DIMACS/CCICADA Workshop
Rutgers University

February 4, 2011

****

Understanding and Improving Propensity Score Methods

Zhiqiang Tan

Department of Statistics
Rutgers University

http://www.stat.rutgers.edu/∼ztan

Outline

Understanding

- Introduction (to a medical study)
- Causal inference & missing-data problems
- Outcome regression vs propensity score weighting

Improvements

- Doubly robust estimation
- Likelihood approach
- Doubly robust likelihood estimator

Effect of right heart catheterization

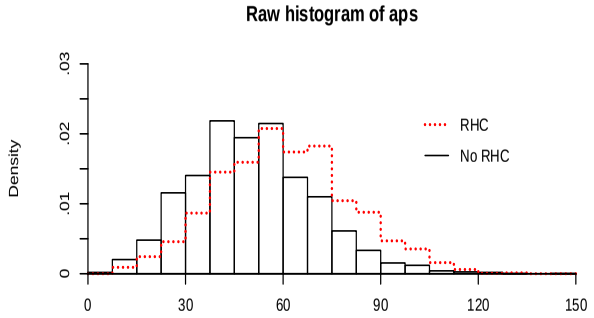
- Right heart catheterization (RHC) is performed daily in hospitals since 1970s.

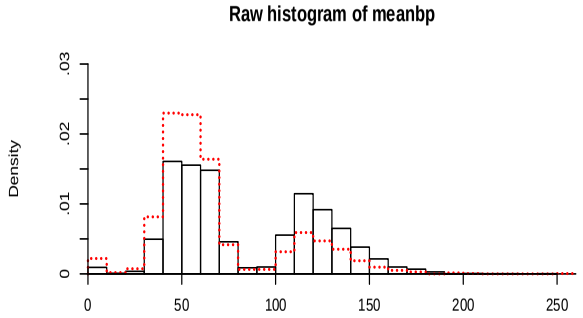
- The benefit of RHC had not been demonstrated in a successful randomized clinical trial.

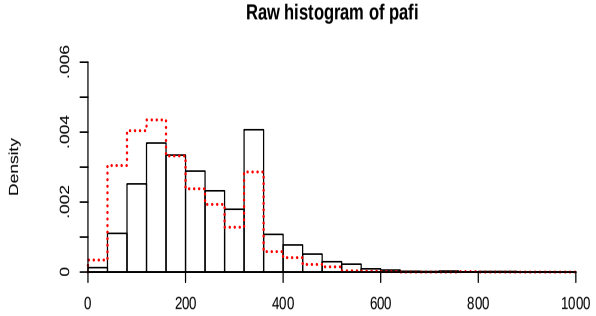
- Connors et al.'s (1996) observational study raised the concern that RHC might not benefit critically ill patients and might in fact cause harm.

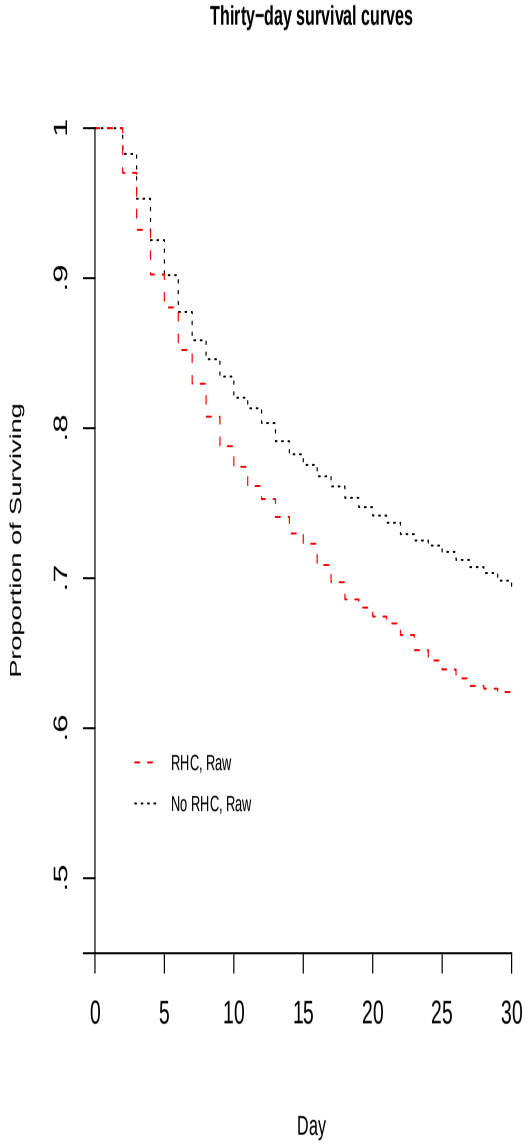
- Data were collected on 5735 critically ill patients admitted to the ICUs of five medical centers:
  - Treatment: No-RHC or RHC
  - Outcome: 30-day survival
  - Covariates: 75 covariates (specified by doctors)

- What is the effect of RHC on survival?


Raw histogram of aps
Density
0
.01
.02
.03
0
30
60
90
120
150
RHC
No RHC


Raw histogram of meanbp
Density
0
.01
.02
.03
0
50
100
150
200
250


Raw histogram of pafi
Density
0
.002
.004
.006
0
200
400
600
800
1000


Thirty-day survival curves
Proportion of Surviving
.5
.6
.7
.8
.9
1
0
5
10
15
20
25
30
Day
RHC, Raw
No RHC, Raw

Causal inference (Rubin 1974, 1978, etc)

- X: pre-treatment covariates (measured)
- (Y_0, Y_1): potential outcome that would be observed had a patient received No-RHC or RHC
- T: treatment variable taking value 0 or 1 if a patient actually received No-RHC or RHC
- $Y = (1 - T)\, Y_0 + T\, Y_1$: observed outcome
- Observed data & missing data

id	X_1	$\cdots$	X_{75}	T	Y	Y_0	Y_1
1	*	*	*	0	10	10	??
2	*	*	*	1	7	??	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	*	*	*	0	12	12	??

- Average causal effects:

$$E(Y_1) \text{ versus } E(Y_0)$$

or other comparisons ...

Missing-data problems

- X: a vector of explanatory variables
- Y: an outcome variable
- R: non-missing indicator
  $= 1$ if Y is observed or 0 if Y is missing
- Data structure

id	X_1	$\cdots$	X_{75}	R	Y
1	$*$	$*$	$*$	0	??
2	$*$	$*$	$*$	1	7
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$*$	$*$	$*$	0	??

- Assumption of ignorability:

$$R \perp Y \,|\, X$$

- Objective: to estimate $\mu = E(Y)$

Survey sampling with a superpopulation

- X: a vector of auxiliary variables
  (measured on the entire population)

- Y: a study variable

- R: sampling indicator
  $= 1$ if Y is sampled or 0 otherwise

- Data structure

id	X_1	$\cdots$	X_{75}	R	Y
1	$*$	$*$	$*$	0	??
2	$*$	$*$	$*$	1	7
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$*$	$*$	$*$	0	??

- By design:

$$R \perp Y \,|\, X$$

- Objective: to estimate $\mu = E(Y)$ $\quad \left[\text{ or } \sum_{i=1}^{N} Y_i \right]$

No-confounding estimation

- Assumption of no (unmeasured) confounding:

$$T \perp (Y_0, Y_1) \mid X,$$

as is the case for randomized clinical trials.

- Identification:

$$E(Y|T=1,X) = E(Y_1|T=1,X) \overset{Y_1 \perp T|X}{=} E(Y_1|X)$$
$$E[E(Y|T=1,X)] = E(Y_1)$$

- Outcome regression (OR) model $\quad [\hookrightarrow$ Imputation$]$

$$E(Y|T=t,X) = m(t,X;\alpha_t), \quad t=0,1.$$

Let $\widehat{m}(t,X) = m(t,X;\widehat{\alpha}_t)$. An estimator of $E(Y_1)$ is

$$\widehat{\mu}_{\text{OR}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(1,X_i).$$

- Propensity score (PS) model (Rosenbaum and Rubin 1983)

$$P(T=1|X) = \pi(X;\gamma).$$

Let $\widehat{\pi}(X) = \pi(X;\widehat{\gamma})$. An estimator of $E(Y_1)$ is

$$\widehat{\mu}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)} Y_i.$$

Illustration Left: $\sharp\{Y=1\}$, Right: $\sharp\{Y=0\}$

	$R=0$ (missing)	$R=1$
$X=0$	(?, ?) 80	(11, 9) 20
$X=1$	(?, ?) 40	(37, 23) 60
	? , ? 120	48, 32 80

- Outcome regression: $E(Y|R,X)$

	$R=0$	$R=1$
$X=0$	?	55.0%
$X=1$	?	61.7%

$$E(Y): \ [(100)\,55.0\% + (100)\,61.7\%]/200 = 58.3\%$$

- Propensity score weighting: $P(R|X)$

	$R=0$	$R=1$
$X=0$	80/100	20/100
$X=1$	40/100	60/100

$$E(Y): \ [11/(20/100) + 37/(60/100)]/200 = 58.3\%$$

Illustration Left: $\sharp\{Y=1\}$, Right: $\sharp\{Y=0\}$

	$T=0$		$T=1$	
$X=0$	(52, 28)	80	(11, 9)	20
$X=1$	(30, 10)	40	(37, 23)	60
	82, 38	120	48, 32	80

- Outcome regression: $E(Y|T,X)$

	$T=0$	$T=1$
$X=0$	65.0%	55.0%
$X=1$	75.0%	61.7%

$$E(Y_0): \ [(100)\,65.0\% + (100)\,75.0\%]/200 = 70.0\%$$
$$E(Y_1): \ [(100)\,55.0\% + (100)\,61.7\%]/200 = 58.3\%$$

- Propensity score weighting: $P(T|X)$

	$T=0$	$T=1$
$X=0$	80/100	20/100
$X=1$	40/100	60/100

$$E(Y_0): \ [52/(80/100) + 30/(40/100)]/200 = 70.0\%$$
$$E(Y_1): \ [11/(20/100) + 37/(60/100)]/200 = 58.3\%$$

Propensity score paradox (Robins and Ritov 1997)

- The likelihood is

$$\prod_{i=1}^{n} P(\{X_i\}) \times \prod_{i=1}^{n} \left[(1-\pi(X_i;\gamma))^{1-T_i}\pi(X_i;\gamma)^{T_i}\right]$$
$$\times \prod_{i=1}^{n} \left[P(Y_i|T_i=0,X_i;\alpha_0)^{1-T_i}P(Y_i|T_i=1,X_i;\alpha_1)^{T_i}\right].$$

- But, $E(Y_1) = \int E(Y|T=1,x)\,\mathrm{d}P(x)$.

- By the (strong) likelihood principle,

  inference should not depend on γ.

- Yet, if $\pi(X)$ is known, then

$$\frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)} Y_i$$

  is unbiased & regular, regardless of $\dim(X)$, whereas $\widehat{\mu}_{\mathsf{OR}}$ suffers the curse of dimensionality.

- If X is categorical, then

$$\frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)} Y_i \stackrel{def}{=} \widehat{\mu}_{\mathsf{IPW}} = \widehat{\mu}_{\mathsf{OR}}$$

  agree with each other.

Propensity score peculiarity

- Outcome regression (OR) model

$$E(Y|T=1,X) = m(1,X;\alpha_1).$$

Compute $\widehat{\alpha}_1$ as a solution to

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(1,X_i;\alpha_1)/\partial\alpha_1}{\text{var}(Y|T=1,X_i)}\, T_i \left(Y_i - m(1,X_i;\alpha_1)\right).$$

Had I used the true value $m(1,X)$:

$$\text{asy.var}\left[\frac{1}{n}\sum_{i=1}^{n} m(1,X_i)\right] \leq \text{asy.var}\left[\frac{1}{n}\sum_{i=1}^{n} \widehat{m}(1,X_i)\right].$$

- Propensity score (PS) model

$$P(T=1|X) = \pi(X;\gamma).$$

Compute $\widehat{\gamma}$ as a solution to

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \pi(X_i;\gamma)/\partial\gamma}{\pi(X_i;\gamma)(1-\pi(X_i;\gamma))}\left(T_i - \pi(X_i;\gamma)\right).$$

Had I used the true value $\pi(X)$:

$$\text{asy.var}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)}Y_i\right] \geq \text{asy.var}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)}Y_i\right].$$

Propensity score peculiarity (cont'd)

- Had I used a bigger OR model: $\alpha_1 \subset \alpha_1^\dagger$

$$\text{asy.var}\,(\widehat{\mu}_{\text{OR}}) \leq \text{asy.var}\left(\widehat{\mu}^\dagger_{\text{OR}}\right).$$

Had I used bigger and bigger OR models:
$\alpha_1 \subset \alpha_1^\dagger \subset \alpha_1^\ddagger \subset \cdots$

$$\begin{aligned}\text{asy.var}\,(\widehat{\mu}_{\text{OR}}) &\leq \text{asy.var}\left(\widehat{\mu}^\dagger_{\text{OR}}\right) \leq \text{asy.var}\left(\widehat{\mu}^\ddagger_{\text{OR}}\right) \leq \cdots \\ &\leq \text{asy.var.bound}\end{aligned}$$

- Had I used a bigger PS model: $\gamma \subset \gamma^\dagger$

$$\text{asy.var}\,(\widehat{\mu}_{\text{OR}}) \geq \text{asy.var}\left(\widehat{\mu}^\dagger_{\text{OR}}\right).$$

Had I used bigger and bigger PS models:
$\gamma \subset \gamma^\dagger \subset \gamma^\ddagger \subset \cdots$

$$\begin{aligned}\text{asy.var}\,(\widehat{\mu}_{\text{IPW}}) &\geq \text{asy.var}\left(\widehat{\mu}^\dagger_{\text{IPW}}\right) \geq \text{asy.var}\left(\widehat{\mu}^\ddagger_{\text{IPW}}\right) \geq \cdots \\ &\geq \text{asy.var.bound}\end{aligned}$$

- The asymptotic variance bound for regular estimators of $\mu_1 = E(Y_1)$ is $n^{-1}E(\tau_1^2)$, where

$$\begin{aligned}\tau_1 &= m(1,X) - \mu_1 + \frac{T}{\pi(X)}\big(Y - m(1,X)\big) \\ &= \frac{T}{\pi(X)}Y - \mu_1 - \left(\frac{T}{\pi(X)} - 1\right)m(1,X).\end{aligned}$$

What does this mean? (Tan 2007)

- Caveats
  - Assumption: each model is correctly specified.
  - Asymptotic variance: fix each model, let $n \to \infty$

- This "shows" $\widehat{\mu}_{\text{IPW}}$ is more variable than $\widehat{\mu}_{\text{OR}}$:

$$\text{asy.var}\left(\widehat{\mu}_{\text{OR}}\right) \leq \text{asy.var.bound} \leq \text{asy.var}\left(\widehat{\mu}_{\text{IPW}}\right).$$

- Does this mean $\widehat{\mu}_{\text{IPW}}$ is not as good as $\widehat{\mu}_{\text{OR}}$?

  Were a statistician asked to choose a correct OR model and a correct PS model ...

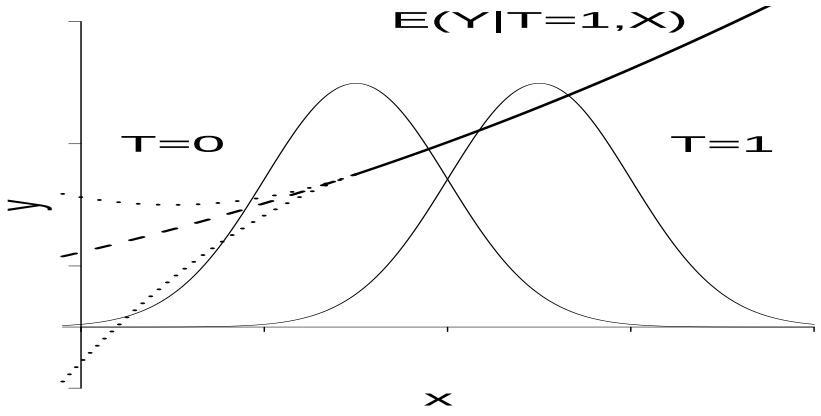
- Which side of the asy.var.bound?

$$\longrightarrow\longrightarrow \text{OR} \longrightarrow\longrightarrow \vdots \longleftarrow\longleftarrow \text{PS} \longleftarrow\longleftarrow$$

$$\vdots$$

$$\textit{aggresive} \qquad \text{asy.var.bound} \qquad \textit{conservative}$$

- The asy.var.bound becomes large or even ∞ whenever $\pi(X) \approx 0$ in some region of X.
  - It is difficult to infer $E(Y_1)$, because few subjects with $\pi(X) \approx 0$ receive $T = 1$ (i.e., Y_1 observed).
  - The difficulty holds for OR and PS, although the symptoms can be different.



- OR model is built on the "truncated" data

$$\{(X_i, Y_i) : T_i = 1, 1 \le i \le n\},$$

and hence makes extrapolation to predict $m(1, X)$ in the region of X such that $\pi(X) \approx 0$.

- PS model is built on the "full" data

$$\{(X_i, T_i) : 1 \le i \le n\}.$$

Inverse PS weighting of Y for treated subjects with $\pi(X) \approx 0$ leads to large variance.

Summary of what's been discussed

OR and PS are two approaches with different characteristics, and one does not dominate the other.

- If an OR model is correctly specified, then $\widehat{\mu}_{\text{OR}}$ is consistent and has asy.var $\leq$ asy.var.bound.

- If a PS model is correctly specified, then $\widehat{\mu}_{\text{PS}}$ is consistent and has asy.var $\geq$ asy.var.bound.

- OR approach suffers the problem of implicitly making extrapolation in the region of X with $\pi(X) \approx 0$, due to data truncation.

- PS approach tends to yield large weights for treated subjects with $\pi(X) \approx 0$, explicitly indicating uncertainty in the estimate.

Doubly robust estimation

- Given a PS model, a class of estimators of $E(Y_1)$ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}(X_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\widehat{\pi}(X_i)} - 1\right)h(X_i),$$

  where $h(X)$ is an arbitrary function.

- Consider the estimator (Robins et al. 1994)

$$\widehat{\mu}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}(X_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\widehat{\pi}(X_i)} - 1\right)\widehat{m}(1, X_i),$$

  depending on both OR and PS models.

  - $\widehat{\mu}_{\text{AIPW}}$ is locally efficient: it attains asy.var.bound if both OR and PS models are correctly specified.

  - $\widehat{\mu}_{\text{AIPW}}$ is doubly robust: it remains consistent if either OR or PS model is correctly specified.

- Comparison of $\widehat{\mu}_{\text{AIPW}}$ vs $\widehat{\mu}_{\text{OR}}$:

$$\text{asy.var}\left(\widehat{\mu}_{\text{AIPW}}\right) \geq \text{asy.var}\left(\widehat{\mu}_{\text{OR}}\right).$$

- Comparison of $\widehat{\mu}_{\text{AIPW}}$ vs $\widehat{\mu}_{\text{IPW}}$:

$$\text{asy.var}\left(\widehat{\mu}_{\text{AIPW}}\right) \overset{??}{\leq} \text{asy.var}\left(\widehat{\mu}_{\text{IPW}}\right).$$

A more desirable DR estimator

- Consider the estimator (Tan 2006)

$$\widetilde{\mu}_{\text{REG}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}(X_i)}Y_i \;-\; \widetilde{\beta}^{(1)}\;\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\widehat{\pi}(X_i)}-1\right)\widehat{m}(1,X_i),$$

  where $\widetilde{\beta}^{(1)}$ is the first element of

$$\widetilde{\beta} = \left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_i\,\widehat{\zeta}_i^{\top}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_i\,\widehat{\eta}_i\right]$$

  and

$$\widehat{\eta}_i = \frac{T_i}{\widehat{\pi}(X_i)}\,Y_i,$$

$$\widehat{\xi}_i = \left(\frac{T_i}{\widehat{\pi}(X_i)}-1\right)\left(\widehat{m}(1,X_i),\;\frac{\widehat{\frac{\partial\pi}{\partial\gamma^{\top}}}(X_i)}{1-\widehat{\pi}(X_i)}\right)^{\top},$$

$$\widehat{\zeta}_i = \frac{T_i}{\widehat{\pi}(X_i)}\left(\widehat{m}(1,X_i),\;\frac{\widehat{\frac{\partial\pi}{\partial\gamma^{\top}}}(X_i)}{1-\widehat{\pi}(X_i)}\right)^{\top}.$$

  - $\widetilde{\mu}_{\text{REG}}$ is locally efficient and doubly robust.
  - $\widetilde{\mu}_{\text{REG}}$ is intrinsically efficient: if PS model is correct, it attains the smallest asy.var among

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}(X_i)}Y_i \;-\; b^{(1)}\;\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\widehat{\pi}(X_i)}-1\right)\widehat{m}(1,X_i),$$

    where $b^{(1)}$ is an arbitrary coefficient.

- Comparison of $\widetilde{\mu}_{\text{REG}}$ vs $\widehat{\mu}_{\text{OR}}$.
  Comparison of $\widetilde{\mu}_{\text{REG}}$ vs $\widehat{\mu}_{\text{IPW}}$.

Likelihood approach (Tan 2006)

- Rewrite the likelihood as

$$L_1 \times L_2 = \prod_{i=1}^{n} \left[(1 - \pi(X_i; \gamma))^{1-T_i} \pi(X_i; \gamma)^{T_i} \right] \times \prod_{i=1}^{n} \left[G_0(\{X_i, Y_i\})^{1-T_i} G_1(\{X_i, Y_i\})^{T_i} \right]$$

  where G_0 is the joint dist'n of (X, Y_0) and
  G_1 is the joint dist'n of (X, Y_1).

- G_0 and G_1 induce the same marginal of X:

$$\int h(x)\, \mathrm{d}G_0(x, y_0) = \int h(x)\, \mathrm{d}G_1(x, y_1)$$

  for each bounded function $h(x)$.

- We retain a finite subset of constraints and ignore other constraints on (G_0, G_1).

- Let $\widehat{h}(x) = \left\{1 - \widehat{\pi}(x),\ (1 - \widehat{\pi}(x))\widehat{m}(1, x),\ \frac{\partial \widehat{\pi}}{\partial \gamma^\top}(x)\right\}^\top$.
  Maximize $L_2(G_0, G_1)$ subject to

$$\int \mathrm{d}G_1 = 1,$$

$$\int (1 - \widehat{\pi}(x))\, \mathrm{d}G_1 = \int (1 - \widehat{\pi}(x)) \mathrm{d}G_0,$$

$$\int (1 - \widehat{\pi}(x))\widehat{m}(1, x)\, \mathrm{d}G_1 = \int (1 - \widehat{\pi}(x))\widehat{m}(1, x) \mathrm{d}G_0,$$

$$\int \frac{\partial \widehat{\pi}}{\partial \gamma}(x)\, \mathrm{d}G_1 = \int \frac{\partial \widehat{\pi}}{\partial \gamma}(x)\, \mathrm{d}G_0.$$

Likelihood approach (cont'd)

- Postulate the following model for (T, X, Y):
  - T is generated according to

$$P(T = 0) = \int (1 - \widehat{\pi}(x))\, \mathrm{d}G_0,$$

$$P(T = 1) = \int \widehat{\pi}(x)\, \mathrm{d}G_1.$$

  - (X, Y) given $T = 0$ or 1 is generated from the biased-sampling distribution (Vardi 1985)

$$\frac{1 - \widehat{\pi}(x)}{\int (1 - \widehat{\pi}(x'))\, \mathrm{d}G_0} \mathrm{d}G_0,$$

$$\frac{\widehat{\pi}(x)}{\int \widehat{\pi}(x')\, \mathrm{d}G_1} \mathrm{d}G_1.$$

- Let $\varpi(X; \lambda) = \widehat{\pi}(X) + \lambda^\top \widehat{h}(X)$ and

$$\widehat{\lambda} = \mathrm{argmax}_\lambda$$

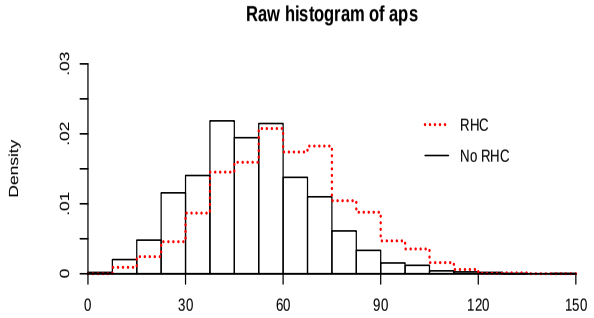
$$\frac{1}{n} \sum_{i=1}^{n} \Big[T_i \log(\varpi(X_i; \lambda)) + (1 - T_i) \log(1 - \varpi(X_i; \lambda)) \Big].$$

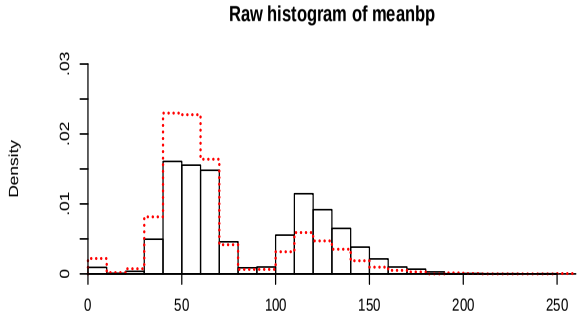
  Then

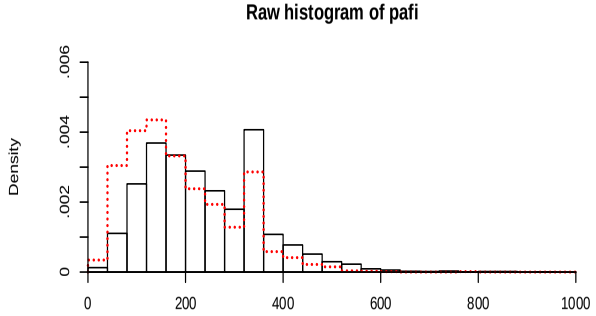
$$\widehat{G}_1(\{X_i, Y_i\}) = \frac{n^{-1}}{\varpi(X_i; \widehat{\lambda})}, \qquad \text{if } T_i = 1,$$

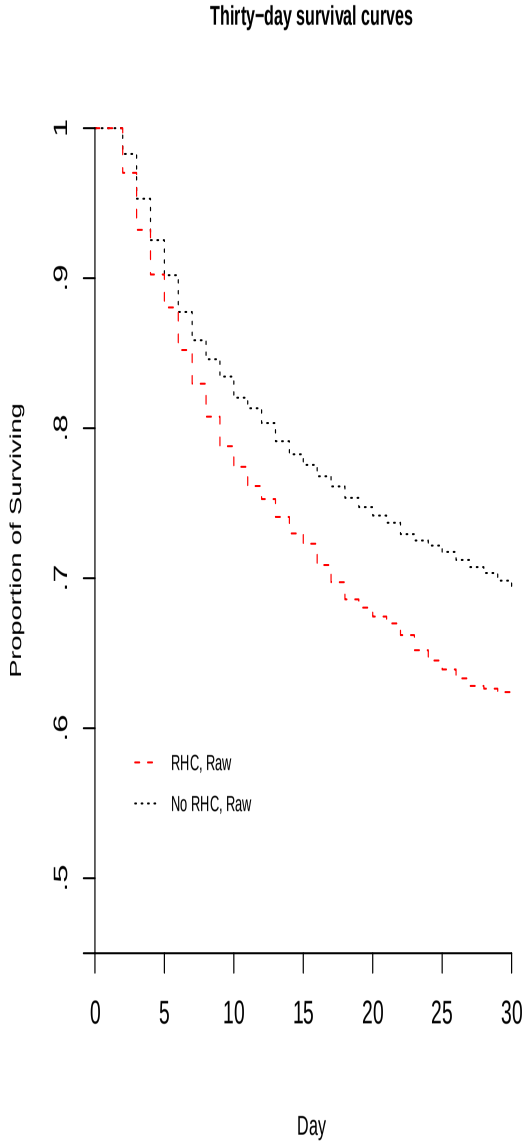
$$\widehat{G}_0(\{X_i, Y_i\}) = \frac{n^{-1}}{1 - \varpi(X_i; \widehat{\lambda})}, \quad \text{if } T_i = 0.$$

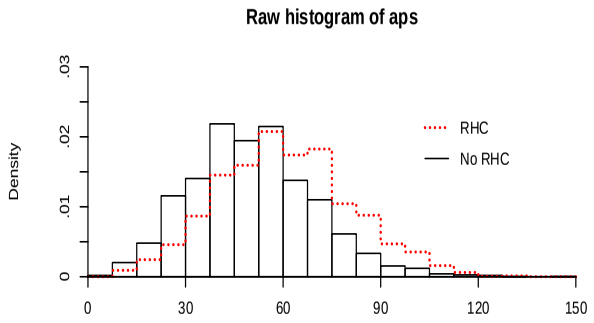
  The dist'n $(\widehat{G}_0, \widehat{G}_1)$ are weighted histograms.

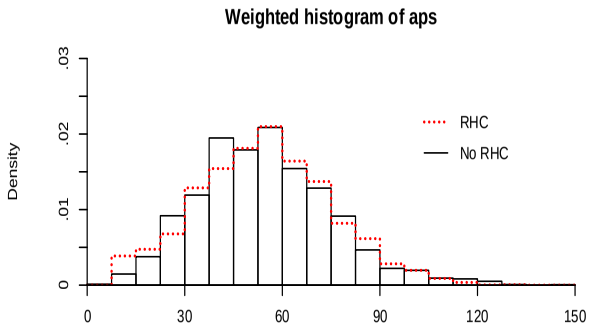

Raw histogram of aps
Density
0
.01
.02
.03
0
30
60
90
120
150
RHC
No RHC

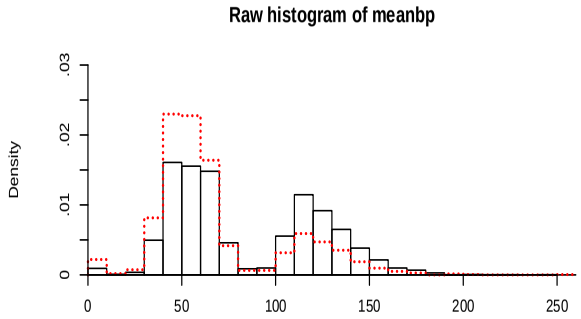

Raw histogram of meanbp
Density
0
.01
.02
.03
0
50
100
150
200
250

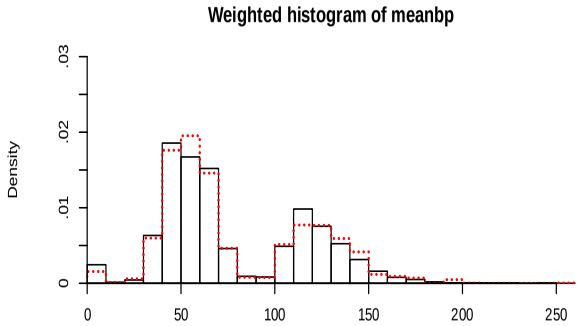

Raw histogram of pafi
Density
0
.002
.004
.006
0
200
400
600
800
1000

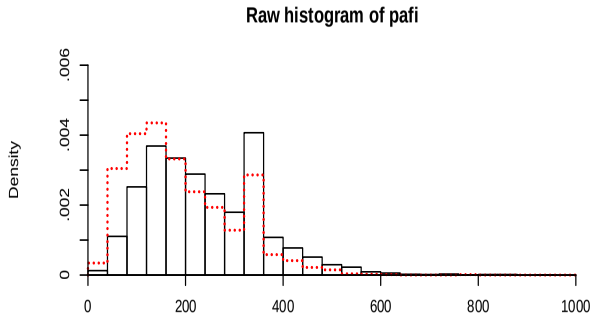

Thirty-day survival curves
Proportion of Surviving
.5
.6
.7
.8
.9
1
0
5
10
15
20
25
30
Day
RHC, Raw
No RHC, Raw

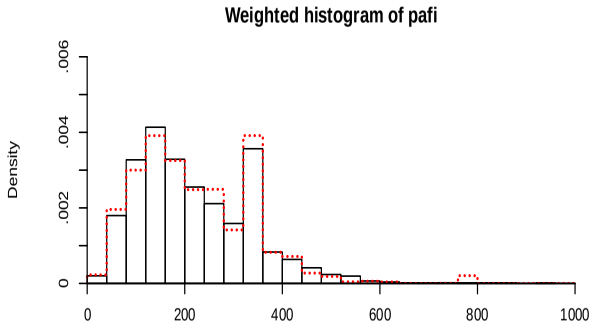

Raw histogram of aps
Density
RHC
No RHC
0
.01
.02
.03
0
30
60
90
120
150

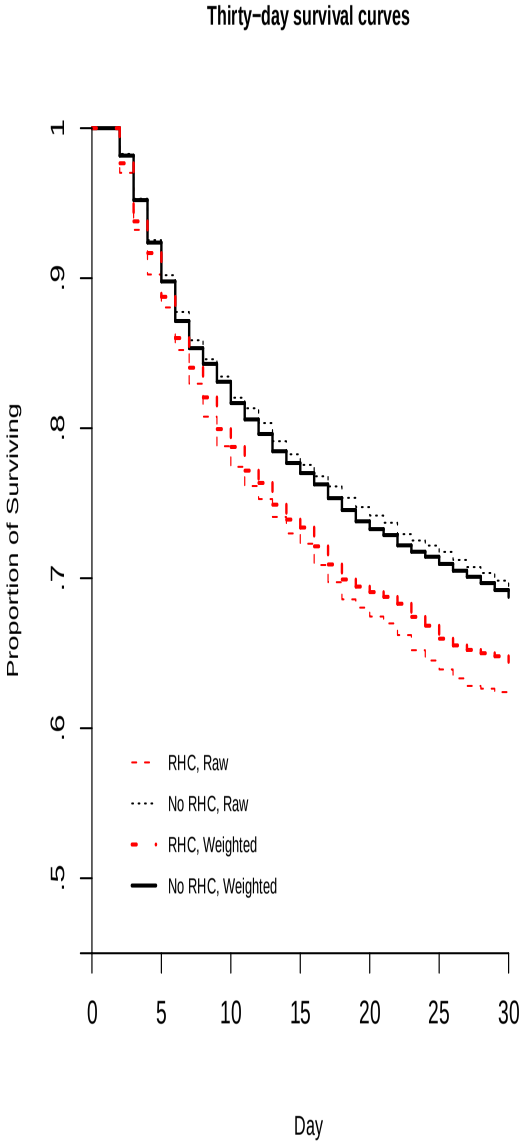

Weighted histogram of aps
Density
RHC
No RHC
0
.01
.02
.03
0
30
60
90
120
150


Raw histogram of meanbp
Density
0
.01
.02
.03
0
50
100
150
200
250


Weighted histogram of meanbp
Density
0
.01
.02
.03
0
50
100
150
200
250


Raw histogram of pafi
Density
0
.002
.004
.006
0
200
400
600
800
1000


Weighted histogram of pafi
Density
0
.002
.004
.006
0
200
400
600
800
1000


Thirty-day survival curves
Proportion of Surviving
.5
.6
.7
.8
.9
1
RHC, Raw
No RHC, Raw
RHC, Weighted
No RHC, Weighted
0
5
10
15
20
25
30
Day

Likelihood appraoch (cont'd)

- The resulting estimator of μ

$$\widehat{\mu}_{\text{LIK}} = \int y \, \mathrm{d}\widehat{G}_1 = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i \, Y_i}{\varpi(X_i; \widehat{\lambda})} \right].$$

  - $\widehat{\mu}_{\text{LIK}}$ is locally efficient: it attains asy.var.bound if both OR and PS models are correctly specified.

  - $\widehat{\mu}_{\text{LIK}}$ is intrinsically efficient: if PS model is correct, it is asymptotically at least efficient as $\widehat{\mu}_{\text{IPW}}$ and $\widehat{\mu}_{\text{AIPW}}$.

- $\widehat{\lambda}$ is a solution to

$$0 = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i - \varpi(X_i; \lambda)}{\varpi(X_i; \lambda)(1 - \varpi(X_i; \lambda))} \widehat{h}(X_i) \right].$$

- $\widehat{\lambda}$ also satisfies

$$1 = \int \mathrm{d}\widehat{G}_1 = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i}{\varpi(X_i; \widehat{\lambda})} \right].$$

  - $\varpi(X_i; \widehat{\lambda})$ with $T_i = 1$, $1 \leq i \leq n$, are bounded from below by n^{-1}.

  - $\widehat{\mu}_{\text{LIK}}$ satisfies sample boundedness: it lies within the range of $\{Y_i : T_i = 1, 1 \leq i \leq n\}$, even if OR and/or PS model are misspecified.

DR Likelihood estimator (Tan 2010)

- Let
$$\widehat{v}(x) = (1, \widehat{m}(1,x))^\top,$$
$$\widehat{h}(x) = \left\{(1-\widehat{\pi}(x))\,\widehat{v}^\top(x),\ \frac{\partial\widehat{\pi}}{\partial\gamma^\top}(x)\right\}^\top.$$
Recall $\varpi(X;\lambda) = \widehat{\pi}(X) + \lambda^\top\widehat{h}(X)$.
- Let $\widetilde{\lambda}$ be a solution to
$$0 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{T_i}{\varpi(X_i;\lambda)} - 1\right)\widehat{v}(X_i)\right],$$
$$0 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i - \varpi(X_i;\lambda)}{\varpi(X_i;\lambda)(1-\varpi(X_i;\lambda))}\frac{\partial\widehat{\pi}}{\partial\gamma}(X_i)\right],$$
subject to $\varpi(X_i;\lambda) > 0$ if $T_i = 1$.
- The resulting estimator of μ
$$\widetilde{\mu}_{\mathsf{LIK}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\varpi(X_i;\widetilde{\lambda})}\right].$$
- $\widetilde{\lambda}$ satisfies
$$1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\varpi(X_i;\widetilde{\lambda})}\right],$$
$$\frac{1}{n}\sum_{i=1}^{n}\widehat{m}(1,X_i) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\varpi(X_i;\widetilde{\lambda})}\widehat{m}(1,X_i)\right].$$
  - $\widetilde{\mu}_{\mathsf{LIK}}$ satisfies sample boundedness.
  - $\widetilde{\mu}_{\mathsf{LIK}}$ is doubly robust.
  - $\widetilde{\mu}_{\mathsf{LIK}}$ is locally efficienct and intrinsically efficient.
- Issues: existence and computation of $\widetilde{\lambda}$.

Asymptotic expansions under misspecified models

- $\widehat{\lambda}$ converges to a constant λ^*. Then

$$\widehat{\mu}_{\mathrm{LIK}} = \widetilde{E}\left[\frac{TY}{\varpi(X;\lambda^*)}\right] - \widehat{C}^\top \widehat{B}^{-1}\, \widetilde{E}\left[\frac{T-\varpi(X;\lambda^*)}{\varpi(X;\lambda^*)(1-\varpi(X;\lambda^*))}\widehat{h}(X)\right] + o_p(n^{-1/2}),$$

where

$$\widehat{B} = \widetilde{E}\left[\frac{(T-\varpi(X;\lambda^*))^2}{\varpi^2(X;\lambda^*)(1-\varpi(X;\lambda^*))^2}\widehat{h}(X)\widehat{h}^\top(X)\right],$$

$$\widehat{C} = \widetilde{E}\left[\frac{T}{\varpi^2(X;\lambda^*)}Y\widehat{h}(X)\right].$$

If $\lambda^* = 0$, then the first-order term reduces to $\widehat{\mu}_{\mathrm{REG}}$.

- $\widetilde{\lambda}$ converges to a constant $\lambda^\dagger$. Then

$$\widetilde{\mu}_{\mathrm{LIK}} = \widetilde{E}\left[\frac{TY}{\varpi(X;\lambda^\dagger)}\right] - \widehat{C}^\top \widetilde{B}^{\top -1}\, \widetilde{E}\left[\begin{pmatrix} \left(\frac{T}{\varpi(X;\lambda^\dagger)}-1\right)\widehat{v}(X) \\ \frac{T-\varpi(X;\lambda^\dagger)}{\varpi(X;\lambda^\dagger)(1-\varpi(X;\lambda^\dagger))}\widehat{h}_2(X) \end{pmatrix}\right] + o_p(n^{-1/2}),$$

where

$$\widetilde{B} = \widetilde{E}\left[\begin{pmatrix} \frac{T}{\varpi^2(X;\lambda^\dagger)}\widehat{h}_1(X)\widehat{v}^\top(X) & \frac{(T-\varpi(X;\lambda^\dagger))^2}{\varpi^2(X;\lambda^\dagger)(1-\varpi(X;\lambda^\dagger))^2}\widehat{h}_1(X)\widehat{h}_2^\top(X) \\ \frac{T}{\varpi^2(X;\lambda^\dagger)}\widehat{h}_2(X)\widehat{v}^\top(X) & \frac{(T-\varpi(X;\lambda^\dagger))^2}{\varpi^2(X;\lambda^\dagger)(1-\varpi(X;\lambda^\dagger))^2}\widehat{h}_2(X)\widehat{h}_2^\top(X) \end{pmatrix}\right]$$

with

$$\widehat{h}_1(x) = (1-\widehat{\pi}(x))\widehat{v}^\top(x) = (1-\widehat{\pi}(x))\,(1, \widehat{m}(1,x))^\top,$$

$$\widehat{h}_2(x) = \frac{\partial\widehat{\pi}}{\partial\gamma}(x).$$

If $\lambda^\dagger = 0$, then the first-order term reduces (almost) to $\widetilde{\mu}_{\mathrm{REG}}$.

DR Likelihood estimator (cont'd)

- Two-step estimator
  - Compute $\widehat{\lambda} = (\widehat{\lambda}_1^\top, \widehat{\lambda}_2^\top)^\top$, partitioned according to $\widehat{h}(x) = \left\{(1-\widehat{\pi}(x))\,\widehat{v}^\top(x),\ \frac{\partial\widehat{\pi}}{\partial\gamma^\top}(x)\right\}^\top$.
  - Compute $\tilde{\lambda}_{\text{step2}} = (\tilde{\lambda}_{1,\text{step2}}^\top, \widehat{\lambda}_2^\top)^\top$, where

    $$\tilde{\lambda}_{1,\text{step2}} = \text{argmax}_{\lambda_1} \frac{1}{n}\sum_{i=1}^{n}\left[T_i \frac{\log\{\varpi(X_i;\lambda_1,\widehat{\lambda}_2)/\varpi(X_i;\widehat{\lambda})\}}{1-\widehat{\pi}(X_i)} - \lambda_1^\top \widehat{v}(X_i)\right].$$

    The objective function is strictly concave and bounded from above under mild conditions.

- For step 2, $\tilde{\lambda}_{1,\text{step2}}$ is a solution to

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{T_i}{\varpi(X_i;\lambda_1,\widehat{\lambda}_2)} - 1\right)\widehat{v}(X_i)\right].$$

- The resulting estimator of μ

$$\widetilde{\widetilde{\mu}}_{\text{LIK}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i\,Y_i}{\varpi(X_i;\widetilde{\widetilde{\lambda}})}\right].$$

  - $\tilde{\mu}_{\text{LIK2}}$ satisfies sample boundedness.
  - $\tilde{\mu}_{\text{LIK2}}$ is doubly robust.
  - $\tilde{\mu}_{\text{LIK2}}$ is locally efficienct and intrinsically efficient.

Simulation study

- The same design as in Kang and Schafer (2007)

  $X = (X_1, X_2, X_3, X_4)^\top$,
  $Y = 210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + \epsilon$,
  $T = 1\{U \leq \text{expit}(-X_1 + .5X_2 - .25X_3 - .1X_4)\}$.

- Missing-data setup: observed data (X, T, TY)

  Causal-inference setup: observed data (X, T, Y)

- 22 estimators are labelled as follows.

(1–3) $\widehat{\mu}_{\text{LIK,OLS}}$, $\widehat{\mu}_{\text{REG,OLS}}$, $\widetilde{\mu}_{\text{REG,OLS}}$;

(4–7) $\widehat{\mu}_{\text{AIPW}}$ (ratio), $\widehat{\mu}_{\text{OLS,ext}}$, $\widehat{\mu}_{\text{WLS}}$, $\widetilde{\mu}_{\text{RV}}$;

(8–12) $\widehat{\mu}_{\text{IPW,ext}}$ (ratio), $\widetilde{\mu}_{\text{IPW,ext2}}$,
$\widehat{\mu}_{\text{AIPW,ext}}$ (ratio), $\widehat{\mu}_{\text{WLS,ext}}$ (ratio), $\widehat{\mu}_{\text{WLS,ext2}}$;

(13–15) $\widehat{\mu}_{\text{TIPW}}$ (ratio), $\widehat{\mu}_{\text{TML}}$, $\widehat{\mu}_{\text{TAIPW}}$ (ratio);

(16–22) $\widehat{\mu}_{\text{AIPW,lik}}$, $\widetilde{\mu}_{\text{LIK2,OLS}}$,
$\widehat{\mu}_{\text{WLS,lik}}$, $\widehat{\mu}_{\text{WLS,lik2}}$, $\widetilde{\mu}_{\text{LIK2,WLS}}$,
$\widetilde{\mu}_{\text{RV,lik}}$, $\widetilde{\mu}_{\text{LIK2,RV}}$.

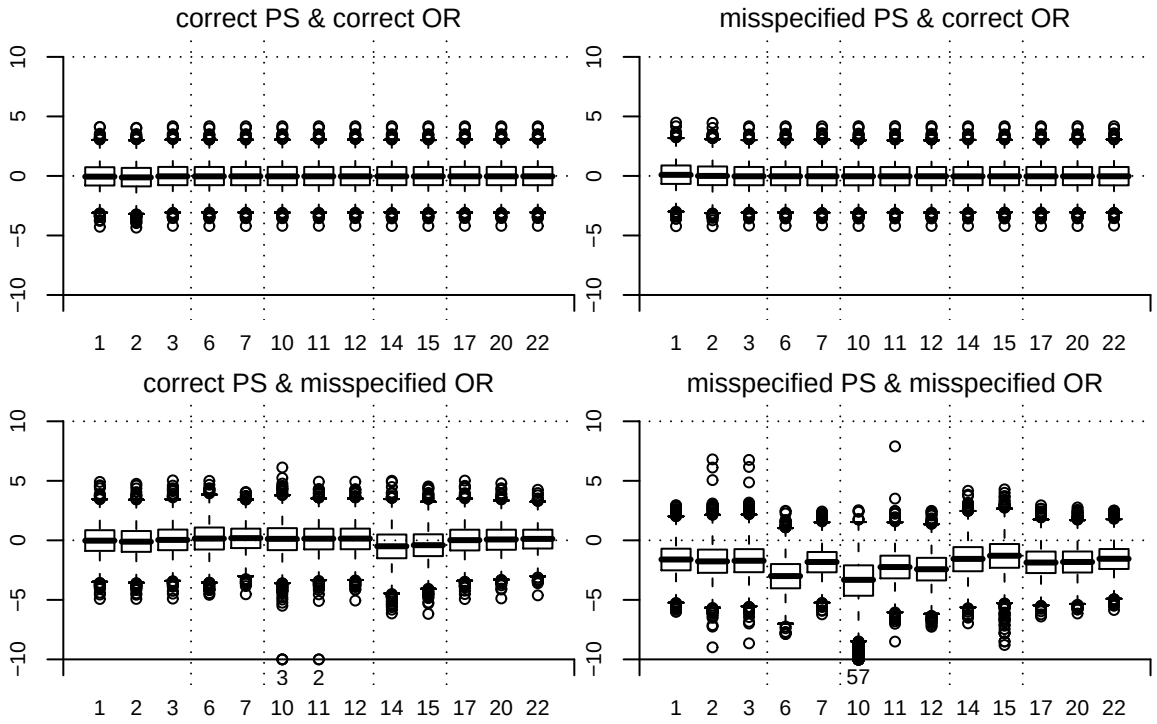


Figure 1: Boxplots of estimators of $\mu_1 - 210$ ($n = 1000$)

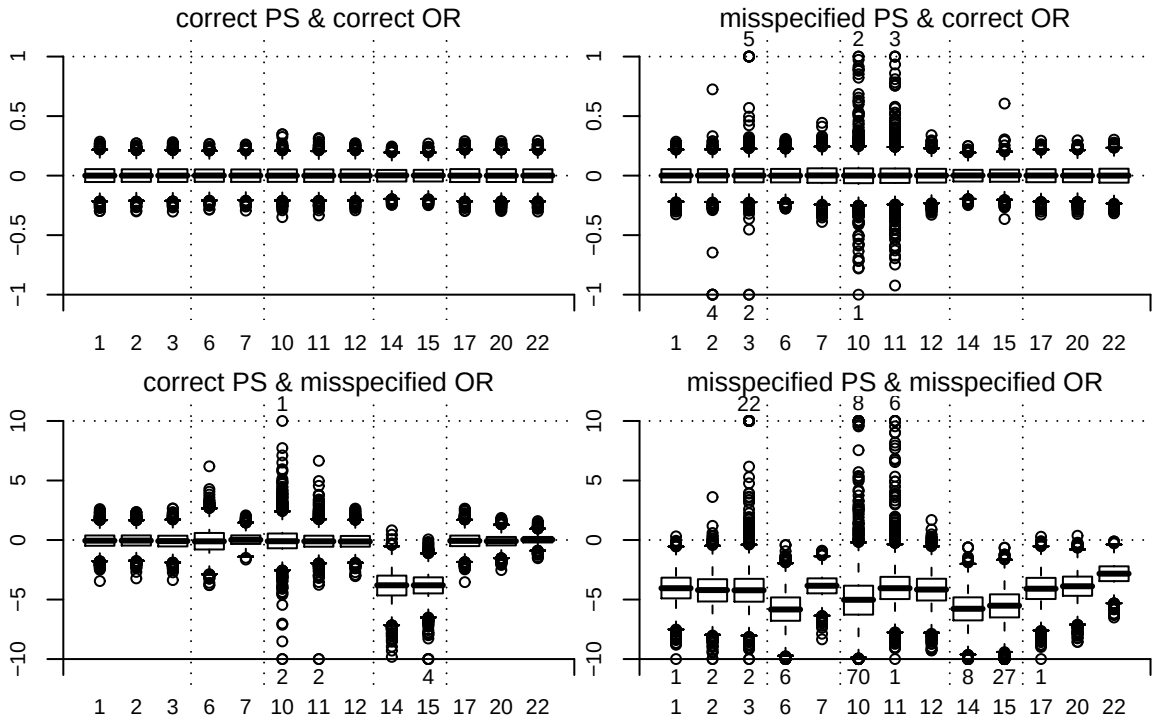


Figure 2: Boxplots of estimators of $\mu_1 - \mu_0$ ($n = 1000$)

Table 1: Numerical comparison of estimators

Estimator	1	2	3	6	7	10	11	12	14	15	17	20	22
						Estimators of μ_1 in missing data setup							
C-PS&M-OR	1·23	1·28	1·20	1·32	1·07	1·39	1·25	1·24	1·51	1·35	1·24	1·17	1·00
	1·21	1·21	1·20	1·36	1·07	1·75	1·33	1·20	1·80	1·49	1·21	1·13	1·00
M-PS&M-OR	1·32	1·50	1·34	1·62	1·12	1·87	1·31	1·30	1·32	1·12	1·27	1·20	1·00
	1·10	1·31	1·27	2·82	1·24	4·86	1·78	1·99	1·21	1·04	1·31	1·27	1·00
						Estimators of $\mu_1 - \mu_0$ in causal inference setup							
C-PS&M-OR	1·97	1·79	2·76	4·50	1·85	5·00	3·27	3·16	19·8	18·1	2·65	1·87	1·00
	3·80	3·58	4·05	9·08	2·33	11·7	5·42	4·17	134	130	4·07	2·58	1·00
M-PS&M-OR	1·77	1·77	2·51	2·75	1·46	2·39	1·87	1·82	2·91	2·42	1·88	1·74	1·00
	2·02	2·22	73·3	3·99	1·77	3·42	2·08	2·16	3·97	3·72	2·07	1·87	1·00

C-PS (or M-PS): correct (or misspecified) propensity score model; M-OR: misspecified outcome regression model; Each cell gives the ratios of mean squared errors for $n = 200$ (upper) and $n = 1000$ (lower).