Matching methods for categorical and continuous treatments

Spencer Woody

SDS 384: Causal Inference Methodology

April 30, 2020

The University of Texas at Austin
Department of Statistics
and Data Sciences
College of Natural Sciences

Introduction

Let $\mathcal{Z}$ be the sample space for the treatment assigment Z.

- Most of our course has only considered binary treatments.

  $$Z \in \mathcal{Z} = \{0, 1\}$$

  Causal estimands are comparisons of counterfactual outcomes $Y_i(Z_i = 1)$ vs $Y_i(Z_i = 0)$

- Now we consider nonbinary treatments
  - **Categorical** (possibly ordinal): $\mathcal{Z} = \{1, 2, \ldots, k\}$, e.g. multiple treatment arms
  - **Continuous**: $\mathcal{Z} \subseteq \mathbb{R}$, e.g. drug dose

Causal estimands in the Rubin Causal Model

Categorical treatment with k categories:

- There are $\binom{k}{2}$ pairwise comparisons of treatment assignment

$$Y_i(Z_i = j) \text{ vs. } Y_i(Z_i = j') \text{ for } j, j' \in \{1, 2, \ldots, k\}$$

Continuous treatment:

- Finite difference comparison

$$Y_i(Z_i = z) \text{ vs. } Y_i(Z_i = z') \text{ for } z \neq z'$$

- *Average dose-response function*

$$\mu(z) = \mathrm{E}[Y_i(z)]$$

Generalized propensity score

Let X be the vector of observed covariates.

Definition: Generalized propensity score[*] (GPS)

Let $r(z, x)$ be the conditional density (or mass function) of the treatment given the covariates:

$$r(z, x) = f_{Z|X}(z \mid x)$$

The generalized propensity score is $R = r(Z, X)$.

Note that R may be a vector, e.g. if Z is categorical.

[*] Imbens (2000); Hirano and Imbens (2004)

Overlap

Assumption: Overlap

$$r(z, x) = f_{Z|X}(z \mid x) > 0 \quad \forall z \in \mathcal{Z}$$

Generalized propensity score

Assumption: Weak unconfoundedness

$Y(z) \perp\!\!\!\perp Z \mid X$ for all $z \in \mathcal{Z}$

Note: this does not require joint independence of all potential outcomes $\{Y(z)\}_{z \in \mathcal{Z}}$

Similar to Rosenbaum and Rubin (1983) for the case of binary Z, Imbens (2000) and Hirano and Imbens (2004) demonstrate:

Theorem: Weak unconfoundedness given the GPS

If weak unconfoundedness holds given X, then, for every z,

$$f_Z(z \mid r(z, X), Y(z)) = f_Z(z \mid r(z, X)).$$

Existing methods mostly rely on GPS

- Imai and van Dyk (2004): Subclassify on GPS, then take average over subclasses
- Hirano and Imbens (2004): Parametric model for $Y \mid Z, R$, then marginalize over R
- Robins et al. (2000): IPTW estimator using GPS

Disadvantage: These methods rely on parametric assumptions

Work on matching for nonbinary treatments is relatively new

Outline

Presenting methodologies from three papers:

(i) Nattino et al. (2020): Compare treatment effects across 3 treatment arms (*categorical*)

(ii) Sävje et al. (2017): Generalized full matching for multiple treatment categories (*categorical*)

(iii) Wu et al. (2020): Use matching to estimate average dose-response (*continuous*)

Nattino et al. (2020)

Nattino et al. (2020)

Goal: Compare effectiveness of trauma centers as measured by *emergency department mortality*, for three classes of trauma center,

- level 1 trauma center (TC I)
- level 2 trauma center (TC II)
- nontrauma center (NTC)

Counterfactual of interest: *"... the key research question is whether TC II is a justified investment of limited trauma care resources. If trauma patients treated at TC II had, instead, been treated at TC I or NTC, would their outcomes have been different?"* – p. 1

Assumptions

Let $Y_\ell^{(z)}$ for $z \in \{1, 2, 3\}$ denote the counterfactual outcome (1 for death, 0 for survival) for unit $\ell = 1, \ldots, N$.

The observed value is $Y_\ell = Y_\ell^{\text{obs}} = \sum_{z=1}^{3} I(Z_\ell = z) Y_\ell^{(z)}$

$\mathbf{X}_\ell$ is a vector of pre-treatment covariates

1. SUTVA: no interference between units, no multiple versions of same treatment
2. Positivity

$$0 < \Pr(Z_\ell = z \mid Y_\ell^{(1)}, Y_\ell^{(2)}, Y_\ell^{(3)}, \mathbf{X}_\ell) < 1 \quad \forall z \in \{1, 2, 3\}$$

3. Strong ignorability

$$Z_\ell \perp\!\!\!\perp Y_\ell^{(1)}, Y_\ell^{(2)}, Y_\ell^{(3)} \mid \mathbf{X}_\ell$$

Three-way matching

Idea: replicate conventional block randomization design, using triplets of units containing all treatment assigments $z = 1, 2, 3$

Let $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$ denote the sets of indices of subjects in subject. We will create $S = \min\{n_1, n_2, n_3\}$ matched triplets. Will match on variables $\mathbf{V}$ (either covariates $\mathbf{X}$ or the GPS).

- Define a distance metric $d^3(i, j, k)$, $i \in \mathcal{I}, j \in \mathcal{J}, k \in \mathcal{K}$ as a function of $\mathbf{V}_i$, $\mathbf{V}_j$ and $\mathbf{V}_k$, with additivity property

$$d^3(i, j, k) = d^2(i, j) + d^2(i, k) + d^2(j, k)$$

- Denote set of possible matches as $\mathcal{M} = \{i, j_i, k_i\}_{i \in \mathcal{I}}$, where the units j_i and k_i are matched to units i
- Goal is to find $\mathcal{M}$ to minimize $D(\mathcal{M}) = \sum_{i \in \mathcal{I}} d^3(i, j_i, k_i)$

Triplet matching algorithm

Rough outline:

(i) Select two treatment groups arbitrarily, and optimally match them into pairs

(ii) Optimally match units in the third treatment group to each of the pairs from step (i) (keeping previous pairs fixed)

(iii) Switch the two fixed treatment groups, and then optimally match units from the third treatment group

(iv) Iterate through step (iii) until total distance cannot be decreased further

This method produces sets of matched triplets, but each step only requires two-way matching

Inference on mortality differences

Denote treatment and outcome vectors for triplet $s = 1, \ldots, S$ as $\mathbf{Z}_s = \{Z_{s1}, Z_{s2}, Z_{s3}\}$ and $\mathbf{Y}_s = \{Y_{s1}, Y_{s2}, Y_{s3}\}$

- Fisher's sharp null hypothesis of no effect at all: $H_0 = Y_{sr}^{(1)} = Y_{sr}^{(2)} = Y_{sr}^{(3)}$ for subject $r = 1, 2, 3$.
- Consider two comparisons:
  (1) NTC vs TC overall ($z = 1$ vs $z = 1, 2$ combined)
  (2) TC II vs TC I ($z = 2$ vs $z = 3$)

Use Fisher randomization based inference

Comparing NTC vs TC overall

- Mantel-Haenszel test statistic is no. of events in NTC

$$\sum_{s=1}^{S} \sum_{r=1}^{3} I(Z_{sr} = 1)Y_{sr}$$

- Under null hypothesis, each subject is equally likely to be the patient assigned to NTC within each triplet. Conditioning on $m_s = \sum_{r=1}^{3} Y_{sr}$, define p_s as $p_s = \Pr(\sum_{r=1}^{3} I(Z_{sr} = 1)Y_{sr} = 1 \mid \sum_{r=1}^{3} = m_s)$. $p_s = 0, 1/3, 2/3, 1$ for $m_s = 0, 1, 2, 3$
- The standardized statistic is

$$T_{\text{MH}} = \frac{\sum_{s=1}^{S} \sum_{r=1}^{3} I(Z_{sr} = 1)Y_{sr} - \sum_{s=1}^{S} p_s}{\sqrt{\sum_{s=1}^{S} p_s(1 - p_s)}}$$

Under the null hypothesis, $T_{\text{MH}} \sim N(0, 1)$ as $S \to \infty$

Comparing TC I vs TC II overall

- McNemar test statistic is no. of events in TC II

$$\sum_{s=1}^{S}\sum_{r=1}^{3} I(Z_{sr}=3)Y_{sr}$$

- Under null hypothesis, each subject is equally likely to be the patient assigned to NTC within each triplet.
  Conditioning on $n_s = \sum_{r\in\{2,3\}} Y_{sr}$, define q_s as
  $q_s = \Pr(\sum_{r=1}^{3} I(Z_{sr}=3)Y_{sr} = 1 \mid \sum_{r\in\{2,3\}} = n_s)$.
  $q_s = 0, 1/2, 1$ for $n_s = 0, 1, 2$
- The standardized statistic is

$$T_{\text{MH}} = \frac{\sum_{s=1}^{S}\sum_{r=1}^{3} I(Z_{sr}=1)Y_{sr} - \sum_{s=1}^{S} q_s}{\sqrt{\sum_{s=1}^{S} q_s(1-q_s)}}$$

Under the null hypothesis, $T_{\text{MN}} \sim N(0,1)$ as $S \to \infty$

Results on trauma center mortality data

- Estimate GPS using multinomial regression
- Match subjects on the basis of the linear predictor of GPS (log-odds)
- Results in 3158 matched triplets

Results: covariate balance after matching

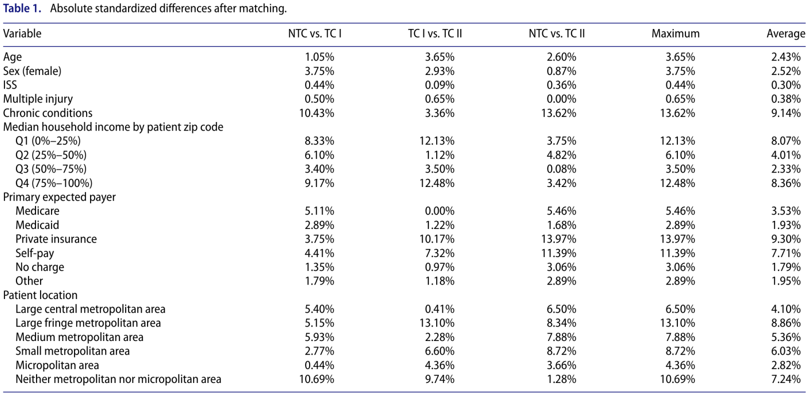
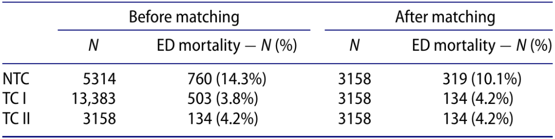


Table 1. Absolute standardized differences after matching.

Variable	NTC vs. TC I	TC I vs. TC II	NTC vs. TC II	Maximum	Average
Age	1.05%	3.65%	2.60%	3.65%	2.43%
Sex (female)	3.75%	2.93%	0.87%	3.75%	2.52%
ISS	0.44%	0.09%	0.36%	0.44%	0.30%
Multiple injury	0.50%	0.65%	0.00%	0.65%	0.38%
Chronic conditions	10.43%	3.36%	13.62%	13.62%	9.14%
Median household income by patient zip code					
Q1 (0%–25%)	8.33%	12.13%	3.75%	12.13%	8.07%
Q2 (25%–50%)	6.10%	1.12%	4.82%	6.10%	4.01%
Q3 (50%–75%)	3.40%	3.50%	0.08%	3.50%	2.33%
Q4 (75%–100%)	9.17%	12.48%	3.42%	12.48%	8.36%
Primary expected payer					
Medicare	5.11%	0.00%	5.46%	5.46%	3.53%
Medicaid	2.89%	1.22%	1.68%	2.89%	1.93%
Private insurance	3.75%	10.17%	13.97%	13.97%	9.30%
Self-pay	4.41%	7.32%	11.39%	11.39%	7.71%
No charge	1.35%	0.97%	3.06%	3.06%	1.79%
Other	1.79%	1.18%	2.89%	2.89%	1.95%
Patient location					
Large central metropolitan area	5.40%	0.41%	6.50%	6.50%	4.10%
Large fringe metropolitan area	5.15%	13.10%	8.34%	13.10%	8.86%
Medium metropolitan area	5.93%	2.28%	7.88%	7.88%	5.36%
Small metropolitan area	2.77%	6.60%	8.72%	8.72%	6.03%
Micropolitan area	0.44%	4.36%	3.66%	4.36%	2.82%
Neither metropolitan nor micropolitan area	10.69%	9.74%	1.28%	10.69%	7.24%

Results: Comparisons between trauma centers

Table 2. Results of the outcome analysis.



	Before matching		After matching	
	N	ED mortality — *N* (%)	*N*	ED mortality — *N* (%)
NTC	5314	760 (14.3%)	3158	319 (10.1%)
TC I	13,383	503 (3.8%)	3158	134 (4.2%)
TC II	3158	134 (4.2%)	3158	134 (4.2%)

- NTC vs TC (TC I and TC II combined): $T_{\text{MH}} = 11.45$, $p < 0.001$
- TC I vs TC II: $T_{\text{MN}} = 0$, $p = 0.500$
- Assess sensitivity to unobserved confounding (Rosenbaum, 1987) gives $\Gamma_{\text{MH}} = 2.34$.

Sävje et al. (2017)

Sävje et al (2017)

- **Hypothesis:** social norms influence citizens' propensity to vote (Gerber, Green, and Larimer, 2008).
- **Goal:** study effectiveness of a postcard intervention in increasing voter turnout. There are six total treatment conditions.
- Introduce *generalized full matching*, which extends full matching to the case of categorical treatment with k levels.

Gerber et al. prescreened voters to be included in the study, so the original results were not generalizeable to the entire population.

Full matching

This paper generalizes full matching[†]:

- Construct groups of units that are as homogeneous as possible
- Require that each group has at least one unit of each treatment condition
- So far, only developed for case of binary treatment

All units are matched to a subclass, hence the term “full”

[†] Rosenbaum (1991); Hansen (2004); Stuart and Green (2008)

Notation

- Denote the sample of n units by $\mathbf{U} = \{1, 2, \ldots, n\}$
- Unit i is assigned to treatment condition $W_i \in \{1, 2, \ldots, k\}$
- The vectors $\mathbf{w}_x = \{i : W_i = x\}$ denote sets of units assigned to a given treatment condition
- Matched groups are denoted by $\mathbf{m}$, and the union of matched groups is $\mathbf{M} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$
- Define an objective function $L : \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of possible matches

Match group constraints

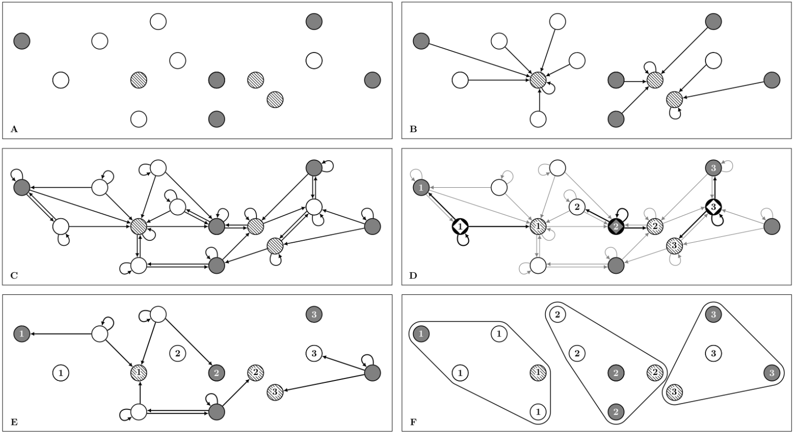
Constrain the set of admissible matches $\mathcal{M}$ as follows:

- Each match group $\mathbf{m}$ must contain c_x no. of units with treatment condition x
- Each match group must contain at least $t \geq \sum_{x=1}^{k} c_x$ no. of units overall
- Union of match groups must contain all units, $\mathbf{M} = \bigcup \mathbf{m} = \mathbf{U}$

Algorithm

. . .

Graphical example



Properties

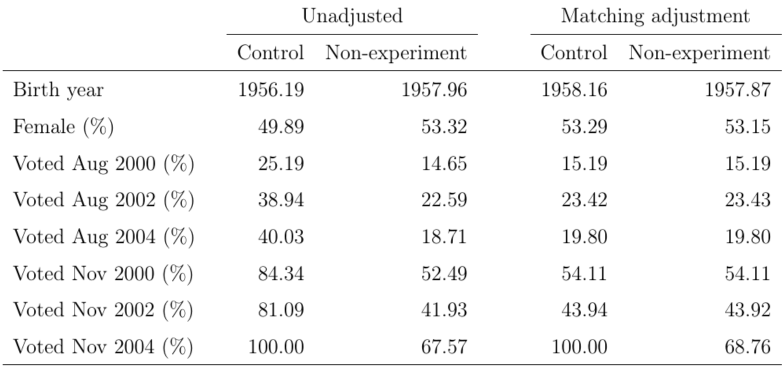
Let $\mathbf{M}_{\text{alg}}$ be the set of matches resulting from the algorithm

Theorem: Sävje et al. (2019)

$$L(\mathbf{M}_{\text{alg}}) \leq \min_{\mathbf{M} \in \mathcal{M}} 4L(\mathbf{M})$$

Covariate balance

Table 3: Covariate balance with and without matching adjustment.



	Unadjusted		Matching adjustment	
	Control	Non-experiment	Control	Non-experiment
Birth year	1956.19	1957.96	1958.16	1957.87
Female (%)	49.89	53.32	53.29	53.15
Voted Aug 2000 (%)	25.19	14.65	15.19	15.19
Voted Aug 2002 (%)	38.94	22.59	23.42	23.43
Voted Aug 2004 (%)	40.03	18.71	19.80	19.80
Voted Nov 2000 (%)	84.34	52.49	54.11	54.11
Voted Nov 2002 (%)	81.09	41.93	43.94	43.92
Voted Nov 2004 (%)	100.00	67.57	100.00	68.76

Construct matched groups based on Mahalonobis distance

Results on voter turnout data (1)

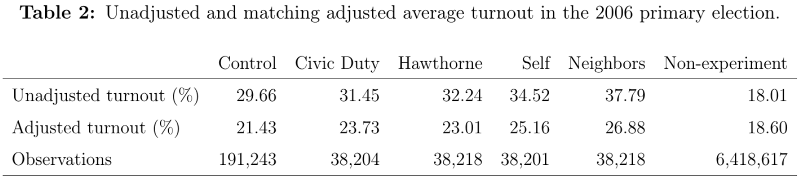


Table 2: Unadjusted and matching adjusted average turnout in the 2006 primary election.

	Control	Civic Duty	Hawthorne	Self	Neighbors	Non-experiment
Unadjusted turnout (%)	29.66	31.45	32.24	34.52	37.79	18.01
Adjusted turnout (%)	21.43	23.73	23.01	25.16	26.88	18.60
Observations	191,243	38,204	38,218	38,201	38,218	6,418,617

The figures [in the second row] should be interpreted as estimates of turnout of the six conditions if scaled up to the whole population

Control and non-experiment groups should be more similar....

Results on voter turnout data (2)

Now restrict to units that voted in 2004 election. . .

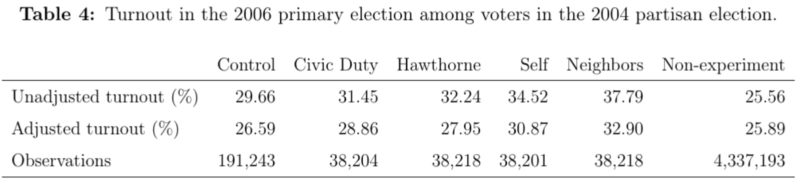


Table 4: Turnout in the 2006 primary election among voters in the 2004 partisan election.

	Control	Civic Duty	Hawthorne	Self	Neighbors	Non-experiment
Unadjusted turnout (%)	29.66	31.45	32.24	34.52	37.79	25.56
Adjusted turnout (%)	26.59	28.86	27.95	30.87	32.90	25.89
Observations	191,243	38,204	38,218	38,201	38,218	4,337,193

Differences between Nattino et al. and Sävje et al.

- Nattio et al.
  - Attempt to mimic block randomization design
  - Adapts existing matched pair algorithm
  - Fisher randomization paradigm
  - Frequentist test and confidence intervals are standard
- Sävje et al.
  - Less conventional experimental design → more researcher degrees of freedom (how to set c_x?)
  - Novel algorithm which generalizes full matching
  - Direct comparison of average outcomes
  - Quantifying uncertainty appears difficult, and is not attempted by the authors

Wu et al. (2020)

Wu et al. (2020)

- **Goal:** Study effect of long-term $PM_{2.5}$ exposure on mortality rates
- Estimand: $\mathrm{E}[Y(w)]$, where Y is mortality rate per 100 Medicare enrollees, and w is $PM_{2.5}$ exposure in μg/m^3

Local weak unconfoundedness

Treatment W_j and covariates $\mathbf{C}_j$

Assumption: Local weak unconfoundedness (Imbens, 2000)

$W_j \perp\!\!\!\perp Y_j(w) \mid \mathbf{C}_j$ for all $w \in \mathcal{W}$

Note: does not require joint independence of all potential outcomes $\{Y_j(w)\}_{w \in \mathcal{W}}$

Define the indicator variable $I_j(\tilde{w}) = 1$ if $W_j = \tilde{w}$ and 0 otherwise.

Assumption: Local weak unconfoundedness (Wu et al.)

$\{I_j(\tilde{w})\}_{\tilde{w} \in [w-\delta, w+\delta]} \perp\!\!\!\perp Y_j(w) \mid \mathbf{C}_j$ for all $z \in \mathcal{Z}$

Note: this does not require joint independence of all potential outcomes $\{Y(z)\}_{z \in \mathcal{Z}}$

That is, the assignment is unconfounded *within a neighborhood* of w (not all $w \in \mathcal{W}$)

Here δ is called the *caliper*.

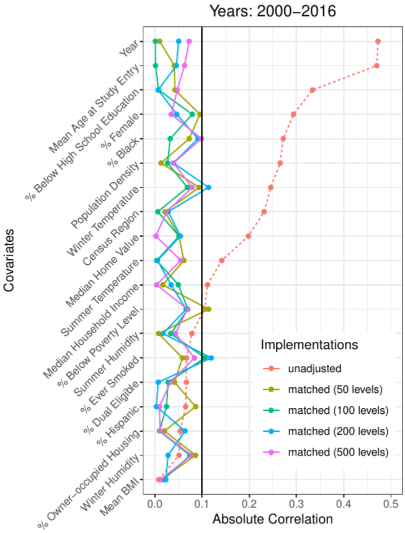
Matching with continuous treatments

- Define a grid of values for w
- **Idea**: Match on both w and the estimated GPS e, i.e. the objective function for matching is

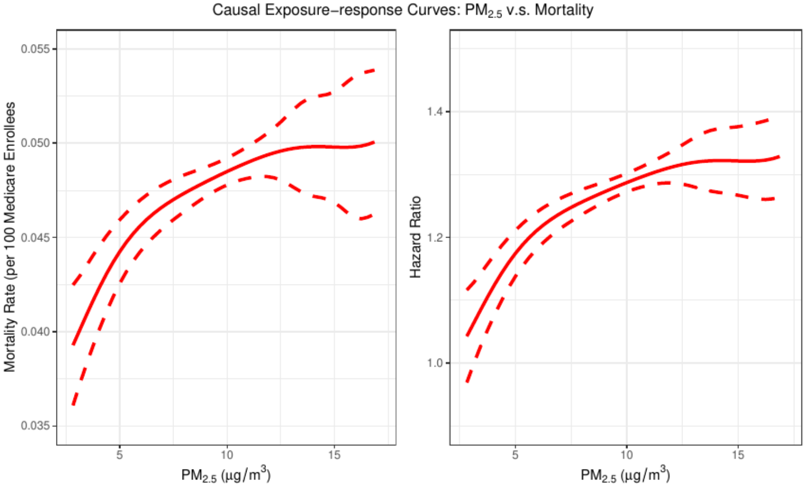
$$m(e_j, w) = \arg \min_{k : w_k \in [w-\delta, w+\delta]} \|\lambda \cdot [e^\star(w_k, \mathbf{c}_k) - e_j^\star] + (1-\lambda) \cdot [w_k^\star - w_j^\star]\|$$

- The counterfactual outcome for unit j at level treatment level w is imputed as $\hat{Y}_j(w) = Y^{\text{obs}}_{m(e(w,\mathbf{c}_j),w)}$, i.e., impute it from the unit close to w (not w_j) and close in propensity score for unit j, e_j
- Must select tuning parameters λ and δ
- Take average within each level of w, then use a kernel smoother to estimate the dose-response curve

Results on $PM_{2.5}$ mortality data



Results on $PM_{2.5}$ mortality data



Confidence bands

Open questions from Wu et al.

- Is the bootstrap a valid way to represent uncertainty?
- This method cannot estimate heterogeneous effects (e.g., subgroups of the population)

Conclusion

Slides at `spencerwoody.github.io/talks`

References I

Alan S Gerber, Donald P Green, and Christopher W Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48, 2008.

Ben B Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618, 2004. doi: 10.1198/016214504000000647. URL https://doi.org/10.1198/016214504000000647.

Keisuke Hirano and Guido W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd, 2004. ISBN 9780470090459. doi: 10.1002/0470090456.ch7. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470090456.ch7.

Kosuke Imai and David A van Dyk. Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467):854–866, 2004. doi: 10.1198/016214504000001187. URL https://doi.org/10.1198/016214504000001187.

Guido W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000. ISSN 00063444. URL http://www.jstor.org/stable/2673642.

Giovanni Nattino, Bo Lu, Junxin Shi, Stanley Lemeshow, and Henry Xiang. Triplet matching for estimating causal effects with three treatment arms: A comparative study of mortality by trauma center level. *Journal of the American Statistical Association*, 0(0):1–10, 2020. doi: 10.1080/01621459.2020.1737078. URL https://doi.org/10.1080/01621459.2020.1737078.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 2000.

Paul R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 03 1987. ISSN 0006-3444. doi: 10.1093/biomet/74.1.13. URL https://doi.org/10.1093/biomet/74.1.13.

Paul R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):597–610, 1991. ISSN 00359246. URL http://www.jstor.org/stable/2345589.

References II

Fredrik Sävje, Michael J. Higgins, and Jasjeet S. Sekhon. Generalized full matching and extrapolation of the results from a large-scale voter mobilization experiment, 2017.

Elizabeth A Stuart and Kerry M Green. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental psychology*, 44(2):395, 2008.

Xiao Wu, Fabrizia Mealli, Marianthi-Anna Kioumourtzoglou, Francesca Dominici, and Danielle Braun. Matching on generalized propensity scores with continuous exposures. *arxiv 1812.06575*, 2020.

Shu Yang, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065, 2016. doi: 10.1111/biom.12505. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12505.