UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Robust and efficient estimation in the presence of a randomly censored covariate

Brian Richardson

Acknowledgements

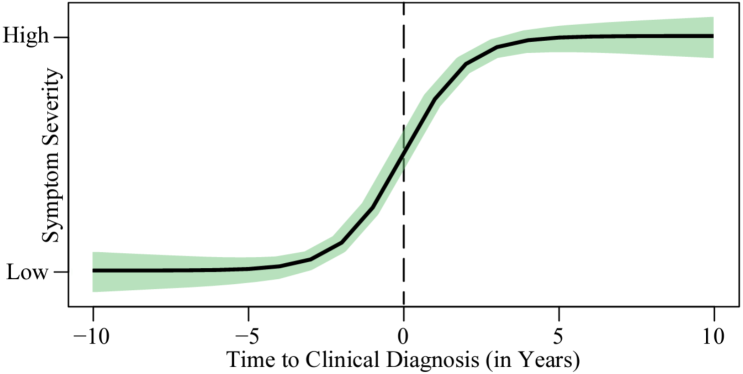
Seong-Ho Lee, PhD

Tanya Garcia, PhD

Yanyuan Ma, PhD

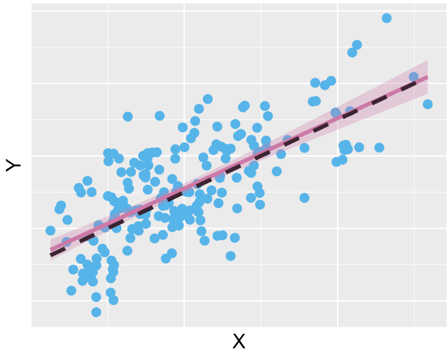
This research was supported by the National Institute of Environmental Health Sciences grant T32ES007018.

Huntington's Disease and Censored Covariates



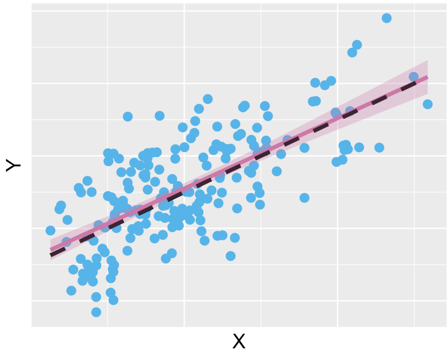
Sarah C Lotspeich et al. "Making Sense of Censored Covariates: Statistical Methods for Studies of Huntington's Disease". In: *Annual Review of Statistics and Its Application* 11 (2024)

Censored Covariates: a Simple Example



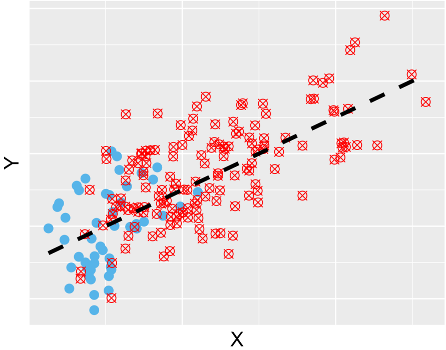
- Regression model: $E(Y) = \beta_0 + \beta_1 X$

Censored Covariates: a Simple Example



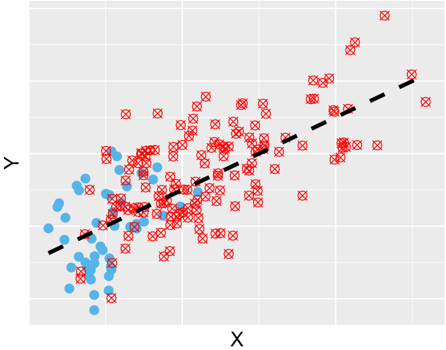
- Regression model:
  $\mathrm{E}(Y) = \beta_0 + \beta_1 X$
- Estimate $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ with least squares/maximum likelihood

Censored Covariates: a Simple Example



Problem: X is censored by a **random censoring time** C

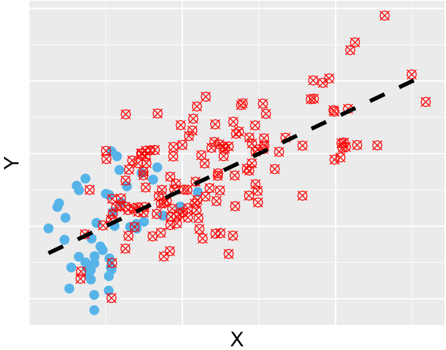
Censored Covariates: a Simple Example



Problem: X is censored by a **random censoring time** C

- $W = \min(X, C)$
- $\Delta = I(X \leq C)$

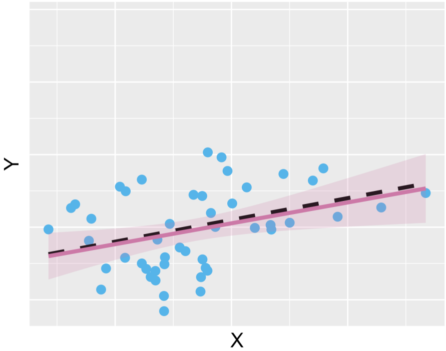
Censored Covariates: a Simple Example



Problem: X is censored by a **random censoring time** C

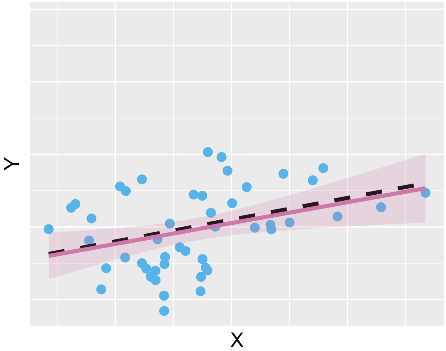
- $W = \min(X, C)$
- $\Delta = I(X \leq C)$
- assume: $C \perp\!\!\!\perp (X, Y)$

Complete Case Analysis



Only use observations where X is *uncensored*

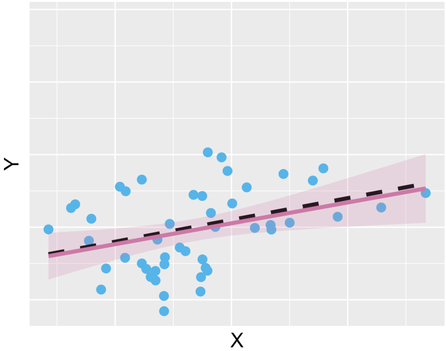
Complete Case Analysis



Only use observations where X is *uncensored*

✓ Consistent

Complete Case Analysis



Only use observations where X is *uncensored*

- ✓ Consistent
- ✗ Inefficient

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

$$\mathbf{S}_{\boldsymbol{\beta}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \qquad \sum_{i=1}^{n} \mathbf{S}_{\boldsymbol{\beta}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_w^\infty f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

$$\mathbf{S}_{\boldsymbol{\beta}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \qquad \sum_{i=1}^{n} \mathbf{S}_{\boldsymbol{\beta}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

- ✓ consistent
- ✓ fully efficient

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

$$\mathbf{S}_{\boldsymbol{\beta}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}),$$

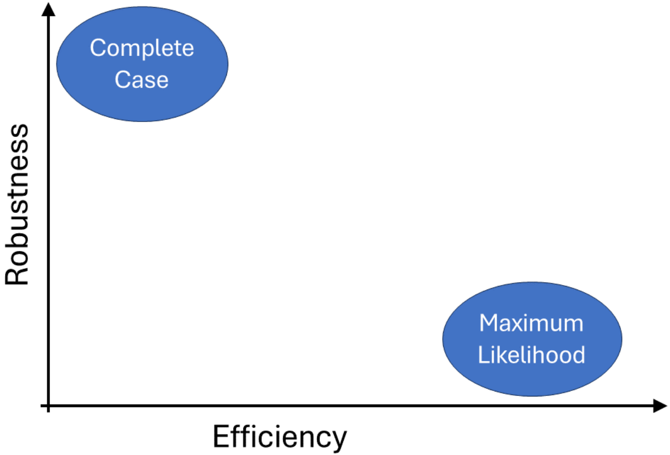
$$\sum_{i=1}^{n} \mathbf{S}_{\boldsymbol{\beta}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

✓ consistent

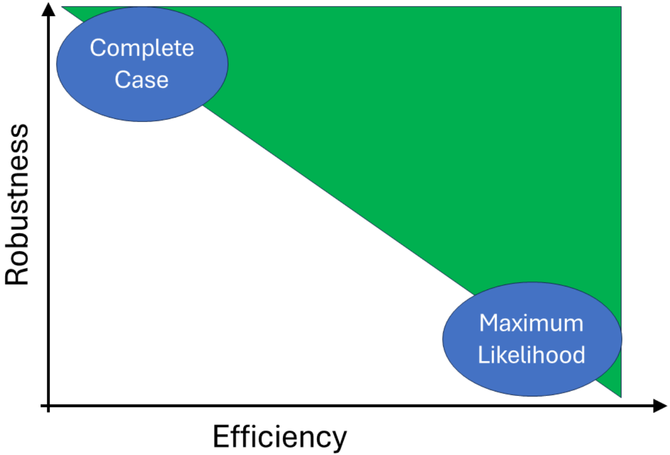
✓ fully efficient

✗ inconsistent when model for **nuisance parameter** f_X is incorrect

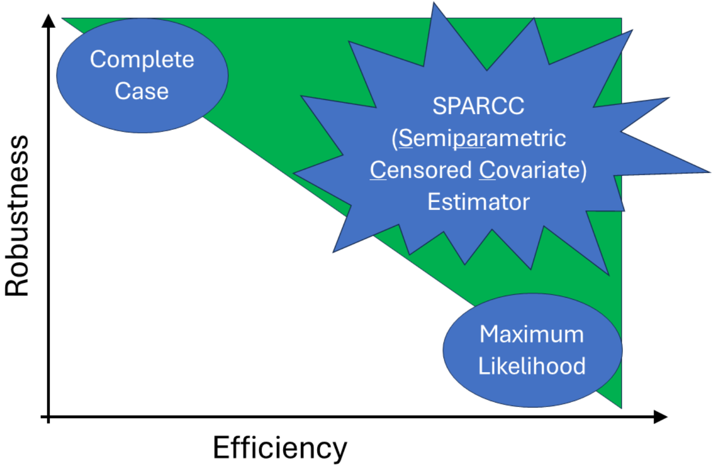
Existing Methods



Existing Opportunity



A New Approach



The Semiparametric Recipe

- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$

The Semiparametric Recipe

- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$

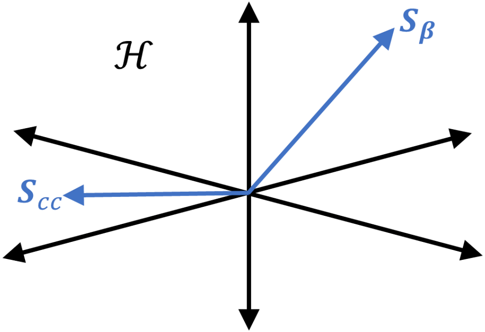
The Semiparametric Recipe

- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$
- **semiparametric**: avoid parametric assumptions for $\boldsymbol{\eta}$

The Semiparametric Recipe

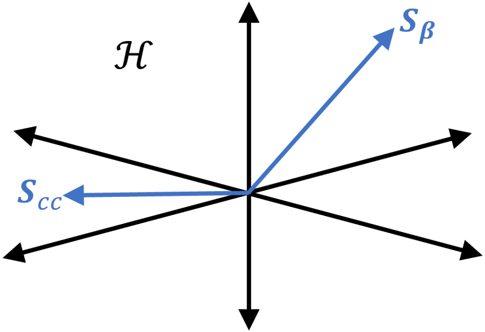
- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$
- **semiparametric**: avoid parametric assumptions for $\boldsymbol{\eta}$
- **goal:** derive the SPARCC estimator

The Semiparametric Recipe



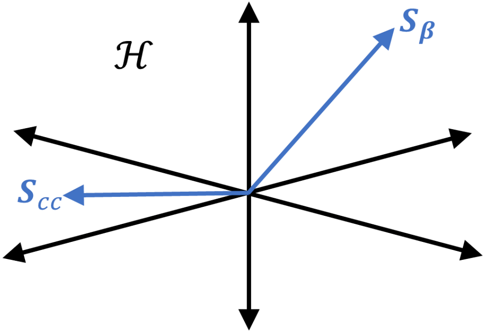
- **Hilbert space** of estimating functions

The Semiparametric Recipe



- **Hilbert space** of estimating functions
- **covariance inner product** $\langle \boldsymbol{h}, \boldsymbol{g} \rangle \equiv \mathrm{E}(\boldsymbol{h}^T \boldsymbol{g})$

The Semiparametric Recipe

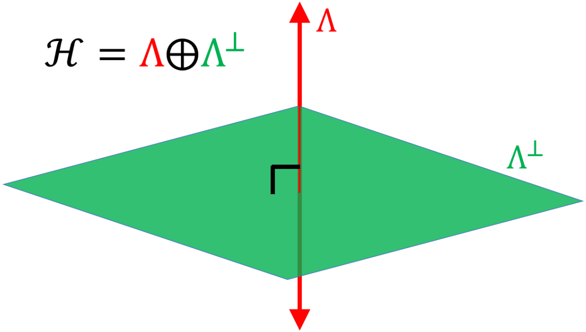


- **Hilbert space** of estimating functions
- **covariance inner product** $\langle \boldsymbol{h}, \boldsymbol{g} \rangle \equiv \mathrm{E}(\boldsymbol{h}^T \boldsymbol{g})$
- **orthogonal** $\Leftrightarrow$ **uncorrelated**

$$\boldsymbol{h} \perp \boldsymbol{g} \iff \langle \boldsymbol{h}, \boldsymbol{g} \rangle = 0$$

The Semiparametric Recipe

$$\mathcal{H} = \Lambda \oplus \Lambda^{\perp}$$

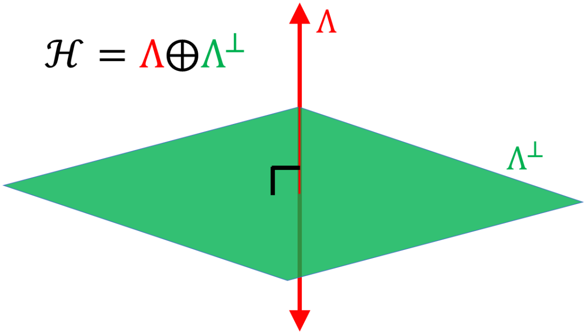


- construct Λ using **nuisance scores**

$$\partial \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$$

The Semiparametric Recipe

$$\mathcal{H} = \Lambda \oplus \Lambda^{\perp}$$

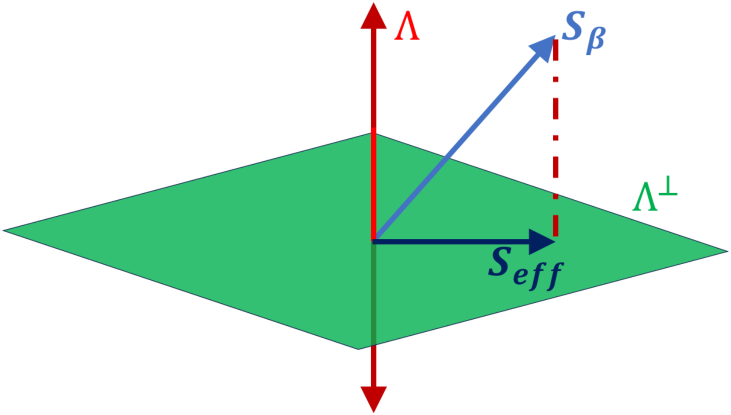


- construct Λ using **nuisance scores**

$$\partial \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$$

- orthogonal complement $\Lambda^{\perp}$

The Semiparametric Recipe



Implementing the SPARCC Estimator

The **SPARCC Estimator** $\widehat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

Implementing the SPARCC Estimator

The **SPARCC Estimator** $\widehat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

Computing $\mathbf{S}_{\text{eff}}$ requires $\boldsymbol{\eta} = (f_X, f_C)$. These distributions can be modeled either **parametrically** or **nonparametrically**

Properties of the SPARCC Estimator

With **parametric** f_X, f_C, $\widehat{\boldsymbol{\beta}}$ is:

✓ **doubly robust**: consistent if one of f_X, f_C is correctly specified,

Properties of the SPARCC Estimator

With **parametric** f_X, f_C, $\widehat{\boldsymbol{\beta}}$ is:

- ✓ **doubly robust**: consistent if one of f_X, f_C is correctly specified,
- ✓ **semiparametric efficient** if f_X, f_C are *both* correctly specified

Properties of the SPARCC Estimator

With **parametric** f_X, f_C, $\widehat{\boldsymbol{\beta}}$ is:

- ✓ **doubly robust**: consistent if one of f_X, f_C is correctly specified,
- ✓ **semiparametric efficient** if f_X, f_C are *both* correctly specified

With **nonparametric** f_X, f_C, $\widehat{\boldsymbol{\beta}}$ is:

- ✓ **consistent**
- ✓ **semiparametric efficient**

Simulation Setup

- large sample size $n = 8000$

Simulation Setup

- large sample size $n = 8000$
- uncensored covariate
  $Z \sim \text{Bernoulli}(0.5)$

Simulation Setup

- large sample size $n = 8000$
- uncensored covariate
  $Z \sim \text{Bernoulli}(0.5)$
- $Y|X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$

Simulation Setup

- large sample size $n = 8000$
- uncensored covariate
  $Z \sim \text{Bernoulli}(0.5)$
- $Y|X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$
- $X|Z \sim \text{beta}(\alpha_{11} + \alpha_{12} Z, \alpha_{13} + \alpha_{14} Z)$

Simulation Setup

- large sample size $n = 8000$
- uncensored covariate
  $Z \sim \text{Bernoulli}(0.5)$
- $Y|X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$
- $X|Z \sim \text{beta}(\alpha_{11} + \alpha_{12} Z, \alpha_{13} + \alpha_{14} Z)$
- $C|Z \sim \text{beta}(\alpha_{21} + \alpha_{22} Z, \alpha_{23} + \alpha_{24} Z)$

Simulation Setup

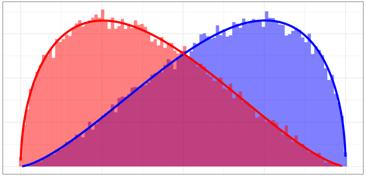
- large sample size $n = 8000$
- uncensored covariate
  $Z \sim \text{Bernoulli}(0.5)$
- $Y|X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$
- $X|Z \sim \text{beta}(\alpha_{11} + \alpha_{12} Z, \alpha_{13} + \alpha_{14} Z)$
- $C|Z \sim \text{beta}(\alpha_{21} + \alpha_{22} Z, \alpha_{23} + \alpha_{24} Z)$
- f_X, f_C possibly **misspecified** as
  marginal beta

Simulation Setup

- large sample size $n = 8000$
- uncensored covariate $Z \sim \text{Bernoulli}(0.5)$
- $Y|X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$
- $X|Z \sim \text{beta}(\alpha_{11} + \alpha_{12} Z, \alpha_{13} + \alpha_{14} Z)$
- $C|Z \sim \text{beta}(\alpha_{21} + \alpha_{22} Z, \alpha_{23} + \alpha_{24} Z)$
- f_X, f_C possibly **misspecified** as *marginal* beta

($Z = 0$, $Z = 1$)

f_X correct:

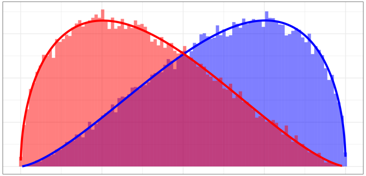


Simulation Setup

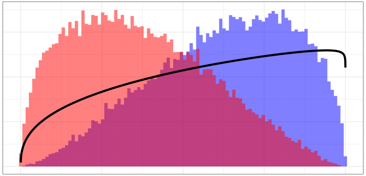
- large sample size $n = 8000$
- uncensored covariate $Z \sim \text{Bernoulli}(0.5)$
- $Y|X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$
- $X|Z \sim \text{beta}(\alpha_{11} + \alpha_{12} Z, \alpha_{13} + \alpha_{14} Z)$
- $C|Z \sim \text{beta}(\alpha_{21} + \alpha_{22} Z, \alpha_{23} + \alpha_{24} Z)$
- f_X, f_C possibly **misspecified** as *marginal* beta

($Z = 0$, $Z = 1$)

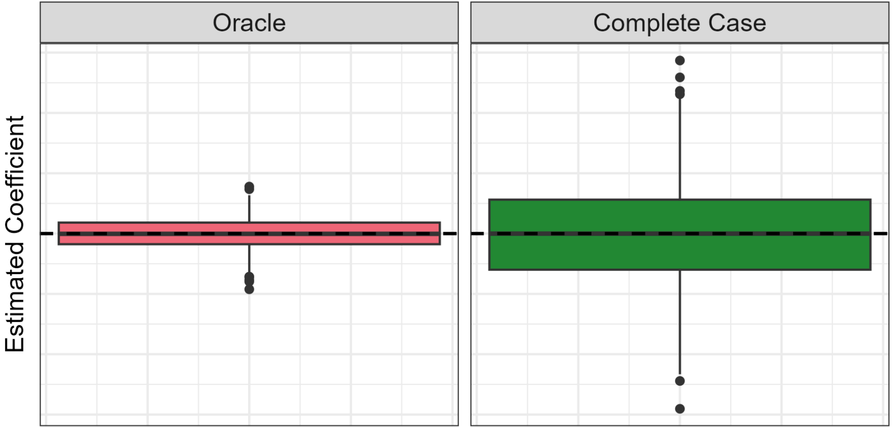
f_X correct:



f_X incorrect:

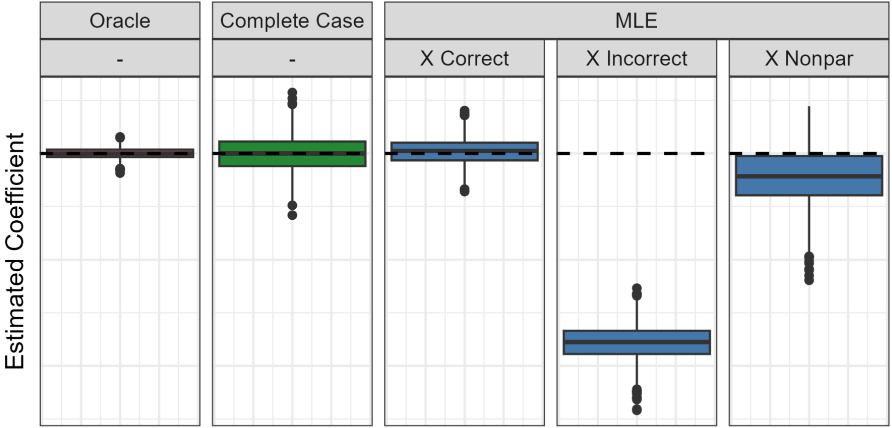


Simulation Results: Robustness



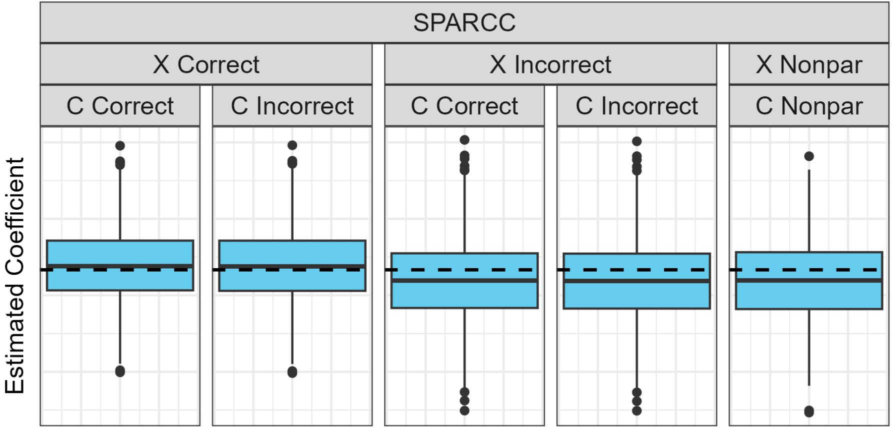
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



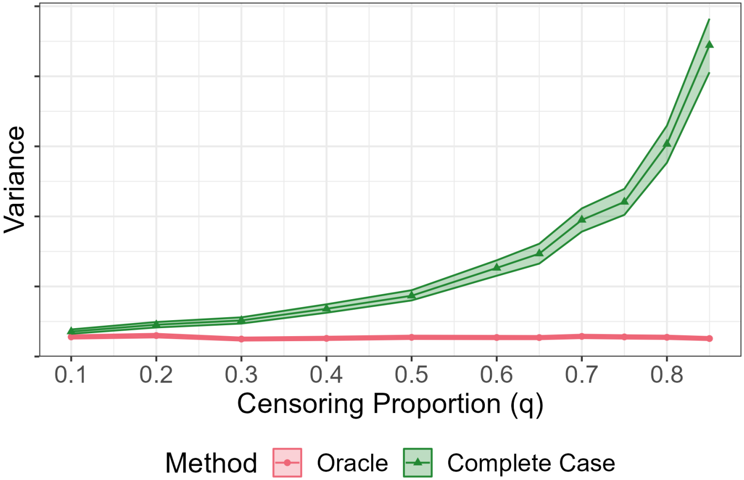
(censoring proportion $q = 0.8$)

Simulation Results: Robustness

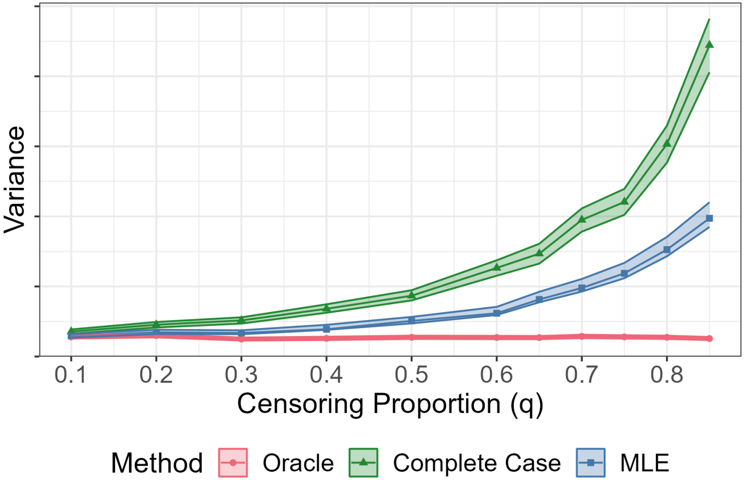


(censoring proportion $q = 0.8$)

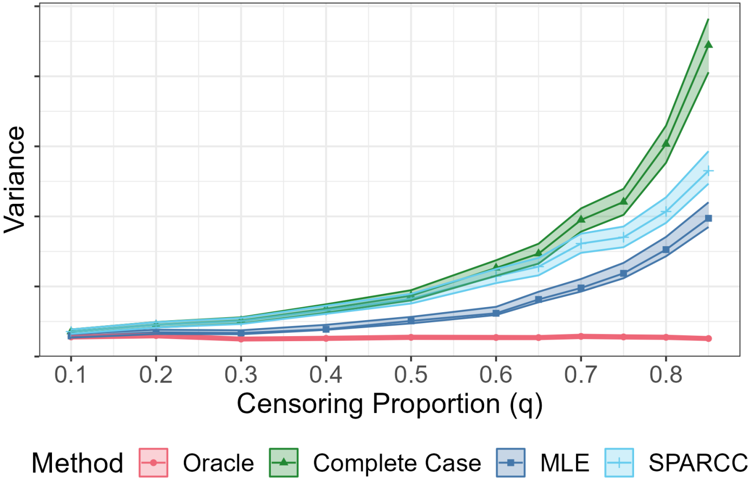
Simulation Results: Efficiency



Simulation Results: Efficiency



Simulation Results: Efficiency



Discussion

- Proposed **SPARCC estimator** balances two desirable properties:

  -robustness: doubly robust to parametric nuisance models or consistent with nonparametric nuisance models
  -efficiency: semiparametric efficient

Discussion

- Proposed **SPARCC estimator** balances two desirable properties:

  -**robustness:** doubly robust to parametric nuisance models or consistent with nonparametric nuisance models

  -**efficiency:** semiparametric efficient

- Not discussed:

  -How is $\mathbf{S}_{\text{eff}}$ computed?

Discussion

- Proposed **SPARCC estimator** balances two desirable properties:

  -**robustness:** doubly robust to parametric nuisance models or consistent with nonparametric nuisance models
  -**efficiency:** semiparametric efficient

- Not discussed:

  -How is $\mathbf{S}_{\text{eff}}$ computed?
  -What can the SPARCC estimator tell us about Huntington's disease symptom progression?

SPARCC: Semiparametric Censored Covariate Estimation

R package available at https://github.com/brian-d-richardson/sparcc

Appendix I: MLE Score Function

$$\mathbf{S}_{\boldsymbol{\beta}}(y, w, \delta, \mathbf{z}, \boldsymbol{\beta}) = \underbrace{\delta \mathbf{S}^{\mathrm{F}}_{\boldsymbol{\beta}}(y, w, \mathbf{z}, \boldsymbol{\beta})}_{\text{uncensored}} + \underbrace{(1-\delta)\frac{\mathrm{E}\{\mathrm{I}(X > w)\mathbf{S}^{\mathrm{F}}_{\boldsymbol{\beta}}(y, X, \mathbf{z}, \boldsymbol{\beta}) \mid y, \mathbf{z}\}}{\mathrm{E}\{\mathrm{I}(X > w) \mid y, \mathbf{z}\}}}_{\text{censored}}$$

Appendix II: Efficient Score Function

$$
\begin{aligned}
\mathbf{S}_{\text{eff}}(y, w, \delta, \mathbf{z}, \boldsymbol{\beta}) \equiv \delta\{\mathbf{S}^{\text{F}}_{\boldsymbol{\beta}}(y, w, \mathbf{z}, \boldsymbol{\beta}) - \boxed{\mathbf{a}(w, z, \boldsymbol{\beta})}\} \\
+ (1-\delta)\frac{\text{E}[\text{I}(X > w)\{\mathbf{S}^{\text{F}}_{\boldsymbol{\beta}}(y, X, \mathbf{z}, \boldsymbol{\beta}) - \boxed{\mathbf{a}(X, \mathbf{z}, \boldsymbol{\beta})}\} \mid y, \mathbf{z}]}{\text{E}\{\text{I}(X > w) \mid y, \mathbf{z}\}},
\end{aligned}
$$

Appendix II: Efficient Score Function

$$\mathbf{S}_{\text{eff}}(y, w, \delta, \mathbf{z}, \boldsymbol{\beta}) \equiv \delta\{\mathbf{S}^{\text{F}}_{\boldsymbol{\beta}}(y, w, \mathbf{z}, \boldsymbol{\beta}) - \boxed{\mathbf{a}(w, z, \boldsymbol{\beta})}\}$$
$$+ (1-\delta)\frac{\text{E}[\text{I}(X > w)\{\mathbf{S}^{\text{F}}_{\boldsymbol{\beta}}(y, X, \mathbf{z}, \boldsymbol{\beta}) - \boxed{\mathbf{a}(X, \mathbf{z}, \boldsymbol{\beta})}\} \mid y, \mathbf{z}]}{\text{E}\{\text{I}(X > w) \mid y, \mathbf{z}\}},$$

where $\mathbf{a}(x, \mathbf{z}, \boldsymbol{\beta})$ satisfies

$$\text{E}\{\text{I}(x \leq C) \mid \mathbf{z}\}\mathbf{a}(x, \mathbf{z}, \boldsymbol{\beta}) + \text{E}\left[\text{I}(x > C)\frac{\text{E}\{\text{I}(X > C)\mathbf{a}(X, \mathbf{z}, \boldsymbol{\beta}) \mid Y, C, \mathbf{z}\}}{\text{E}\{\text{I}(X > C) \mid Y, C, \mathbf{z}\}} \mid x, \mathbf{z}\right]$$
$$= \text{E}\left[\text{I}(x > C)\frac{\text{E}\{\text{I}(X > C)\mathbf{S}^{\text{F}}_{\boldsymbol{\beta}}(Y, X, \mathbf{z}, \boldsymbol{\beta}) \mid Y, C, \mathbf{z}\}}{\text{E}\{\text{I}(X > C) \mid Y, C, \mathbf{z}\}} \mid x, \mathbf{z}\right]$$

Thank you! Any questions?

Brian Richardson
🖂: brichson@ad.unc.edu