



GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

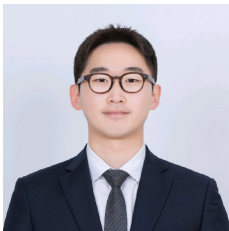


Doubly robust estimation under a randomly censored covariate

Brian Richardson

Acknowledgements

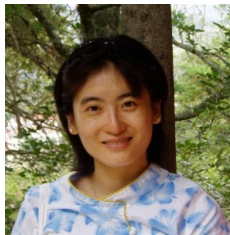
Seong-Ho Lee, PhD



Tanya Garcia, PhD

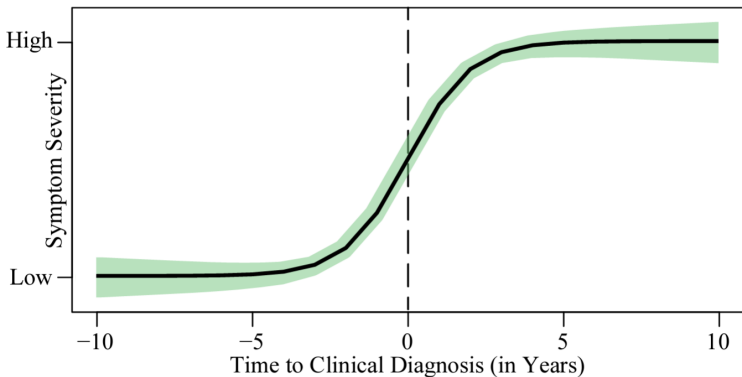


Yanyuan Ma, PhD



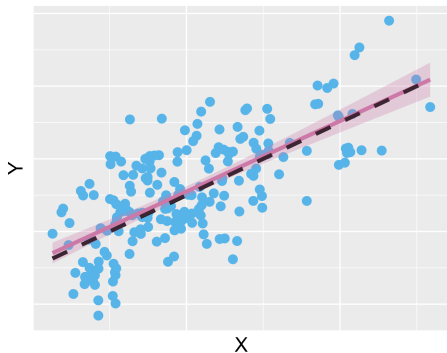
This research was supported by the National Institute of Environmental Health Sciences grant T32ES007018.

Huntington's Disease and Censored Covariates



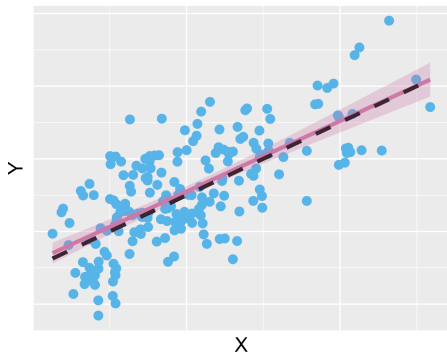
Sarah C Lotspeich et al. "Making Sense of Censored Covariates: Statistical Methods for Studies of Huntington's Disease". In: *Annual Review of Statistics and Its Application* 11 (2024)

Censored Covariates: a Simple Example



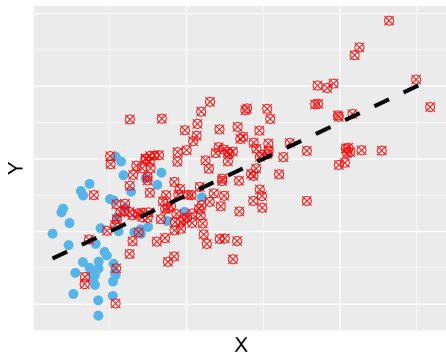
- Regression model:
 $E(Y) = \beta_0 + \beta_1 X$

Censored Covariates: a Simple Example



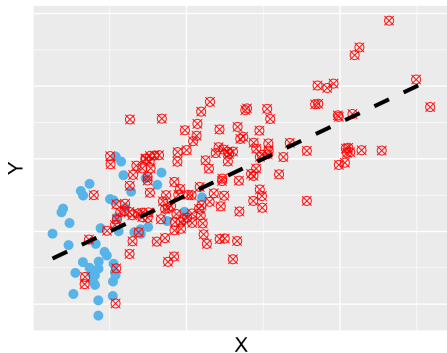
- Regression model:
 $E(Y) = \beta_0 + \beta_1 X$
- Estimate $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ with least squares/maximum likelihood

Censored Covariates: a Simple Example



Problem: X is censored by a random censoring time C

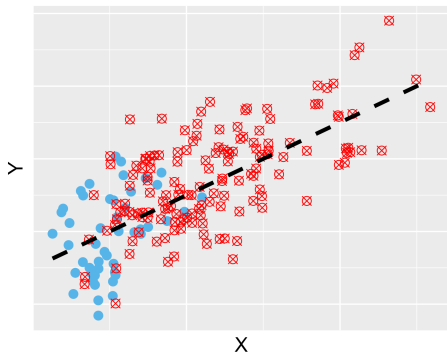
Censored Covariates: a Simple Example



Problem: X is censored by a random censoring time C

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

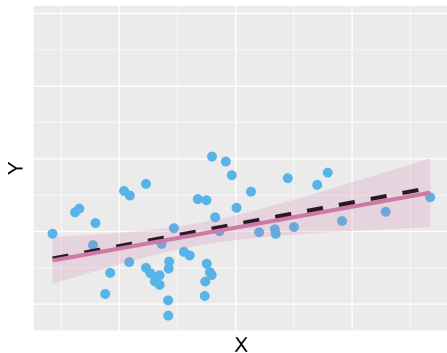
Censored Covariates: a Simple Example



Problem: X is censored by a random censoring time C

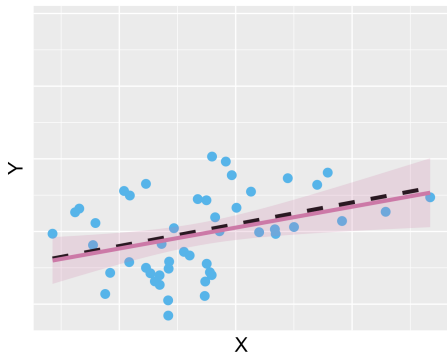
- $W = \min(X, C)$
- $\Delta = I(X \leq C)$
- assume: $C \perp\!\!\!\perp (X, Y)$

Complete Case Analysis



Only use observations where X is *uncensored*

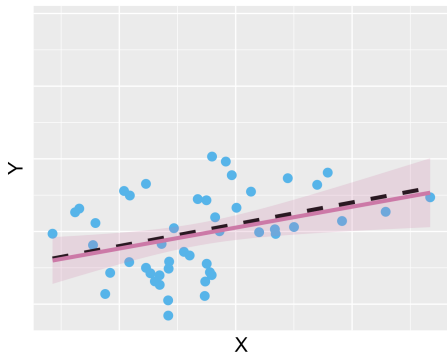
Complete Case Analysis



Only use observations where X is *uncensored*

✓ Consistent

Complete Case Analysis



Only use observations where X is *uncensored*

- ✓ Consistent
- ✗ Inefficient

Maximum Likelihood Estimation (MLE)

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

Maximum Likelihood Estimation (MLE)

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

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

Maximum Likelihood Estimation (MLE)

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

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

- ✓ consistent
- ✓ fully efficient

Maximum Likelihood Estimation (MLE)

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

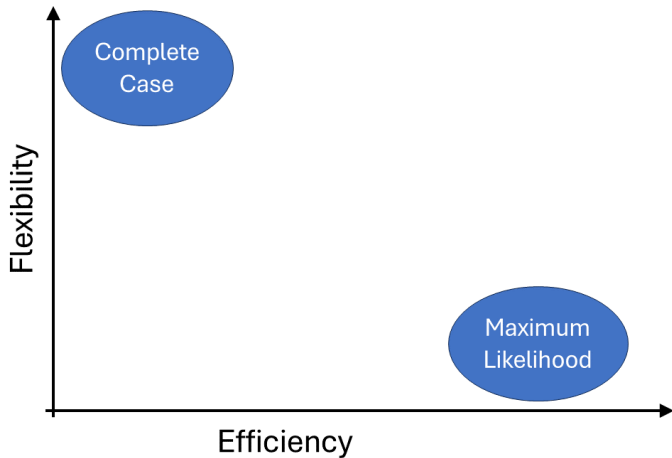
$$\mathbf{S}_\beta(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \quad \sum_{i=1}^n \mathbf{S}_\beta(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

✓ consistent

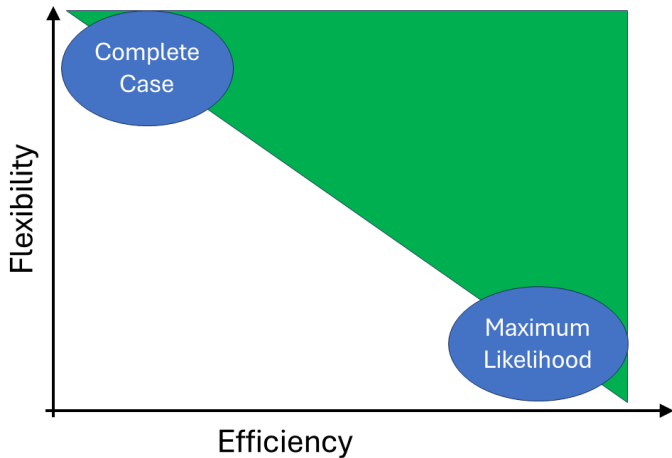
✓ fully efficient

✗ inconsistent when model for
nuisance parameter f_X is
incorrect

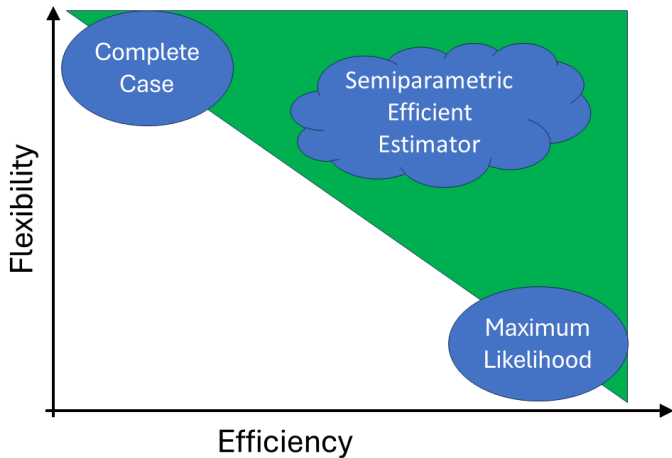
Existing Methods



Existing Opportunity



A New Approach



The Semiparametric Recipe

- **of interest:** parameter β characterizing $Y|X$

The Semiparametric Recipe

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

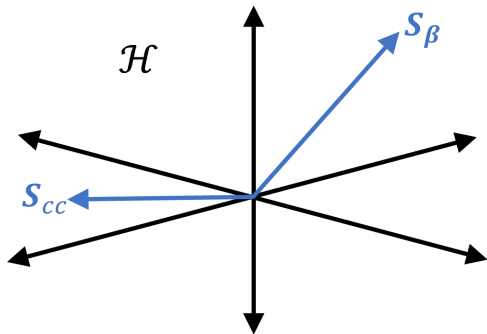
The Semiparametric Recipe

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

The Semiparametric Recipe

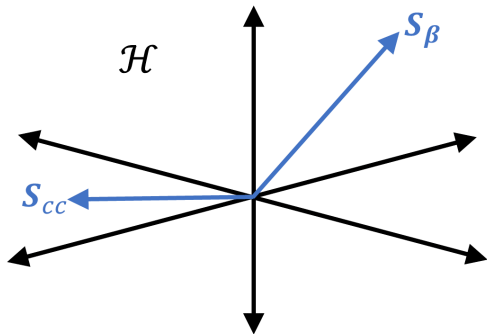
- **of interest:** parameter β characterizing $Y|X$
- **nuisance:** distributions $(f_X, f_C) = \boldsymbol{\eta}$
- **semiparametric:** avoid parametric assumptions for $\boldsymbol{\eta}$
- **goal:** find semiparametric efficient estimating function

The Semiparametric Recipe



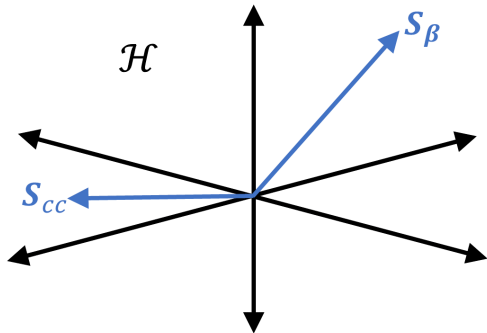
- Hilbert space of estimating functions

The Semiparametric Recipe



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

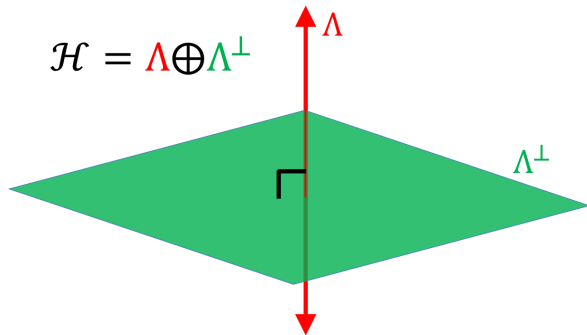
The Semiparametric Recipe



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

$$h \perp g \iff \langle h, g \rangle = 0$$

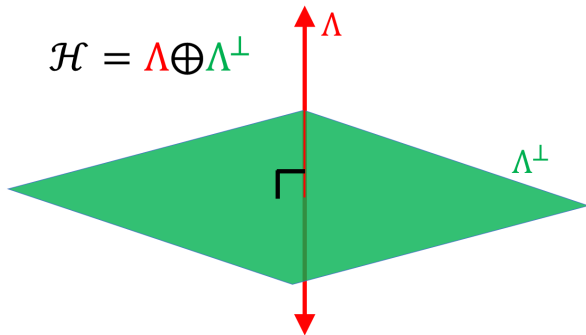
The Semiparametric Recipe



- construct Λ using **nuisance scores**

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

The Semiparametric Recipe

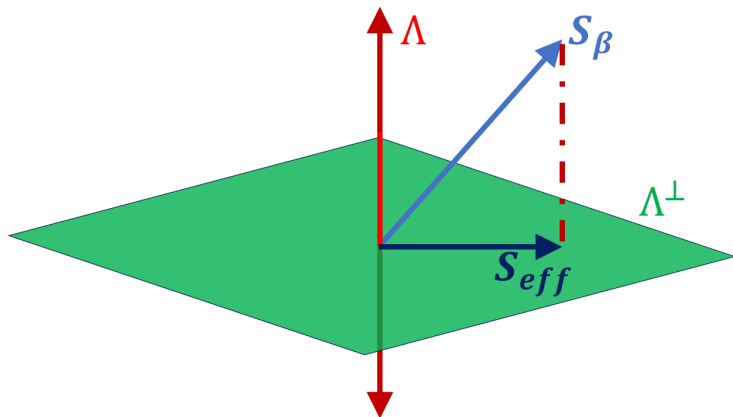


- construct Λ using **nuisance scores**

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

- orthogonal complement Λ^\perp

The Semiparametric Recipe



Properties of the Proposed Estimator

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

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

Properties of the Proposed Estimator

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

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

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

Properties of the Proposed Estimator

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

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

- ✓ **Doubly Robust:** $\hat{\beta}_{\text{eff}}$ 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 $\hat{\beta}_{\text{eff}}$ 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

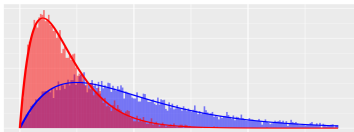
- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$
- large sample size $n = 10,000$
- high censoring rate $q = P(X > C) = 0.75$

Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$
- large sample size $n = 10,000$
- high censoring rate $q = P(X > C) = 0.75$
- $X, C \sim$ gamma distributions

Simulation Setup

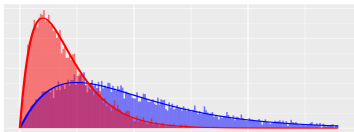
- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$
- large sample size $n = 10,000$
- high censoring rate $q = P(X > C) = 0.75$
- $X, C \sim$ gamma distributions
- X, C possibly **misspecified** as exponential



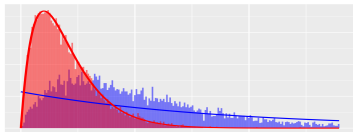
X, C correct

Simulation Setup

- $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$
- large sample size $n = 10,000$
- high censoring rate $q = P(X > C) = 0.75$
- $X, C \sim$ gamma distributions
- X, C possibly **misspecified** as exponential

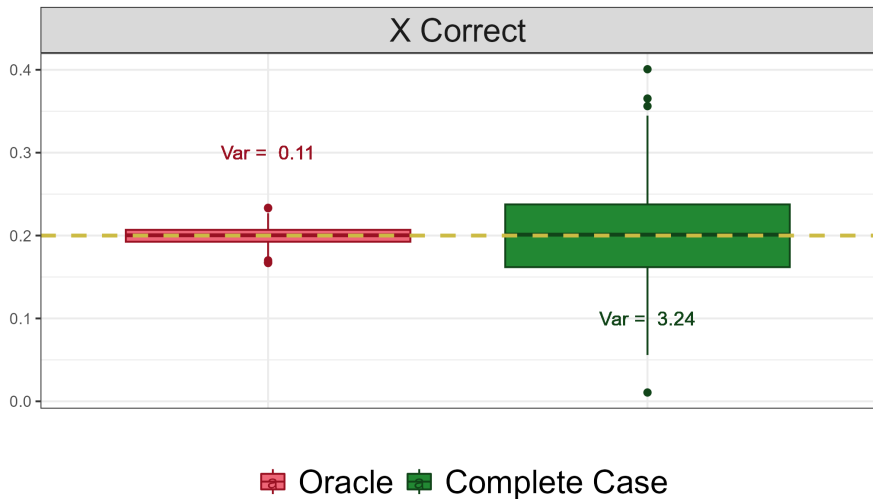


X, C correct

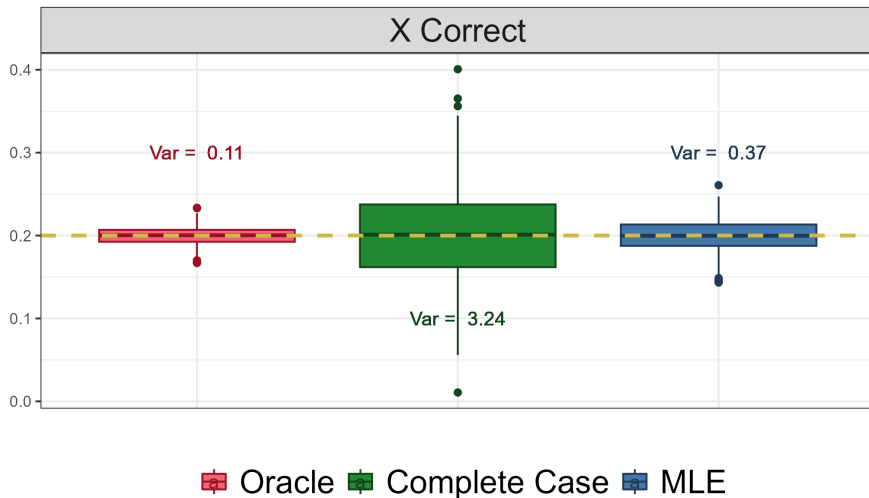


X incorrect, C correct

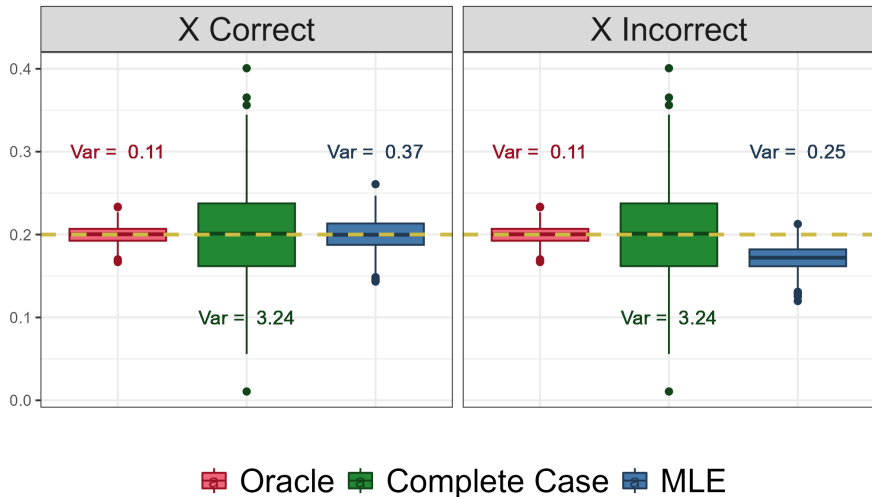
Simulation Results



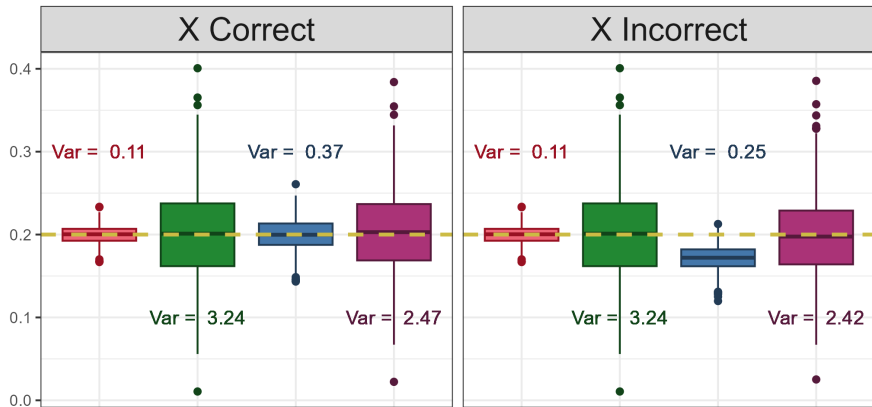
Simulation Results



Simulation Results

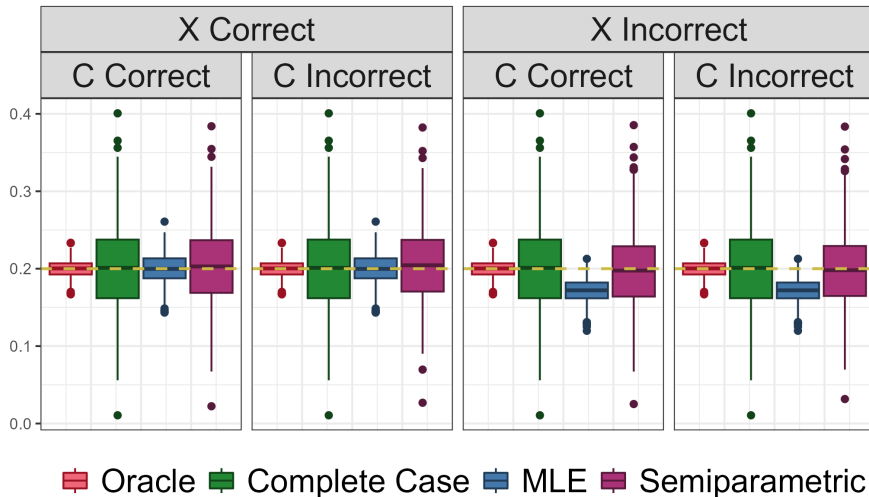


Simulation Results



Oracle Complete Case MLE Semiparametric

Simulation Results



Generalizations

The methods presented here extend to:

- Nonlinear $E(Y|X) = m(X, \beta)$

Generalizations

The methods presented here extend to:

- Nonlinear $E(Y|X) = m(X, \beta)$
- Additional uncensored covariates \mathbf{Z}
 - $E(Y|X, \mathbf{Z}) = m(X, \mathbf{Z}, \beta)$
 - Nuisance distributions become $f_{X|\mathbf{Z}}, f_{C|\mathbf{Z}}, f_{\mathbf{Z}}$

SPARCC: Semiparametric Censored Covariate Estimation



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

Appendix I: MLE Score Function

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

Appendix II: Efficient Score Function

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

Appendix II: Efficient Score Function

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

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

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

Thank you! Any questions?

Brian Richardson

✉: brichson@ad.unc.edu