The Doubly Robust Estimator (DRE) for the Average Causal Effect

Hanyu Wu

Department of Biostatistics, Peking University

2023.11.9

Two Fomulae for $\tau = \mathbb{E}(Y(1) - Y(0))$

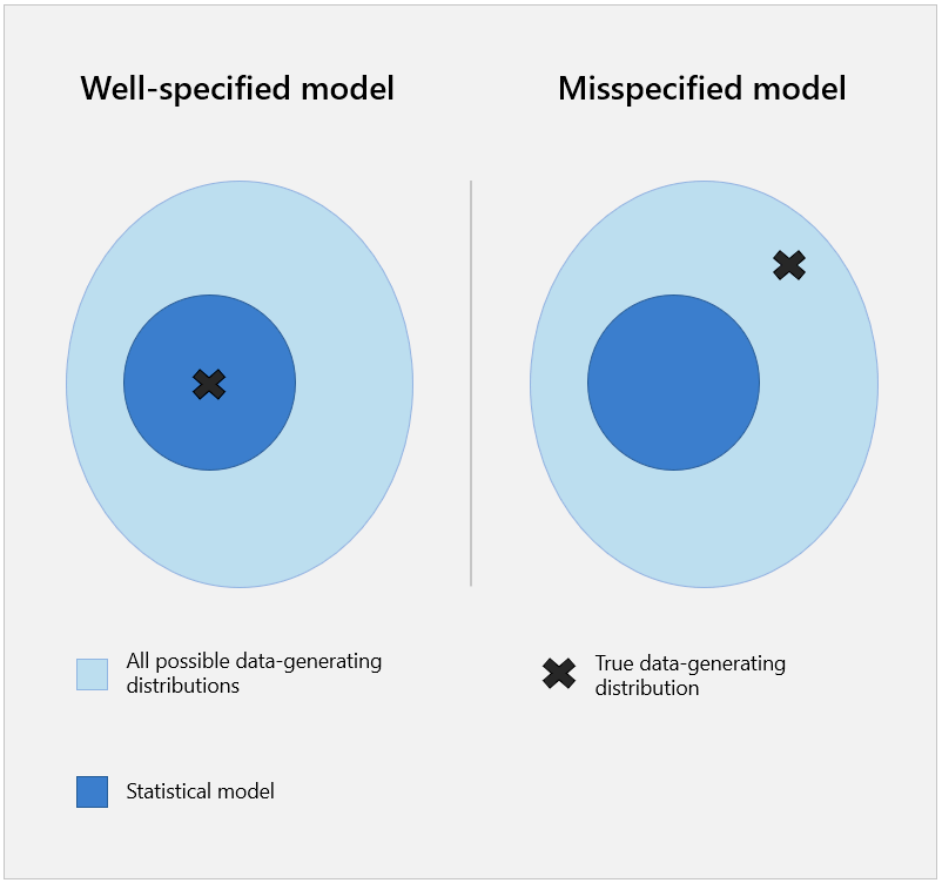
- $\tau_1 = \mathbb{E}(\mu_1(X)) - \mathbb{E}(\mu_0(X))$, where $\mu_i(X) = \mathbb{E}(Y(i)|X)$.
- $\tau_2 = \mathbb{E}\left(\frac{ZY}{e(X)}\right) - \mathbb{E}\left(\frac{(1-Z)Y}{1-e(X)}\right)$, where $e(X) = \mathbb{P}(Z = 1|X)$.
- Under strong ignorability and overlap conditions, $\tau_1 = \tau_2 = \tau$.

Motivation of Doubly Robust

- Model misspecification: happens when the set of probability distributions considered by the statistician does not include the distribution that generated the observed data.

Motivation of Doubly Robust

- Model misspecification: happens when the set of probability distributions considered by the statistician does not include the distribution that generated the observed data.



Motivation of Doubly Robust

- τ_1 requires fitting a specified outcome model given Z and X.
- τ_2 requires fitting a specified treatment model given X.

Motivation of Doubly Robust

- τ_1 requires fitting a specified outcome model given Z and X.
- τ_2 requires fitting a specified treatment model given X.
- Motivation: make a combination of τ_1 and τ_2 to create an estimator, that is consistent if either $\mu_i(X)$ and $e(X)$ is correctly specified.

Working Model

- Working model for outcome: $\mu_1(X, \beta_1)$, $\mu_0(X, \beta_0)$, indexed by parameters β_1 and β_0, e.g. regression coefficients for linear model.

- If the outcome model is correctly specified, then $\mu_i(X, \beta_1) = \mu_i(X)$, $i = 1, 2$.

Working Model

- Working model for outcome: $\mu_1(X, \beta_1)$, $\mu_0(X, \beta_0)$, indexed by parameters β_1 and β_0, e.g. regression coefficients for linear model.
- If the outcome model is correctly specified, then $\mu_i(X, \beta_1) = \mu_i(X)$, $i = 1, 2$.
- Working model for propensity score: $e(X, \alpha)$.
- If the propensity score model is correctly specified, then $e(X, \alpha) = e(x)$.

Working Model

- Working model for outcome: $\mu_1(X, \beta_1)$, $\mu_0(X, \beta_0)$, indexed by parameters β_1 and β_0, e.g. regression coefficients for linear model.
- If the outcome model is correctly specified, then $\mu_i(X, \beta_1) = \mu_i(X)$, $i = 1, 2$.
- Working model for propensity score: $e(X, \alpha)$.
- If the propensity score model is correctly specified, then $e(X, \alpha) = e(x)$.
- In practice, both models may be misspecifie.

Doubly Robust Estimator

Augment the outcome model by IPW terms of the residual:

$$\widetilde{\mu}_1^{dr} = \mathbb{E}\left[\frac{Z(Y - \mu_1(X, \beta_1))}{e(X, \alpha)} + \mu_1(X, \beta_1)\right]$$

$$\widetilde{\mu}_0^{dr} = \mathbb{E}\left[\frac{(1 - Z)(Y - \mu_0(X, \beta_0))}{1 - e(X, \alpha)} + \mu_0(X, \beta_0)\right]$$

Doubly Robust Estimator

Augment IPW estimator by the outcome model:

$$\widetilde{\mu}_1^{dr} = \mathbb{E}\left[\frac{ZY}{e(X,\alpha)} - \frac{Z - e(X,\alpha)}{e(X,\alpha)}\mu_1(X,\beta_1)\right]$$

$$\widetilde{\mu}_0^{dr} = \mathbb{E}\left[\frac{(1-Z)Y}{1-e(X,\alpha)} - \frac{e(X,\alpha) - Z}{1-e(X,\alpha)}\mu_0(X,\beta_0)\right]$$

Theoretical Properties of DR Estimator

Theorem (12.1)

Assume ignorability $Z \perp\!\!\!\perp \{Y(1), Y(0)\}$ *and overlap* $0 < e(X) < 1$. *Then*

1. *If either* $e(X, \alpha) = e(X)$ *or* $\mu_1(X, \beta_1) = \mu_1(X)$, *then* $\widetilde{\mu}_1^{dr} = \mathbb{E}\{Y(1)\}$.
2. *If either* $e(X, \alpha) = e(X)$ *or* $\mu_0(X, \beta_1) = \mu_0(X)$, *then* $\widetilde{\mu}_0^{dr} = \mathbb{E}\{Y(0)\}$.
3. *If either* $e(X, \alpha) = e(X)$ *or* $\{\mu_1(X, \beta_1) = \mu_1(X), \mu_0(X, \beta_1) = \mu_0(X)\}$,

then $\widetilde{\mu}_1^{dr} - \widetilde{\mu}_0^{dr} = \tau$.

Proof of Theorem 12.1

$$\begin{aligned}
\widetilde{\mu}_1^{dr} - \mathbb{E}\{Y(1)\} &= \mathbb{E}\left[\frac{Z\{Y(1) - \mu_1(X, \beta_1)\}}{e(X, \alpha)} - \{Y(1) - \mu_1(X, \beta_1)\}\right] \\
&= \mathbb{E}\left[\frac{Z - e(X, \alpha)}{e(X, \alpha)}\{Y(1) - \mu_1(X, \beta_1)\}\right] \\
&= \mathbb{E}\left(\mathbb{E}\left[\frac{Z - e(X, \alpha)}{e(X, \alpha)}\{Y(1) - \mu_1(X, \beta_1)\}|X\right]\right) \\
&\quad \text{(law of total expectation)} \\
&= \mathbb{E}\left[\mathbb{E}\left\{\frac{Z - e(X, \alpha)}{e(X, \alpha)}|X\right\} \times \mathbb{E}\{Y(1) - \mu_1(X, \beta_1)|X\}\right] \\
&= \mathbb{E}\left[\frac{e(X) - e(X, \alpha)}{e(X, \alpha)} \times \{\mu(X) - \mu_1(X, \beta_1)\}\right]
\end{aligned}$$

Therefore, $\widetilde{\mu}_1^{dr} = 0$ if either $e(X, \alpha) = e(X)$ or $\mu_1(X, \beta_1) = \mu_1(X)$. The proof for μ_0 is similar.

Summary

According to Theroem 12.1, $\widetilde{\tau}^{dr} = \widetilde{\mu}_1^{dr} - \widetilde{\mu}_0^{dr}$ is a doubly robust (DR) estimator of τ. In reality, one definitely cannot guarantee whether one model can accurately explain the relationship among variables. The combination of outcome regression with weighting by propensity score ensures that the estimators are robust to misspecification of one of these models.

Doubly Robust Estimator for ATE

Definition (12.1)

Based on the data $(X_i, Z_i, Y_i)_{i=1}^n$, we can obtain a doubly robust estimator for τ by the following steps:

1. obtain the fitted values if the propensity scores: $e(X_i, \hat{\alpha})$;
2. obtain the fitted values of the outcome means: $\mu_1(X_i, \hat{\beta}_1)$ and $\mu_0(X_i, \hat{\beta}_0)$;
3. construct the doubly robust estimator: $\hat{\tau}^{dr} = \hat{\mu}_1^{dr} - \hat{\mu}_0^{dr}$, where

$$\hat{\mu}_1^{dr} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i\{Y_i - \mu_1(X_i, \beta_1)\}}{e(X_i, \alpha)} + \mu_1(X_i, \beta_1)\right]$$

and

$$\hat{\mu}_0^{dr} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-Z_i)\{Y_i - \mu_0(X_i, \beta_0)\}}{1 - e(X_i, \alpha)} + \mu_0(X_i, \beta_0)\right]$$

Two Forms of $\hat{\tau}^{dr}$

Write $\hat{\tau}^{dr}$ as:

$$\hat{\tau}^{dr} = \hat{\tau}^{reg} + \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i\{Y_i - \mu_1(X_i, \hat{\beta}_1)\}}{e(X_i, \hat{\alpha})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)\{Y_i - \mu_0(X_i, \hat{\beta}_0)\}}{1 - e(X_i, \hat{\alpha})}$$

$$\hat{\tau}^{dr} = \hat{\tau}^{ipw} - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i - e(X_i, \hat{\alpha})}{e(X_i, \hat{\alpha})}\mu_1(X_i, \hat{\beta}_1) - \frac{1}{n}\sum_{i=1}^{n}\frac{e(X_i, \hat{\alpha}) - Z_i}{1 - e(X_i, \hat{\alpha})}\mu_0(X_i, \hat{\beta}_0)\right)$$

where

$$\hat{\tau}^{reg} = \frac{1}{n}\sum_{i=1}^{n}\mu_1(X_i, \hat{\beta}_1) - \frac{1}{n}\sum_{i=1}^{n}\mu_0(X_i, \hat{\beta}_0)$$

$$\hat{\tau}^{ipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e(X_i, \hat{\alpha})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1 - e(X_i, \hat{\alpha})}$$

Intuition Perspectives of DR Estimator

- The original motivation for $\widetilde{\mu}_1^{dr}$ and $\widetilde{\mu}_0^{dr}$ was quite theoretical, which relies on the *semiparametric efficiency theory* in advanced mathematical statistics.

- The book gives two more intuitive perspectives to construct $\widetilde{\mu}_1^{dr}$, while $\widetilde{\mu}_0^{dr}$ is similar by symmetry.

I. Reducing the Variance of the IPW Estimator

- The IPW estimator for μ_1 based on

$$\mu_1 = \mathbb{E}\left\{\frac{ZY}{e(X)}\right\}$$

  completely ignores the outcome model of Y. It has the advantages of being consistent without assuming any outcome model.

- However, **if the covariates are predictive to the outcome**, the residual based on a working outcome model usually has smaller variance than the outcome even if this working outcome model is wrong.

I. Reducing the Variance of the IPW Estimator

- With a possibly mis-specified outcome model $\mu_1(X, \beta_1)$, consider a trivial decomposition

$$\mu_1 = \mathbb{E}\{Y(1)\} = \mathbb{E}\{Y(1) - \mu_1(X, \beta_1)\} + \mathbb{E}\{\mu_1(X, \beta_1)\}$$

- View the residual $Y(1) - \mu_1(X, \beta_1) := Y^*(1)$ as a pseudo potential outcome under the treatment, and apply the IPW formula to $Y^*(1)$.

I. Reducing the Variance of the IPW Estimator

- Given the true propensity score $e(X)$,

$$\begin{aligned}\mu_1 &= \mathbb{E}\{Y^*(1)\} + \mathbb{E}\{\mu_1(X, \beta_1)\} \\ &= \mathbb{E}\left\{\frac{ZY^*}{e(X)}\right\} + \mathbb{E}\{\mu_1(X, \beta_1)\} \quad (\text{Theorem11.2}) \\ &= \mathbb{E}\left\{\frac{Z(Y - \mu_1(X, \beta_1))}{e(X)}\right\} + \mathbb{E}\{\mu_1(X, \beta_1)\} \\ &= \mathbb{E}\left\{\frac{Z(Y - \mu_1(X, \beta_1))}{e(X)} + \mu_1(X, \beta_1)\right\}\end{aligned}$$

- The above equation holds if the propensity score is correct.

Remark: DRE in Sample Survey

- Connections: Z (indicator of missing), $\pi(X)$ (conditional probability of observed)

- Missing data occurs in surveys for many reasons. Two methods to deal with missing data: **imputation** and **IPW estimator**.

Remark: DRE in Sample Survey

- Imputation: assume a data generation model $\mu(W, \gamma_1)$ and

$$\hat{\mu}_1^{imp} = \frac{1}{n}\sum_{i=1}^{n}\{Z_i Y_i + (1 - Z_i)\mu(W_i, \hat{\gamma}_1)\}$$

- IPWCC (complete-case): posit a model $\pi(w, \gamma_2)$ and

$$\hat{\mu}_1^{ipw} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{Z_i Y_i}{\pi(W_i, \hat{\gamma}_2)}$$

Remark: DRE in Sample Survey

- Doubly robust augmented inverse probability weighted complete-case (AIPWCC) estimator:

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\frac{Z_i Y_i}{\pi(W_i, \hat{\gamma}_2)} - \frac{Z_i - \pi(W_i, \hat{\gamma}_2)}{\pi(W_i, \hat{\gamma}_2)} \mu(W_i, \hat{\gamma}_1) \right]$$

II. Reducing the Bias of the Outcome Regression Estimator

- Start with an outcome regression estimator based on

$$\widetilde{\mu}_1 = \mathbb{E}\{\mu_1(X, \beta_1)\}$$

which may not be the same as μ_1 since the outcome model may be wrong, and the bias is $\mathbb{E}\{\mu_1(X, \beta_1) - Y(1)\}$.

II. Reducing the Bias of the Outcome Regression Estimator

- Start with an outcome regression estimator based on

$$\widetilde{\mu}_1 = \mathbb{E}\{\mu_1(X, \beta_1)\}$$

which may not be the same as μ_1 since the outcome model may be wrong, and the bias is $\mathbb{E}\{\mu_1(X, \beta_1) - Y(1)\}$.

- Consider an IPW estimator of the bias:

$$B = \mathbb{E}\left\{\frac{Z\{\mu_1(X, \beta_1) - Y\}}{e(X)}\right\} \tag{1}$$

So the de-biased estimator is $\widetilde{\mu}_1 - B$, which is identical to the DR estimator.

The Product Structure

- Recall the proof of Theorem 12.1, the key for the double robustness property is the product structure in

$$\widetilde{\mu}_1^{dr} - \mathbb{E}\{Y(1)\} = \mathbb{E}\left[\frac{e(X) - e(X,\alpha)}{e(X,\alpha)} \times \{\mu(X) - \mu_1(X,\beta_1)\}\right]$$

- This delicate structure renders the doubly robust estimator possibly doubly fragile when both the propensity score and the outcome models are misspecified. **The product of two errors multiply to yield potentially much larger errors**.

The Product Structure

- Kang and Schafer (2007) found that the finite-sample performance of the doubly robust estimator can be even more wild than the simple regression imputation and IPW estimators in simulation.

Performance of bias-corrected regression estimators over 1000 *samples from the artificial population*

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	*BC-OLS*	−0.08	−3.4	2.48	1.68
		Incorrect	*BC-OLS*	0.25	7.5	3.28	2.17
	Incorrect	Correct	*BC-OLS*	−0.08	−3.3	2.48	1.70
		Incorrect	*BC-OLS*	−5.12	−43.0	12.96	3.54
(b) $n = 1000$	Correct	Correct	*BC-OLS*	0.00	−0.1	1.17	0.79
		Incorrect	*BC-OLS*	0.06	3.4	1.75	1.02
	Incorrect	Correct	*BC-OLS*	−0.02	−1.4	1.49	0.80
		Incorrect	*BC-OLS*	−21.03	−13.5	157.21	5.32

- In spite of this, the doubly robust estimator has been a standard strategy in causal inference.

Double Machine Learning

- Consider the partially linear regression (PLR) model

$$Y = D\theta_0 + g_0(X) + U,\ \mathbb{E}[U|X, D] = 0,$$

$$D = m_0(X) + V,\ \mathbb{E}[V|X] = 0.$$

- The bias has two sources, **regularization** and **overfitting**. Double Machine Learning aims to correct both.

Regularization Bias

- Motivation: Frisch-Waugh-Lovell theorem and moment condition.

- Consider Neyman orthogonal score

$$\Phi(W, \theta, \eta) = (D - m_0(Z)) \times (Y - g_0(Z) - (D - m_0(Z))\theta)$$

where nuisance parameter $\eta^T = (g^T, m^T)$, W is our piece of data.

- Neyman orthogonality condition: $\mathbb{E}\Phi(W, \theta, \eta) = 0$.

Overfitting Bias

- Sample-splitting approach:
  - Randomly partition our data into two subsets.
  - Fit our Machine Learning models for D and Y on the first subset.
  - Estimate θ_0 in the second subset, using the models obtained in step 2.

Overfitting Bias

- Sample-splitting approach:
  - Randomly partition our data into two subsets.
  - Fit our Machine Learning models for D and Y on the first subset.
  - Estimate θ_0 in the second subset, using the models obtained in step 2.
- Cross-fitting approach:
  - Exchange the two subset and repeat.
  - Obtain our final estimator θ_0 as an average of $\theta_{0,1}$ and $\theta_{0,2}$.

Overview of DML

DEFINITION 3.1. (DML1) *(a) Take a K-fold random partition* $(I_k)_{k=1}^K$ *of observation indices* $[N] = \{1, \ldots, N\}$ *such that the size of each fold* I_k *is* $n = N/K$. *Also, for each* $k \in [K] = \{1, \ldots, K\}$, *define* $I_k^c := \{1, \ldots, N\} \setminus I_k$. *(b) For each* $k \in [K]$, *construct an ML estimator*

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

of η_0, *where* $\hat{\eta}_{0,k}$ *is a random element in* T, *and where randomness depends only on the subset of data indexed by* I_k^c. *(c) For each* $k \in [K]$, *construct the estimator* $\check{\theta}_{0,k}$ *as the solution of the following equation:*

$$\mathbb{E}_{n,k}[\psi(W; \check{\theta}_{0,k}, \hat{\eta}_{0,k}] = 0, \tag{3.1}$$

where ψ *is the Neyman orthogonal score, and* $E_{n,k}$ *is the empirical expectation over the kth fold of the data; that is,* $E_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$. *If achievement of exact 0 is not possible, define the estimator* $\check{\theta}_{0,k}$ *of* θ_0 *as an approximate* ϵ_N*-solution:*

$$\|E_{n,k}[\psi(W; \check{\theta}_{0,k}, \hat{\eta}_{0,k})]\| \leq \inf_{\theta \in \Theta} \|E_{n,k}[\psi(W; \theta, \hat{\eta}_{0,k})]\| + \epsilon_N, \quad \epsilon_N = o(\delta_N N^{-1/2}), \tag{3.2}$$

where $(\delta_N)_{N \geq 1}$ *is some sequence of positive constants converging to zero. (4) Aggregate the estimators:*

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \check{\theta}_{0,k}. \tag{3.3}$$

Thanks!