STA 640 — Causal Inference

Chapter 3.1. Observational Studies - Outcome Regression

Fan Li

Department of Statistical Science
Duke University

Observational Studies

- In observational studies, we do not control or know the assignment mechanism
- Presence of measured and unmeasured confounders: unbalanced between groups
- Measured confounders: also called covariates or pre-treatment variables. Covariate imbalance is the norm than exception
- Some structural (often untestable) assumptions must be made, e.g. on the treatment assignment, for identifying causal effects
- Model assumptions are also made
- For simplicity, we first focus on treatment at one time point

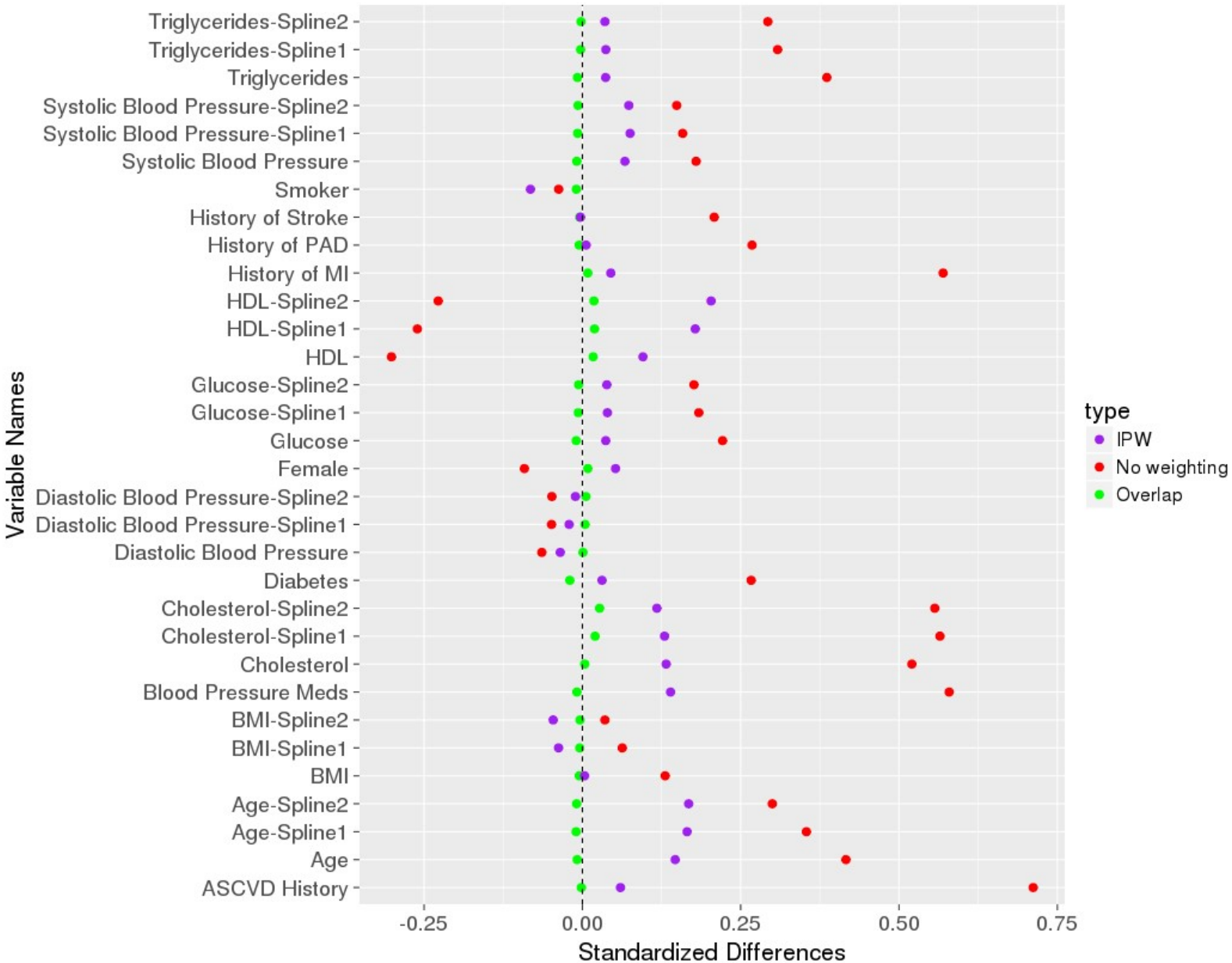
Example: Framingham Heart Study

- One of the longest-running observational studies
- Three cohorts first recruited in the early 1950s, 1970s and 2000s
- Total of over 15,000 participants with 5-60 years of follow-up
- Examined every 2-8 years
- Pioneer of cardiovascular risk prediction

Example: Framingham Heart Study

- **Patients:** cross-sectional population from the offspring cohort with a visit 6 (1995-1998)
- **Treatment:** statin use at visit 6 vs. no statin use
- **Outcomes:** CV death, myocardial infarction (MI), stroke Follow-up for outcomes begins on the date of visit 6, and is available for approximately 10 years
- **Confounders:** sex, age, body mass index, diabetes, history of MI, history of PAD, history of stroke, history of atherosclerotic cardiovascular disease, current smoking, systolic and diastolic blood pressure, blood pressure medications, total cholesterol, HDL cholesterol, triglycerides and fasting glucose

Visualize Covariate Balance in Framingham Study



Estimands

- Conditional Average treatment effect (CATE):

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x].$$

- Average treatment effect (ATE):

$$\tau^{\text{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

- Average treatment effect for the treated (ATT) or for the control (ATC):

$$\tau^{\text{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0)|Z_i = 1], \quad \tau^{\text{ATC}} = \mathbb{E}[Y_i(1) - Y_i(0)|Z_i = 0].$$

- For binary outcomes, causal odds ratio (OR):

$$\tau^{OR} = \frac{\Pr(Y_i(1) = 1)/\Pr(Y_i(1) = 0)}{\Pr(Y_i(0) = 1)/\Pr(Y_i(0) = 0)}.$$

- For survival outcomes, causal hazard ratio (HR) (Cox proportional hazards model.

Estimands

- The relation between ATE, ATT, ATC:

$$\tau^{\text{ATE}} = \Pr(Z_i = 1)\tau^{\text{ATT}} + \Pr(Z_i = 0)\tau^{\text{ATC}}$$

- In randomized experiments, ATE is equivalent to ATT (and ATC) because treatment and control groups are comparable in expectation
- In observational studies, ATE is usually different from ATT and ATC
- The above relation does not hold for ratio estimands
- To identify the causal estimands, additional assumptions are required.

Unconfoundedness

Assumption 1: Unconfoundedness (or ignorability)

$$\{Y_i(0), Y_i(1)\} \perp Z_i | X_i$$

Equivalently, $\Pr(Z_i | Y_i(0), Y_i(1), X_i) = \Pr(Z_i | X_i)$

- Assumes that within subpopulations defined by values of observed covariates, the treatment assignment is random
- Rules out unobserved confounders, also known as the assumption of "no unmeasured confounders"
- Randomized experiments satisfy unconfoundedness
- Untestable in most observational studies, but sometimes can be indirectly tested, and sensitivity can be checked

Overlap

Assumption 2: Overlap (or positivity or probabilistic assignment):

$$0 < \Pr(Z_i = 1|X_i) < 1, \quad \text{for all} i.$$

- $e(x) = \Pr(Z_i = 1|X_i = x)$ is called the propensity score (Rosenbaum and Rubin, 1983)
- This implies that, for all possible values of the covariates there are both treated and control units
- Important both conceptually and operationally (variance inflation)
- Overlap can be directly checked from the data
- Unconfoundedness and positivity jointly define the "strong ignorability" assumption (Rosenbaum and Rubin, 1983)

The Importance of Overlap

- Overlap is as important as unconfoundedness in observational studies, **both** are required for identifying the causal effects
- But overlap is often taken for granted and is less understood
- With little overlap in covariates between treatments, reliable causal inference is practically impossible. Only extrapolation, large bias and variance, sensitive
- A notable example of ignoring overlap: recent literature on “multiple causes" in machine learning (Wang and Blei, 2019)
  - Claim to bypass unconfoundedness when there are multiple causes (treatments), using latent factor models
  - However, it implicitly and violates the overlap assumption (large number of causes will inevitably lead to lack of overlap between treatments), and simply won’t work as the number of causes increases
  - Excellent counter examples were given in D’Amour (2019)

Assumptions for Estimating ATT

- When the estimand is ATT, unconfoundedness and overlap can be weakened (Heckman, Ichimura, Todd, 1997)
  - Unconfoundedness for untreated: $Y_i(0) \perp Z_i \mid \mathbf{X}_i$,
  - Weak overlap: $\Pr(Z_i = 1|\mathbf{X}_i = \mathbf{x}) < 1$ for any $\mathbf{x}$.
- Note: the definition of ATT depends on the sample ($Z_i = 1$), so it is not a sample or population estimand, but a mix.
- All ATT analysis makes an implicit assumption: the treated sample ($Z_i = 1$) is representative of the treatment population

Nonparametric Identification

- Under unconfoundedness and overlap, we have

$$\Pr(Y(z)|X) = \Pr(Y|X, Z = z)$$

The observed distribution of Y in treatment arm $Z = z$ equals the distribution of the potential outcome $Y(z)$.

- Thus we have the following two identification formuli for ATE:

$$\begin{aligned} \tau^{\text{ATE}} &= E\{\mu_1(X) - \mu_0(X)\} \\ &= E\left\{\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right\} \end{aligned}$$

where $\mu_z(X) = E(Y(z)|X) = E(Y|Z = z, X)$ is the outcome model under treatment $z (z = 0, 1)$.

Two/Three Estimation Strategies

- Two corresponding estimation strategies for ATE:
  1. Outcome modeling (or regression)) (Rubin, 1979):

$$\tau^{reg} = N^{-1} \sum_{i=1}^{N} \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}$$

  2. Inverse probability weighting (IPW) (Rosenbaum, 1987):

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

  where $\hat{e}$ is the estimated PS, $\hat{\mu}_z$ is the estimated outcome with X_i

- A third estimator is to combine outcome modeling and propensity score (IPW): augment one by another, known as the doubly-robust (DR) estimator (more in Chapter 3.5)

$$\tau^{dr} = \tau^{reg} + N^{-1} \sum_{i=1}^{N} \left\{ \frac{Z_i R_i}{\hat{e}(X_i)} - \frac{(1 - Z_i) R_i}{1 - \hat{e}(X_i)} \right\}$$

  where $R_i = Y_i - \mu_{Z_i}(X_i)$ denotes the residual from outcome modeling.

Outcome Modeling

- Outcome modeling approach: Specify a regression model for the (potential) outcome on the treatments and covariates $\mu_z(X)$
- Fit $\mu_z(X)$ to the observed data and obtain fitted potential outcomes $\hat{\mu}_z(X_i)$ $(z = 0, 1)$, i.e. impute the missing potential outcomes
- All estimands can be estimated based on $\hat{\mu}_z(X_i)$:
  - For ATE, the outcome-modeling estimator is

$$\hat{\tau}^{\text{ATE}} = \sum_{i=1}^{N} Z_i(Y_i - \hat{\mu}_0(X_i)) + (1 - Z_i)(\hat{\mu}_1(X_i) - Y_i) \Big/ N$$

  - For ATT, the outcome-modeling estimator is

$$\hat{\tau}^{\text{ATT}} = \sum_{i=1}^{N} Z_i \{Y_i - \hat{\mu}_0(X_i)\} \Big/ N_1$$

  - For CATE $\tau(x)$, the outcome-modeling estimator is

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

Outcome Modeling: Linear models

- A common choice of outcome model is the linear regression:

$$\mu_z(x) = \alpha_z + \beta_z X_i + \epsilon_{z,i}$$

- For linear models, fitting one model for each z is equivalent to adding an interaction term Xz; not the case for nonlinear models
- Then the outcome-modeling estimator of ATE has the exact form as the ANCOVA estimator in randomized experiments

$$\widehat{\tau}^{\text{ATE}} = \left\{\bar{Y}_1 - \widehat{\beta}_1(\bar{X}_1 - \bar{X})\right\} - \left\{\bar{Y}_0 - \widehat{\beta}_0(\bar{X}_0 - \bar{X})\right\}$$

  $\widehat{\beta}_z$: OLS estimate of the coefficient of Z in the regression $\mu_z(x)$

- Unlike randomized experiments, the estimator is not consistent if the linear model is misspecified
- Variance: delta method or bootstrap
- (Posterior) inference in Bayesian paradigm is automatic (more in later Chapter on Bayesian)

A Toy Example

- Population: patients with heart attack
- Treatment Z: 1 surgery; 0 medical
- Outcome Y: prognosis after 1 month
- Single covariate X - severity: 0 severe; 1 mild-average. All other covariates are matched between groups
- Sample: N patients admitted in a hospital
- Goal: (1) estimate the effect of surgery comparing to medication; (2) predict the prognosis of a new patient
- It happens to be: in the observed sample, all patients with $X = 0$ get $Z = 1$, and all patients with $X = 1$ get $Z = 0$

A Toy Example

- An idiosyncratic way is to write down a linear regression model for the observed data:

$$Y \sim a + bZ + cX$$

- For goal (1): fit the model to the sample, and the OLS coefficient of b is the "treatment effect"
- For goal (2): for a new patient, plug in Z and X to get a predicted Y
- Question: what if the new patient is with $(Z = 1, X = 1)$ or $(Z = 0, X = 0)$?
- What is odd here?

A Toy Example

- There is no interaction $Z * X$ in the model
- No interaction: effectively, but implicitly, assuming the effect of Z is additive (equivalently homogenous effects)
- Moreover, there is a complete lack of overlap in X between the two groups in the observed data: $Z * X = 0$ for all units
- Therefore, even if there is an interaction term, there is no information in the data to estimate the coefficient
- Regression itself does not take the lack of overlap into account; via extrapolation based on an untestable assumption (homogeneity), the previous model gives a—most likely wrong—point prediction
- Take home message: Regression (or any model) comes with a package, you need to know and acknowledge what assumptions—explicit or implicit—come with that model

Outcome Modeling: Importance of Overlap

- One can use a wide range of outcome models besides linear models for $\mu_z(X)$, frequentist or Bayesian, machine learning, etc.
- Key decision on model specification:
  - Two separate models for each treatment group vs. one unified model with treatment indicator? Equivalence with linear models, but not so for non-linear models
  - Case dependent, for the latter, crucial to include treatment-covariate interactions
- If the imbalance of the covariates between the two groups is large, the model-based results heavily relies on extrapolation in the region with little overlap, which is sensitive to the model specification assumption
- We do not know the exact nature of dependence of the assignment on the covariates, resulting in increased sensitivity to model and *a priori* assumptions

Strategies to Reduce Model Sensitivity

- To mitigate model dependence, two strategies:
  (1) design - balance covariates, (2) analysis- flexible models
- Best strategy is to use both jointly (double-robust or double learning): first balance covariates in the design stage, then use flexible models in the analysis stage
- Balance covariates (often involve propensity scores)
  - Stratification
  - Matching
  - Weighting
- Flexible models (revisit in the Chapter on HTE)
  - Semi-parametric models (e.g., power series)
  - Machine learning methods (e.g., tree-based methods (CART, random forest), boosting)
  - Bayesian non- and semi-parametric models (e.g., Gaussian Processes, BART, Dirichlet Processes mixtures)

Outcome Modeling Approach: Remarks

- Pros
  - General: once the potential outcomes are imputed based on the model, any estimand can be estimated
  - Many advanced and flexible models, particularly in complex data, can be used
- Cons
  - No amount of fancy models can solve the fundamental problem of causal inference: structural missing data (at least half of the data)
  - Over reliance on outcome model sometimes obscure the crucial role of covariate overlap and balance
- Ensuring covariate overlap and balance is the key for outcome modeling in causal inference – the key distinction between prediction and causal tasks

References

D'Amour, A. (2019). On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. AISTAT. arXiv:1902.10286.

Heckman, JJ, Ichimura H, Todd PE. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The review of economic studies, 64(4), 605-654.

Rubin D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association; 74:318-324.

Rosenbaum, P. R. (1987). Model-based direct adjustment. Journal of the American statistical Association, 82(398), 387-394.

Rosenbaum P, Rubin D. (1983) The central role of the propensity score in observational studies for causal effects. Biometrika; 70(1):41-55.

Wang, Y., Blei, D. M. (2019). The blessings of multiple causes. Journal of the American Statistical Association, 114(528), 1574-1596.