STA 640 — Causal Inference

Chapter 3.5. Doubly Robust Estimation

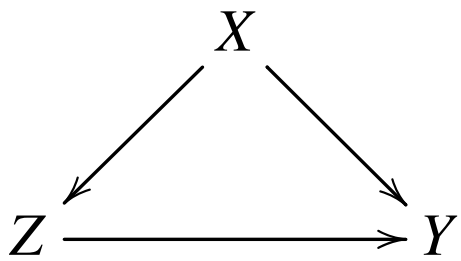
Fan Li

Department of Statistical Science

Duke University

Confounding

- Confounding X (or common cause; see DAG) is the main complication/hurdle between association and causation



- Outcome-regression: need models for $Y \sim X$ within $Z = z$
- Propensity score methods: need a model for $Z \sim X$
- Combination of both to achieve “double robustness"
- For now, focus on the ATE estimand τ

Outcome Regression (OR) Estimator: Recap

- Denote the true mean model for the potential outcomes as $E(Y(z)|X) = m_z(X)$, for $z = 0, 1$
- Then specify regression models (with parameters, say, α) to estimate the true models $\hat{m}_z(X) = m_z(X; \hat{\alpha})$, where $\hat{\alpha}$ are the estimated α
- The regression ATE estimator is $N^{-1} \sum_{i=1}^{N} \{\hat{m}_1(X_i) - \hat{m}_0(X_i)\}$ or

$$\hat{\tau}_{adj} = N^{-1} \sum_{i=1}^{N} \{Z_i(Y_i - \hat{m}_0(X_i)) + (1 - Z_i)(\hat{m}_1(X_i) - Y_i)\}$$

- If the specified regression model is true, then $\hat{\tau}_{adj}$ is consistent and efficient, but not otherwise
- Can be sensitive when overlap is weak

Propensity Score Weighting Estimator: Recap

- Denote the true propensity score as $e(X)$
- Specify a model, e.g. logistic model, to estimate the propensity score (with parameters, say, β): $\hat{e}(X) = e(X; \hat{\beta})$, where $\hat{\beta}$ are the estimated parameters of β
- The IPW estimator of ATE is

$$\hat{\tau}_{ipw} = \frac{\sum_{i=1}^{N} Y_i Z_i / e(X_i)}{\sum_{i=1}^{N} Z_i / e(X_i)} - \frac{\sum_{i=1}^{N} Y_i (1 - Z_i)/(1 - e(X_i))}{\sum_{i=1}^{N} (1 - Z_i)/(1 - e(X_i))}$$

- If the postulated PS model is true, then $\hat{\tau}_{ipw}$ is consistent
- Can be inefficient
- Question: Is it possible to combine the virtues of regression and IPW estimator? DR estimator

Doubly Robust Estimator

- Easy to verify: with true $m_z(X)$ and $e(X)$

$$
\begin{aligned}
\tau &= \mathbb{E}\left\{\frac{ZY}{e(X)} - \frac{Z-e(X)}{e(X)} m_1(X)\right\} \\
&\quad - \mathbb{E}\left\{\frac{(1-Z)Y}{1-e(X)} + \frac{Z-e(X)}{1-e(X)} m_0(X)\right\} \\
&= \mathbb{E}\left[m_1(X_i) + \frac{Z_i\{Y_i - m_1(X_i)\}}{e(X_i)}\right] \\
&\quad - \mathbb{E}\left[m_0(X_i) + \frac{(1-Z_i)\{Y_i - m_0(X_i)\}}{1-e(X_i)}\right] \\
&= \mu_1 - \mu_0 = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]
\end{aligned}
$$

Doubly Robust Estimator

- In the previous formula, replace the true PS $e(x)$ and outcome $m_z(x)$ by the estimated ones from postulated models $\hat{e}(x)$ and $\hat{m}_z(x)$, we obtain two augmented estimators:

$$
\begin{aligned}
\hat{\tau}_{\text{dr}} &= \hat{\mu}_{1,\text{dr}} - \hat{\mu}_{0,\text{dr}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{Z_i - \hat{e}(X_i)}{\hat{e}(X_i)}\hat{m}_1(X_i)\right\} - \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} + \frac{Z_i - \hat{e}(X_i)}{1-\hat{e}(X_i)}\hat{m}_0(X_i)\right. \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{m}_1(X_i) + \frac{Z_i\{Y_i - \hat{m}_1(X_i)\}}{\hat{e}(X_i)}\right] - \frac{1}{N}\sum_{i=1}^{N}\left[\hat{m}_0(X_i) + \frac{(1-Z_i)\{Y_i - \hat{m}_0(X_i)\}}{1-\hat{e}(X_i)}\right.
\end{aligned}
$$

- The two estimators are mathematically equivalent, but different statistical implications: the first estimator augments an IPW estimator by outcome regression (OR); the second augments an OR estimator by IPW
- The first estimator is usually refereed to as the doubly-robust (DR) estimator

Doubly Robust Estimator

- Now focus on the DR estimator
- Notice that $\hat{\mu}_{1,\text{dr}}$ and $\hat{\mu}_{0,\text{dr}}$ have the same structure

$$\begin{aligned}
\hat{\mu}_{1,\text{dr}} &= N^{-1} \sum_{i=1}^{N} \left\{ \frac{Z_i Y_i}{\hat{e}_i} - \frac{(Z_i - \hat{e}_i)\hat{m}_1}{\hat{e}_i} \right\} \\
\hat{\mu}_{0,\text{dr}} &= N^{-1} \sum_{i=1}^{N} \left\{ \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} + \frac{(Z_i - \hat{e}_i)\hat{m}_0}{1 - \hat{e}_i} \right\} \\
&= N^{-1} \sum_{i=1}^{N} \left\{ \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} - \frac{\{(1 - Z_i) - (1 - \hat{e}_i)\}\hat{m}_0}{1 - \hat{e}_i} \right\}
\end{aligned}$$

- We will study the properties of $\hat{\mu}_{1,\text{dr}}$ as an estimator for $\mu_1 = \mathbb{E}[Y(1)]$; $\hat{\mu}_{0,\text{dr}}$ is symmetric

Doubly Robust Estimator

- What does $\hat{\mu}_{1,\text{dr}}$ estimate? Simple algebra shows that $\mu_{1,\text{dr}}$ converges to

$$\begin{aligned}
&\mathbb{E}\left\{\frac{ZY}{e(X)} - \frac{Z - e(X)}{e(X)} m_1(X)\right\} \\
&= \mathbb{E}\left\{Y(1) + \frac{Z - e(x)}{e(X)} Y(1) - \frac{Z - e(X)}{e(X)} m_1(X)\right\} \\
&= \mathbb{E}\{Y(1)\} + \mathbb{E}\left[\frac{\{Z - e(X)\}}{e(X)}\{Y(1) - m_1(X)\}\right] \qquad (1)
\end{aligned}$$

- Thus, in order for $\hat{\mu}_{1,\text{dr}}$ estimate $\mathbb{E}\{Y(1)\}$, the second term in (4) must be 0
- Regression estimator augmented by weighted residuals
- When does the second term = 0?

Doubly Robust Estimator

When does $R = \mathbb{E}\left[\frac{\{Z-e(X)\}}{e(X)}\{Y(1) - m_1(X)\}\right] = 0$?

- Scenario 1: Postulated propensity score model $e(X;\beta)$ is incorrect, but postulated regression model $m_1(X;\alpha)$ is correct

$$
\begin{aligned}
R &= \mathbb{E}\left[\mathbb{E}\left\{\frac{\{Z - e(X)\}}{e(X)}\{Y(1) - m_1(X)\} \mid X\right\}\right] \\
&= \mathbb{E}\left[\frac{\{e_{\text{true}}(X) - e(X)\}}{e(X)}\,\mathbb{E}\{Y(1) - m_1(X) \mid X\}\right]
\end{aligned}
$$

- Easy to see that $\mathbb{E}\{Y(1) - m_1(X) \mid X\} =$ $\mathbb{E}\{Y(1) \mid X\} - m_1(X) = m_{1,\text{true}}(X) - m_1(X) = 0$

Doubly Robust Estimator

When does $\mathbb{E}\left[\frac{\{Z-e(X)\}}{e(X)}\{Y(1)-m_1(X)\}\right]=0$?

- Scenario 2: Postulated propensity score model $e(X;\beta)$ is correct, but postulated regression model $m_1(X;\alpha)$ is incorrect

$$
\begin{aligned}
R &= \mathbb{E}\left[\mathbb{E}\left\{\frac{\{Z-e(X)\}}{e(X)}\{Y(1)-m_1(X)\} \mid X\right\}\right] \\
&= \mathbb{E}\left[\frac{\{e_{\text{true}}(X)-e(X)\}}{e(X)}\,\mathbb{E}\{Y(1)-m_1(X) \mid X\}\right]
\end{aligned}
$$

- Equals to zero because $e_{\text{true}}(X)-e(X)=0$

Doubly Robust Estimator

- In both cases, the second term goes to 0 in large samples, and thus $\hat{\mu}_{1,\text{dr}}$ is consistent (asymptotically unbiased) for $E(Y(1))$.
- Similarly, $\hat{\mu}_{0,\text{dr}}$ is consistent for $E(Y(0))$, and hence $\hat{\mu}_{\text{dr}}$ is consistent for the ATE.
- Obviously, if both models are correct, $\hat{\mu}_{\text{dr}}$ is consistent for estimating the ATE.

Doubly Robust Estimator

Double Robustness: $\hat{\tau}_{\text{dr}}$ is a consistent estimator of the ATE if either the propensity score model or the potential outcome model is, but not necessary both are, correctly specified

- DR is a large sample property
- Offers protection against model mis-specification: give you two chances to get it right (and wrong)!
- If $e(X)$ and $m_z(X)$ are modeled correctly, $\hat{\tau}_{\text{dr}}$ will have smaller variance than the IPW estimator (in large samples)
- If the outcome model $m_z(X)$ is correct, $\hat{\tau}_{\text{dr}}$ has larger variance (in large samples) than the direct regression estimator
- ... but gives protection in the event it is not

Double robustness, double jeopardy?

- Finite sample performance of DR estimator critically depends on the degree of overlap
- Poor overlap: inherit the problem of IPW with extreme weights, and amplify the residuals from the outcome model
- With poor overlap, DR with non-trimmed IPW is usually worse than the outcome model estimator (see simulations in the next page)
- With poor overlap, one should use DR with trimmed weights or use OW (no augmentation)

Simulations: finite sample performance

- Simulation setting (Li et al. 2019)
  - 3 continuous and 3 binary covariates X
  - True propensity score: $\text{logit}(e(X)) = \alpha X'$
    - Set $\alpha = d(0.2, 0.3, 0.4, -0.25, -0.3, -0.3)$, $d = 1$ good overlap, $d = 3$ poor overlap
    - Choose the intercept α_0 to control the proportion of treated $\Pr(Z = 1)$: 0.4 (more balanced design), 0.1 (less balanced)
  - True outcome model (linear): $Y \sim \texttt{lm}(\beta X' + \tau Z)$, $\tau = 1$, homogeneous treatment effect, with $\beta = (-0.5, -0.8, -1.2, 0.8, 0.8, 1)$
  - Different model misspecfication
- Apply IPW and DR with or without trimming, OW. Report mean absolute bias (MAB) and root mean squared error (RMSE)

Simulation: role of overlap in DR

Table. RMSE and MAB for estimators with correct model specifications under various degree of overlap.

Estimator	Treatment prevalence = 0.4				Treatment prevalence = 0.1			
	$d = 1$		$d = 3$		$d = 1$		$d = 3$	
	RMSE	**MAB**	**RMSE**	**MAB**	**RMSE**	**MAB**	**RMSE**	**MAB**
Outcome regression	0.05	0.04	0.07	0.06	0.11	0.08	0.22	0.17
IPW								
No trimming	0.09	0.07	0.83	0.60	0.31	0.24	1.46	1.26
$\alpha = 0.05$	0.08	0.06	0.10	0.08	0.12	0.10	0.13	0.11
$\alpha = 0.1$	0.07	0.05	0.08	0.07	0.11	0.09	0.12	0.10
Doubly-robust								
No trimming	0.06	0.04	0.31	0.14	0.11	0.09	0.62	0.33
$\alpha = 0.05$	0.05	0.04	0.08	0.06	0.10	0.08	0.12	0.10
$\alpha = 0.1$	0.05	0.04	0.08	0.06	0.10	0.08	0.12	0.10
Overlap weighting	0.05	0.04	0.07	0.06	0.08	0.06	0.10	0.08

- Under poor overlap, even with correct specification of both PS and outcome models, DR without trimmed weight perform worse than outcome model or OW

Simulation: role of overlap in DR

Table 3. RMSE and MAB for the DR estimator with one of the two models being mis-specified.

Mis-specified model	Treatment prevalence = 0.4				Treatment prevalence = 0.1			
	$d = 1$		$d = 3$		$d = 1$		$d = 3$	
	RMSE	**MAB**	**RMSE**	**MAB**	**RMSE**	**MAB**	**RMSE**	**MAB**
Omit variables								
PS model	0.05	0.04	0.11	0.09	0.11	0.09	0.37	0.25
Outcome model	0.06	0.05	0.70	0.30	0.15	0.12	1.74	0.81
Wrong polynomial								
PS model	0.05	0.04	0.12	0.09	0.11	0.09	0.40	0.27
Outcome model	0.07	0.06	1.18	0.50	0.20	0.16	2.05	0.87

- Outcome model is much more dominant than PS model in causal estimates
- Under poor overlap, correct specification of the PS model does little to correct the bias from a misspecified outcome model

An Efficiency Perspective

If $e(X)$ and $m_z(X)$ are modeled correctly, $\hat{\tau}_{\text{dr}}$ will have smaller variance than the IPW estimator (in large samples)

- Related to semi-parametric theory (Tsiatis, 2007)
- Again, for estimating μ_1 (mirror image for estimating μ_0)
- Intuitively, write $O_i = (Y_i, Z_i, X_i)$, the IPW estimator for μ_1 is based on solving $\sum_{i=1}^{N} h_{\text{ipw}}(O_i; \mu_1) = 0$, where

$$h_{\text{ipw}}(O; \mu_1) = \frac{ZY}{e(X)} - \mu_1$$

- Call $h_{\text{ipw}}(O; \mu_1)$ the estimating function of the IPW estimator (technically the influence function, but we will not use that term here)

An Efficiency Perspective

- Can characterize the class of all estimating functions for "regular" estimators for μ_1 by

$$\mathcal{H} = \left\{ h \mid h(O;\mu_1) = \frac{ZY}{e(X)} - \mu_1 + \Lambda \right\}$$

  where $\Lambda = \{Z - e(X)\}L(X)$ for arbitrary X

- IPW is a special case by setting $h(X) = 0$

- The most efficient (corresponding to estimator with smallest asymptotic variance) is identified by setting $L(X) = m_1(X) = \mathbb{E}[Y|Z = 1, X]$

An Efficiency Perspective

- This is exactly the estimating function for DR, or frequently referred to as the “augmented IPW" (AIPW) estimator

$$h_{\text{dr}}(O;\mu_1) = \frac{ZY}{e(X)} - \frac{Z-e(X)}{e(X)} m_1(X) - \mu_1$$

- In other words, $\hat{\mu}_{1,\text{dr}}$ solves $\sum_{i=1}^{N} h_{\text{dr}}(O_i;\hat{\mu}_{1,\text{dr}}) = 0$
- Because DR is based on the optimal estimating function, it improves the large-sample efficiency of IPW

DR: Semiparametric Efficiency

- The DR estimator is semiparametric in the sense that it does not require correct specification of the entire data-generating mechanism
- Robins, Rotnitzky, Zhao (1994, JASA) proves the DR estimator is the (locally) semiparametric efficient estimator—i.e. having the smallest asymptotic variance—within the class of inversely weighted estimators
- Robins et al. proposed the DR estimator in the context of missing (or incomplete) data. Causal inference can be viewed as a special case, where we want to infer the full data $(Y(1), Y(0), X)$ for each group from the observed, incomplete data (Y, X, Z).

DR estimator: Variance

- Lunceford and Davidian (2004) provides an estimator to approximate the variance of $\hat{\tau}_{\text{dr}}$:

$$s^2_{\text{dr}} = \sum_i (\hat{\tau}_i - \hat{\tau}_{\text{dr}})^2 / N^2, \tag{2}$$

where

$$\begin{aligned} \hat{\tau}_i \quad = \quad & \left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(Z_i - \hat{e}_i)\hat{m}_1(\mathbf{X}_i)}{\hat{e}_i} \right] \\ & - \left[\frac{(1 - Z_i) Y_i}{(1 - \hat{e}_i)} + \frac{(Z_i - \hat{e}_i)\hat{m}_0(\mathbf{X}_i)}{(1 - \hat{e}_i)} \right] \end{aligned}$$

- Does not take into account of the variability in estimating $e(X)$ and $m_z(X)$, and so may be biased in small samples
- The above formula works well in simulations with large samples; one can also use bootstrap to estimate the variance

DR estimator: Sandwich Variance via M-estimation

- A more accurate variance estimator proceeds by the M-estimation theory and is given by the sandwich variance
- Notice that $\hat{\tau}_{\text{dr}} = \hat{\nu}_1 - \hat{\nu}_0$, which jointly solve

$$\sum_{i=1}^{N} \Psi_i(\theta) = \sum_{i=1}^{N} \begin{bmatrix} \nu_1 - \{Z_i Y_i - (Z_i - e_i)\mu_1(X_i;\alpha_1)\}/e_i \\ \nu_0 - \{(1-Z_i)Y_i + (Z_i - e_i)\mu_0(X_i;\alpha_0)\}/(1-e_i) \\ Z_i S_1(Y_i, X_i;\alpha_1) \\ (1-Z_i) S_0(Y_i, X_i;\alpha_0) \\ S_\beta(X_i;\beta) \end{bmatrix} = 0$$

  where S_1, S_0, S_β are score function of the outcome models and PS model, $\theta = (\nu_1, \nu_0, \alpha_0', \alpha_1', \beta')'$
- If $\widehat{Var}(\hat{\theta})$ is obtained, can use Delta method to compute $\widehat{Var}(\hat{\tau}_{\text{dr}})$

DR estimator: Sandwich Variance via M-estimation

- Notice that $\hat{\tau}_{\text{dr}} = \hat{\nu}_1 - \hat{\nu}_0$, which jointly solve $\sum_{i=1}^{N} \Psi_i(\hat{\theta}) = 0$
- First-order Taylor expansion

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \Psi_i(\hat{\theta}) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \Psi_i(\theta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \frac{\partial}{\partial \theta'} \Psi_i(\theta_0)(\hat{\theta} - \theta_0)$$

- Moving terms

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \left\{ -\frac{1}{n} \sum_{i=1}^{N} \frac{\partial}{\partial \theta'} \Psi_i(\theta_0) \right\}^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \Psi_i(\theta_0) \right\}$$

- From CLT and Slutsky's Theorem (sandwich variance)

$$\widehat{Var}(\hat{\theta}) = \left\{ \sum_{i=1}^{N} \frac{\partial}{\partial \theta'} \Psi_i(\hat{\theta}) \right\}^{-1} \left\{ \sum_{i=1}^{N} \Psi_i^{\otimes 2}(\hat{\theta}) \right\} \left\{ \sum_{i=1}^{N} \frac{\partial}{\partial \theta} \Psi_i'(\hat{\theta}) \right\}^{-1}$$

DR: Cautionary Remarks

- Empirical evidence (including my own experience) suggests outcome model plays a more prominent role than the PS model:
  - Misspecification of the outcome model leads to much larger bias than misspecification of the PS model
  - Even if the PS model is correct, a misspecified outcome model usually lead to notable bias
- Empirically, the extra protection from the propensity score augmentation is somewhat limited
- Kang and Schafer (2007, Stat Sci) gives an example where moderate misspecification of both PS and outcome model (very likely scenario in practice) leads to large bias and variance of DR

DR: Alternative Construction

- But the form of DR estimator is not unique, and there are other constructions that lead to better finite sample performance
  - the previous estimator is a moment construction, sometimes called a “one-step" or plug-in estimator
  - Regression on full covariates and inverse propensity score as a “clever covariate" (Bang and Robins, 2005)
  - Weighted regression (Robins et al. 2007)
  - Calibrated weighting
  - . . .
- Same large-sample behaviour (DR and efficiency), but different finite-sample performance

Regression on Clever Covariate

Bang and Robins, 2005

- Consider estimating μ_1, we fit the GLM with canonical link, adding a clever covariate to the OR estimator

$$g\{\mathbb{E}[Y|X, Z = 1]\} = m_1(X; \alpha_1) + \phi_1 e^{-1}(X; \hat{\beta})$$

- Write $s_1(X_i; \hat{\alpha}_1, \hat{\phi}_1) = g^{-1}\{m_1(X; \hat{\alpha}_1) + \hat{\phi}_1 e^{-1}(X; \hat{\beta})\}$, the DR estimator is given by

$$\begin{aligned}\hat{\mu}_1 &= \sum_{i=1}^{N} s_1(X_i; \hat{\alpha}_1, \hat{\phi}_1) \\ &= \sum_{i=1}^{N} s_1(X_i; \hat{\alpha}_1, \hat{\phi}_1) + \sum_{i=1}^{N} \frac{Z_i}{e(X_i; \hat{\beta})}\{Y_i - s_1(X_i; \hat{\alpha}_1, \hat{\phi}_1)\}\end{aligned}$$

- This second term is precisely the score equations of the GLM and equals to 0

Regression on Clever Covariate

- Specifically, the estimators $\hat{\alpha}_1$ and $\hat{\phi}_1$ jointly solve the score equation

$$0 = \sum_{i=1}^{N} \frac{Z_i}{e(X_i;\hat{\beta})}\{Y_i - s_1(X_i;\hat{\alpha}_1,\hat{\phi}_1)\}$$

- Then $\hat{\mu}_1$ has the same form of the AIPW estimator using moment construction; consistency requires $m_1(X)$ or $e(X)$ to be correct
- Same steps can be repeated to estimate μ_0, but the "clever covariate" now becomes $\{1 - e(X;\hat{\beta})\}^{-1}$
- Difference between regression on PS
  - clever covariate is inverse PS, is connected with DR, and includes a single term for adjustment
  - regression on PS is on PS itself, motivated by bias removal, and includes flexible functional forms of $e(X)$

Weighted Regression

Robins et al, 2007

- Assume OR model $g\{\mathbb{E}[Y|X, Z = 1]\} = m_1(X; \alpha_1)$, but use weighted regression with inverse probability weights $e^{-1}(X; \hat{\beta})$
- Write $s_1(X_i; \hat{\alpha}_1) = g^{-1}\{m_1(X; \hat{\alpha}_1)\}$, obtain an OR estimator

$$\hat{\mu}_1 = \sum_{i=1}^{N} s_1(X_i; \hat{\alpha}_1) = \sum_{i=1}^{N} s_1(X_i; \hat{\alpha}_1) + \sum_{i=1}^{N} \frac{Z_i}{e(X_i; \hat{\beta})}\{Y_i - s_1(X_i; \hat{\alpha}_1)\}$$

- Again $\hat{\alpha}_1$ solves the score equation

$$0 = \sum_{i=1}^{N} \frac{Z_i}{e(X_i; \hat{\beta})}\{Y_i - s_1(X_i; \hat{\alpha}_1)\}$$

- Weighted regression estimator $\hat{\mu}_1$ is also DR

DR: Connection to Sample Survey

- In causal inference literature, the DR property was first pointed out by Scharfstein et al. (1999)
- But the DR type of estimators have been proposed and used in survey literature pre-dating causal inference
- In that context, DR is usually interpreted as IPW estimator augmented by a regression (AIPW)
- Have already shown

$$\tau = \mathbb{E}\left[m_1(X_i) + \frac{Z_i\{Y_i - m_1(X_i)\}}{e(X_i)}\right] - \mathbb{E}\left[m_0(X_i) - \frac{(1-Z_i)\{Y_i - m_0(X_i)\}}{1-e(X_i)}\right]$$

- Therefore, one can also interpret DR as a regression estimator augmented by IPW

DR: Connection to Sample Survey

- In survey, sampling probabilities–analogy of propensity scores $e(X)$–and hence survey weights are usually known for every unit
- Cassel et al. (1976, Biometrika) proposed essentially the same DR estimator (with known $e(X)$) for estimating the mean of Y in a finite population where X is measure for every unit
- The goal is to gain efficiency by using an outcome model but augment it by the survey weights to protect against the bias if the outcome model is misspecified – "design consistency" (analogous to DR)
- Here the proposed outcome model $m_z(X;\alpha) = \alpha X$ and $\hat{\alpha}$ is either the LS complete-case estimator or the weighted LS complete-case estimator with weights $1/e(X)$

DR as a Diagnostic Tool

- DR as a diagnostic tool (Robins and Rotnitzky, 2001)– compare DR with IPW and direct regression estimates of the same data
  - If DR is close to IPW but not REG (OR), then the outcome model is likely wrong
  - If DR is close to REG (OR) but not IPW, then the PS model is likely wrong
  - If DR is far away from both IPW and REG (OR), trouble! Change models, try more flexible ones
- In theory, one can even formalize a test for “similarity" between IPW and DR, or DR and OR, but little empirical investigations on the performance of the test (Robins and Rotnitzky, 2001)
- An empirical example is given in Mercatanti and Li (2014)

DR Estimator for ATT

- The DR concept applies beyond IPW (ATE), e.g. to ATT
- An DR estimator for ATT (Mercatanti and Li, 2014; Moodie et al. 2018)

$$\hat{\tau}_{\text{dr}}^{\text{ATT}} = \sum_{i=1}^{N} \left[Y_i Z_i - \frac{Y_i(1 - Z_i)\hat{e}_i + \hat{m}_0(\mathbf{X}_i)(Z_i - \hat{e}_i)}{1 - \hat{e}_i} \right] \Big/ N_1,$$

- DR for ATT only requires modeling $Y(0)$, or $Y|Z = 0 \sim X$, in contrast to DR for ATE
- This is because $E[Y(1)|Z = 1]$ can be determined completely non-parametrically, and do not require any modeling
- M-estimation variance estimator (Moodie et al. 2018, Stat): sandwich variance estimator that accounts for variability in estimating PS and outcome model
- Of course, one can always use bootstrap

DR for Balancing Weights?

- IPW (ATE) and ATT are members of the class of balancing weights (Li, Morgan, Zaslavksy, 2018). Both still involve inverse PS weighting, and may have excessive variance
- Do we have DR with other balancing weights, e.g. overlap weights?
- No DR for overlap weights – because the ATO estimand involves $e(X)$: if $e(X)$ is misspecified, ATO is problematic; nonetheless, one can still use augmentation to improve efficiency
- Find the efficient estimating functions for *weighted average treatment effect* (WATE; Hirano et al. 2003)

DR for Balancing Weights?

Mao et al. 2019

- Consider the tilting function $h(X)$ that defines the target population as $h(X)f(X)$
- Recall the balancing weights are defined as :

$$\begin{cases} w_1(x) \propto \frac{f(x)h(x)}{f_1(x)} = \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)}, \\ w_0(x) \propto \frac{f(x)h(x)}{f_0(x)} = \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{1-e(x)}. \end{cases} \tag{3}$$

- An augmented weighting estimator is

$$\begin{aligned} \hat{\tau}^h =& \frac{\sum_{i=1}^N h(X_i)\{\hat{m}_1(X_i) - \hat{m}_0(X_i)\}}{\sum_{i=1}^N h(X_i)} + \frac{\sum_{i=1}^N w_1(X_i)Z_i\{Y_i - \hat{m}_1(X_i)\}}{\sum_{i=1}^N w_1(X_i)Z_i} \\ & - \frac{\sum_{i=1}^N w_0(X_i)(1-Z_i)\{Y_i - \hat{m}_0(X_i)\}}{\sum_{i=1}^N w_0(X_i)(1-Z_i)} \end{aligned}$$

DR for Balancing Weights?

Mao et al. 2019

- Regression estimator augmented by weighted residuals

$$\hat{\tau}^h = \frac{\sum_{i=1}^N h(X_i)\{\hat{m}_1(X_i) - \hat{m}_0(X_i)\}}{\sum_{i=1}^N h(X_i)} + \frac{\sum_{i=1}^N w_1(X_i)Z_i\{Y_i - \hat{m}_1(X_i)\}}{\sum_{i=1}^N w_1(X_i)Z_i} - \frac{\sum_{i=1}^N w_0(X_i)(1-Z_i)\{Y_i - \hat{m}_0(X_i)\}}{\sum_{i=1}^N w_0(X_i)(1-Z_i)}$$

- Set $h(X) = 1$ and $w_1(X) = 1/e(X)$, $w_0(X) = 1/(1-e(X))$, and this estimator is the DR estimator for ATE

- But can choose $h(X)$ as a function of PS and use this augmented estimator for alternative causal estimands, such as ATO

- Addresses the issue of lack of overlap, and further improves the efficiency of OW estimator by exploiting outcome regression

DR for Balancing Weights

Mao et al. 2019

- $\hat{\mu}_1$ in the augmented estimator converges to

$$\frac{\mathbb{E}[h(X)Y(1)]}{\mathbb{E}[h(X)]} + \{\mathbb{E}[h(X)]\}^{-1}\,\mathbb{E}\left[\frac{h(X)\{Z - e(X)\}}{e(X)}\{Y(1) - m_1(X)\}\right]$$

- Thus, in order for $\hat{\mu}_1$ estimate $\mathbb{E}[h(X)Y(1)]/\mathbb{E}[h(X)]$, $h(X)$ has to be "correctly" defined, and the second term in (4) must be 0

- $h(X) = e(X)\{1 - e(X)\}$ for overlap weights, and requires correct model for $e(X)$

- With correct $e(X)$, misspecification of $m_1(X)$ still leads to a consistent estimator (singly robust)

- Misspecified $e(X)$ still leads to a consistent estimator for a well-defined WATE, $\mathbb{E}[h^*(X)Y(1)]/\mathbb{E}[h^*(X)]$

Balancing Weighting under PS misspecification

Zhou, Matsouaka, Thomas, 2020:

- Zhou et al. provide theoretical and simulation evidence that overlap weighting is less sensitive to misspecification of PS than ATT and IPW weights
- The analytical form of the asymptotic bias of the Hajek estimator of WATE when PS is misspecified (denoting the misspecified PS model as $\tilde{e}$):

$$\text{Abias}(\hat{\tau}_h) = \frac{\frac{e(x)}{\tilde{e}(x)}\tilde{h}(x)m_1(x)}{\frac{e(x)}{\tilde{e}(x)}\tilde{h}(x)} - \frac{\frac{1-e(x)}{1-\tilde{e}(x)}\tilde{h}(x)m_0(x)}{\frac{1-e(x)}{1-\tilde{e}(x)}\tilde{h}(x)} - \tau_h$$

- Clearly with inverse weights (e.g. IPW, ATT), PS close to 0 and 1 will exacerbate the asymptotic bias, but not so for OW or similar weights focusing on the middle (e.g. matching weights)

DR Beyond Weighting

- The essence of DR estimators is to combine outcome regression (OR) and propensity score (PS) models
- There are other forms
  - Bias corrected matching (Abadie and Imbens, 2011)
  - OR and stratification (Gutman and Rubin, 2015)
  - Regression with PS as an additional covariate (Rod Little et al.; Zhou, Elliott, Little, 2019): the outcome model is the sum of (1)a nonparametric function (e.g. penalized splines) of PS, and (2) a parametric function of the propensity score and covariates
  - Bayesian version (Hahn et al. 2020)
- The DR property is not exact as the augmented IPW estimator, often weaker. Essence is the same

Practical Guide

- Always check overlap and balance first
- In general, combining a nonparametric PS method (weighting, matching, stratification) and an outcome model is better than either of the two separate methods
- Outcome model is always the key: flexible models (e.g., semiparametric ones via splines, power series, machine learning methods) are preferable to the simple linear regression
- Among all the DR (or double) methods, which one is the best? Depends, no panacea

Realistic Simulations

- Well-designed simulations can shed good insights
- Difficult to generalize conclusions from arbitrary simulations
- Comparative performance between different methods might differ greatly in different data set
- Real-data-based simulations are usually more relevant for any specific application
- Main point: retain the data structure, covariates and treatment variable, but simulate the outcome-generating mechanism, loosely based on the outcome model in mind

Case Study: Effects of Debit Card on Spending

- Mercatanti and Li (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Annals of Applied Statistics*. 8(4), 2405-2508.
- Goal: estimate the effect of possessing debit cards on household spending (ATT)
- Data: Italy Survey on Household Income and Wealth (SHIW)
- Unit: households; treatment Z: possess at least one debit card; outcome Y: monthly consumption; covariates X: social economic statuses, banking info, geographical...
- Linear models likely not fit well; use flexible semiparametric models via power series, four different estimators: regression, matching with regression, IPW, DR

Case Study: Effects of Debit Card on Spending

Estimated PATT in thousands of Italian Lira (standard errors in parenthesis).

		$\hat{\tau}_{reg}$		$\hat{\tau}_{wt}$		$\hat{\tau}_{mix}$		$\hat{\tau}_{dr}$	
span	AOT	PATT	$\frac{\text{PATT}}{\text{AOT}}$	PATT	$\frac{\text{PATT}}{\text{AOT}}$	PATT	$\frac{\text{PATT}}{\text{AOT}}$	PATT	$\frac{\text{PATT}}{\text{AOT}}$
1993-95	2092.9	90.2	0.043	102.3	0.049	100.6	0.048	97.2	0.046
		(41.8)		(47.1)		(50.4)		(42.7)	
1995-98	2027.6	199.1	0.098	160.7	0.079	208.7	0.103	202.2	0.100
		(87.6)		(73.4)		(69.8)		(93.2)	
1998-00	2116.4	148.1	0.069	137.7	0.065	122.8	0.058	142.1	0.067
		(68.5)		(73.1)		(60.7)		(70.5)	

- AOT: average outcome of treated / PATT: population ATT
- *reg*: outcome regression; *wt*: IPW
- *mix*: bias-corrected matching; *dr*: doubly robust estimator

References

Bang, H., Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962-973.

Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., Davidian, M. (2011). Doubly robust estimation of causal effects. American journal of epidemiology, 173(7), 761-767.

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

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427), 846-866.

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.

Kang, J. D., Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science, 22(4), 523-539.

References

Tan, Z. (2007). Comment: Understanding OR, PS and DR. Statistical Science, 22(4), 560-568.

Robins, J., Sued, M., Lei-Gomez, Q., Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when “inverse probability" weights are highly variable. Statistical Science, 22(4), 544-559.

Cassel, C. M., Sarndal, C. E., Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63(3), 615-620.

Seaman, S. R., Vansteelandt, S. (2018). Introduction to double robust methods for incomplete data. Statistical science. 33(2), 184.

Scharfstein, D. O., Rotnitzky, A., Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448), 1096-1120.

Robins, J. M., Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, “Inference for semiparametric models: Some questions and an answer". Statistica Sinica, 11(4), 920-936.

Hirano, K., Imbens, G. W., Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4), 1161-1189.

References

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. Statistica Sinica, 2, 949-968.

Zhou, T., Elliott, M. R., and Little, R. J. (2019). Penalized Spline of Propensity Methods for Treatment Comparison. Journal of the American Statistical Association, 114(525):1-19.

Zhou, Y., Matsouaka, R. A., Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. arXiv:2006.04038.

Hahn, PR, Murray J, and Carvalho C. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. Bayesian Analysis.

Mercatanti, A, and Li F. (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Annals of Applied Statistics*. 8(4), 2405-2508.

Mao, H., Li, L., Greene, T. (2019). Propensity score weighting analysis and treatment effect discovery. Statistical methods in medical research, 28(8), 2439-2454.

Moodie, E.E., Saarela, O. and Stephens, D.A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. Stat, 7(1), p.e205.

References

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189(1), 1-23.

Belloni, A., Chernozhukov, V. (2011). $l1$-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics, 39(1), 82-130.

Belloni, A., Chernozhukov, V., Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29-50.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1-C68.

van Der Laan, M. J., Rubin, D. (2006). Targeted maximum likelihood learning. The international journal of biostatistics, 2(1).

van der Laan, M. J., Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. The international journal of biostatistics, 6(1).