A robust approach for location estimation in a missing data setting

Julieta Molina, Mariela Sued, Marina Valdora, Victor Yohai,

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires-CONICET

May, 2015

The problem - Classical Approach - Rotnitzky & Robins ...

- Scalar response Y.
- Goal: Estimate $\mathrm{E}(Y)$.
- Problem: Y is not always observed, there are missing observations.
- Some help: Always observed vector of covariables $\mathbf{X} = (X_1, \cdots, X_d)$.

Data:

- Explanatory variables $\mathbf{X} = (X_1, \cdots, X_d)$ always observed.
- Response indicator A, always observed.
- Response variable Y observed only when $A = 1$.

Missing at Random (MAR) Assumption

Given $\mathbf{X}$, Y is independent of A.

Propensity Score: $\pi(\mathbf{X}) = \mathrm{P}(A \mid \mathbf{X})$.

Under the MAR assumption,

$$\mathrm{E}(Y) = \mathrm{E}\left\{ \frac{AY}{\pi(\mathbf{X})} \right\}$$

Inverse Probability Weighted (IPW) estimator of $\mathrm{E}(Y)$

- $\mathrm{E}(Y) = \mathrm{E}\left\{\frac{AY}{\pi(\mathbf{X})}\right\}$
- Propensity Score $\pi(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X})$
- If $\widehat{\pi}_n \to \pi$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\widehat{\pi}(\mathbf{X}_i)} \approx \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\pi(\mathbf{X}_i)} \to \mathrm{E}\left\{\frac{AY}{\pi(\mathbf{X})}\right\} = \mathrm{E}(Y)\ .$$

Estimating $\pi(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X})$ non parametrically.

Cheng (1994), Hahn (1998); Hirano, Imbens and Ridder (2003); Imbens, Newey Ridder (2007).

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\widehat{\pi}(\mathbf{X}_i)} - \mathrm{E}(Y)\right\} \to \mathcal{N}(0, V)$$

Estimating $\pi(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X})$

In practice, non parametric regression estimation of $\pi(\mathbf{X})$ is infeasible.

Estimating $\pi(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X})$ parametrically

- Parametric Model : $\pi(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X}) = \pi(\mathbf{X}, \boldsymbol{\gamma}_0)$.
- Example: Logistic Regression $\pi(\mathbf{X}, \gamma) = \text{expit}(\boldsymbol{\gamma}'\mathbf{X})$, where $\text{expit}(t) = e^t/(1 + e^t)$
- $\widehat{\gamma}_n$ M.L.E. based on $(A_i, \mathbf{X}_i)_{1 \le i \le n}$.
- $\widehat{\pi}(\cdot) = \pi(\cdot, \widehat{\gamma}_n)$.

$$\widehat{\mu}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\pi(\mathbf{X}_i, \widehat{\gamma}_n)}$$

- Yao L, Sun Z, Wang Q. (2010)

On the positivity assumption

The propensity score is bounded away from zero:

$$\mathrm{P}(\pi(\mathbf{X}) > \varepsilon) = 1, \text{ for some } \varepsilon > 0$$

Arxives: Models for the Propensity Score that Contemplate the Positivity Assumption and their Application to Missing Data and Causality.
with Molina & Valdora

IPW with *tuned weights* : Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19), 2937-2960.

Outcome regression estimator of $\mathrm{E}(Y)$

- Given $\mathbf{X}$, Y is independent of A.
- $\mathrm{E}(Y|\mathbf{X}) = \mathrm{E}(Y|\mathbf{X}, A = 1)$.
- $\mathrm{E}(Y) = \mathrm{E}\left\{\mathrm{E}\left(Y \mid \mathbf{X}, A = 1\right)\right\}$
- $g(\mathbf{X}) = \mathrm{E}(Y \mid \mathbf{X}) = \mathrm{E}(Y \mid \mathbf{X}, A = 1)$
- If $\widehat{g}_n \to g$

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{g}_n(\mathbf{X}_i) \approx \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i) \to \mathrm{E}\left\{g(\mathbf{X})\right\} = \mathrm{E}(Y) .$$

- Non parametric regression - Cheng (1994).

Estimating $g(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x})$

In practice, non parametric regression estimation of $g(\mathbf{x})$ is infeasible.

Modeling $g(\mathbf{X}) = \mathrm{E}(Y \mid \mathbf{X})$

$$\mathrm{E}(Y \mid \mathbf{X}) = \mathrm{E}(Y \mid A = 1, \mathbf{X}) = g(\boldsymbol{\beta}_0, \mathbf{X})$$

Example: linear regression $g(\boldsymbol{\beta}, \mathbf{x}) = \boldsymbol{\beta}^t \mathbf{x}$.

Estimation:

- Use $(\mathbf{X}_i, Y_i)$ with $A_i = 1$ to fit the regression model, estimating $\boldsymbol{\beta}_0$ with $\widehat{\boldsymbol{\beta}}_n$ (least squares) .
- Let $\widehat{\mu}_{reg} = n^{-1} \sum_{i=1}^{n} g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X}_i)$.

$$\frac{1}{n} \sum_{i=1}^{n} g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X}_i) \rightarrow \mathrm{E}(Y) .$$

- survey in Kan and Schafer (2007)

Doubly Protected Estimation (DP)

$$\mathrm{E}(Y) = \mathrm{E}\left\{\frac{AY}{p(\mathbf{X})}\right\} - \mathrm{E}\left[\left\{\frac{A}{p(\mathbf{X})} - 1\right\} r(\mathbf{X})\right]$$

if $p(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X})$ **OR** $r(\mathbf{X}) = \mathrm{E}(Y \mid \mathbf{X})$

$$\widehat{\mu}_{DP} = \mathbb{P}_n\left\{\frac{AY}{\pi(\mathbf{X}, \widehat{\gamma}_n)}\right\} - \mathbb{P}_n\left[\left\{\frac{A}{\pi(\mathbf{X}, \widehat{\gamma}_n)} - 1\right\} g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X})\right]$$

$\widehat{\mu}_{DP}$ converges to $E(Y)$

if $\mathrm{P}(A = 1 \mid \mathbf{X}) = \pi(\mathbf{X}, \gamma_0)$ **OR** $\mathrm{E}(Y \mid \mathbf{X}) = g(\boldsymbol{\beta}_0, \mathbf{X})$

Robustness

"The mean is not robust"

The mean will be replaced by any robust location estimator

- Median
- Trimmed means
- M- estimators

Robust Goal

Estimate

$$\mu_0 = T_L(F_0)$$

where $Y \sim F_0$ and T_L is location functional, weakly continuous at F_0.

Possible location functionals:

- Median - Percentiles - Trimmed means - M- estimators
- In fact, we can consider any parameter $\mu_0 = T(F_0)$, for T continuos.

Robust approaches for location:

- Bianco, A., Boente, G., González-Manteiga, W. & Pérez-González, A. (2010). Estimation of the marginal location under a partially linear model with missing responses. *Computational Statistics & Data Analysis*.

How can we estimate $T_L(F_0)$?

Plug- in procedure: Consider $T_L(\widehat{F_0})$, with $\widehat{F_0} \to F_0 = F_Y$

We need to estimate $F_0 = F_Y$

IPW estimator of $F_0(y) = \mathrm{P}(Y \leq y)$

- Under MAR, $\mathrm{E}(Y) = \mathrm{E}\left\{\frac{AY}{\pi(\mathbf{X})}\right\}$, $\pi(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{x})$
- $\mathrm{E}\{\ell(Y)\} = \mathrm{E}\left\{\frac{A\ell(Y)}{\pi(\mathbf{X})}\right\}$
- $\ell(Y) = \ell_y(Y) = \mathrm{I}_{\{Y \leq y\}}$
- $F_0(y) = \mathrm{E}\{\ell_y(Y)\} = \mathrm{E}\left\{\frac{A\mathrm{I}_{\{Y \leq y\}}}{\pi(\mathbf{X})}\right\}$
- $\widehat{F_0}(y) : \mathbb{P}_n \frac{A\mathrm{I}_{\{Y \leq y\}}}{\widehat{\pi}(\mathbf{X})}$
- $\widetilde{F_0}$: corrected $\widehat{F_0}$
- In practice, $\widehat{\pi}(\mathbf{X}) = \pi(\mathbf{X}, \widehat{\gamma}_n)$

Robust regression estimator of $F_0(y) = \mathrm{P}(Y \leq y)$ - SY(2013)

- Model: $Y = \boldsymbol{\beta}_0' \mathbf{X} + \mathbf{u}$, $\mathbf{u}$ independent of $(\mathbf{X}, A)$.
- $\widehat{\boldsymbol{\beta}}_n$: fit the regression model using $(\mathbf{X}_i, Y_i)$ with $A_i = 1$.
- Predicted values: $\widehat{\boldsymbol{\beta}}_n \mathbf{X}_j$, $1 \leq j \leq n$.
- Residuals: $\widehat{\mathbf{u}}_i = Y_i - \widehat{\boldsymbol{\beta}}_n \mathbf{X}_i$, $i : A_i = 1$.
- Pseudo-responses (Sued & Yohai 2013):

$$\widehat{y}_{ij} := \widehat{\boldsymbol{\beta}}_n \mathbf{X}_j + \widehat{\mathbf{u}}_i \;, 1 \leq j \leq n \;, i : A_i = 1 \;.$$

- $\widehat{F_0}$: empirical distribution at $\widehat{y}_{ij}$, $1 \leq j \leq n$, $i : A_i = 1$.

Robust regression estimator of $F_0(y) = \mathrm{P}(Y \leq y)$ - SSY

- Model: $Y = \boldsymbol{\beta}_0' \mathbf{X} + \mathbf{u}$, $\mathbf{u}$ independent of $(\mathbf{X}, A)$.

- Predicted values: $\widehat{\boldsymbol{\beta}}_n \mathbf{X}_j$, $1 \leq j \leq n$.

- Residuals: $\widehat{\mathbf{u}}_i = Y_i - \widehat{\boldsymbol{\beta}}_n \mathbf{X}_i$, $i : A_i = 1$.

- Pseudo-responses (Stati, Sued & Yohai ?):

$$\widetilde{y}_{ij} = \begin{cases} Y_j & \text{if} \quad j \in A, i : A_i = 1 \\ \widehat{\boldsymbol{\beta}}_n \mathbf{X}_j + \widehat{u}_i & \text{if} \quad j \notin A, i : A_i = 1 \;. \end{cases}$$

- $\widetilde{F_0}$: empirical distribution at $\widetilde{y}_{ij}$, $1 \leq j \leq n$, $i : A_i = 1$.

Regression estimator of $F_0(y)$ SY - a different point of view

- $F_0(y) = \mathrm{E}\left\{\mathrm{E}\left(\mathrm{I}_{\{Y\le y\}} \mid \mathbf{X}\right)\right\} = \mathrm{E}\{r_y(\mathbf{X})\}$,
- where $r_y(\mathbf{X}) = \mathrm{E}\left(\mathrm{I}_{\{Y\le y\}} \mid \mathbf{X}\right) = \mathrm{P}(Y \le y \mid \mathbf{X})$

- $\widehat{F_0}(y) = \mathbb{P}_n \widehat{r}_y(\mathbf{X})$, $\widehat{r}_y(\mathbf{X})$?

- $Y = \boldsymbol{\beta}_0' \mathbf{X} + \mathbf{u}$, $\mathbf{u} \coprod (\mathbf{X}, A)$. $r_y(\mathbf{X}) = \mathrm{P}(\mathbf{u} \le y - \boldsymbol{\beta}_0 \mathbf{X})$,

- Residuals: $\widehat{\mathbf{u}}_i = Y_i - \widehat{\boldsymbol{\beta}}_n \mathbf{X}_i$, $i : A_i = 1$.

$$\widehat{r}_y(\mathbf{X}) = \frac{1}{\sum_{i=1}^n A_i} \sum_{i=1}^n A_i \mathrm{I}_{\{\widehat{\mathbf{u}}_i \le y - \widehat{\boldsymbol{\beta}}_n \mathbf{X}\}}$$

$$\widehat{F_0}(y) = \mathbb{P}_n \widehat{r}_y(\mathbf{X}) = \frac{1}{n\sum_{i=1}^n A_i} \sum_{i,j=1}^n A_i \mathrm{I}_{\{\widehat{\boldsymbol{\beta}}_n \mathbf{X}_j + \widehat{\mathbf{u}}_i \le y\}}$$

Generalized Linear Regression estimator of $F_0(y)$

- $F_0(y) = \mathrm{E}\left\{\mathrm{E}\left(\mathrm{I}_{\{Y \leq y\}} \mid \mathbf{X}\right)\right\} = \mathrm{E}\{r_y(\mathbf{X})\}$,
- where $r_y(\mathbf{X}) = \mathrm{E}\left(\mathrm{I}_{\{Y \leq y\}} \mid \mathbf{X}\right) = \mathrm{P}(Y \leq y \mid \mathbf{X})$
- $\widehat{F_0}(y) = \mathbb{P}_n \widehat{r}_y(\mathbf{X})$, $\widehat{r}_y(\mathbf{X})$?
- Generalized Linear Model $Y \mid \mathbf{X} \sim G_{\boldsymbol{\beta}_0 \mathbf{X}}$

$$\widehat{r}_y(\mathbf{X}) = G_{\widehat{\boldsymbol{\beta}}_n \mathbf{X}}(y)$$

$$\widehat{F_0}(y) = \mathbb{P}_n G_{\widehat{\boldsymbol{\beta}}_n \mathbf{X}}(y) = \frac{1}{n} \sum_{i=1}^{n} G_{\widehat{\boldsymbol{\beta}}_n \mathbf{X}_i}(y) \, ,$$

Robustifiable!

Doubly Protected Estimation of $F_0(y)$

$$\mathrm{E}\{\ell(Y)\} = \mathrm{E}\left\{\frac{A\ell(Y)}{p(\mathbf{X})}\right\} - \mathrm{E}\left[\left\{\frac{A}{p(\mathbf{X})} - 1\right\} r(\mathbf{X})\right]$$

$$\text{if } p(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X}) \textbf{ OR } r(\mathbf{X}) = \mathrm{E}(\ell(Y) \mid \mathbf{X})$$

- Consider $\ell(Y) = \ell_y(Y) = \mathrm{I}_{\{Y \leq y\}}$

$$F_0(y) = \mathrm{E}\{\ell_y(Y)\} = \mathrm{E}\left\{\frac{A\ell_y(Y)}{p(\mathbf{X})}\right\} - \mathrm{E}\left[\left\{\frac{A}{p(\mathbf{X})} - 1\right\} r_y(\mathbf{X})\right]$$

$$\text{if } p(\mathbf{X}) = \mathrm{P}(A = 1 \mid \mathbf{X}) \textbf{ OR}$$
$$r_y(\mathbf{X}) = \mathrm{E}(\ell_y(Y) \mid \mathbf{X}) = \mathrm{P}(Y \leq y \mid \mathbf{X})$$

$$\widehat{F}(y)_{DP} = \mathbb{P}_n\left\{\frac{A\ell_y(Y)}{\widehat{\pi}(\mathbf{X})}\right\} - \mathbb{P}_n\left[\left\{\frac{A}{\widehat{\pi}(\mathbf{X})} - 1\right\} \widehat{r}_y(\mathbf{X})\right]$$

Doubly Protected Estimation of F_0

$$\widehat{F}(y)_{DP} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i \mathrm{I}_{\{Y_i \le y\}}}{\widehat{\pi}(\mathbf{X}_i)}\right\} - \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{A_i}{\widehat{\pi}(\mathbf{X}_i)} - 1\right\}\widehat{r}_y(\mathbf{X}_i)\right]$$

- Zhang (2012) Normal linear model holds after a Box - Cox transformation of Y.
- GLM: $Y \mid \mathbf{X} \sim G_{\boldsymbol{\beta}_0 \mathbf{X}}$, (parametric)

$$\widehat{r}_y(\mathbf{X}) = G_{\widehat{\boldsymbol{\beta}}_n \mathbf{X}}(y)\ ,$$ Robustifiable!

- Our proposal: semiparametric model $Y = g(\boldsymbol{\beta}_0, \mathbf{X}) + \mathbf{u}$, $\mathbf{u}$ independent of $(\mathbf{X}, A)$.

$$\widehat{r}_y(\mathbf{X}) = \frac{1}{\sum_{i=1}^{n} A_i}\sum_{i=1}^{n} A_i \mathrm{I}_{\{\widehat{\mathbf{u}}_i \le y - g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X})\}}$$

with a robust fit of the regression model.

Doubly Protected Estimation of F_0 - GLM Regression

$$\widehat{F}(y)_{DP} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i \mathrm{I}_{\{Y_i \le y\}}}{\pi(\mathbf{X}_i, \widehat{\gamma}_n)}\right\} - \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{A_i}{\pi(\mathbf{X}_i, \widehat{\gamma}_n)} - 1\right\} G_{\widehat{\boldsymbol{\beta}}_n \mathbf{X}_i}(y)\right]$$

consistent if if $\mathrm{P}(A = 1 \mid \mathbf{X}) = \pi(\mathbf{X}, \gamma_0)$ **OR** $Y \mid \mathbf{X} \sim G_{\boldsymbol{\beta}_0 \mathbf{X}}$

- GLM: $Y \mid \mathbf{X} \sim G_{\boldsymbol{\beta}_0 \mathbf{X}}$ is parametric.
- No protection against outliers, but robustifiable.

Our proposal: Robust Doubly Protected Estimation of F_0

$$\widehat{F}(y)_{RDP} =$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i \mathrm{I}_{\{Y_i \leq y\}}}{\pi(\mathbf{X}_i, \widehat{\gamma}_n)}\right\} - \frac{1}{n\sum A_j}\sum_{i,j=1}^{n}\left[\left\{\frac{A_i}{\pi(\mathbf{X}_i, \widehat{\gamma}_n)} - 1\right\} A_j \mathrm{I}_{\{g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X}_i) + \widehat{\mathbf{u}}_j}$$

consistent if if $\mathrm{P}(A = 1 \mid \mathbf{X}) = \pi(\mathbf{X}, \gamma_0)$ **OR** $Y = g(\boldsymbol{\beta}_0, \mathbf{X}) + \mathbf{u}$

- Semiparametric model: $Y = g(\boldsymbol{\beta}_0, \mathbf{X}) + \mathbf{u}$, $\mathbf{u} \sim$ unspecified.
- Protection against outliers

Doubly Protected Estimation (DP)

- Consider $g(Y) = g_y(Y) = I_{\{Y \le y\}}$
- $\mathrm{E}\{g(Y)\} = F(y)$, where $F = F_Y$
- $r_y(\mathbf{X}) = \mathrm{E}\{g_y(Y) \mid \mathbf{X}\} = \mathrm{P}(Y \le y \mid \mathbf{X})$
- $Y = \beta_0^t + u$, $u \sim \mathcal{N}(0, \sigma^2)$,
- $r_y(\mathbf{X}) = \mathrm{P}(u \le y - \beta_0^t \mathbf{X}) = \Phi\{(y - \beta_0^t \mathbf{X})/\sigma\}$.

$$\widehat{r}_y(\mathbf{X}) = \Phi\left(\frac{y - \widehat{\beta}_n^t \mathbf{X}}{\widehat{\sigma}}\right)$$

$$\widehat{F}(y)_{DP} = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(\mathbf{X})}\right\} - \mathbb{P}_n\left[\left\{\frac{A}{\widehat{\pi}(\mathbf{X})} - 1\right\}\widehat{r}_y(\mathbf{X})\right]$$

Monte Carlo Simulation: Data Generation

1. $\mathbf{X} = (X_1, X_2)$, X_1, X_2 i.i.d., $X_i \sim \mathcal{N}(0, 1)$
2. $\mathrm{P}(A = 1 | X_1, X_2) = \text{expit}((1, X_1, X_2)'\gamma_0)$, where $\text{expit}(t) = e^t/(1 + e^t)$ and $\gamma_0 = (0, 0.1, -1.1)$.
3. $Y = \beta_0'\mathbf{X} + \mathbf{u}$, $\mathbf{u}$ independent of $(\mathbf{X}, A)$, $\beta_0 = (-3, 10)$. Normal errors: $\mathbf{u} \sim \mathcal{N}(0, 1)$, Cauchy errors: $\mathbf{u} \sim t_1$.

- $\mathrm{P}(A = 1) = 0.5$
- $\mathrm{E}(Y) = \text{med}(Y) = 0$
- Wrong Model: X_2 is omitted.
- Contaminated samples: 10% of the data are replaced by outliers $(\mathbf{X}_0, A_0, Y_0)$ where
  - $\mathbf{X}_0 = (2, 0)$,
  - $\mathrm{P}(A_0 = 1 | \mathbf{X} = \mathbf{X}_0) = \text{expit}((1, 2, 0)'\gamma_0)$,
  - $Y_0 \in \{-100, -90, \ldots, -20, -10, 0, 10, 20, \ldots, 90, 100\}$.

Estimators of $\mu_0 = \mathbb{E}(Y) = \text{med}(Y)$

- DP: $\mu_0 = \mathbb{E}(Y)$can be estimated with

$$\widehat{\mu}_{DP} = \mathbb{P}_n \left\{ \frac{AY}{\pi(\mathbf{X}, \widehat{\gamma}_n)} \right\} - \mathbb{P}_n \left[\left\{ \frac{A}{\pi(\mathbf{X}, \widehat{\gamma}_n)} - 1 \right\} g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X}) \right]$$

  For each possible estimator $\widetilde{F}$ of F_Y, we estimate the median of Y taking

$$\inf\{y : \widetilde{F}(y) \geq 0.5\}$$

- RDP: For $m = \sum_{j=1}^{n} A_j$, estimate F_Y with

$$\widehat{F}(y)_{DP} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i \mathrm{I}_{\{Y_i \leq y\}}}{\widehat{\pi}(\mathbf{X}_i)} \right\} - \frac{1}{nm} \sum_{i,j=1}^{n} \left[\left\{ \frac{A_i}{\widehat{\pi}(\mathbf{X}_i)} - 1 \right\} A_j \mathrm{I}_{\{\widehat{\mathbf{u}}_j + g(\widehat{\boldsymbol{\beta}}_n, \mathbf{X}_i) \leq y\}} \right]$$

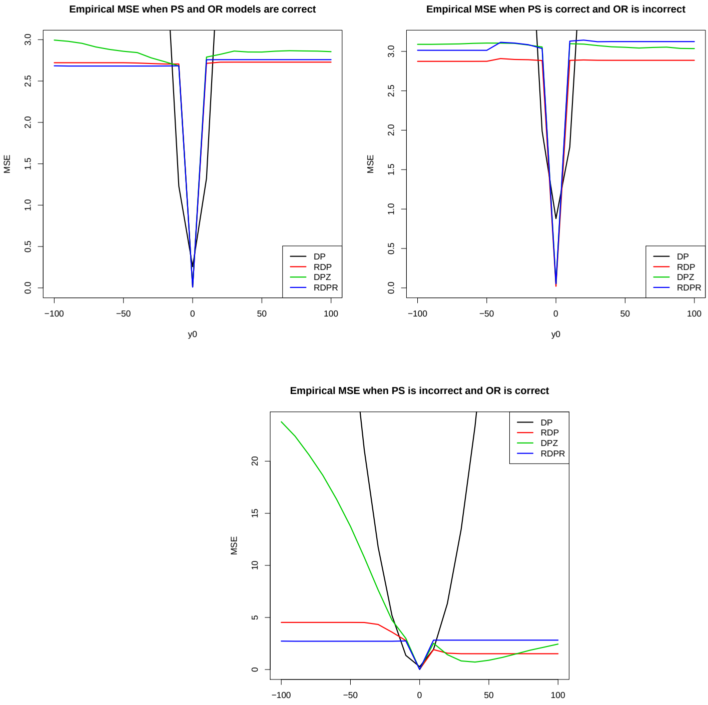
- DPZ: assuming normal errors in the regression model, estimate F_Y with

$$\widehat{F}(y)_{DPZ} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i \mathrm{I}_{\{Y_i \leq y\}}}{\widehat{\pi}(\mathbf{X}_i)} \right\} - \frac{1}{n} \sum_{i=1}^{n} \left[\left\{ \frac{A_i}{\widehat{\pi}(\mathbf{X}_i)} - 1 \right\} \Phi \left(\frac{y - \widehat{\boldsymbol{\beta}}_n^t \mathbf{X}_i}{\widehat{\sigma}} \right) \right]$$

- RDPZ: We estimate F_Y with as in the previous line, with a robust fit of the proposed linear model.

Monte Carlo Simulation: Results $n = 100$, $Nrep = 1000$, $Nrepcont = 100$

Method	PS	OR	Normal errors	t3 errors	Max MSE
IPW	correct		-0.16	-0.16	112.87
IPW	incorrect		-4.40	-4.39	203.80
RIPW	correct		0.05	0.05	4.90
RIPW	incorrect		-4.33	-4.31	35.43
R		correct	0.04	0.04	14.37
R		incorrect	-0.01	-0.00	13.41
SY		correct	0.05	0.06	11.51
SY		incorrect	-4.37	-4.35	53.94
RZ		correct	0.05	0.07	151.43
RZ		incorrect	-4.40	-4.39	195.45
RRZ		correct	0.05	0.05	3.90
RRZ		incorrect	-4.38	-4.35	53.27
DR	correct	correct	0.03	0.03	115.95
DR	correct	incorrect	0.03	0.05	114.82
DR	incorrect	correct	-0.11	-0.11	153.17
DR	incorrect	incorrect	-4.40	-4.39	204.08
RDP	correct	correct	0.02	0.04	4.46
RDP	correct	incorrect	0.01	0.02	6.13
RDP	incorrect	correct	0.11	0.14	6.00
RDP	incorrect	incorrect	-4.34	-4.34	36.01
DPZ	correct	correct	0.06	0.09	6.03
DPZ	correct	incorrect	0.05	-0.01	9.62
DPZ	incorrect	correct	0.17	0.22	27.02
DPZ	incorrect	incorrect	-4.33	-4.32	35.79
RDPZ	correct	correct	0.06	0.08	4.78
RDPZ	correct	incorrect	0.05	0.13	10.23
RDPZ	incorrect	correct	0.19	0.24	4.91
RDPZ	incorrect	incorrect	-4.32	-4.32	35.75


Empirical MSE when PS and OR models are correct
MSE
y0
DP
RDP
DPZ
RDPR
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−100
−50
0
50
100
Empirical MSE when PS is correct and OR is incorrect
MSE
y0
DP
RDP
DPZ
RDPR
Empirical MSE when PS is incorrect and OR is correct
MSE
0
5
10
15
20
DP
RDP
DPZ
RDPR

References

- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19), 2937-2960
- Yao L, Sun Z, Wang Q. Estimation of average treatment effects based on parametric propensity score model. Journal of Statistical Planning and Inference 2010; 140 (3): 806-816.
- Zhang, Z., Chen, Z., Troendle, J. F., & Zhang, J. (2012). Causal inference on quantiles with an obstetric application. Biometrics, 68(3), 697-706.