Doubly Robust Regression Analysis for Data Fusion

Katherine Evans†, BaoLuo Sun‡, James Robins*

and Eric J. Tchetgen Tchetgen**

† *Verily Life Sciences*

‡ *Department of Statistics and Applied Probability,*

National University of Singapore

* *Departments of Epidemiology and Biostatistics,*

Harvard T.H. Chan School of Public Health

** *Department of Statistics,*

The Wharton School of the University of Pennsylvania

Supplementary Material

This Supplementary Material contains proofs of the results, as well as additional simulation results.

S1 Derivation of DR linear space

The observed data likelihood is given by

$$L(O) = f(R|V;\eta)\left\{\int f(Y|V,L;\theta)dF(L|V;\alpha)\right\}^{R} f(L|V;\alpha)^{1-R} f(V;\epsilon),$$

where we consider α and ϵ to be possibly infinite-dimensional nuisance parameters and $O = (R, RY, (1-R)L, V)$. The nuisance tangent space is $\Lambda_\eta \oplus \Lambda_\alpha \oplus \Lambda_\epsilon$, where

$$\Lambda_\epsilon = \{B_1 S_\epsilon(V) : E[S_\epsilon(V)] = 0\}$$
$$\Lambda_\alpha = \{B_2 E[S_\alpha(V,L)|O] = B_2 \{RE[S_\alpha(V,L)|Y,V] + (1-R)S_\alpha(V,L)\} :$$
$$E[S_\alpha(V,L)|V] = 0\}$$
$$\Lambda_\eta = \left\{ B_3 \left[\frac{\partial}{\partial \eta} \log f(R|V;\eta) \right] \right\}.$$

Let $\Lambda^\perp$ be the observed-data linear space that is orthogonal to $\Lambda_\epsilon \oplus \Lambda_\alpha$. Then for given $h(O) \in \Lambda^\perp_{\epsilon,\alpha}$ we have

$$E\left[h(O)S_\epsilon(V)\right] = 0 \quad \forall S_\epsilon(V) \in \Lambda_\epsilon,$$
$$E\left\{h(O)E[S_\alpha(V,L)|O]\right\} = E\left\{E[h(O)S_\alpha(V,L)|O]\right\}$$
$$= E\left\{h(O)S_\alpha(V,L)\right\} = 0 \quad \forall S_\alpha(V,L).$$

From the results of Robins et al. (1995) and Hasminskii and Ibragimov (1983), $\Lambda^\perp_{\epsilon,\alpha}$ is given by

$$\Lambda^\perp_{\epsilon,\alpha} = \{Bh(O) : E[h(R,V)|V] = 0 \text{ or } E[h(O)|L,V] = 0\}$$
$$= \left\{ B\left[\frac{R}{\pi(V)}\left[g(Y,V) + k(V)\right] - \frac{1-R}{1-\pi(V)} E[g(Y,V) + k(V)|V,L] \right] : \right.$$
$$\left. g, k \text{ arbitrary, } g(0,x) = 0 \right\}.$$

Therefore, when the data source process is modeled, a typical element in the ortho-complement $\Lambda^{\perp}$ to the nuisance tangent space is given by

$$\left\{h(O) - \Pi\left[h(O)|\Lambda_{\eta}\right] : h(O) \in \Lambda_{\epsilon,\alpha}^{\perp}\right\},$$

where Π denotes the projection operator. For a fixed choice of function $g(Y,V)$, the space of elements in $\Lambda^{\perp}$ is a translation of a linear space away from the origin. Specifically, this linear space is given by $V(g) = x_0 + M$, with the element

$$x_0 = \left\{\frac{R}{\pi(V)} g(Y,V) - \frac{1-R}{1-\pi(V)} E[g(Y,V)|V,L]\right\} - \Pi\left[\{\cdot\}|\Lambda_{\eta}\right]$$

and linear subspace

$$M = \left\{\left[\frac{R}{\pi(V)} - \frac{1-R}{1-\pi(V)}\right] k(V)\right\} - \Pi\left[\{\cdot\}|\Lambda_{\eta}\right] = \Pi[\Omega(V)|\Lambda_{\eta}^{\perp}].$$

It is clear that $\Lambda_{\eta} \subset \Omega(V)$. By Theorem 10.1 of (Tsiatis, 2007), the optimal influence function (in terms of smallest variance) for fixed $g(Y,V)$ is given by

$$\mathbb{IF}^*(g) = \left\{\frac{R}{\pi(V)} g(Y,V) - \frac{1-R}{1-\pi(V)} E[g(Y,V)|V,L]\right\} - \Pi\left[\{\cdot\}|\Omega(V)\right].$$

Let

$$\left[\frac{R}{\pi(V)} - \frac{1-R}{1-\pi(V)}\right] k^0(V) \in \Omega(V)$$

be the projection $\Pi\left[\{\cdot\}|\Omega(V)\right]$. Then $k^0(V)$ needs to satisfy

$$E\left\{\left\{\frac{R}{\pi(V)}\left[g(Y,V)-k^0(V)\right]-\frac{1-R}{1-\pi(V)}\left[k^0(V)-E[g(Y,V)|V,L]\right]\right\}\times\right.$$
$$\left.\left\{\left[\frac{R}{\pi(V)}-\frac{1-R}{1-\pi(V)}\right]k(V)\right\}\right\}$$
$$=E\left\{k(V)\left\{\frac{1}{\pi(V)}\left[E[g(Y,V)|V]-k^0(V)\right]+\right.\right.$$
$$\left.\left.\frac{1}{1-\pi(V)}\left[k^0(V)-E[g(Y,V)|V]\right]\right\}\right\}=0\quad\forall k(V).$$

By assumption (A2), since $\delta<\pi(V)<1-\delta$ almost surely, $k^0(V)=E[g(Y,V)|V]$ and the DR linear space is given by

$$\mathcal{L}_{DR}=\left\{\mathbb{IF}^*(g):g(Y,V)\text{ arbitrary}\right\},$$

where

$$\mathbb{IF}^*(g)=\left\{\frac{R}{\pi(V)}\left[g(Y,V)-E[g(Y,V)|V]\right]\right.$$
$$\left.+\frac{1-R}{1-\pi(V)}\left[E[g(Y,V)|V]-E[g(Y,V)|V,L]\right]\right\}.$$

S2 Proofs of Results

In the following, expectations are evaluated at the true parameter values.

Proof of Result 1.

$$E_{\eta,\theta}\left\{U_g(\theta;\eta)\middle|V,L\right\}=E_{\eta,\theta}\left\{\frac{R}{\pi(V)}g(Y,V)-\frac{1-R}{1-\pi(V)}E_\theta[g(Y,V)|V,L]\middle|V,L\right\}$$
$$=E_\theta[g(Y,V)|V,L]-E_\theta[g(Y,V)|V,L]=0.$$

□

Proof of Result 2 (DR property). **Case 1: $\pi(V)$ is correct but $\tilde{t}(L|V)$ is incorrect**

Unbiasedness of DR estimating function follows from Result 1 by taking $g'(V,L) = g(V,L) + k(V)$; the proof does not involve $\tilde{t}(L|V)$.

Case 2: $\tilde{\pi}(V)$ is incorrect but $t(L|V)$ is correct

$$
\begin{aligned}
E_{\theta,\eta,\alpha}\left\{U_g^{DR}(\theta;\eta,\alpha)\middle|V\right\} &= E_{\theta,\eta,\alpha}\left\{\frac{R}{\tilde{\pi}(V)}\left\{g(Y,V) - E_{\theta,\alpha}[g(Y,V)|V]\right\}\right. \\
&\quad\left.+\frac{1-R}{1-\tilde{\pi}(V)}\left\{E_{\theta,\alpha}[g(Y,V)|V] - E_{\theta}[g(Y,V)|V,L]\right\}\middle|V\right\} \\
&=\frac{\pi(V)}{\tilde{\pi}(V)}\left\{E_{\theta,\alpha}[g(Y,V)|V] - E_{\theta,\alpha}[g(Y,V)|V]\right\} \\
&\quad+\frac{1-\pi(V)}{1-\tilde{\pi}(V)}\left\{E_{\theta,\alpha}[g(Y,V)|V] - E_{\theta,\alpha}[g(Y,V)|V]\right\} = 0.
\end{aligned}
$$

□

Proof of Result 3. The proof is based on the following lemma which is part of Theorem 5.3 in Newey and McFadden (1994).

Lemma S1.

If $\exists\tilde{h}(V)$ satisfying

$$
-E\left[h(V)\nabla_\theta M(\theta)\right] = E\left[M^2(\theta)h(V)\tilde{h}(V)^T\right] \quad \forall h(V),
$$

then the estimator indexed by $\tilde{h}(V)$ is most efficient.

Proof of Lemma S1. If $h(V)$ and $\tilde{h}(V)$ satisfy the equality in lemma S1 then the difference of the asymptotic variances of the respective estimators indexed by them is as follows:

$$\begin{aligned} & E\left[M^2(\theta)h(V)\tilde{h}(V)^T\right]^{-1} E\left[M^2(\theta)h(V)h(V)^T\right] E\left[M^2(\theta)\tilde{h}(V)h(V)^T\right]^{-1} \\ & - E\left[M^2(\theta)\tilde{h}(V)\tilde{h}(V)^T\right]^{-1} \\ =& E\left[M^2(\theta)h(V)\tilde{h}(V)^T\right]^{-1} E\left[UU^T\right] E\left[M^2(\theta)\tilde{h}(V)h(V)^T\right]^{-1}, \end{aligned}$$

where $U = h(V) - E\left[M^2(\theta)h(V)\tilde{h}(V)^T\right] E\left[M^2(\theta)\tilde{h}(V)\tilde{h}(V)^T\right]^{-1} \tilde{h}(V)$ and $E\left[UU^T\right]$ is positive semi-definite. $\square$

We show that if $\tilde{h}(V)$ satisfies the equality in lemma S1 then $\tilde{h}(V) = h^{opt}(V)$.

$$\begin{aligned} & - E\left[h(V)\nabla_\theta M(\theta)\right] = E\left[M^2(\theta)h(V)h^{opt}(V)^T\right] \quad \forall h(V), \\ \iff & E\left\{h(V)\left[M^2(\theta)h^{opt}(V) + \nabla_\theta M(\theta)\right]^T\right\} = 0 \quad \forall h(V), \\ \iff & E\left\{h(V)E\left[M^2(\theta)h^{opt}(V) + \nabla_\theta M(\theta)\middle|V\right]^T\right\} = 0 \quad \forall h(V), \\ \implies & E\left\{E\left[M^2(\theta)h^{opt}(V) + \nabla_\theta M(\theta)\middle|V\right]^{\otimes 2}\right\} = 0, \\ \implies & E\left[M^2(\theta)h^{opt}(V) + \nabla_\theta M(\theta)\middle|V\right] = 0, \\ \iff & h^{opt}(V) = -E\left[\nabla_\theta M(\theta)|V\right] E\left[M^2(\theta)|V\right]^{-1}. \end{aligned}$$

Due to Hájek's representation theorem (Hájek, 1970), the most efficient regular estimator is asymptotically linear and so the existence condition in lemma S1 holds when we consider only RAL estimators. □

S3 Additional Simulation Results

Figure 1: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_0, whose true value of 0.5 is marked by the horizontal line, when $\alpha_3 = 2$.

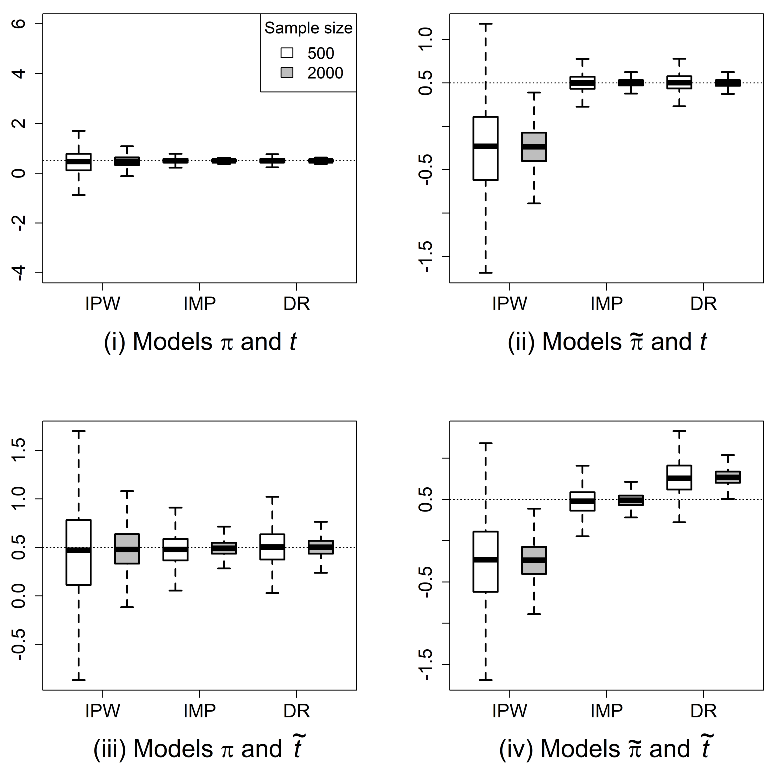


Figure 2: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_1, whose true value of -0.5 is marked by the horizontal line, when $\alpha_3 = 2$.

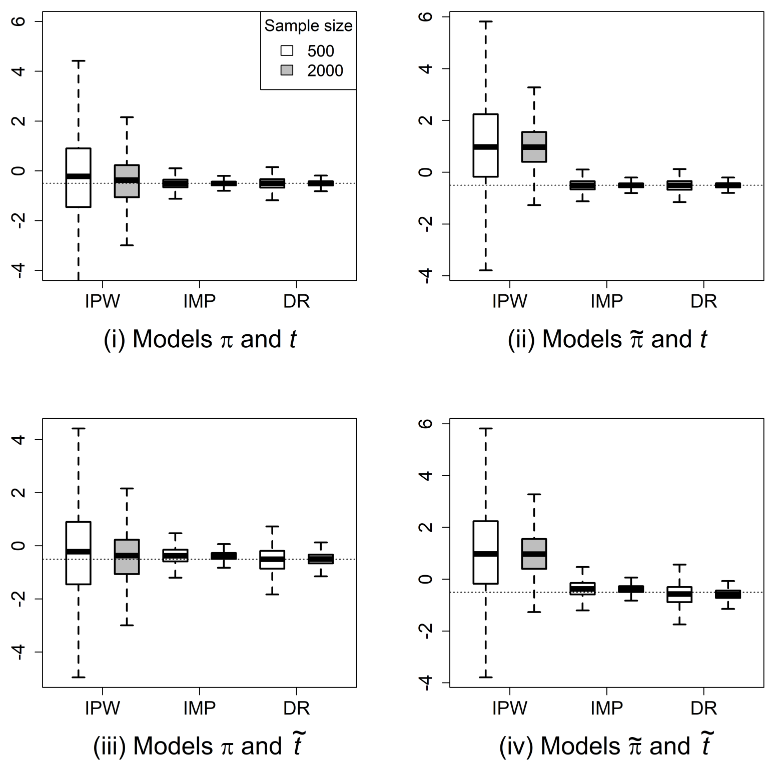


Figure 3: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_2, whose true value of 1.0 is marked by the horizontal line, when $\alpha_3 = 2$.

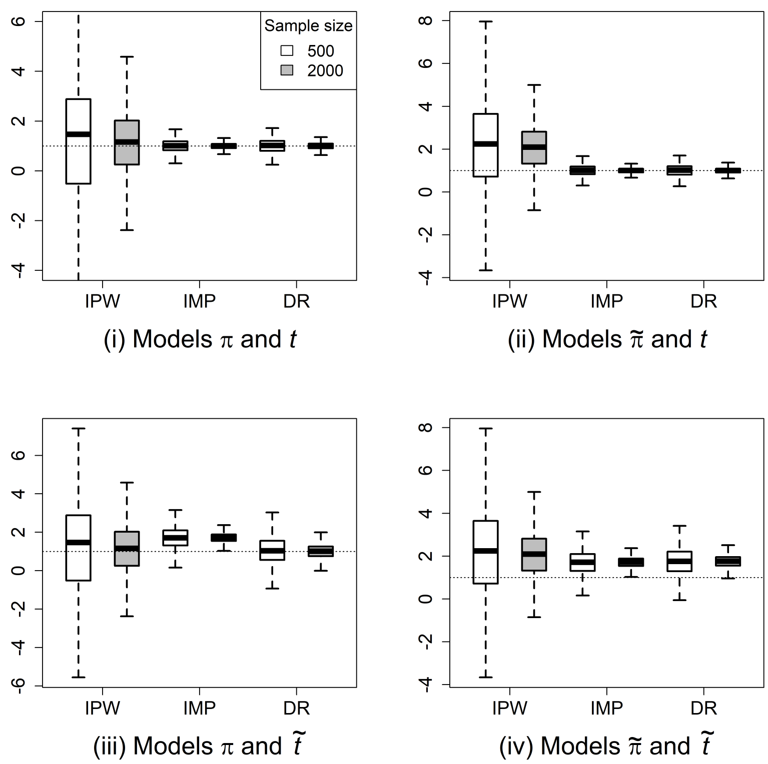


Figure 4: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_0, whose true value of 0.5 is marked by the horizontal line, when $\alpha_3 = 0.5$.

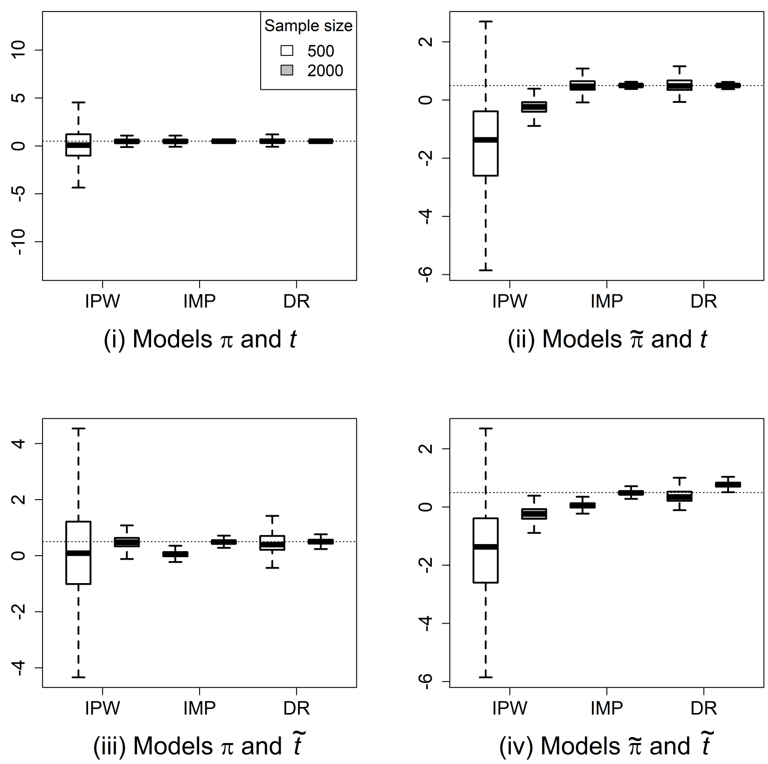


Figure 5: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_1, whose true value of -0.5 is marked by the horizontal line, when $\alpha_3 = 0.5$.

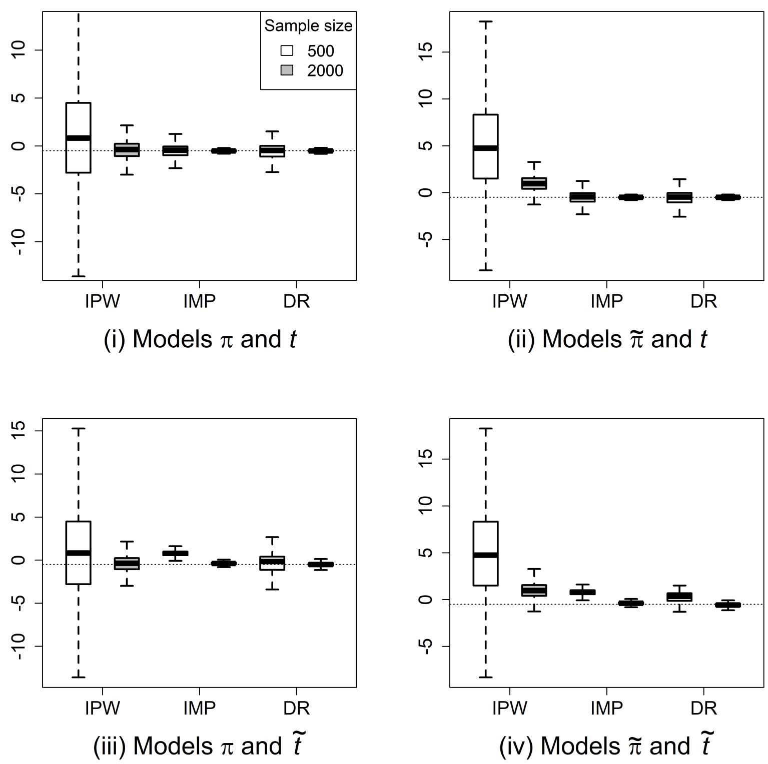
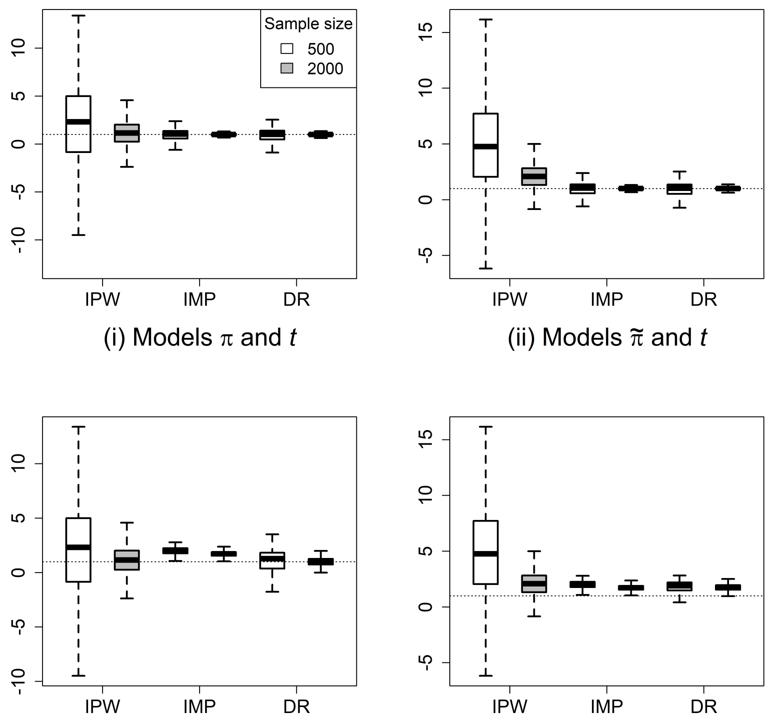


Figure 6: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_2, whose true value of 1.0 is marked by the horizontal line, when $\alpha_3 = 0.5$.



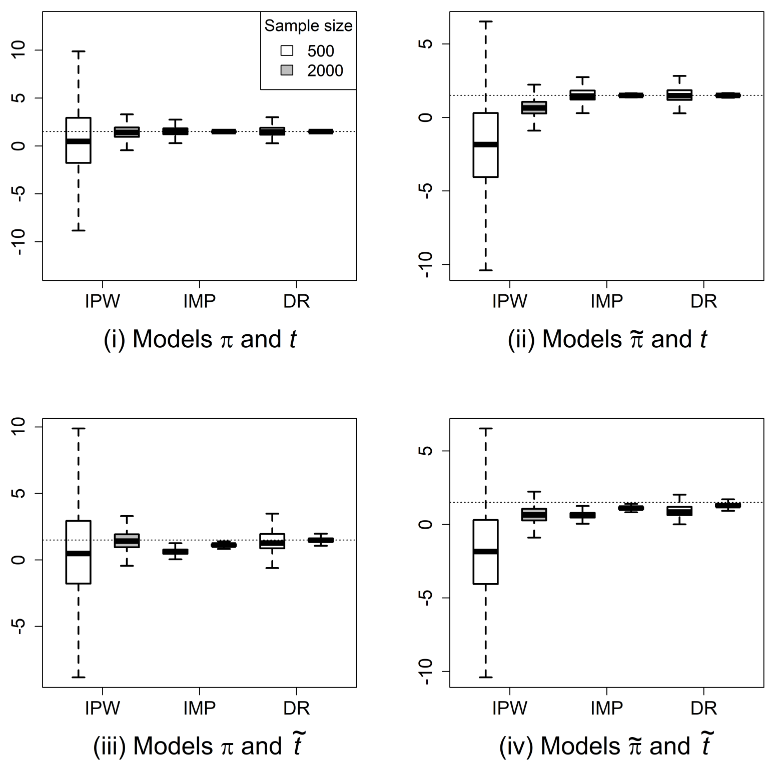
(i) Models π and t

(ii) Models $\tilde{\pi}$ and t

(iii) Models π and $\tilde{t}$

(iv) Models $\tilde{\pi}$ and $\tilde{t}$

Figure 7: Boxplots of inverse probability weighted (IPW), imputation-based (IMP) and doubly-robust (DR) estimators of the regression coefficient β_3, whose true value of 1.5 is marked by the horizontal line, when $\alpha_3 = 0.5$.



Bibliography

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330.

Hasminskii, R. and Ibragimov, I. (1983). On asymptotic efficiency in the presence of an infinitedimensional nuisance parameter. In *Probability theory and mathematical statistics*, pages 195–229. Springer.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 409–424.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.