Causal Analysis

Impact Evaluation and Causal Machine Learning with Applications in R

Chapter 4: Selection on Observables (2)

Table of Contents

Common Support (1)

- Common support ensures overlap in the distribution of propensity scores between treated and nontreated groups.
- It does **not** require the *shape* of the distributions to be the same (in contrast to random assignment, where distributions are expected to be the same across groups).
- In the population, this is guaranteed by the assumption $0 < p(X) < 1$. In finite samples, this assumption does **not** automatically ensure common support.

Common support

The propensity score distributions for treated and nontreated groups must overlap in terms of the range of values.

Common Support (2)

Implication of common support for average effect estimation:

- For ATET: every treated observation should have a nontreated match with a similar estimated propensity scores.
- For ATE: each treated **and** nontreated observation should have matches in the other group.
- Graphical checks: histogram or density plots by treatment group can assess overlap in estimated propensity scores.

Violation of Common Support

- Common support is violated if no match with a sufficiently similar propensity score exists in the other group.
- A common solution: trimming observations with extreme or unmatched propensity scores.

Trimming

The removal of observations whose propensity scores lack overlap between treated and nontreated groups to enforce common support.

- Improves **internal validity** by reducing bias in the trimmed sample.
- Reduces **external validity**, as estimates apply only to a subpopulation.
- Always report the number or share of trimmed observations.

Trimming Criteria: Density-Based

- Heckman, Ichimura, Smith, and Todd (1998) propose dropping observations for which the estimated density of estimated propensity score is (close to) zero in (at least) one of the treatment groups.
- Estimate densities using kernel methods (e.g., Rosenblatt, 1956; Parzen, 1962).
- Threshold can be data-driven (e.g., based on a quantile of estimated densities).

Trimming Criteria: Extreme Propensity Scores

- Crump et al. (2009) propose trimming extreme propensity scores.
- Suggested rule: retain only observations with $\hat{p}(X_i) \in [0.1, 0.9]$ (minimizes the variance of ATE estimation under certain conditions).
- Further alternatives: $[0.05, 0.95]$ or $[0.01, 0.99]$ - potential trade-off concerning cut-off choice in terms of external validity and variance.
- Dehejia and Wahba (1999) propose discarding all treated observations with $\hat{p}(X_i)$ higher than the highest value among the nontreated when estimating the ATET.
- For ATE estimation, also discard nontreated observations with $\hat{p}(X_i)$ lower than the lowest value among treated.

Sample-Size Dependent Trimming

- Trimming rules should be adapted to the sample size:
  - In large samples, observations with extreme propensity score values are more likely to find a match.
  - Therefore, trimming becomes less necessary as $n \to \infty$.
- One possible approach: limit how much influence (weight) any observation can have in a given sample when estimating $E[Y(1)]$ or $E[Y(0)] \to$ drop too influential observations.

Example: IPW estimation (see equation (4.42))

The weight for a treated unit is: $\text{Weight}_i = \frac{D_i/\hat{p}(X_i)}{\sum_{i=1}^{n} D_i/\hat{p}(X_i)}$. This weight increases as $\hat{p}(X_i)$ decreases. Fixing the maximum weight to 0.05 implies that no single unit should contribute more than 5% to $E[Y(1)]$, see e.g. Huber, Lechner, and Wunsch (2013).

Match Quality (Covariate Balance)

- Match quality concerns whether the estimated propensity score balances X across groups.
- Poor balance implies risk of treatment selection bias in causal effect estimation.

Match quality

The extent to which propensity score adjustment (e.g., matching, IPW) equalizes the distribution of X across treated and nontreated groups.

- Reasons for poor match quality:
  - Misspecified propensity score model $\Rightarrow$ use a more flexible model.
  - Inadequate matching algorithm $\Rightarrow$ try a different algorithm.
  - Lack of common support $\Rightarrow$ apply trimming.

Verifying Covariate Balance: t-Test

A standard approach for assessing covariate balance is the two-sample t-test (Welch, 1947):

- Test applied to each covariate X_k in matched or IPW-weighted samples. X_k^m denotes covariate k among matched (or weighted) observations.
- Tests the null hypothesis: $E[X_k^m|D=1] = E[X_k^m|D=0]$
- Test statistics:

$$\frac{\bar{X}_k^{m1} - \bar{X}_k^{m0}}{\sqrt{\frac{\widehat{Var}(X_k^{m1})}{n^{m1}} + \frac{\widehat{Var}(X_k^{m0})}{n^{m0}}}} \tag{4.47}$$

  - $\bar{X}_k^{m1}, \bar{X}_k^{m0}$: sample means in matched treated ($m1$) and nontreated ($m0$) groups.
  - n^{m1}, n^{m0}: number of matched treated and nontreated observations.
  - $\widehat{Var}(X_k^{m1}), \widehat{Var}(X_k^{m0})$: sample variances of X_k in the matched groups.

Verifying Covariate Balance: Distributional Test

Kolmogorov-Smirnov test

- Instead of testing for mean differences only, we may test for distributional differences.
- Null hypothesis: each covariate X_k has the same distribution in matched treated and nontreated groups.
- Successful balancing implies that the entire distribution of a covariate (not just the mean) is equal across matched groups.

Verifying Covariate Balance: Joint Testing

Joint regression-based test (Smith and Todd, 2005):

- Regress X_k on constant, D, $\hat{p}(X)$, higher-order terms, and interactions of D and $\hat{p}(X)$ in the toal (rather than matched) sample.
- If $\hat{p}(X)$ balances well, coefficients on D and its interactions with $\hat{p}(X)$ should be close to zero.
- Null hypothesis: coefficients on D and its interactions with the propensity score are jointly zero.
- Use an F-test to jointly test that all such coefficients are zero. High p-value indicates good balance conditional on $\hat{p}(X)$.

Verifying Covariate Balance: Accounting for Multiple Testing

- Testing many covariates (e.g., by t-tests) introduces a multiple hypothesis testing problem.
- With $\alpha = 0.05$ and 100 covariates, one may expect ≈ 5 false rejections even if null holds for all.
- A few significant rejections do not necessarily imply imbalance, as multiple testing increases the risk of obtaining false positives.
- To avoid this issue, apply joint tests (e.g., F-tests) across all covariates.
- Regress each X_k on a constant and D among matched observations.
- Test whether all D coefficients are jointly zero.

Verifying Covariate Balance: Pseudo-R^2 Check

Alternative joint test is to re-estimate the propensity score after matching:

- Estimate $\Pr(D = 1|X^m)$ using the matched sample only, as in Sianesi (2004).
- Check whether covariates still predict treatment assignment after matching.
- The goodness of fit is measured by the pseudo-R^2 from this re-estimated propensity score model.
- Interpretation: a pseudo-R^2 close to zero indicates good covariate balance.
- Intuition: low predictive power of X^m for D implies $X^m \perp D$, i.e., balance.

Verifying Covariate Balance: Standardized Differences

Issue with hypothesis tests: test statistics depend on sample size.

- t-statistic is a function of the matched sample sizes n^{m1} and n^{m0}.
- As sample size grows, even negligible mean differences can lead to rejection of the balancing hypothesis.

Standardized difference test (Rosenbaum and Rubin, 1985):

- Measures mean differences relative to the pooled variance in the original samples.
- Test statistics:
$$100 \cdot \frac{\bar{X}_k^{m1} - \bar{X}_k^{m0}}{\sqrt{\frac{\widehat{Var}(X_k^1) + \widehat{Var}(X_k^0)}{2}}} \tag{4.48}$$
  - Insensitive to the number of matched observations.
  - $\widehat{Var}(X_k^1)$ and $\widehat{Var}(X_k^0)$ refer to variances in the *original* treated and nontreated samples (not matched).
- Thresholds (e.g., 10 or 20) used to judge balance; absolute differences above the threshold suggest imbalance.

Table of Contents

Multivalued Treatments: Identification

- The selection-on-observables framework extends to multivalued treatments $D \in \{0, 1, 2, \ldots, J\}$.
- Treatment effects can be identified by pairwise comparisons: $D = d$ vs. $D = d'$.

Identifying assumptions (Imbens, 2000)

$$\{Y(d), Y(d')\} \perp D|X, \quad \Pr(D = d|X) > 0,$$
$$\Pr(D = d'|X) > 0, \quad X(d) = X(d') = X \tag{4.49}$$

Under these assumptions you can:

- Identify ATE and ATET for d vs. d'.
- Apply identification strategies from the binary case (regression, matching, IPW, DR). How?
  1. Replace D with $I\{D = d\}$ and $1 - D$ with $I\{D = d'\}$.
  2. Use $\Pr(D = d|X)$ as $p(X)$ and $\Pr(D = d'|X)$ as $1 - p(X)$

Estimation with Multivalued Treatments

- As in the binary case, estimation via regression, matching, IPW, or DR can be $\sqrt{n}$-consistent and semiparametrically efficient if plug-in parameters are estimated nonparametrically (Cattaneo, 2010).
- Propensity score matching can be applied after estimating propensity scores $\Pr(D = d|X)$ and $\Pr(D = d'|X)$ (Lechner, 2001).

Notice:

- If assumptions (4.49) hold for all d, d', then $\{Y(0), ..., Y(J)\} \perp D \mid X$.
- This full selection-on-observable assumption is stronger then a pair-wise version and must be carefully assessed in applications.

Continuous Treatments: Setup

- If D is continuously distributed (e.g., training hours), then $\Pr(D = d|X)$ becomes the conditional density $f(D = d|X)$.

Generalized propensity score

Conditional density of treatment given covariates: $f(D = d \mid X)$

- Replace $\Pr(D = d|X) > 0$ with $f(D = d|X) > 0$ in (4.49).
- If these conditions hold for all d, effects across the full treatment range are identified.

Continuous Treatments: Estimation

- Estimate causal effects via parametric or nonparametric regression of Y on D and X:
  - Estimate $\mu_d(x)$, $\mu_{d'}(x)$, and $E[\mu_d(x) - \mu_{d'}(x)]$ to estimate ATE of $D = d$ versus $D = d'$.
  - Estimate $\frac{\partial \mu_d(x)}{\partial d}$ to estimate marginal treatment effect $\frac{\partial E[Y(d)]}{\partial d}$.
- Alternatively, regress Y on D and $\hat{f}(D|X)$ (Hirano and Imbens, 2005), or use stratification (Imai and van Dyk, 2004).
- IPW estimation replaces indicators like $I\{D = d\}$ with kernel weights (Flores et al., 2012; Galvao and Wang, 2015):
  - Define kernel weight: $\mathcal{K}((D - d)/h)/h$ with bandwidth h and symmetric kernel $\mathcal{K}$.
  - Weight observations by closeness of D to d.
  - ATE identified by: $\Delta = \lim_{h \to 0} E\left[\frac{Y \cdot \mathcal{K}\left(\frac{D-d}{h}\right)/h}{f(D=d|X)} - \frac{Y \cdot \mathcal{K}\left(\frac{D-d'}{h}\right)/h}{f(D=d'|X)}\right]$
- Also doubly robust (DR) approaches (Kennedy et al., 2017) can be applied to estimate effects of continuous treatments.

Distributional Treatment Effects

- Selection on observables permits assessing effects on the entire outcome distribution, not just averages.
- To do so, replace Y with indicator function $I\{Y \leq y\}$, e.g. in IPW expressions.

Distributional effect

$F_{Y(1)}(y) - F_{Y(0)}(y)$: effect on the treatment on the share of subjects with outcome $\leq y$, where $F_{Y(d)}(y) = E[I\{Y(d) \leq y]$ denotes the cumulative distribution function of potential outcome $Y(d)$.

Example

$F_{Y(0)}(4{,}000) = 0.5$: 50% would earn $\leq$ €4,000 without treatment.

- DiNardo, Fortin, and Lemieux (1996) and Chernozhukov, Fernández-Val, and Melly (2013) discuss the estimation of potential outcome distributions.

Quantile Treatment Effects (1)

- QTEs correspond to treatment effects at specific ranks (e.g., median, quartiles) of the outcome distribution.
- Useful for studying effect heterogeneity across ranks of outcome distribution (e.g., low- versus high-income groups).

Requirements for QTEs

- Outcome Y must be continuously distributed.
- Distribution must be strictly increasing across ranks of interest.

Example

If no one earns between 2,000 and 2,500 EUR, ranks in this interval are undefined.
$\Rightarrow$ No one-to-one mapping between quantiles and ranks.

Quantile Treatment Effects (2)

- The quantile function of a potential outcome $Y(d)$ is the inverse of its CDF:

$$F^{-1}_{Y(d)}(\tau), \quad \tau \in (0,1), \quad d \in \{0,1\}$$

- τ: rank in the outcome distribution (e.g., $\tau = 0.5$ is the median).

Identification via IPW (Firpo, 2007)

Under selection on observables, quantiles can be identified by solving: $F^{-1}_{Y(d)}(\tau) = \min_y E\left[\frac{D}{\Pr(D=d|X)} \cdot (Y-y) \cdot (\tau - I\{Y-y<0\})\right]$

- The loss function $(Y-y) \cdot (\tau - I\{Y-y<0\})$ targets quantiles (not means).

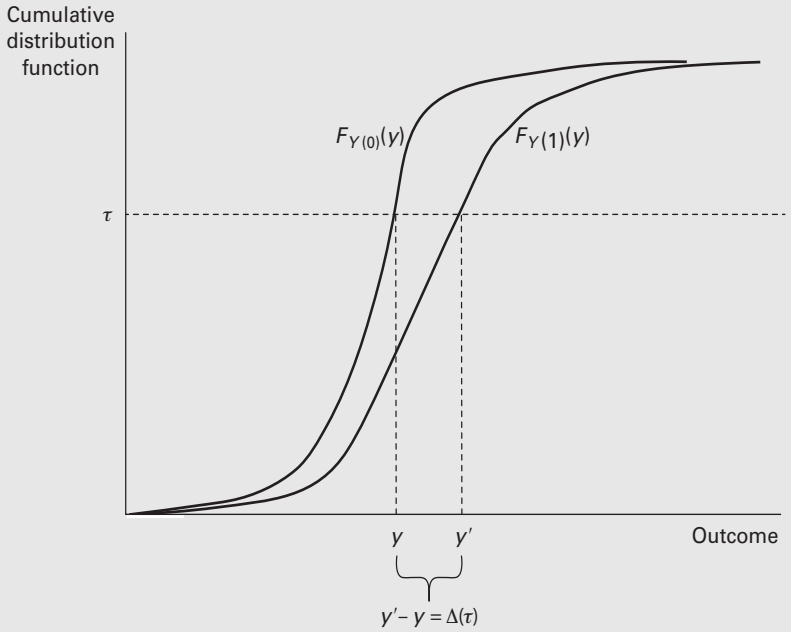
Quantile Treatment Effects (3)

QTE

The quantile treatment effect at rank τ is: $\Delta(\tau) = F^{-1}_{Y(1)}(\tau) - F^{-1}_{Y(0)}(\tau)$

It compares outcome values at the same rank under treatment and nontreatment.

Example



The figure shows: at rank τ, quantiles $y = F^{-1}_{Y(0)}(\tau)$ and $y' = F^{-1}_{Y(1)}(\tau)$
$\Rightarrow$ QTE $= y' - y$.

Table of Contents

Dynamic Treatment Effects

- Dynamic treatment effects: effects of sequences of treatments over time (Robins, 1986, Robins, Hernan, and Brumback, 2000).
- Examples: courses (language course followed by IT course), medical interventions (surgery followed by physiotherapy).
- Control for time-varying confounders affecting outcome and treatment over different periods.
- Sequential selection-on-observables assumes random assignment conditional on past information.

Average treatment effect of a sequence

$$\Delta(\underline{d}_2, \underline{d}'_2) = E[Y_2(\underline{d}_2) - Y_2(\underline{d}'_2)] \tag{4.53}$$

Example

ATE of (1=language, 2=IT) vs. no training: $\underline{d}_2 = (1, 2)$, $\underline{d}'_2 = (0, 0)$

Sequential Selection on Observables

- Covariates X_t evolve over time and may be affected by past treatments/outcomes.
- X_0: pre-treatment.
- X_1: observed after D_1, before D_2, may include Y_1.
- Confounding is dynamic: must control for post-treatment covariates.

Sequential selection-on-observables assumption

$$Y_2(\underline{d}_2) \perp D_1 | X_0 \text{ and } Y_2(\underline{d}_2) \perp D_2 | D_1, X_0, X_1$$
$$\Pr(D_1 = d_1 | X_0) > 0, \ \Pr(D_2 = d_2 | D_1, X_0, X_1) > 0 \qquad (4.54)$$

Graphical Illustration

- No unobserved confounding of (D_1, Y_2) given X_0 and (D_2, Y_2) given (D_1, X_0, X_1).
- D_2 may depend on D_1: IT training more likely after language course.

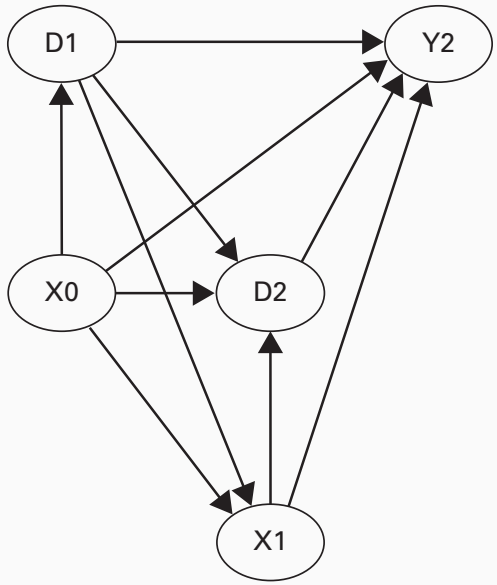


Figure 4.12: Sequential conditional independence with posttreatment confounders

Regression and IPW

- Nested regression approach:

$$\begin{aligned}\Delta(\underline{d}_2, \underline{d}_2') = E[E[E[Y_2|\underline{D}_2 = \underline{d}_2, X_0, X_1]|D_1 = d_1, X_0] \\ - E[E[Y_2|\underline{D}_2 = \underline{d}_2', X_0, X_1]|D_1 = d_1', X_0]]\end{aligned} \tag{4.55}$$

- IPW approach using sequential propensity scores (Lechner, 2009):

$$\begin{aligned}\Delta(\underline{d}_2, \underline{d}_2') = E\Bigg[\frac{Y \cdot I\{D_1 = d_1\} I\{D_2 = d_2\}}{p^{d_1}(X_0) p^{d_2}(D_1, X_0, X_1)} \\ - \frac{Y \cdot I\{D_1 = d_1'\} I\{D_2 = d_2'\}}{p^{d_1'}(X_0) p^{d_2'}(D_1, X_0, X_1)}\Bigg],\end{aligned} \tag{4.56}$$

where $p^{d_1}(X_0) = \mathsf{Pr}(D_1 = d_1|X_0)$ and $p^{d_2}(D_1, X_0, X_1) = \mathsf{Pr}(D_2 = d_2|D_1, X_0, X_1)$ are the propensity scores in the two periods.

Doubly Robust Approach

- DR approach combining outcome models and sequential propensity scores (Robins, 2000):

$$\Delta(\underline{d}_2, \underline{d}_2') = E[\psi^{\underline{d}_2} - \psi^{\underline{d}_2'}], \tag{4.57}$$

$$\text{where } \psi^{\underline{d}_2} = \frac{I\{D_1 = d_1\} I\{D_2 = d_2\}(Y_2 - \mu^{Y_2}(\underline{d}_2, X_0, X_1))}{p^{d_1}(X_0) p^{d_2}(d_1, X_0, X_1)}$$

$$+ \frac{I\{D_1 = d_1\}(\mu^{Y_2}(\underline{d}_2, X_0, X_1) - \nu^{Y_2}(\underline{d}_2, X_0))}{p^{d_1}(X_0)} + \nu^{Y_2}(\underline{d}_2, X_0),$$

with $\mu^{Y_2}(\underline{d}_2, X_0, X_1) = E[Y_2 | \underline{D}_2 = \underline{d}_2, X_0, X_1]$ and $\nu^{Y_2}(\underline{d}_2, X_0) = E[E[Y_2 | \underline{D}_2 = \underline{d}_2, X_0, X_1] | D_1 = d_1, X_0]$ being the (nested) conditional mean outcomes.

Stronger Assumptions

- If $D_2 \perp Y_2(\underline{d}_2) | D_1, X_0$, we may drop X_1 from models.
- Reduced data requirement: no posttreatment covariates needed to be controlled for.

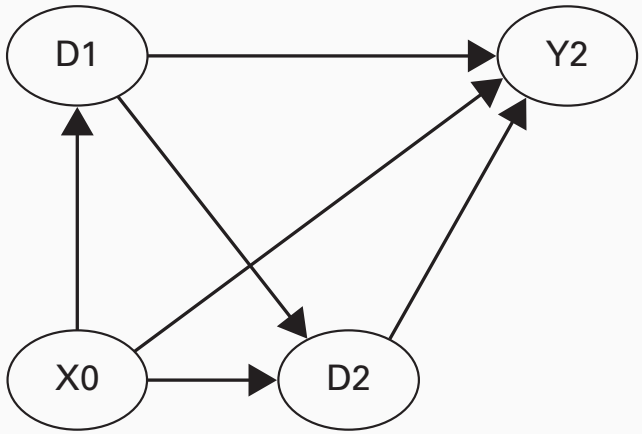


Figure 4.13: Sequential conditional independence without posttreatment confounders

Simplified DR Approach

- DR expression simplifies to:

$$\psi^{\underline{d}_2} = \frac{I\{D_1 = d_1\}I\{D_2 = d_2\}(Y_2 - \mu^{Y_2}(\underline{d}_2, X_0))}{p^{d_1}(X_0)p^{d_2}(d_1, X_0)} + \mu^{Y_2}(\underline{d}_2, X_0) \qquad (4.58)$$

- Equivalent to multivalued discrete treatment evaluation.

When Post-treatment Covariates Matter

- Long time gaps between treatments may invalidate simpler assumptions.
- Individual characteristics may change over time and act as confounders jointly affecting the second treatment and the outcome (e.g., health, labor behavior).
- In such cases, controlling for X_1 is necessary to make identifying assumptions more credible.

Table of Contents

Causal Mediation Analysis

How can we assess the causal mechanisms through which the treatment affects the outcome?

Mediation analysis (Robins and Greenland, 1992, Pearl, 2001)

Disentangles a total treatment effect into one or several indirect effects (via intermediate variables), as well as a direct effect.

- Indirect effects operate through one or several intermediate variables that are commonly referred to as mediators.
- Direct effect includes any causal mechanisms not operating through the mediators.

Controlled Direct Effect

Controlled direct effect

Net effect of D_1 obtained by controlling for the mediator D_2, when setting it to the same value for everyone in the population.

- The sizes of the direct effects may differ across values of D_2 if there are interaction effects between D_1 and D_2.

Example

- Assessing the effect on earnings of a sequence of training programs (job application training, IT course).
- Direct effect of the job application training net of participation in the IT course is obtained by setting the latter to zero.
- This implies assessing the treatment effect $\Delta(\underline{d}_2, \underline{d}'_2)$ with sequences $\underline{d}_2 = (1, 0)$ and $\underline{d}'_2 = (0, 0)$.

Natural Direct and Indirect Effects (1)

Natural direct effect

Effect of D_1, conditional on the value of the mediator D_2 that is naturally chosen as a reaction to D_1.

Natural indirect effect

Effect operating through the choice of D_2 in reaction to D_1.

- The naturally chosen value of D_2 under a specific value of D_1 may vary across individuals (e.g., as a function of their preferences).
- Depending on the empirical context, either controlled or natural effects may be more relevant.

Natural Direct and Indirect Effects (2)

- Let D_1 be a binary treatment and D_2 a binary mediator.
- Extend potential outcome notation:
  - $D_2(d_1)$: potential mediator as a function of $d_1 \in \{0, 1\}$.
  - $E[Y_2(d_1, D_2(d_1'))]$: potential outcome conditional on $D_1 = d_1$ and the potential mediator under $D_1 = d_1'$ (with $d_1, d_1' \in \{0, 1\}$).
- Total ATE of D_1 on Y_2:

$$\Delta(D_1) = E[Y_2(1, D_2(1)) - Y_2(0, D_2(0))] \tag{4.59}$$

Natural Direct and Indirect Effects (3)

- The total effect of D_1 is the sum of the natural direct and indirect effects defined based on opposite treatment states:

$$\begin{aligned}\Delta(D_1) = & \underbrace{E[Y_2(1, D_2(1)) - Y_2(0, D_2(1))]}_{=\theta(1)} + \underbrace{E[Y_2(0, D_2(1)) - Y_2(0, D_2(0))]}_{=\delta(0)} \\ = & \underbrace{E[Y_2(1, D_2(0)) - Y_2(0, D_2(0))]}_{=\theta(0)} + \underbrace{E[Y_2(1, D_2(1)) - Y_2(1, D_2(0))]}_{=\delta(1)} \quad (4.60)\end{aligned}$$

  - $\theta(1)$ and $\theta(0)$: natural direct effects of D_1.
  - $\delta(1)$ and $\delta(0)$: natural indirect effects.

- $\theta(1)$ and $\theta(0)$ (and $\delta(1)$ and $\delta(0)$) may differ if there are interaction effects between D_1 and D_2.
- Direct and indirect effect sum up to the total effect if interaction effects between D_1 and D_2 are either accounted for in the direct or the indirect effect, but not both at the same time.

Identifying Assumptions

- $Y_2(1, D_2(0))$ and $Y_2(0, D_2(1))$ are never observed.
  $\Rightarrow$ Identification of natural direct and indirect effects requires stronger assumptions than for controlled or dynamic effects.

Assumption 1 (conditional independence of the treatment):

$$\{Y_2(\underline{d}_2), D_2(d_1')\} \perp D_1 | X_0 \text{ for } \underline{d}_2 = (d_1, d_2) \text{ and } d_1, d_1', d_2 \in \{0, 1, \ldots, J\} \quad (4.61)$$

Assumption 2 (conditional independence of the mediator):

$$Y_2(\underline{d}_2) \perp D_2 | D_1, X_0 \text{ for } \underline{d}_2 = (d_1, d_2) \text{ and } d_1, d_2 \in \{0, 1, \ldots, J\}$$

Assumption 3 (common support):

$$\Pr(D_1 = d_1 | X_0) > 0 \text{ and } \Pr(D_2 = d_2 | D_1, X_0) > 0 \text{ for } d_1, d_2 \in \{0, 1 \ldots, J\}$$

- Here we allow for a multivalued, discrete treatment.

Regression and IPW

Identification based on nested conditional mean outcomes (may be implemented by regression; Imai, Keele, and Yamamoto, 2010)

$$E[Y_2(d_1, D_2(d_1'))] = E[E[\mu^{Y_2}(d_1, D_2, X_0)|D_1 = d_1', X_0]], \tag{4.62}$$

where $\mu^{Y_2}(D_1, D_2, X_0) = E[Y_2|D_1, D_2, X_0]$ is the conditional mean outcome, and d_1, d_1' are specific values of the first treatment.

IPW-based identification (Hong, 2010)

$$E[Y_2(d_1, D_2(d_1'))] = E\left[\frac{I\{D_1 = d_1\} \cdot p^{D_2}(d_1', X_0) \cdot Y_2}{p^{d_1}(X_0) \cdot p^{D_2}(d_1, X_0)}\right], \tag{4.63}$$

where $p^{d_2}(D_1, X_0) = \Pr(D_2 = d_2|D_1, X_0)$ is the propensity score of the mediator.

DR Approach

DR identification (combines IPW with conditional mean outcomes; Tchetgen Tchetgen and Shpitser, 2012)

$$E[Y_2(d_1, D_2(d_1'))] = E[\psi^{d_1,d_1'}],$$

$$\begin{aligned}\text{with } \psi^{d_1,d_1'} =& \frac{I\{D_1 = d_1\} \cdot p^{D_2}(d_1', X_0)}{p^{d_1}(X_0) \cdot p^{D_2}(d_1, X_0)} \cdot [Y_2 - \mu^{Y_2}(d_1, D_2, X_0)] \\ &+ \frac{I\{D_1 = d_1'\}}{p^{d_1'}(X_0)} \cdot [\mu^{Y_2}(d_1, D_2, X_0) - E[\mu^{Y_2}(d_1, D_2, X_0) \Big| D_1 = d_1', X_0]] \\ &+ E[\mu^{Y_2}(d_1, D_2, X_0) | D_1 = d_1', X_0]\end{aligned} \tag{4.64}$$

Avoiding Mediator Probability/Density

Alternative IPW-based identification (Huber, 2014a)

$$E[Y_2(d_1, D_2(d_1'))] = E\left[\frac{I\{D_1 = d_1\} \cdot p^{d_1'}(D_2, X_0) \cdot Y_2}{p^{d_1}(D_2, X_0) \cdot p^{d_1'}(X_0)}\right] \tag{4.65}$$

- When the mediator is continuously distributed and/or consists of several variables, estimating $p^{D_2}(d_1, X_0)$ may be cumbersome.
- $p^{d_2}(D_1, X_0)$ can be avoided by including an alternative treatment propensity score $p^{d_1}(D_2, X_0) = \Pr(D_1 = d_1 | D_2, X_0)$.

Path-Wise (Partial Indirect) Effect (1)

- Assuming as-good-as-random assignment of mediator D_2 given only treatment D_1 and baseline covariates X_0 may be too strong.
- In many cases, posttreatment covariates X_1 also need to be controlled for.
  $\Rightarrow$ Replace $Y_2(\underline{d}_2) \perp D_2 | D_1, X_0$ with $Y_2(\underline{d}_2) \perp D_2 | D_1, X_0, X_1$.
  $\Rightarrow$ Additional assumption: No confounders that jointly affect (i) D_1 and X_1, given X_0 and (ii) X_1 and D_2 or Y_2, given D_1, X_0.
- However, additional assumptions are not sufficient for the nonparametric identification of natural direct and indirect effects - is only obtained if X_0 and D_1 are sufficient to control for confounders of D_2 and Y_2 (Avin, Shpitser, and Pearl, 2005).
- However, under the additional assumptions, we can identify the path-wise effect of D_1 on Y_2 directly operating via D_2, i.e., $D_1 \rightarrow D_2 \rightarrow Y_2$.

Path-Wise (Partial Indirect) Effect (2)

- Path-wise (or partial indirect) effect with a binary treatment based on IPW:

$$\delta^p(d_1) = E\left[\frac{Y_2 \cdot I\{D_1 = d_1\}}{\Pr(D_1 = d_1|D_2, X_0, X_1)} \cdot \frac{\Pr(D_1 = d_1|X_0, X_1)}{\Pr(D_1 = d_1|X_0)}\right.$$
$$\left. \times \left(\frac{\Pr(D_1 = 1|D_2, X_0, X_1)}{\Pr(D_1 = 1|X_0, X_1)} - \frac{1 - \Pr(D_1 = 1|D_2, X_0, X_1)}{1 - \Pr(D_1 = 1|X_0, X_1)}\right)\right], \quad (4.66)$$

with $\delta^p(d_1)$ denoting the pathwise effect of $D_1 \to D_2 \to Y_2$.

- $\delta^p(d_1)$ represents only a partial indirect effect because it omits any indirect impact that operates via X_1 (i.e., $D_1 \to X_1 \to D_2 \to Y_2$).
- For this reason, it does not coincide with the natural indirect effect $\delta(d_1)$.

Full Natural Indirect Effects with Posttreatment Confounders (1)

- To identify the full natural indirect effect, we need to impose further assumptions, such as:

No interaction effects between D_1 and D_2 (Robins, 2003)

- Effect of the treatment does not depend on that of the mediator and vice versa.
- For a binary treatment: $Y(1, m) - Y(0, m) = Y(1, m') - Y(0, m')$ for any distinct mediator values $m \neq m'$.
- Unattractive in many empirical contexts, as it severely restricts effect heterogeneity.

Homogeneous treatment-mediator interaction effect (Imai and Yamamoto, 2013)

- Relaxes the assumption above, but assumes interaction effect to be the same for different subjects.

Full Natural Indirect Effects with Posttreatment Confounders (2)

Zero average interaction effect (Tchetgen Tchetgen and VanderWeele, 2014)

- Average interaction effects of X_1 and D_2 on Y_2 are zero.

Independence or known association of $X_1(1)$ and $X_1(0)$ (Robins and Richardson, 2010; Albert and Nelson, 2011)

- Potential values of X_1 under treatment and nontreatment are independent or the form of their statistical association is known.

Homogeneous path effects given X_0 (Xia and Chan, 2021)

- Given X_0, average effects operating via the paths $D_1 \rightarrow Y_2$ and $D_1 \rightarrow X_1 \rightarrow Y_2$ are homogeneous across values of $M(0)$.

$\Rightarrow$ All assumptions impose specific constraints (to be scrutinized in empirical contexts).

Table of Contents

Outcome Attrition and Posttreatment Sample Selection

- Problem: outcome of interest is observed only for a nonrandom subsample in the data.
- **Outcome attrition:**
  - Outcome is measured in a follow-up survey, but some participants cannot be reinterviewed (e.g., due to relocation or refusal).
- **Posttreatment sample selection:**
  - The outcome is observed only conditional on some other posttreatment variable (e.g., wages only if employed).
- Sample selection and outcome attrition can create bias in causal effect estimation—even if treatment is randomized.
- Are there conditions that permit us to fix this problem?

Missing at Random (MAR) Assumption (1)

- Impose a selection-on-observables assumption with respect to outcome attrition/sample selection:

Missing at random (MAR) assumption (Rubin, 1976)

Outcome attrition/sample selection is as good as random, conditional on observed information (e.g., covariates, treatment).

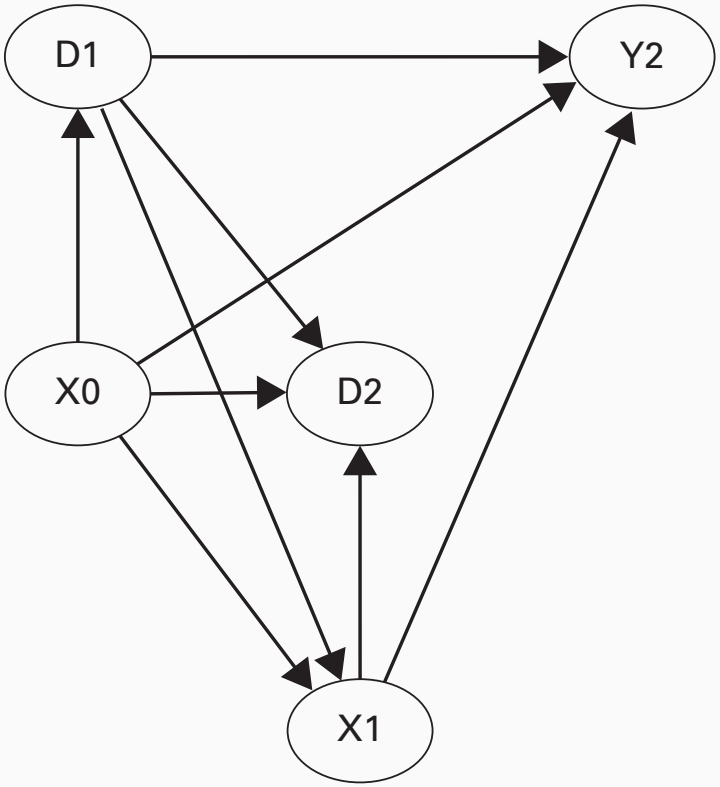
- Under the following assumptions, we can assess the ATE of the (possibly multivalued) treatment D_1:

$$Y_2(d_1) \perp D_1 \,|\, X_0 \quad \text{and} \quad Y_2(d_1) \perp D_2 \,|\, D_1, X_0, X_1$$
$$\Pr(D_1 = d_1 \,|\, X_0) > 0 \quad \text{and} \quad \Pr(D_2 = d_2 \,|\, D_1, X_0, X_1) > 0 \qquad (4.67)$$

  - D_2 is a binary indicator of whether the outcome is observed.
  - Y is known only if $D_2 = 1$, but unknown if $D_2 = 0$.
  - The potential outcome $Y_2(D_1)$ is a function of D_1 only, but not of the indicator for its observability, D_2.

Missing at Random (MAR) Assumption (2)

Figure 4.14: Causal paths under sequential conditional independence



- Causal graph satisfies the assumptions on the previous slide.

Using MAR to Identify the ATE

- The conditions in expression (4.67) are similar to those suggested for dynamic treatment effects.
- Key difference: We now assume that $Y_2(d_1) \perp D_2 | D_1, X_0, X_1$.
  - This implies that D_2 does not affect Y, as it is not a treatment.
  - Also implies: D_2 is not associated with unobserved characteristics affecting Y, conditional on covariates X_0, X_1 and treatment D_1.
- Apply the identification results for dynamic treatment effects for assessing treatment effects under outcome attrition or sample selection.
  - Simply set $D_2 = 1$ in any conditional mean outcome and propensity score.

Simplified Confounding

- The framework simplifies if conditioning on X_1 is not required.
- Then, outcome attrition/sample selection is as good as random given treatment D_1 and baseline covariates X_0 alone.
- In this case:
  - We can drop X_1 from the assumptions in expression (4.67).
  - Any conditional mean outcomes or propensity scores in the expressions for the identification of the ATE no longer require X_1.
- Simplification is unrealistic when there is a substantial time lag between the treatment and the measurement of the outcome.
- In these scenarios, posttreatment confounders affecting both D_2 (e.g., employment) and Y_2 (e.g., wage) likely exist.