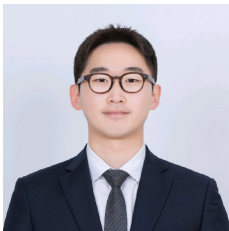# Doubly robust estimation under a randomly censored covariate

Brian Richardson

Acknowledgements

Seong-Ho Lee, PhD · Tanya Garcia, PhD · Yanyuan Ma, PhD
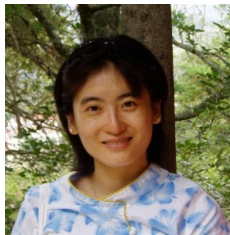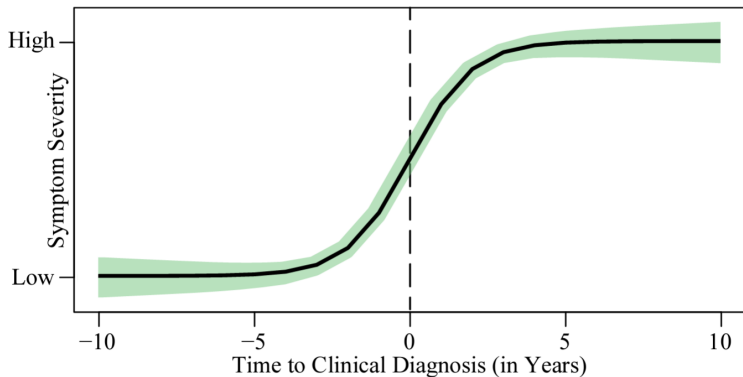
Huntington's Disease and Censored Covariates

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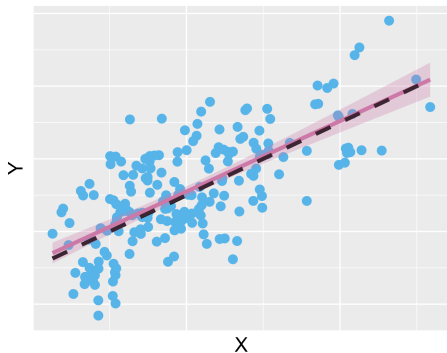
- Regression model:
  $$E(Y) = \beta_0 + \beta_1 X$$

Censored Covariates: a Simple Example
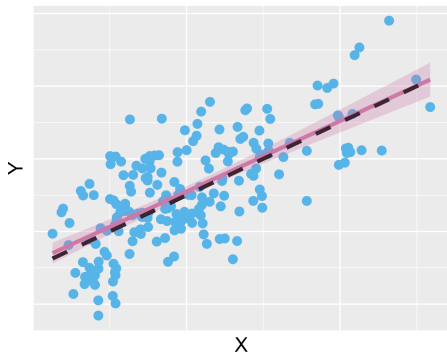
- Regression model:
  $$\mathrm{E}(Y) = \beta_0 + \beta_1 X$$

- Estimate $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ with least squares/maximum likelihood

Censored Covariates: a Simple Example
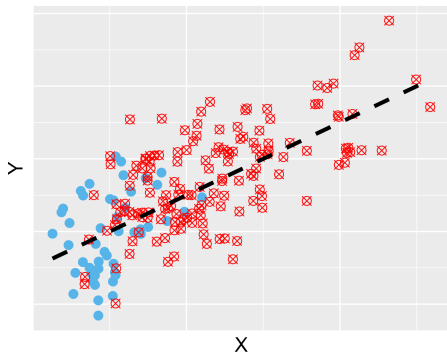
Problem: X is censored by a **random censoring time** C

Censored Covariates: a Simple Example
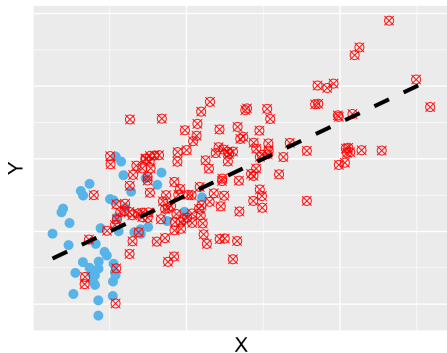
Problem: X is censored by a **random censoring time** C

- $W = \min(X, C)$

- $\Delta = I(X \leq C)$

Censored Covariates: a Simple Example
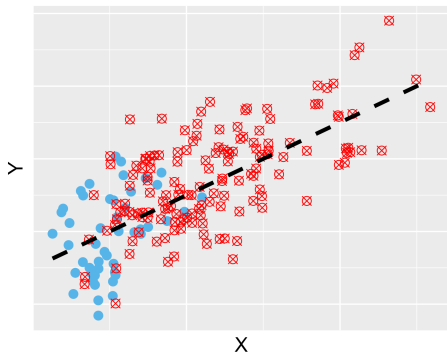
Problem: X is censored by a **random censoring time** C

- $W = \min(X, C)$

- $\Delta = I(X \leq C)$

- assume: $C \perp\!\!\!\perp (X, Y)$

Complete Case Analysis
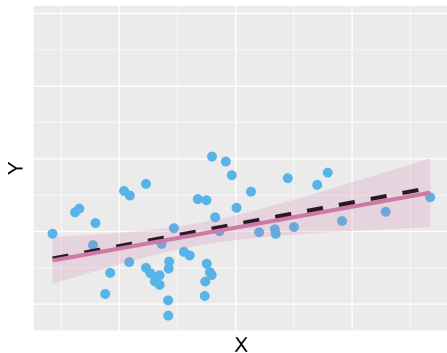
Only use observations where X is *uncensored*

Complete Case Analysis
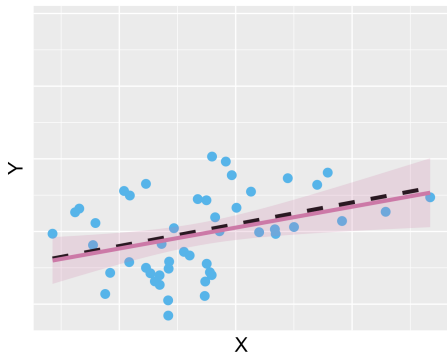
Only use observations where X is *uncensored*

✓ Consistent

Complete Case Analysis
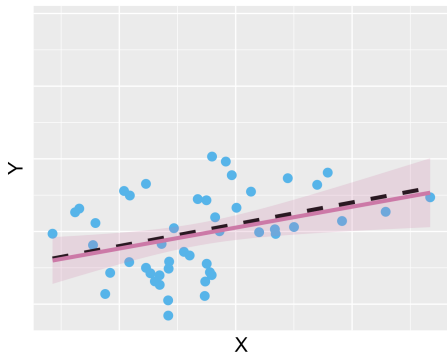
Only use observations where X is *uncensored*

✓ 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)

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

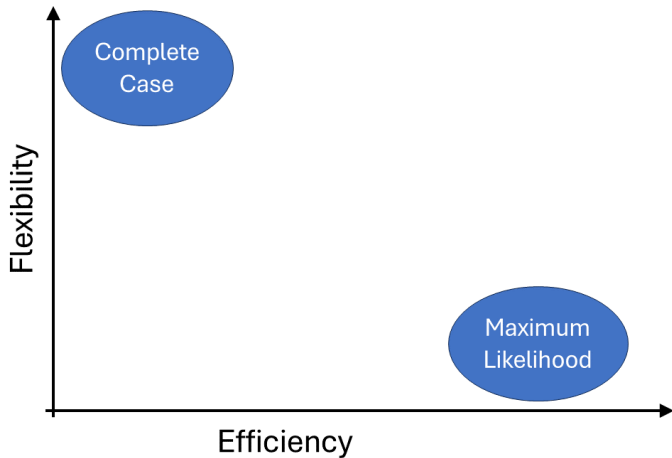$$\mathbf{S}_{\boldsymbol{\beta}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \qquad \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\beta}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

✓ consistent

✓ fully efficient

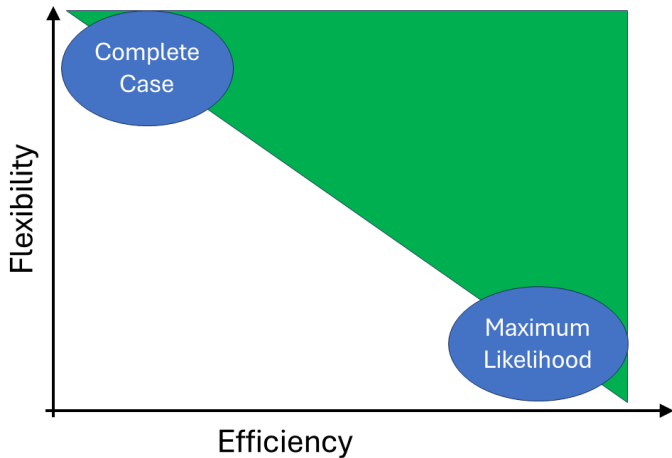✗ inconsistent when model for **nuisance parameter** f_X is incorrect

Existing Methods

Existing Opportunity

A New Approach
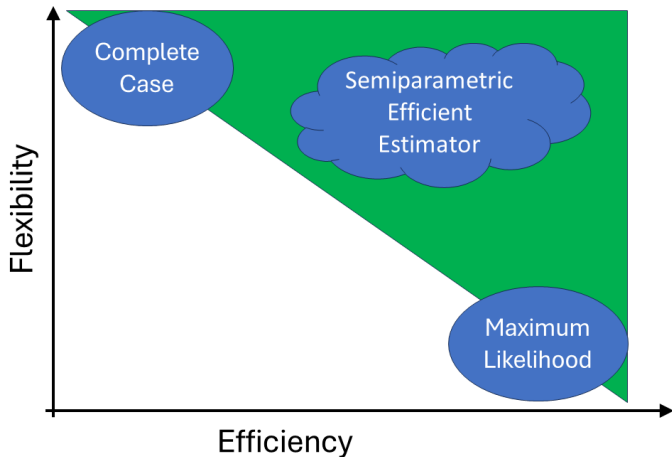
The Semiparametric Recipe

- **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 $\boldsymbol{\eta}$

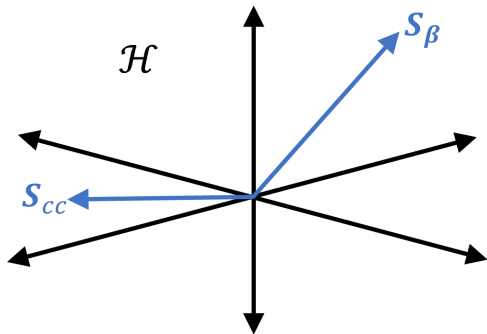# The Semiparametric Recipe

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

- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$

- **semiparametric**: avoid parametric assumptions for $\boldsymbol{\eta}$

- **goal:** find semiparametric efficient estimating function

The Semiparametric Recipe

- **Hilbert space** of estimating functions

The Semiparametric Recipe
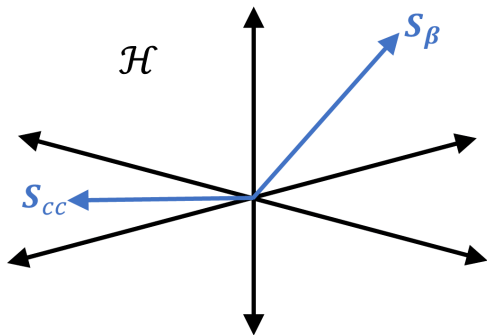
- **Hilbert space** of estimating functions

- covariance inner product $\langle \boldsymbol{h}, \boldsymbol{g} \rangle \equiv \mathrm{E}(\boldsymbol{h}^T \boldsymbol{g})$

The Semiparametric Recipe
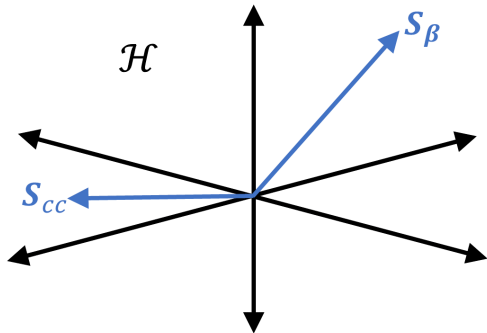
- **Hilbert space** of estimating functions

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

- **orthogonal** ⇔ **uncorrelated**

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

The Semiparametric Recipe

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

Λ

$\Lambda^{\perp}$
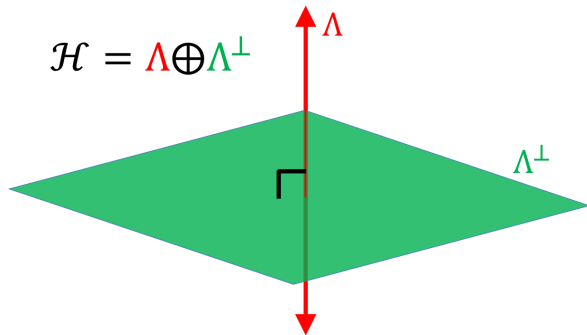
- construct Λ using **nuisance scores**

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

The Semiparametric Recipe

$$\mathcal{H} = \Lambda \oplus \Lambda^{\perp}$$
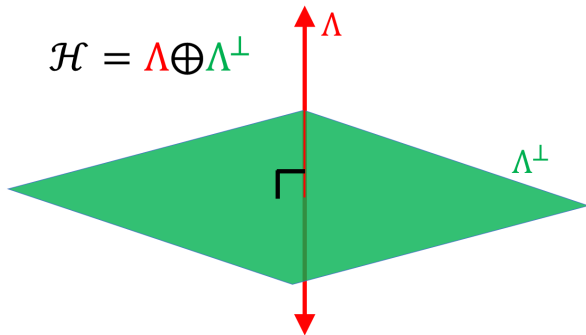
- construct Λ using **nuisance scores**

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

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

The Semiparametric Recipe
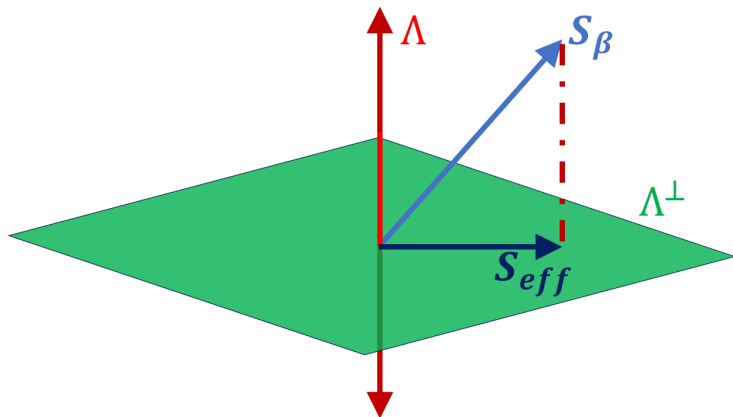
Properties of the Proposed Estimator

The **semiparametric efficient estimator** $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ is the solution to

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

Properties of the Proposed Estimator

The **semiparametric efficient estimator** $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ is the solution to

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

✓ **Doubly Robust**: $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ is consistent if at least one of f_X, f_C is correctly specified

Properties of the Proposed Estimator

The **semiparametric efficient estimator** $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ is the solution to

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

✓ **Doubly Robust**: $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ is consistent if at least one of f_X, f_C is correctly specified

✓ **Locally Efficiency**: If f_X, f_C are *both* correctly specified, then $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ achieves the **semiparametric efficiency bound**

Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

- large sample size $n = 10,000$

Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

- large sample size $n = 10,000$

- high censoring rate $q = \mathrm{P}(X > C) = 0.75$
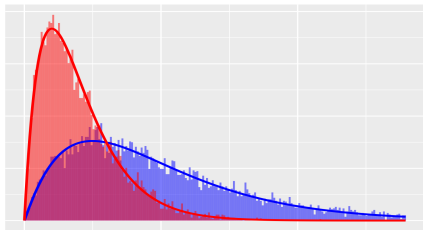
Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

- large sample size $n = 10,000$

- high censoring rate $q = \mathrm{P}(X > C) = 0.75$

- $X, C \sim$ gamma distributions

Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

- large sample size $n = 10,000$

- high censoring rate $q = \mathrm{P}(X > C) = 0.75$

- $X, C \sim$ gamma distributions
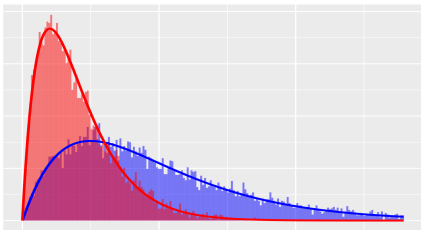
- X, C possibly **misspecified** as exponential
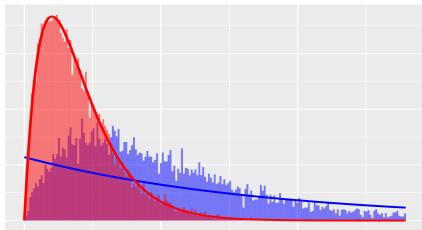
Simulating Model Misspecification
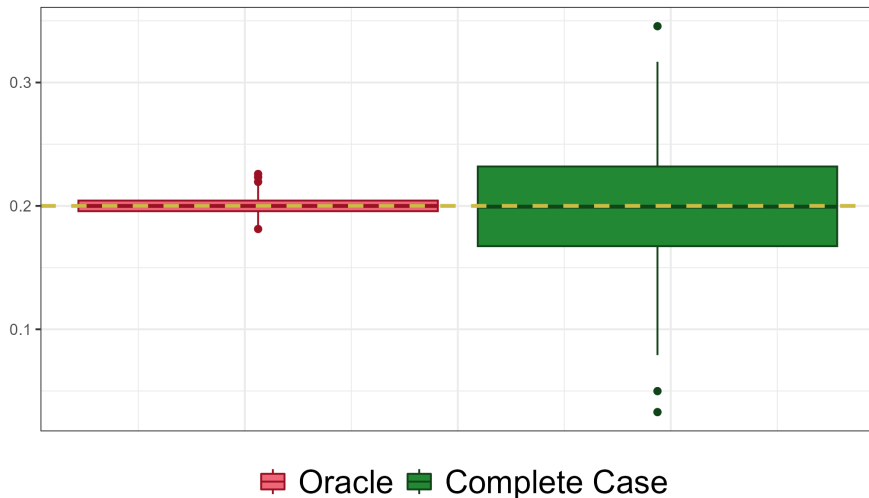
X, C correct

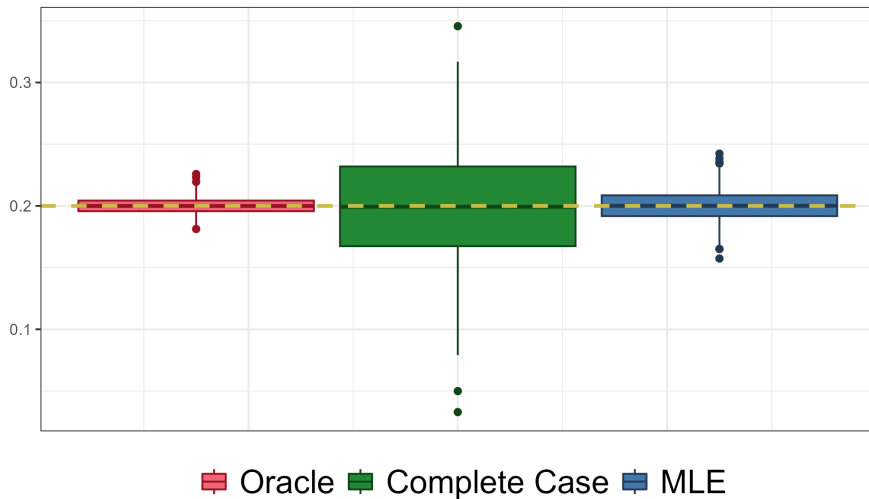# Simulating Model Misspecification

X, C correct

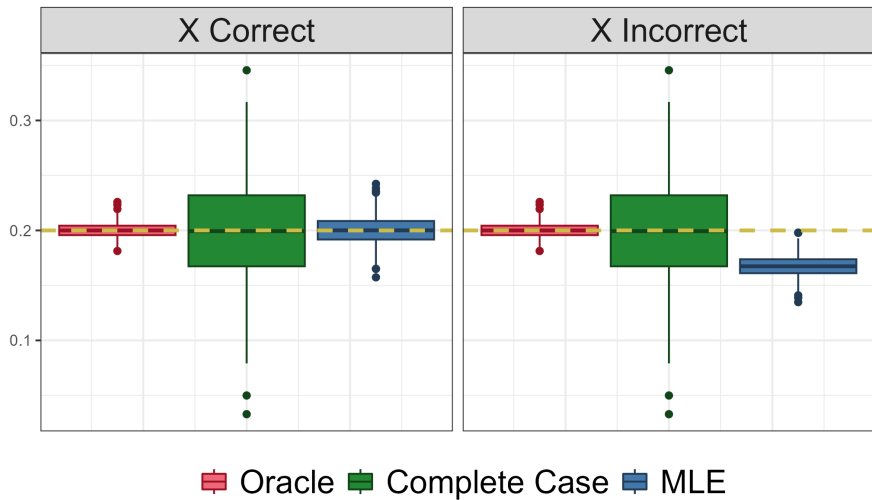X incorrect, C correct

Simulation Results

Simulation Results

Simulation Results

Simulation Results
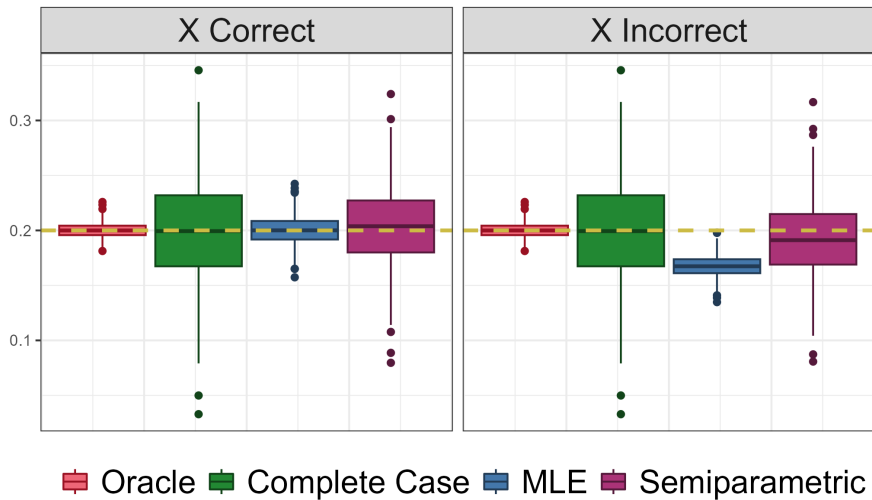
Simulation Results
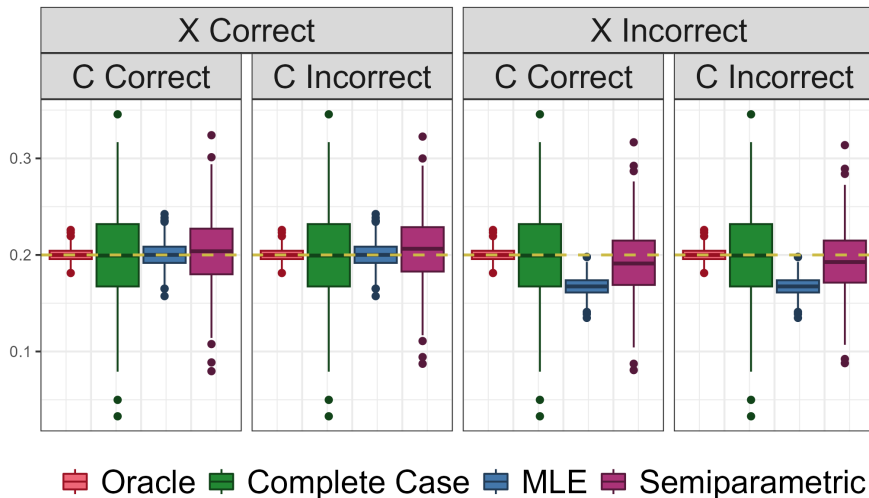
Generalizations

The methods presented here extend to:

- Nonlinear $\mathrm{E}(Y|X) = m(X, \boldsymbol{\beta})$

Generalizations

The methods presented here extend to:

- Nonlinear $\mathrm{E}(Y|X) = m(X, \boldsymbol{\beta})$

- Additional uncensored covariates $\mathbf{Z}$

  - $\mathrm{E}(Y|X, \mathbf{Z}) = m(X, \mathbf{Z}, \boldsymbol{\beta})$

  - Nuisance distributions become $f_{X|\mathbf{Z}}, f_{C|\mathbf{Z}}, f_{\mathbf{z}}$

SPARCC: <u>S</u>emiparametric <u>C</u>ensored <u>C</u>ovariate Estimation

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

Appendix I: MLE Score Function

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

Appendix II: Efficient Score Function

$$\mathbf{S}_{\mathrm{eff}}(y, w, \delta, \mathbf{z}, \boldsymbol{\beta}) \equiv \delta\{\mathbf{S}^{\mathrm{F}}_{\boldsymbol{\beta}}(y, w, \mathbf{z}, \boldsymbol{\beta}) - \boxed{\mathbf{a}(w, z, \boldsymbol{\beta})}\}$$

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

Appendix II: Efficient Score Function

$$\mathbf{S}_{\text{eff}}(y, w, \delta, \mathbf{z}, \boldsymbol{\beta}) \equiv \delta\{\mathbf{S}_{\boldsymbol{\beta}}^{\text{F}}(y, w, \mathbf{z}, \boldsymbol{\beta}) - \boxed{\mathbf{a}(w, z, \boldsymbol{\beta})}\}$$

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

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

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

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

Thank you! Any questions?

Brian Richardson
✉: brichson@ad.unc.edu