DECART Summer School 2018:

Causal Inference Module

Special Topics

Topic 1: Doubly Robust Estimator

Problem Setup

- Consider inverse propensity weighted (IPW) estimator for the mean of $Y_i(1)$ in the study population $\mu_t = E[Y(1)]$

  as discussed in the earlier session, we have

  $$\hat{\mu}_t = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{e}(X_i)}$$

  where $\hat{e}(X_i)$ is the propensity score estimated using logistic regression, we will write this as $\hat{e}(X_i) = \mathrm{e}\left(X_i; \hat{\beta}\right)$ to reflect the fact that this is a parametric model.
- Why does this work?
  - By the law of large numbers, this should estimate the mean of a term in the sum with $\hat{\beta}$ replaced by the quantity it estimates.

Consistency of IPW Estimator

- If $e(X;\beta) = e(X)$, the true propensity score

$$E\left[\frac{AY}{e(X)}\right] = E\left[\frac{AY(1)}{e(X)}\right] = E\left[E\left\{\frac{AY(1)}{e(X)}|Y(1), X\right\}\right]$$

$$= E\left\{\frac{Y(1)}{e(X)}E(A|Y(1), X)\right\} = E\left\{\frac{Y(1)}{e(X)}E(A|X)\right\}$$

$$= E\left\{\frac{Y(1)}{e(X)}e(X)\right\} = E(Y(1))$$

Consistency of IPW Estimator

- If $e(X;\beta) = e(X)$, the true propensity score

$$E\left[\frac{AY}{e(X)}\right] = E\left[\frac{AY(1)}{e(X)}\right] = E\left[E\left\{\frac{AY(1)}{e(X)}|Y(1),X\right\}\right]$$

$$= E\left\{\frac{Y(1)}{e(X)}E(A|Y(1),X)\right\} = E\left\{\frac{Y(1)}{e(X)}E(A|X)\right\}$$

$$= E\left\{\frac{Y(1)}{e(X)}e(X)\right\} = E(Y(1))$$

It is worth noting:

i) The consistency depends on the fact the model used to estimate $e(X)$ is correctly specified.

ii) The estimator only makes use of outcome data with A=1, ignores the information from subjects with A=0

Improve efficiency through data augmentation -- AIPW Estimator

- Modified estimator (Augmented Inverse Propensity Weighted estimator):

$$\hat{\mu}_t = n^{-1}\sum_{i=1}^{n}\left[\frac{A_i}{e(X_i;\hat{\beta})}Y_i - \frac{\{A_i - e(X_i;\hat{\beta})\}}{e(X_i;\hat{\beta})}m_t(X_i;\hat{\alpha})\right]$$

> $e(X;\beta)$ is a postulated model for the true propensity score $e(X) = E(A|X)$ (fitted by logistic regression)

> $m_t(X;\alpha)$ is postulated model for the true regression $E(Y|A=1,X)$ (fitted by least square)

By the law of large numbers, this should estimate the mean of a term in the sum with $\hat{\beta}$ and $\hat{\alpha}$ replaced by the quantity they estimate.

Double Robustness

$$E\left[\frac{A}{e(X;\beta)}Y - \frac{\{A-e(X;\beta)\}}{e(X;\beta)}m_t(X;\alpha)\right]$$
$$= E\left[\frac{A}{e(X;\beta)}Y(1) - \frac{\{A-e(X;\beta)\}}{e(X;\beta)}m_t(X;\alpha)\right]$$
$$= E\left[Y(1) + \frac{\{A-e(X;\beta)\}}{e(X;\beta)}\{Y(1) - m_t(X;\alpha)\}\right]$$
$$= E[Y(1)] + E\left[\frac{\{A-e(X;\beta)\}}{e(X;\beta)}\{Y(1) - m_t(X;\alpha)\}\right]$$

the second term $=\mathrm{E}\left\{\{Y(1) - m_t(X;\alpha)\}E\left[\frac{\{A-e(X;\beta)\}}{e(X;\beta)}|Y(1),X\right]\right\}$

$$=\mathrm{E}\left\{\frac{\{A-e(X;\beta)\}}{e(X;\beta)}E[\{Y(1) - m_t(X;\alpha)\}|A,X]\right\}$$

- If propensity model is correctly specified, $e(X;\beta) = E(A|X)$, then the second term = 0
- If outcome model is correctly specified, $m_t(X;\alpha) = E(Y(1)|X)$, then the second term = 0

Double Robustness

- When either one model is correctly specified, we obtain an unbiased estimator.

- When both models are correctly specified, the resulting estimator is not only unbiased but also more efficient.(incorporate more information)

- Offers protection against mismodeling.

Topic 2: Time Varying Confounding and Marginal Structural Model

Marginal Structural Models

- If the treatment can be quantified on at least an interval scale, we may consider models of the form:

  (MSM1) $E[Y(a)] = \beta_0 + \beta_1 a$, or

  (MSM2) $E[Y(a)] = \beta_0 + \beta_1 a + \beta_2 a^2$, or

  (MSM3) $E[Y(a)] = f(a)$ for some functional form $f(\cdot)$
- Under (MSM1), ATE contrasting treatment 1 to treatment 0 is β_1 (essentially what we did in ipw example)
- Under (MSM3), $ATE = f(1) - f(0)$
- These are called marginal structural models.
- Structural because the models are based on the counterfactual outcomes $Y(a)$
- Marginal because the models are based on the marginal distributions of each $Y(a)$

Marginal Structural Models with Effect Modification

- V = baseline factors
- Models of the form $E[Y(a)|V] = f(a, V)$ can be used to model modification of the causal effect of A by the factors in V
- For example,

$$E[(Y(a))|V] = \beta_0 + \beta_1 a + \beta_2 V + \beta_3 a \times V$$

Causal effect is then $\beta_1 + \beta_3 V$

- Estimate model parameters by fitting regression model

$$E(Y|V, A) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 A \times V$$

using weighted regression with weights W^A or SW^A

- Consider stabilized weights as $SW^A(V) = f(A|V)/f(A|L, V)$

Multiple Levels of Treatment

- Assume treatment A has k levels, $a = 1, 2, \ldots, k$
- Could use **multinomial logistic regression** to estimate $f(A|L) = \Pr(A|L)$ for each $A = a$.
- Then define inverse probability of treatment weights for treatment A as:

$$W^A = 1/f(A|L)$$

- Stabilized weights: $SW^A = f(A)/f(A|L)$
- Then use weighted regression to estimate parameters of a marginal structural model for the effect of the treatment;
- e.g. $E[Y(a)] = \beta_0 + \beta_1 a$, estimate β_0 and β_1 based on weighted regression of Y on A using weights W^A or SW^A

Evaluating the Causal Effect

- For linear MSMs, this is equivalent to a 2-step procedure where we first obtain $\hat{E}(Y(a))$ as $\frac{\sum_i 1_{[Ai=a]} Y_i}{\sum_i W_i^a}$ for each a, and then regress the $\hat{E}(Y(a))$ on a.

- **Problem:** Some treatment levels may be much more common than others, but the IPW weights give equal overall weight to each value of A in the regression

- Solution is to use stabilized weights: $SW^A = f(A)/f(A|X)$
- The stabilized weights give more weight to treatment values a which are more common in the dataset

Continuous Treatment

- When the treatment is continuous (e.g. dosage):

  $$\Pr(A = a|X) = 0 \text{ for all } A \text{ and } X.$$

  cannot use standard propensity weighting approach.
- stabilized weights are OK: $SW^A = f(A)/f(A|X)$ where $f(A)$ and $f(A|X)$ now represents the density of A and the conditional density of A given X.
- To estimate $f(A|X)$, one regresses A on X, heavily dependent on the assumed conditional distribution of the error term, some choices in literature:

  i) normal distribution

  ii) truncated normal distribution

  iii) t distribution

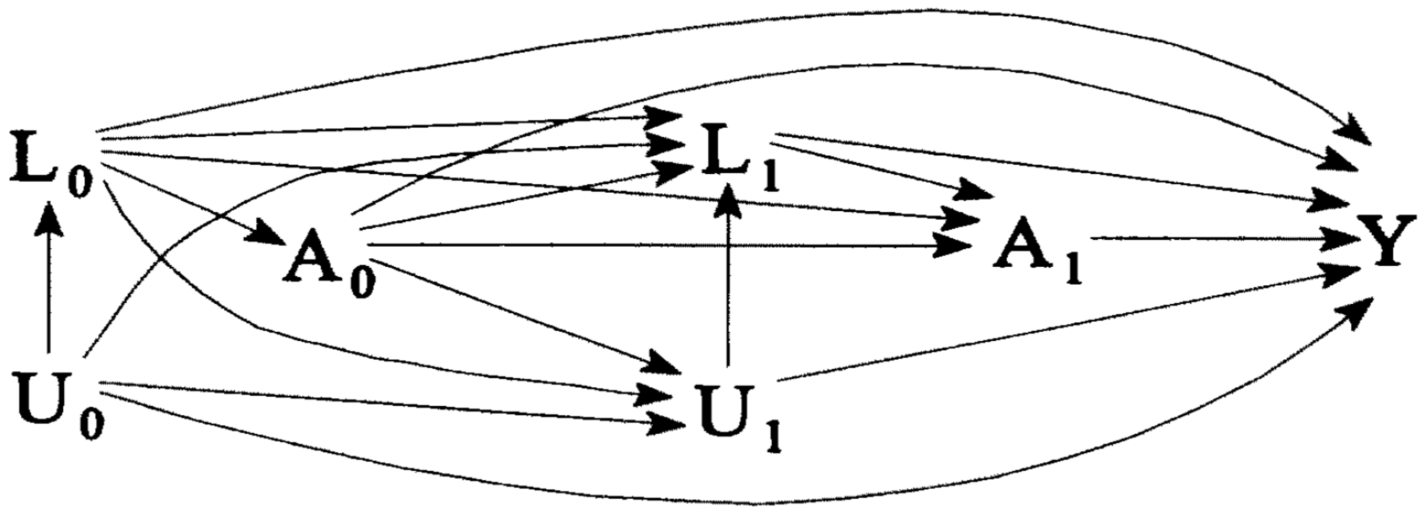
  vi) quantile binning

Time-Varying Treatment: Notation

- Consider a study with K followup visits, indexed by $k = 0, 1, \ldots, K$
- A_k: treatment at kth visit
- L_k: covariates measured at kth visit
- We assume the outcome Y is evaluated at visit $K + 1$
- $\bar{A}_k = (A_0, A_1, \ldots, A_k)$: the treatment history through the kth visit
- $\bar{L}_k = (L_1, L_2, \ldots, L_k)$: the covariate history through the kth visit
- $Y(\bar{a}_K) = Y(a_1, a_2, \ldots, a_K)$: the counterfactual outcome under the treatment history $a_1, a_2, \ldots, a_{K.}$

HEALTH UNIVERSITY OF UTAH

Time Varying Treatment Framework with No Unmeasured Confounders



NUCA for this case: $Y(\bar{a}_K) \perp A_k \mid \bar{L}_k, \bar{A}_{k-1}$

Marginal Structural Models for Time Dependent Treatments

- For continuous Y: $E\left(Y(\bar{a}_K)\right) = f(\bar{a}_K, V)$ If $a_k = 0\ or\ 1$, a simple MSM is

$E\left(Y(\bar{a}_K)\right) = \beta_0 + \beta_1 cum(\bar{a}_K)$, where $cum(\bar{a}_K) = \sum_k a_k$.

When effect modification is of interest:

$E\left(Y(\bar{a}_K)\right) = \beta_0 + \beta_1 cum(\bar{a}_K) + \beta_2 V + \beta_3 V \times cum(\bar{a}_K)$

- For dichotomous Y, $\text{logit}\left(\Pr(Y(\bar{a}_K) = 1)\right) = f(\bar{a}_K, V)$

Weight for Longitudinal MSMs:

$$\mathrm{w} = \prod_{k=0}^{K} \frac{1}{Pr(A_k|\bar{A}_{k-1}, \bar{L}_k)}$$

$$\mathrm{sw} = \prod_{k=0}^{K} \frac{Pr(A_k|\bar{A}_{k-1}, V\ )}{Pr(A_k|\bar{A}_{k-1}, \bar{L}_k)}$$

- where $\bar{A}_{-1}$ is defined to be 0, and V includes a set of baseline covariates including modifiers of the treatment effect, subset of baseline L_0.
- Use the pooled logistic regression to estimate these probabilities
- May also calculate stabilized censoring weight following similar procedure to account for drop-out, final weight is then the product of the treatment weight and the censoring weight.

Why traditional regression methods fail in the time-varying case

- For the case of time-varying treatment, the confounders would also be time-varying. There may be treatment-confounder feedback.

- If time-varying treatments and confounders, and confounders are affected by prior treatment

> Adjusting for confounder at time t masks (partially?) the effect of treatment prior to time t.

> IP weighting controls confounding because they can handle treatment-confounder feedback

Topic 3: Revisit - Yule-Simpson's Paradox

Table 1: Yule-Simpson's Paradox

Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	**Survive**	**Die**	**Survive Rate**
Treatment	18	12	60%
Control	7	3	70%
Female	**Survive**	**Die**	**Survive Rate**
Treatment	2	8	20%
Control	9	21	30%

Example from Pearl 2000

Revisit - Yule-Simpson's Paradox

- Notation:

  Treatment Assignment T: 0 - control, 1 -treat

  Outcome Y: 0 - die, 1 - survive

  Covariate X: 0 – female, 1 – male.

- The unadjusted treatment effect (ATE) is

$$\widehat{ATE}_{unadj} = \hat{P}(Y=1|T=1) - \hat{P}(Y=1|T=0) = 0.50 - 0.40 = \mathbf{+0.10}$$

- The IPW (adjusted) estimator is

$$\widehat{ATE}_{adj} = \frac{\frac{1}{\hat{P}(T=1|X=0)}\times 2 + \frac{1}{\hat{P}(T=1|X=1)}\times 18}{80} - \frac{\frac{1}{\hat{P}(T=0|X=0)}\times 9 + \frac{1}{\hat{P}(T=0|X=1)}\times 7}{80}$$

$$= \frac{\frac{2}{10/40} + \frac{18}{30/40}}{80} - \frac{\frac{9}{30/40} + \frac{7}{10/40}}{80} = (0.40 - 0.50) = \mathbf{-0.10}$$

Two estimates in opposite directions, whom should we trust?

No easy answers from "association" perspective

We have to think of "causality"

This is a good place where we can make use of the theories and tools we learnt from Causal Diagrams

Think about back-door criterion, for models (b) and (c) the correct answer is provided by the unadjusted estimator, while in structures (a), it would be the adjusted estimator

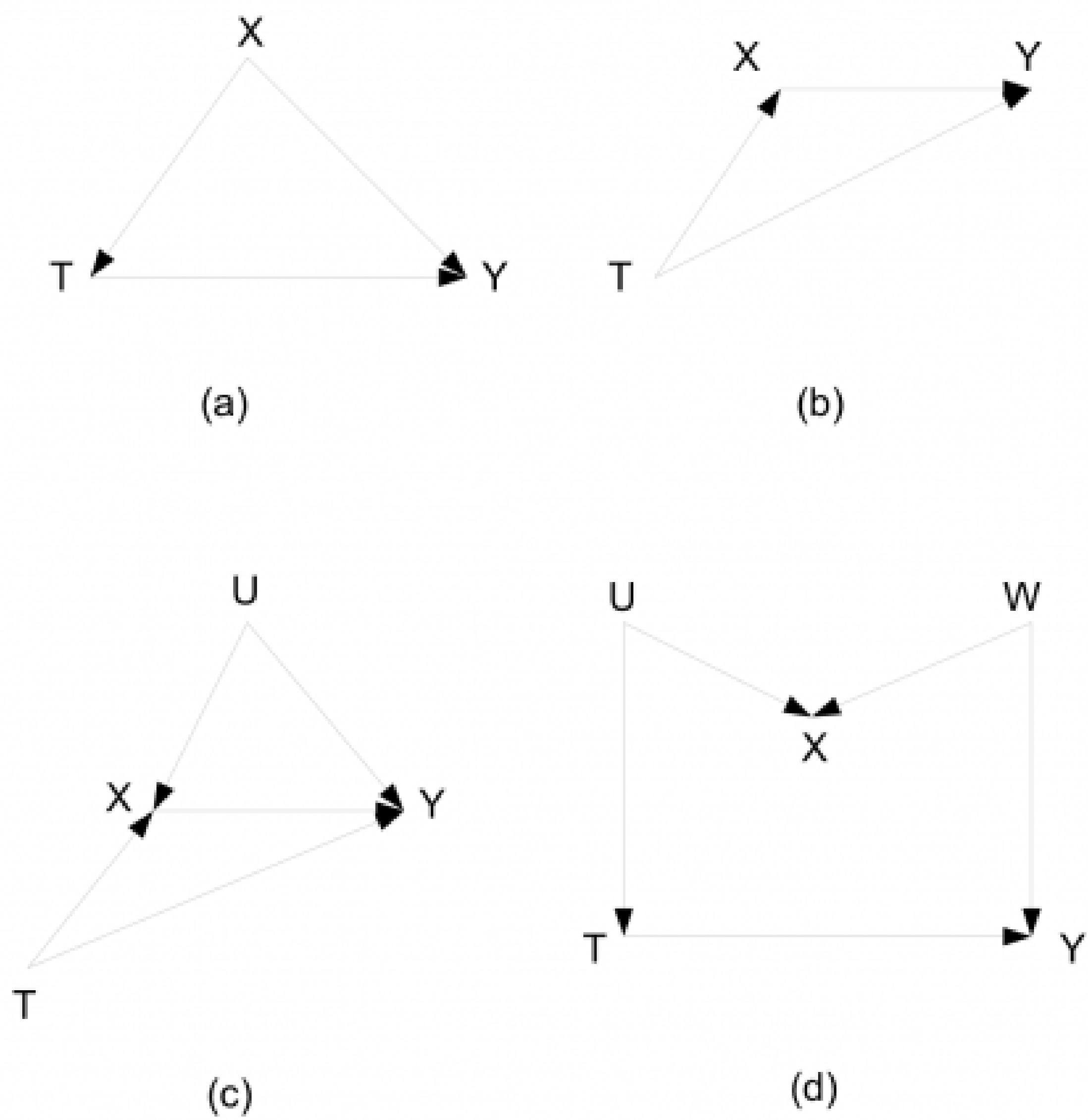
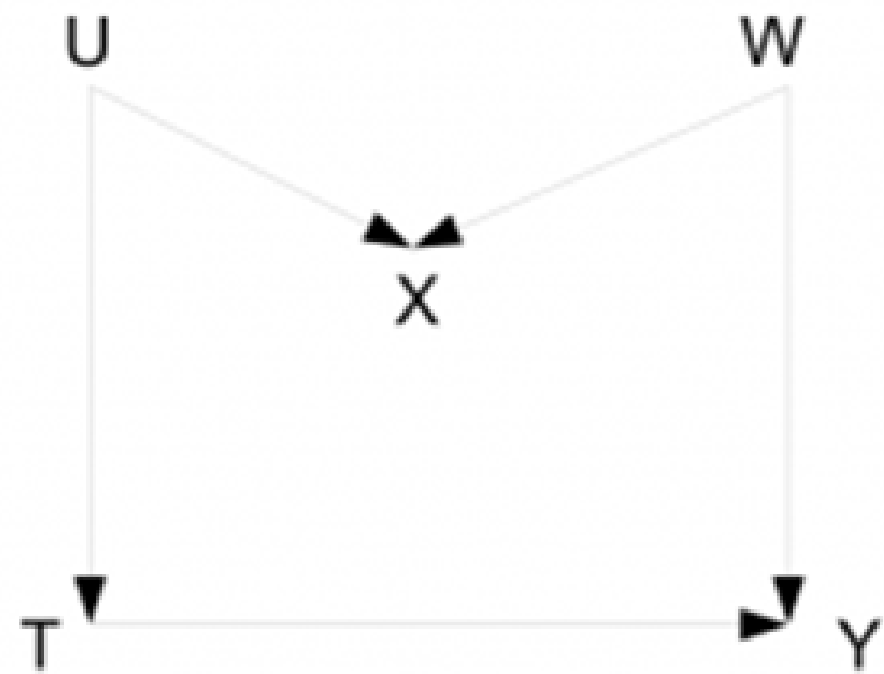


Figure 4 Simpson's paradox: possible DAGs

Bias Introduced under M-structure

- X is a pretreatment variable;
- There is a V-Structure (collider {U,X,W}), controlling X actually opens up back-door path T to Y (U and W are not independent any more !)
- Should we rely on the unadjusted estimator for causality?
- Some empirical studies suggest that the cost for not adjusting for the confounding (X) may be dominating in a lot of cases.



(d)