensity Score Matching

- Famous result by Rosenbaum & Rubin (1983) suggests that we can reduce the problem to a single dimension.

Propensity Score Matching

- Famous result by Rosenbaum & Rubin (1983) suggests that we can reduce the problem to a single dimension.
- Let $e(X_i) = \Pr[D_i = 1|X_i] = \mathrm{E}[D_i|X_i]$ (R&R's notation), the "propensity score".

Propensity Score Matching

- Famous result by Rosenbaum & Rubin (1983) suggests that we can reduce the problem to a single dimension.
- Let $e(X_i) = \Pr[D_i = 1|X_i] = \mathrm{E}[D_i|X_i]$ (R&R's notation), the "propensity score".
- Suppose CIA.
- Then, for $j = 0, 1$,

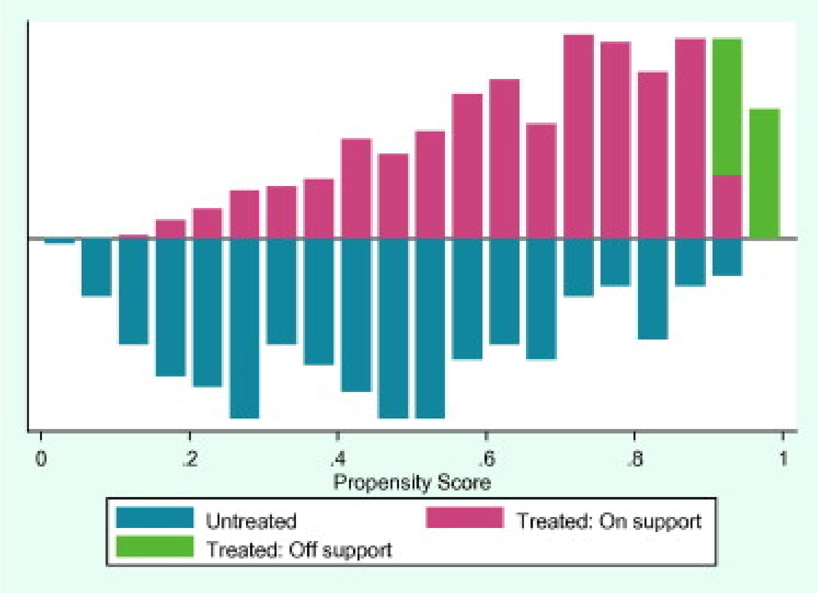
$$\begin{aligned}
\Pr[D_i = 1|Y_{ji}, e(X_i)] &= \mathrm{E}[D_i|Y_{ji}, e(X_i)] \\
&= \mathrm{E}_X\left\{\mathrm{E}[D_i|Y_{ji}, e(X_i), X_i]|Y_{ji}, e(X_i)]\right\} \\
&= \mathrm{E}_X\left\{\mathrm{E}[D_i|Y_{ji}, X_i]|Y_{ji}, e(X_i)]\right\} \text{ by sufficiency of } X_i \\
&= \mathrm{E}_X\left\{\mathrm{E}[D_i|X_i]|Y_{ji}, e(X_i)]\right\} \text{ by CIA} \\
&= \mathrm{E}_X\left\{e(X_i)|Y_{ji}, e(X_i)]\right\} = e(X_i) = \Pr[D_i = 1|e(X_i)].
\end{aligned}$$

- Thus under CIA, $D_i \perp\!\!\!\perp Y_{ji}|e(X_i)$.
- The problem of high-dimensional X is reduced to working with scalar $e(X_i)$.

Propensity Score Matching

- We don't know $e(X_i)$ but can estimate $\widehat{e}(X_i)$ (in olden days, logistic regression, but now fancier, less restrictive methods).
- If $\widehat{e}(X_i)$ is accurate, conditioning on it via matching, subclassification, or regression modeling is sufficient to fully exploit CIA or CMI.
- In practice, though, some questions as to how well this actually works, or at least whether it throws away much potential efficiency (King & Nielsen 2018).

Propensity Score Matching



Propensity Score Weighting

- An alternative use for propensity scores is to construct the inverse-probability-of-treatment weighted estimator, based on,

$$\rho = \mathrm{E}[Y_{1i} - Y_{0i}] = \mathrm{E}_X\left\{\mathrm{E}\left[\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\middle| X_i\right]\right\} \text{ (by CIA)}$$
$$= \mathrm{E}\left[\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right]$$

- The analogue estimator would be,

$$\hat{\rho} = \frac{1}{N}\sum_i \frac{Y_i D_i}{e(X_i)} - \frac{1}{N}\sum_i \frac{Y_i(1-D_i)}{1-e(X_i)}$$

- A stabilized ("Hajek"-type) estimator is,

$$\hat{\rho}_S = \frac{\sum_i D_i Y_i / e(X_i)}{\sum_i D_i / e(X_i)} - \frac{\sum_i (1-D_i) Y_i / (1-e(X_i))}{\sum_i (1-D_i)/(1-e(X_i))}$$

- Similar formulas available for ATT and ATC (MHE, pp. 82-83; Busso et al. 2014).

Propensity Score Weighting

- Classical inverse-propensity score weighting (IPW) does not necessarily optimize with respect to bias minimization, although has attractive efficiency properties (Hirano et al. 2003).
- For a broad class of outcome DGPs, one can improve the bias-reduction property of IPW estimators by targeting covariate balance more directly (Hainmueller, 2011; Imai & Ratkovic, 2014; Athey, Imbens & Wager, 2018; Arbour & Dimmery, 2019).

Propensity Scores More Generally

- Propensity score methods are appealing because they establish a *predictive estimation target*—namely, the propensity score itself.
- This allows for "regularization," which is helpful in high dimensional settings (Samii et al. 2016; Eckles and Bakshy 2017).
- More on this later in the semester.

Coarsened Exact Matching

A direct approximation to exact matching:

- Coarsen your covariates.
- Create stratification cells using covariates.
- For ATT, within each stratification cell, weight control group members so that their weighted total equals number of treated group in that cell.
- Allows you to pre-specify the amount of imbalance you are willing to accept ahead of time.

Coarsened Exact Matching

Region	Age bracket	Gender	No. Tr.	No. Con.	Con. Wgt.
North	Under 35	Female	5	53	5/53
North	Under 35	Male	4	44	1/11
North	Over 35	Female	4	32	1/8
North	Over 35	Male	5	31	5/31
South	Under 35	Female	4	41	4/41
South	Under 35	Male	0	36	0
South	Over 35	Female	4	32	1/8
South	Over 35	Male	4	21	4/21
⋮	⋮	⋮	⋮	⋮	⋮

Matching procedure has us discard the 36 control units in the South, Under 35, Male cell.

Comparing Approaches

- King and Nielsen (2018) find MDM and CEM to be more robust than 1-to-1 nearest neighbor matching on pscores estimated via logistic regression, Jann (2017) finds the real problem in King & Nielsen data was 1-to-1 nearest neighbor matching not pscores.
- Diamond and Sekhon (2012) find that GenMatch offers significant improvement over propensity-score and Mahalanobis-distance matching.
- Busso et al. (2014) find that stabilized weighting outperforms propensity-score and Mahalanobis-distance matching in a range of scenarios.
- Hainmueller (2011) finds that reweighting based on minimizing KL-divergence is optimal in certain respects.
- Imai & Ratkovic (2014), Athey, Imbens, & Wager (2018), and Arbour & Dimmery (2019) find improvement from IPW with covariate balance targets.

Remarks

- Generalization to multi-valued but discrete treatments are straightforward: just match all covariate distributions to that of one of the treatment groups.

Remarks

- Generalization to multi-valued but discrete treatments are straightforward: just match all covariate distributions to that of one of the treatment groups.
- Generalizations to continuous treatments are possible:
  - For propensity score matching or weighting, see Hirano & Imbens (2004), Imai & van Dyk (2004), and Imbens (2000) for methods that rely on parametric treatment models, and then Fong, Hazlett & Imai (2018) and Arbour & Dimmery (2019) for non-parametric approaches.
  - One can apply CEM straightforwardly too.

Remarks

- Focus here has been low dimensional cases (e.g., X has at most a dozen or two elements).
- If CIA is only plausible with high dimensional X, you have a problem.
- As the number of requires X variables increases, overlap (e.g., $0 < Pr[D_i = 1|X_i] < 1$ for the ATE, and similar for ATT and ATC) becomes more difficult to satisfy.
- D'Amour et al. (2017) propose a "curse of dimensionality" for the overlap condition: as the number of needed X variables grows, the degree to which they can be allowed to differ across treatment and control becomes very small.

Remarks

> *For most researchers, the math obscures the assumptions. Without an experiment, a 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. (Sekhon, 2009, p. 503)*

- At the end of the day, these simple matching and weighting methods should be seen as ways to "mop up" imbalances *given* otherwise exogenous variation that makes CIA plausible.
- Recall the crucial thought experiment necessary to test CIA:

  > *How could it be that two units that are identical with respect to all meaningful background factors nonetheless receive different treatment?*

- Your answer to this question is your source of identification. Matching only allows you to exploit it with fewer modeling assumptions.