Semiparametric Estimation for Non-randomly Truncated Data

Yuyao Wang [1], Andrew Ying[2] and Ronghui Xu[1] [3]

[1]Department of mathematics, [3]Herbert Wertheim School of Public Health and Halicioglu Data Science Institute, UC San Diego

[2]Department of Statistics and Data Science, The Wharton School, Univeristy of Pennsylvania

UC San Diego Mathematics

Background

- In prospective cohort studies, only subjects with event times greater than enrollment times are included.
- Subjects with early event times tend not to be captured, leading to selection bias.
- Conventional methods adjusting for left truncation typically hinges heavily on the assumption that the left truncation time and the event time are independent.
- When the left truncation time depends on additional covariates, inverse probability of truncation weighting can be used, but it is sensitive to model misspecification.

Our Contributions

- Leverage semiparametric theory to find the efficient influence curve (EIC) of $\psi := \mathbb{E}[\nu(T)]$, where ν is a known function.
- Construct EIC-based estimators that enjoy model double-robustness and rate double robustness.
- Apply our estimator to analyze a data set related to Alzheimer's disease research.

Notation

- Q, T: left truncation time and event time.
- Z: covariates.
- F, G: full data CDF of $T|Z$ and $Q|Z$ respectively.
- $\beta := \mathbb{P}(Q < T)$
- $\bar{M}_Q(t) := \bar{N}_Q(t) - \bar{A}_Q(t)$ is a backward martingale with respect to $\bar{\mathcal{F}}_t$, where

  $\bar{N}_Q(t) := I(t \le Q < T)$,

  $\bar{A}_Q(t) := \int_t^\infty I(Q \le s < T)\frac{dG(s|Z)}{G(s|Z)}$,

  $\bar{\mathcal{F}}_t := \sigma\{Z, I(Q < T), I(s \le T), I(s \le Q) : s \ge t\}$.
- $\mathbb{E}^*(\cdot)$: expectation under observed data distribution.

Methods

- The efficient influence curve of ψ is

$$\varphi(Q,T,Z;\psi,F,G) = \beta \cdot \left\{\frac{\nu(T)-\psi}{G(T|Z)} - \int \frac{m(v,Z;F) - \psi F(v|Z)}{G(v|Z)(1-F(v|Z))} d\bar{M}_Q(v)\right\},$$

  where $m(v,Z;F) = \mathbb{E}[\nu(T)I(T<v)|Z]$ is the trimmed conditional mean of $\nu(T)$ in full data.
-

$$\hat{\psi}_{\mathrm{DR}} \longleftarrow \text{ Solving } \sum_{i=1}^n \varphi(Q_i, T_i, Z_i; \psi, \hat{F}, \hat{G}) = 0,$$

  where $\hat{F}$ and $\hat{G}$ are the first stage estimators for F and G.
- Plugging constant first stage estimators $\hat{F} \equiv 0$ or $\hat{G} \equiv 1$ into $\hat{\psi}_{\mathrm{DR}}$ leads to estimators that only depend on one of the first stage estimators:

$$\hat{\psi}_{\mathrm{IPW1}} = \left(\frac{1}{n}\sum_{i=1}^n \frac{\nu(T_i)}{\hat{G}(T_i|Z_i)}\right) \Bigg/ \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\hat{G}(T_i|Z_i)}\right),$$

  and

$$\hat{\psi}_{\mathrm{IPW2}} = \left(\frac{1}{n}\sum_{i=1}^n \frac{\mu_\nu(Z_i;\hat{F})}{\hat{S}_T(Q_i|Z_i)}\right) \Bigg/ \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\hat{S}_T(Q_i|Z_i)}\right),$$

  where $\hat{S}_T(Q_i|Z_i) = 1 - \hat{F}(Q_i|Z_i)$ and $\mu_\nu(Z;\hat{F}) = \int \nu(t) d\hat{F}(t|Z)$.

Double Robustness

$\mathbb{E}^*[\varphi(Q,T,Z;\psi_0,F,G)] = 0$ if either $F = F_0$ or $G = G_0$, where F_0, G_0 denote the true CDF's.

- **(Model double robustness)** $\hat{\psi}_{\mathrm{DR}}$ is consistent and asymptotically normal when one of $\hat{F}$ and $\hat{G}$ converges to the truth at root-n rate and the other one converges.
- **(Rate double robustness)** $\hat{\psi}_{\mathrm{DR}}$ is consistent and asymptotically normal and achieves the semiparametric efficiency bound if both models for F or G are correct and the product of the convergence rates for $\hat{F}$ and $\hat{G}$ is $o(n^{-1/2})$.

Simulation Results

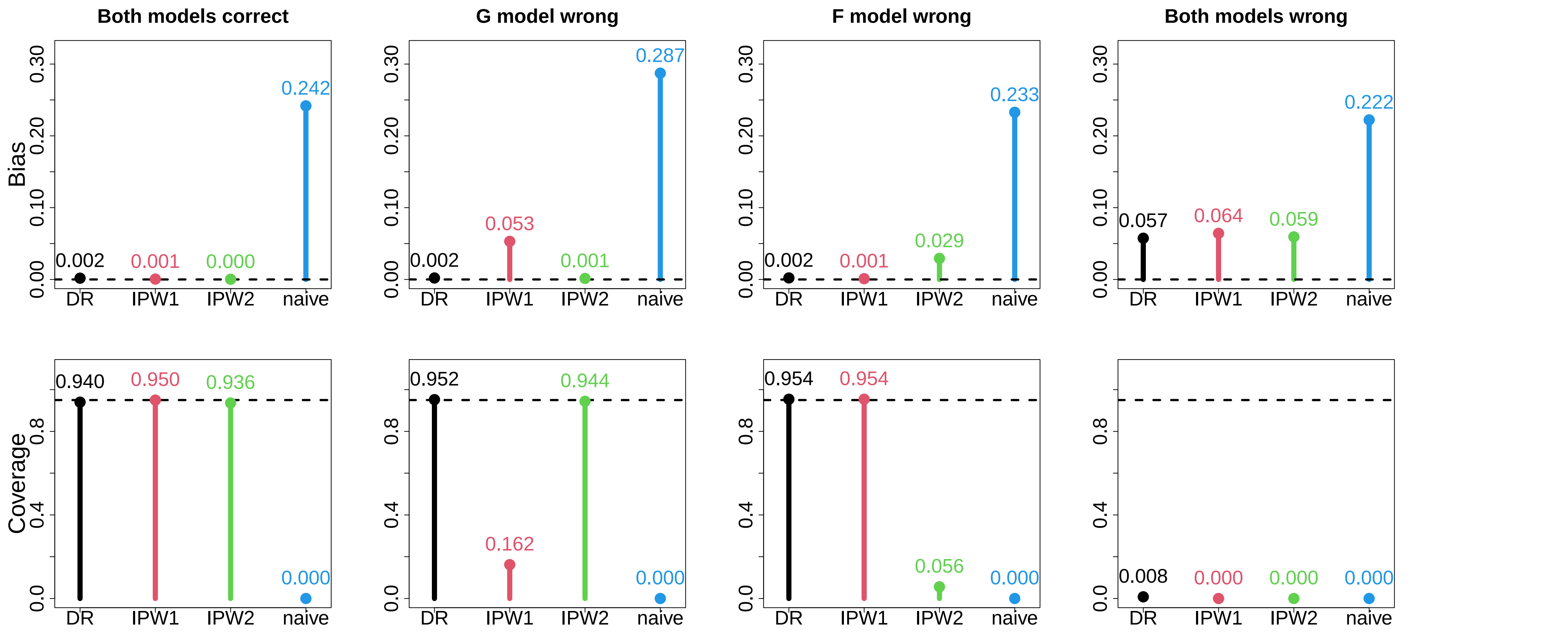
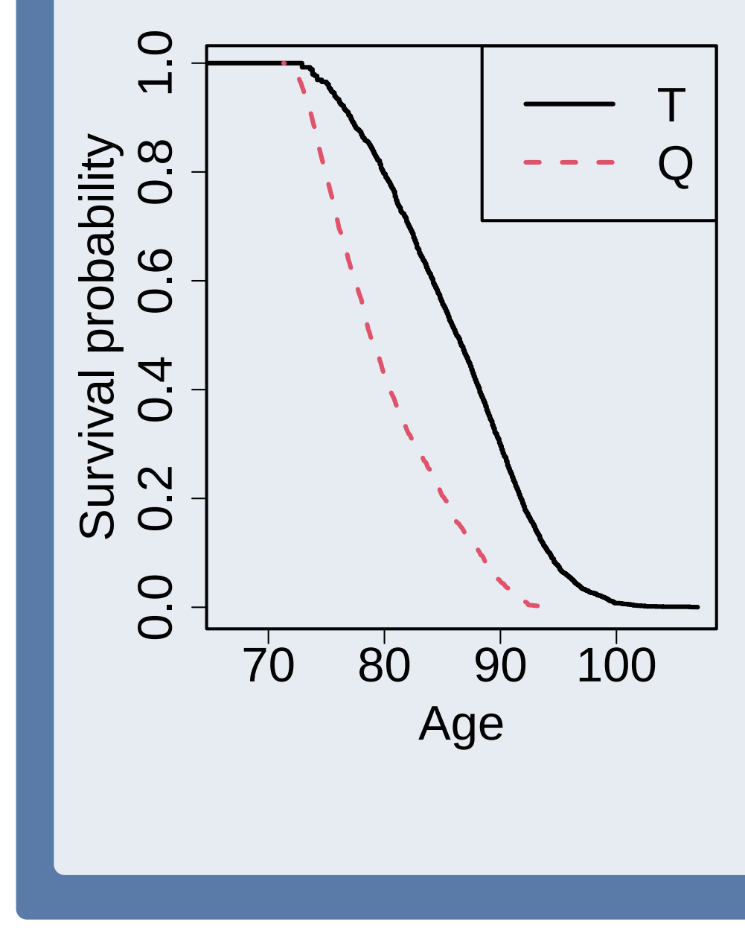


Figure: The absolute bias and the coverage probability of 95% confidence intervals with $\hat{F}$ and $\hat{G}$ fitted using Cox models under different data generation scenarios. Under the scenarios where Cox models are wrong for estimating F and/or G, T and/or Q are simulated from a mixture of Cox model with quadratic and interaction terms and AFT model with quadratic and interaction terms.

Application to HHP/HAAS Data

We consider the data collected between 1965 and 2012 from Honolulu Heart Program (HHP) and Honolulu Asia Aging Study (HAAS). Age is the time scale of interest. The data set contains 2318 Japanese men that were alive at the start of HAAS, so age at death (late-life mortality) is left truncated. The covariates include education, apolipoprotein E genotype, mid-life acohol and cigarettes consumption, systolic blood pressure, ventricular rate, and grip strength. Since age at the start of HAAS can be predicted by education, alcohol consumption, systolic blood pressure, ventricular rate and grip strength, which are also risk factors of mortality, left truncation is nonrandom for this data set. The first stage estimators $\hat{F}$ and $\hat{G}$ are obtained using Cox models with all covariates included.

Table: Estimates using difference methods from the HHP/HAAS data.



Estimand	Method	Estimate	SE / boot SE	95% CIs / 95% boot CIs
$\mathbb{E}(T)$	DR	86.121	0.232 / 0.218	(85.666, 86.576) / (85.693, 86.549)
	IPW	85.911	0.211	(85.499, 86.324)
	IPW1	86.097	0.192	(85.722, 86.473)
	IPW2	86.079	0.196	(85.694, 86.464)
	naive	87.942	0.103	(87.741, 88.144)
$\mathbb{P}(T > 80)$	DR	0.800	0.016 / 0.016	(0.768, 0.832) / (0.770, 0.831)
	IPW	0.787	0.015	(0.757, 0.817)
	IPW1	0.800	0.016	(0.770, 0.831)
	IPW2	0.800	0.016	(0.769, 0.831)
	naive	0.904	0.007	(0.890, 0.918)

Discussion

- The double robustness of our estimator can be easily extended to estimating the average treatment effect with non-randomly truncated time-to-event data in randomized trials.
- The extension to the case with censoring is nontrivial and is an interesting future direction.