Non-parametric methods for doubly robust estimation of continuous treatment effects by Kennedy, Ma, McHugh, and Small (2017)

Hyemin Yeon

Department of Statistics
Iowa State University

April 5, 2023

Contents

Prediscussion

- Examples of continuous treatments
- General ideas
- Theoretical accomplishments/limitations
- Practical advantages/issues
- Potential future directions

Contents

Continuous treatments

- E.g., dose, duration, frequency, ...
- Main example: Hospital readmisisons reduction program
  - Treatment: average nursing hours per patent day (levels of nurse staffing)
  - Outcome: chance of readmission penalty
  - Covariates: skilled nursing facility, teaching intensity, urban location, number of beds, ...
- Existing methods
  - are not flexible due to parametric model assumptions for does-response regression, or
  - rely on correct model specification, e.g., of conditional treatment density or of conditional mean outcome (not doubly robust)

Overview

- Causal Inference + Nonparametric Regression
- Theory & Methods > Computation
  - Will have brief discussion of proofs
- Theory (Connection to ISU STAT courses)
  - Asymptotic theory in nonparametric regression: Fan (1993), Fan and Gijbel (1996), STAT546
  - Semiparametric theory: Tsiatis (2006), STAT621 in 2023 Spring
  - Empirical process theory: van der Waart and Wellner (1996)

Contents

Notation

- $\boldsymbol{Z} = (\boldsymbol{L}, A, Y)$ has support $\mathcal{Z} = \mathcal{L} \times \mathcal{A} \times \mathcal{Y}$.
  $\boldsymbol{L}$: covariates
  A: (continuous) treatment/exposure
  Y: outcome/response
  Y^a: potential outcome under treatment $a \in \mathcal{A}$.

- $p(\boldsymbol{z}) = p(y|\boldsymbol{l}, a)p(a|\boldsymbol{l})p(\boldsymbol{l})$
  $\mu(\boldsymbol{l}, a) \equiv \mathrm{E}[Y|\boldsymbol{L} = \boldsymbol{l}, A = a]$
  $\pi(a|\boldsymbol{l}) \equiv \partial \mathrm{P}(A \leq a|\boldsymbol{L} = \boldsymbol{l})/\partial a$
  $\omega(a) \equiv \partial \mathrm{P}(A \leq a)/\partial a$

- Goal: estimate $\theta(a) \equiv \mathrm{E}[Y^a]$.

Potential outcome and continuous treatment

- We cannot observe Y_i, instead we can observe Y_i^a if $A_i = a$.
- A direct application of the standard nonparametric regression of $\{A_i\}_{i=1}^n$ on $\{Y_i^{A_i}\}_{i=1}^n$ does not make sense.
- General idea:
  - to find pseudo-outcome $\hat{Y}_i$, and
  - to regress $\{A_i\}_{i=1}^n$ on $\{\hat{Y}_i\}_{i=1}^n$ through the nonparametric regression

Assumptions for identification

(A1) Consistency (or SUTVA):

$$A = a \implies Y = Y^a$$

(A2) Positivity:

$$\pi(a|\boldsymbol{l}) \geq \pi_{\min} > 0, \quad \forall \boldsymbol{l} \in \mathcal{L}$$

(A3) Ignorability (or no unmeasured confounders):

$$\mathrm{E}[Y^a|\boldsymbol{L}, A] = \mathrm{E}[Y^a|\boldsymbol{L}]$$

Causal effect curve

- $\theta(a) \equiv \mathrm{E}[Y^a] = \mathrm{E}[\mu(\boldsymbol{L}, a)] = \int_{\mathcal{L}} \mu(\boldsymbol{l}, a) d\mathrm{P}(\boldsymbol{l})$

$\because$

$$\begin{aligned}
\mu(\boldsymbol{l}, a) &= \mathrm{E}[Y | \boldsymbol{L} = \boldsymbol{l}, A = a] \\
&\underset{(A1)}{=} \mathrm{E}[Y^a | \boldsymbol{L} = \boldsymbol{l}, A = a] \\
&\underset{(A3)}{=} \mathrm{E}[Y^a | \boldsymbol{L} = \boldsymbol{l}] \\
\implies \theta(a) \equiv \mathrm{E}[Y^a] &= \mathrm{E}[\mathrm{E}[Y^a | L]] = \mathrm{E}[\mu(L, a)]
\end{aligned}$$

Contents

Idea 1

- Find $\xi = \xi(\cdot; \pi, \mu) : \mathcal{Z} \to \mathbb{R}$ such that

$$\mathrm{E}[\xi(\boldsymbol{Z}; \pi, \mu)|A = a] = \theta(a)$$

  if either $\pi = \pi_0$ or $\mu = \mu_0$.

- Use any non-parametric regression method to estimate $\theta(a)$ by regressing $\xi(\boldsymbol{Z}; \hat{\pi}, \hat{\mu})$ on treatment A

Idea 2

- Use semiparametric theory

$$\begin{aligned} \mathrm{E}[\xi(\boldsymbol{Z};\pi,\mu)] &= \mathrm{E}[\mathrm{E}[\xi(\boldsymbol{Z};\pi,\mu)|A]] = \mathrm{E}[\theta(A)] = \mathrm{E}[\mu(L,A)] \\ &= \int_{\mathcal{A}}\int_{\mathcal{L}} \mu(l,a)\omega(a)d\mathrm{P}(\boldsymbol{l})d(a) \equiv \psi \end{aligned}$$

- Candidate for ξ: influence function ϕ for ψ

Theorem 1

Theorem 1

Under a non-parametric model, the efficient influence function ϕ *for* $\psi \equiv \int_{\mathcal{A}} \int_{\mathcal{L}} \mu(l, a)\omega(a) d\mathrm{P}(\boldsymbol{l}) d(a)$ *is*

$$\xi(\boldsymbol{Z}; \pi, \mu) - \psi + \int_{\mathcal{A}} \left\{ \mu(\boldsymbol{L}, a) - \int_{\mathcal{L}} \mu(\boldsymbol{l}, a) d\mathrm{P}(\boldsymbol{l}) \right\} \omega(a) da,$$

where

$$\xi(\boldsymbol{Z}; \pi, \mu) \equiv \frac{Y - \mu(\boldsymbol{L}, A)}{\pi(A|\boldsymbol{L})} \int_{\mathcal{L}} \pi(A|\boldsymbol{l}) d\mathrm{P}(\boldsymbol{l}) + \int_{\mathcal{L}} \mu(\boldsymbol{l}, A) d\mathrm{P}(\boldsymbol{l}).$$

- Then, $\mathrm{E}[\xi(\boldsymbol{Z}; \mu, \pi)|A = a] = \theta(a)$ if either $\pi = \pi_0$ or $\mu = \mu_0$.

Sketch of proof (Theorem 1)

- $p(\boldsymbol{z}; \varepsilon)$: a parametric submodel with parameter $\varepsilon \in \mathbb{R}$
- $\psi(\varepsilon) = \int_{\mathcal{W}} \theta(a; \varepsilon)\omega(a; \varepsilon)da$
- The efficient influence function for ψ is the unique function $\phi(\boldsymbol{Z})$ that satisfies

$$\mathrm{E}\left[\phi(\boldsymbol{Z})\frac{\partial \log p(\boldsymbol{Z}; \varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0}\right] = \frac{\partial \psi(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0}. \tag{1}$$

  - Compute RHS of (1).
  - Compute LHS of (1) for $\phi(\boldsymbol{Z})$ in Theorem 1.
  - Check if both are equal.
- Manipulating (conditional) densities/expectations/log-likelihoods.
- Related to Theorems 3.2, 4.2, 4.4 of Tsiatis (2006).

General estimation procedure

- Estimated pseudo-outcomes: $\{\hat{\xi}(\boldsymbol{Z}_i; \hat{\pi}, \hat{\mu})\}_{i=1}^{n}$, where

$$\hat{\xi}(\boldsymbol{Z}_i; \hat{\pi}, \hat{\mu}) \equiv \frac{Y_i - \hat{\mu}(\boldsymbol{L}_i, A_i)}{\hat{\pi}(A_i|\boldsymbol{L})} \left\{ n^{-1} \sum_{i'=1}^{n} \hat{\pi}(A_i|L_{i'}) \right\} + n^{-1} \sum_{i'=1}^{n} \mu(L_{i'}, A_i)$$

- Step 1: Estimate the nuisance functions π and μ by $\hat{\pi}$ and $\hat{\mu}$
  - E.g., logistic regression, super learner

- Step 2: Regress the estimated pseudo-outcomes $\{\hat{\xi}(\boldsymbol{Z}_i; \hat{\pi}, \hat{\mu})\}_{i=1}^{n}$ on treatments $\{A_i\}_{i=1}^{n}$ by using any nonparametric regression methods
  - E.g., kernel-based smoothing, targeted maximum likelihood estimator (TMLE), any machine learning methods

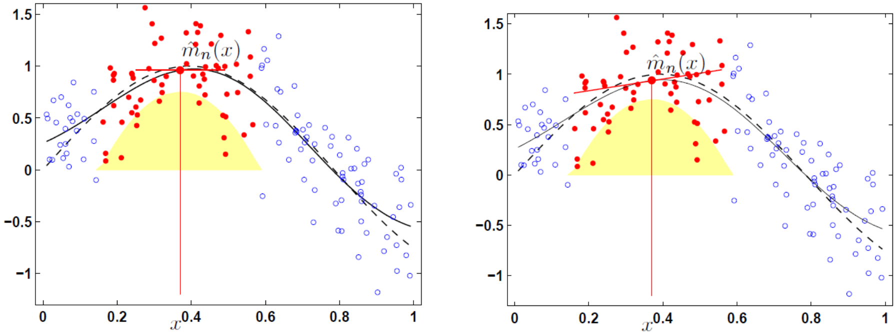
Local linear estimator

- $\hat{\theta}_h(a) = \boldsymbol{g}_{ha}(a)^\top \hat{\boldsymbol{\beta}}_h(a)$, where $\boldsymbol{g}_{ha}(t) = \begin{bmatrix} 1 & (t-a)/h \end{bmatrix}^\top$ and

$$\hat{\boldsymbol{\beta}}_h(a) = \underset{\boldsymbol{\beta} \in \mathbb{R}^2}{\operatorname{argmin}}\, n^{-1} \sum_{i=1}^{n} K_{ha}(A_i) \left\{ \hat{\xi}(\boldsymbol{Z}_i; \hat{\pi}, \hat{\mu}) - \boldsymbol{g}_{ha}(A_i)^\top \boldsymbol{\beta} \right\}^2$$

  for $K_{ha}(t) = h^{-1} K((t-a)/h)$
- K: a kernel function
- h: a (scalar) bandwidth parameter

Review of local polynomial regression



(a) (b)

- Figure 4.2 of the lecture note of STAT546 offered by Professor Kris De Brabanter

Assumptions

- $\|\hat{\pi} - \pi\|_{\text{sup}} = o_{\mathrm{P}}(1)$ and $\|\hat{\mu} - \mu\|_{\text{sup}} = o_{\mathrm{P}}(1)$.

(a) (Double robustness) Either $\pi = \pi_0$ or $\mu = \mu_0$.

(b) (Bandwidth) $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$.

(c) (Kernel) K is a continuous symmetric probability density with support $[-1, 1]$.

(d) (Continuity) $\theta \in C^2(\mathcal{A})$, $\omega \in C^0(\mathcal{A})$, $\partial \mathrm{P}(\xi(\boldsymbol{Z}; \boldsymbol{\pi}, \boldsymbol{\mu}) \leq z)/\partial a \in C^0(\mathcal{A})$.

(e) (Class of bounded functions)
$\hat{\pi}, \hat{\mu}, 1/\hat{\pi}, 1/\hat{\mu}, \pi, \mu$ are uniformly bounded.
$\hat{\pi}, \hat{\mu}, \pi, \mu \in \mathcal{F}$ with $\mathcal{F}$ having finite uniform entropy integrals.

Consistency

Theorem 2

Under some assumptions, we have

$$|\hat{\theta}_h(a) - \theta(a)| = O_{\mathrm{P}}\left((nh)^{-1/2} + h^2 + r_n(a)s_n(a)\right)$$

where

$$\sup_{t:|t-a|\le h}\left\{\int_{\mathcal{L}}\{\hat{\pi}(t|\boldsymbol{l}) - \pi(t|\boldsymbol{l})\}^2 d\mathrm{P}(\boldsymbol{l})\right\}^{1/2} = O_{\mathrm{P}}(r_n(a)),$$

$$\sup_{t:|t-a|\le h}\left\{\int_{\mathcal{L}}\{\hat{\mu}(\boldsymbol{l}, t) - \mu(\boldsymbol{l}, t)\}^2 d\mathrm{P}(\boldsymbol{l})\right\}^{1/2} = O_{\mathrm{P}}(s_n(a))$$

Comments on consistency

(1) $(nh)^{-1/2} + h^2$.

- To balance two terms, $h \sim n^{-1/5}$ and $(nh)^{-1/2} \sim h^2 \sim n^{-2/5}$.
- Optimal rate of convergence for standard non-parametric regression

(2) $r_n(a)s_n(a)$

- Product of "local" rates of convergence
- $r_n(a) = o(1), s_n(a) = O(1) \implies r_n(a)s_n(a) = o(1)$
- $r_n(a) = n^{-2/5}, s_n(a) = n^{-1/10} \implies r_n(a)s_n(a) = O(n^{-1/2}) = o(n^{-2/5})$.

Asymptotic normality

Theorem 3

Under Theorem 2 assumptions, if $r_n(a)s_n(a) = o_{\mathrm{P}}((nh)^{-1/2})$, *then*

$$\sqrt{nh}\{\hat{\theta}_h(a) - \theta(a) - b_h(a)\} \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, m_0(K^2)\frac{\sigma^2(a)}{\omega_0(a)}\right)$$

- $b_h(a) = \theta''(a)m_2(K)h^2/2 + o(h^2)$
- $\sigma^2(a) \equiv \mathrm{var}[\xi(\boldsymbol{Z};\pi,\mu)|A=a]$
  $= \mathrm{E}\left[\frac{\mathrm{var}[Y|\boldsymbol{L},A=a]+\{\mu_0(\boldsymbol{L},a)-\mu(\boldsymbol{L},a)\}^2}{\{\pi(a|\boldsymbol{L})/\mathrm{E}[\pi(a|\boldsymbol{L})]\}^2/\{\pi(a|\boldsymbol{L})/\omega_0(a)\}^2}\right] - \{\theta(a) - \mathrm{E}[\mu(\boldsymbol{L},a)]\}^2$
- $m_j(K^k) = \int u^j K(u)^k du$
- Same form as non-causal nonparametric regression

Comments on asymptotic normality

- bias correction vs undersmoothing
  (US) $h = o(n^{-1/5}) \implies b_h(a) = o((nh)^{-1/2})$
  $\implies$ negligible bias through undersmoothing
  (BC) Or, bias correction by estimating $b_h(a)$ by $\hat{b}_h(a)$

- Need to estimate $\sigma^2(a)$ by $\hat{\sigma}^2(a)$

- Confidence intervals:

$$CI_{\mathrm{us}} = \left[\hat{\theta}_h(a) - z_{1-\alpha/2}\frac{\hat{\sigma}(a)}{\sqrt{nh}}, \hat{\theta}_h(a) + z_{1-\alpha/2}\frac{\hat{\sigma}(a)}{\sqrt{nh}}\right]$$

$$CI_{\mathrm{bc}} = \left[\hat{\theta}_h(a) - \hat{b}_h(a) - z_{1-\alpha/2}\frac{\hat{\sigma}(a)}{\sqrt{nh}}, \hat{\theta}_h(a) - \hat{b}_h(a) + z_{1-\alpha/2}\frac{\hat{\sigma}(a)}{\sqrt{nh}}\right]$$

Sketch of proof (Theorems 2-3)

- Decomposition:

$$\hat{\theta}_h(a) - \theta(a) = \tilde{\theta}_h(a) - \theta(a) + R_{1n} + R_{2n},$$

where $\tilde{\theta}_h(a)$ is the local linear estimator based on $\{\xi(Z_i; \pi, \mu)\}_{i=1}^n$

- $\tilde{\theta}_h(a) - \theta(a)$: from the standard non-parametric regression

- R_{1n}: from $\hat{P}_n - P$ by empirical process theory,
$\hat{P}_n(A) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \in A)$

- R_{2n}: from $\|\hat{\pi} - \pi\|_{\sup}$, $\|\hat{\mu} - \mu\|_{\sup}$, and $\hat{P}_n - P$

Data-driven bandwidth selection

- Leave-one-out cross-validation:

$$\hat{h}_{\mathrm{opt}} = \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{n} \left\{ \frac{\hat{\xi}(\boldsymbol{Z}_i; \hat{\pi}, \hat{\mu}) - \hat{\theta}_h(A_i)}{1 - \hat{W}_h(A_i)} \right\}^2 ,$$

  where $\hat{W}_h(A_i)$ is the i-th diagonal of the smoothing or hat matrix.
- Treat the pseudo-outcomes $\{\hat{\xi}(\boldsymbol{Z}_i; \hat{\pi}, \hat{\mu})\}_{i=1}^{n}$ as real outcomes

Contents

Set-ups

- $\boldsymbol{L} = \begin{bmatrix} L_1 & L_2 & L_3 & L_4 \end{bmatrix}^\top \sim \mathrm{N}(0, I_4)$

- $(A/20)|\boldsymbol{L} \sim \mathrm{Beta}(\lambda(\boldsymbol{L}), 1 - (\lambda(\boldsymbol{L}))$
  $\operatorname{logit}(\lambda(\boldsymbol{L}) = -0.8 + 0.1L_1 + 0.1L_2 - 0.1L_3 + 0.2L_4$

- $Y|\boldsymbol{L}, A \sim \mathrm{Ber}(\mu(\boldsymbol{L}, A))$
  $\operatorname{logit}(\mu(\boldsymbol{L}, A)) =$
  $1 + \begin{bmatrix} 0.2 & 0.2 & 0.3 & -0.1 \end{bmatrix} \boldsymbol{L} + A(0.1 - 0.1L_1 + 0.1L_3 - 0.13^2 A^2)$

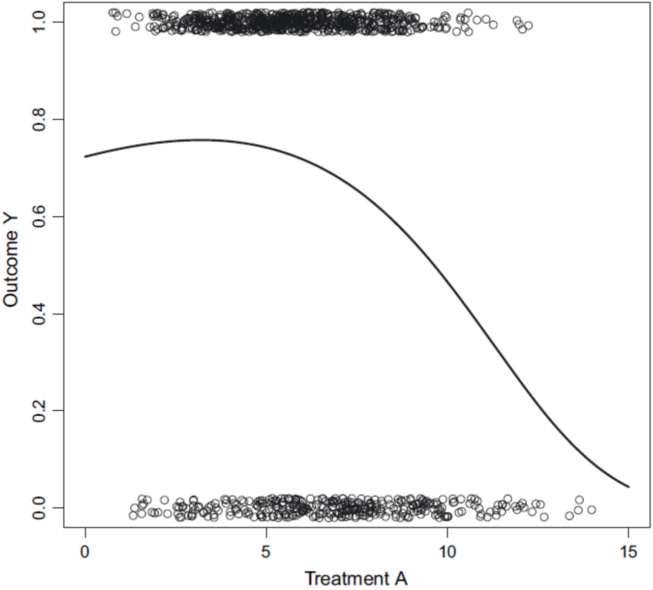


Fig. 2. Plot of effect curve $\theta(a)$ induced by the simulation set-up (———) with treatment and outcome data (○) from one simulated data set with $n = 1000$

Methods

(1) Plug-in regression: $\hat{m}(a) = n^{-1} \sum_{i=1}^{n} \hat{\mu}(L_i, a)$

(2) Inverse-probability-weighted (IPW) approach
by Rubin and van der Laan (2006)
only use $\hat{\pi}$ with $\hat{\mu} = 0$

(3) The proposed doubly robust approach
use both $\hat{\pi}$ and $\hat{\mu}$

- $\hat{\pi}$, $\hat{\mu}$: using logistic regression
- bandwidth selection
    - LOOCV
    - oracle choice: $\text{argmin}_{h \in \mathcal{H}} n^{-1} \sum_{i=1}^{n} \{\hat{\theta}_h(A_i) - \theta(A_i)\}^2$

Table 1. Integrated mean bias and root-mean-squared error (in parentheses) after 500 simulations

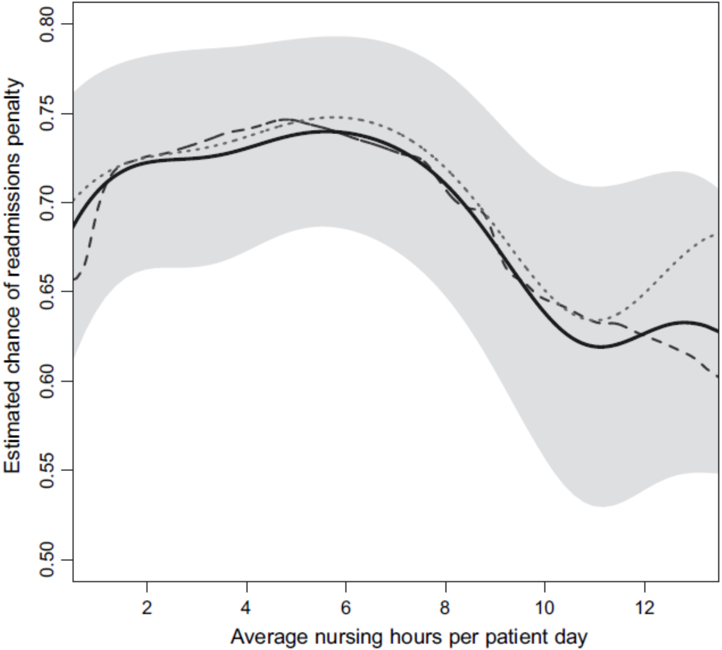
n	*Method*	*Results when correct model is as follows:*			
		Neither	*Treatment*	*Outcome*	*Both*
100	Regression	2.67 (5.54)	2.67 (5.54)	0.62 (5.25)	0.62 (5.25)
	Inverse probability weighted	2.26 (8.49)	1.64 (8.57)	2.26 (8.49)	1.64 (8.57)
	Inverse probability weighted†	2.26 (7.36)	1.58 (7.37)	2.26 (7.36)	1.58 (7.37)
	Doubly robust	2.23 (6.27)	1.01 (6.28)	1.12 (5.92)	1.10 (6.50)
	Doubly robust†	2.12 (5.48)	1.00 (5.36)	1.03 (5.08)	1.02 (5.65)
1000	Regression	2.62 (3.07)	2.62 (3.07)	0.06 (1.53)	0.06 (1.53)
	Inverse probability weighted	2.38 (3.97)	0.86 (2.94)	2.38 (3.97)	0.86 (2.94)
	Inverse probability weighted†	2.11 (3.44)	0.70 (2.34)	2.11 (3.44)	0.70 (2.34)
	Doubly robust	2.03 (3.11)	0.75 (2.39)	0.74 (2.53)	0.68 (2.25)
	Doubly robust†	1.84 (2.67)	0.64 (1.88)	0.61 (1.78)	0.58 (1.78)
10000	Regression	2.65 (2.70)	2.65 (2.70)	0.02 (0.47)	0.02 (0.47)
	Inverse probability weighted	2.36 (3.42)	0.33 (1.09)	2.36 (3.42)	0.33 (1.09)
	Inverse probability weighted†	2.24 (3.28)	0.35 (0.85)	2.24 (3.28)	0.35 (0.85)
	Doubly robust	1.81 (2.35)	0.26 (0.86)	0.20 (1.21)	0.25 (0.78)
	Doubly robust†	1.76 (2.27)	0.31 (0.68)	0.24 (1.10)	0.29 (0.64)

†Uses the oracle bandwidth.

Contents

Real data application

- nurse staffing $\rightarrow$ hospital readmissions penalties
- A: nurse staffing hours
- Y: whether the hospital was penalized because of excess readmissions
- $\theta(a)$: proportion of hospitals that would have been penalized
  if all hospitals had changed their nurse staffing hours to level a
- $\pi(a|\boldsymbol{l})$: $A = \lambda(\boldsymbol{L}) + \gamma(\boldsymbol{L})\varepsilon$, $\varepsilon \sim (0, 1)$,
  use the Super Learner for λ, γ and KDE for the density of ε
- $\mu(a, \boldsymbol{l})$: use the Super Learner


Estimated chance of readmissions penalty
0.50
0.55
0.60
0.65
0.70
0.75
0.80
2
4
6
8
10
12
Average nursing hours per patient day

Contents

Future directions

- What if θ is not (pathwise) differentiable?
- Is there an uniform distributional convergence?
- Hypothesis testing for θ?

Continuous variables in Causal Inference

- Kennedy, Ma, McHugh, and Small (2017)
  Local linear regression on continuous treatment
- Kennedy, Lorch, and Small (2019)
  Continuous instrument variables
- Westling, Gilbert, and Carone (2020)
  Isotonic regression on continuous treatment
- Westling (2022)
  Hypothesis testing with continuous treatment
  H_0: the causal effect curve is flat ($\theta(a) = constant$)

-
-
-
-
-
-

The End

THANK YOU