Causal Inference (Hernán & Robins) Chapter 11-14

Richard A.J. Post

TU Eindhoven

Leiden, 03-12-2018

TU/e

Outline

1. Recap first meeting
   - Causal inference assumptions
   - Standardization & IPTW
2. Chapter 11
3. Chapter 12
   - Estimating IP weights via modelling
   - IP weights
   - Marginal structural models
4. Chapter 13
   - Standardization via modelling
   - IP weighting or standardization?
   - Intermezzo - Parametric g-formula
5. Time-dependent confounding
6. Chapter 14
   - Structural nested models
   - G-estimation

Conditions for causal inference

TU/e

1. (C) Consistency, $Y = Y^A$
2. (CE) Conditional exchangeability, $Y^a \perp\!\!\!\perp A|L \; \forall a$
3. (P) Positivity, $\mathbb{P}(A = a|L) > 0$ a.s. $\forall a$

Under randomization $Y^a \perp\!\!\!\perp A \; \forall a$, thus

$$\begin{aligned}\mathbb{E}[Y^{a=1} - Y^{a=0}] &= \mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] \\ &\overset{\text{CE}}{=} \mathbb{E}[Y^{a=1}|A=1] - \mathbb{E}[Y^{a=0}|A=0] \\ &\overset{\text{C}}{=} \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]\end{aligned}$$

and the average causal effect (ACE) can be directly estimated from the data.

Standardization & IPTW

Under (C), (CE) and (P):

$$\mathbb{E}[Y^a] = \mathbb{E}_L[\mathbb{E}[Y|L, A = a]] = \mathbb{E}\left[\frac{Y\mathbb{1}_{A=a}}{f(a|L)}\right]$$

Thus, under (C), (CE), (P) and binary A and L the ACE can be estimated from the data using:

- The empirical estimates of $\hat{f}_n(a|L)$ and $\bar{y}_{\text{pseudo},A=a}$, where y_{pseudo} is an observation from the pseudo population created (inverse probability of treatment weighting).
- The empirical estimates of $\hat{f}_n(l)$ and $\bar{y}_{L=l,A=a}$ (standardization).

The two methods use different empirical estimates, thus the estimate of the ACE will be different (though asymptotically equivalent).

Chapter 11

In a setting where $L = (L_1, L_2, ..., L_k)$, and thus 2^k strata, or where A and L are continuous rather than binary, the observed data will not be rich enough to estimate the quantities of interest empirically.

To draw causal inference in those cases we will rely on parametric estimators, e.g.

$$\mathbb{E}[Y|A] = \theta_0 + \theta_1 A,$$

and will suffer from model misspecification additionally to the fundamental problem of causal inference. The nonparametric estimators discussed before are examples of *saturated* models.

Estimating IP weights via modelling

TU/e

Probability for treatment level a given covariates $\boldsymbol{l}$ can be obtained by fitting the appropriate regression model. Finally the estimates of the IP weights equal

$$W^A = \hat{f}(A|\boldsymbol{L})^{-1}.$$

- In the case of a binary A, we model $A|\boldsymbol{L}$ using a logistic regression model:

$$\ln\left(\frac{p_{\boldsymbol{L}=\boldsymbol{l}}}{1-p_{\boldsymbol{L}=\boldsymbol{l}}}\right) = \boldsymbol{\theta}^T\boldsymbol{L}$$

- If A is categorical $A|\boldsymbol{L}$ could be modelled using a categorical or ordinal regression model.
- Otherwise one should specify a continuous model for $A|\boldsymbol{L}$.

While using parametric estimation of IP weights in the presence of zero cells, we are effectively assuming random nonpositivity (Fine point 12.2).

IP weights

TU/e

IP weighting creates a sample of a pseudo-population in which the arrow from the confounders L to the treatment A is removed. The probability of receiving treatment level A does not depend on L. Using the weights W^A:

- The new sample is twice as large as the original sample, $\mathbb{E}[W^A] = 2$, one could use $0.5 \cdot W^A$ instead.
- The probability of receiving treatment level A is equal for everyone in the pseudo-population.

In the case of selection bias (censoring or missing data, C) the causal effect can be estimated under CE for the joint 'treatment' (A, C) conditional on L: $(Y^{a=1,c=0} \perp\!\!\!\perp (A, C)|L)$. To do so one can use IP weights $W^{A,C} = W^A \cdot W^C$, where $W^C = \frac{1}{\mathbb{P}(C=0|L,A)}$.

Stabilized IP weights

TU/e

Alternatively, one could create a pseudo-population in which different people have different probabilities of treatment, as long as the probability of treatment does not depend on the value of L. E.g. a common choice is to use weights $SW^A = \frac{f(A)}{f(A|L)}$, referred to as *stabilized weights*. Using the weights SW^A:

- The original sample size is preserved, $\mathbb{E}[SW^A] = 1$.
- Typically result in narrower 95% confidence intervals (compared to using $0.5W^A$).

Marginal structural models (MSM)

TU/e

Models for the marginal mean of a counterfactual outcome are referred to as *marginal structural mean models*:

$$\mathbb{E}[Y^a] = \beta_0 + \beta_1 a.$$

The parameters for treatment in structural mean models correspond to ACE. For estimation of the ACE we fit the model

$$\mathbb{E}[Y|A] = \theta_0 + \theta_1 A$$

to the sample of the pseudo-population created by using the IP-weights. Now, a consistent estimate $\hat{\theta}_1$ of the associational parameter in the pseudo-population is also a consistent estimator of the causal effect β_1 in the population (proof: $\mathbb{E}[Y|A] \overset{\text{C}}{=} \mathbb{E}[Y^A|A] \overset{\text{CE}}{=} \mathbb{E}[Y^A]$ $\square$).

Marginal structural models (MSM)

TU/e

Following the same reasoning one can come up with different MSM:

- For a dichotomous outcome; $\text{logit}(\mathbb{P}(D^a = 1)) = \beta_0 + \beta_1 a$.
- For a survival outcome (marginal structural Cox model); $\lambda_{T^a} = \lambda_0(t)\exp(\beta_1 a)$, see (Hernan et al., 2000).

MSMs do not include covariates when the target parameter is the average causal effect in the population. However, one may include covariates to assess effect modification:

$$\mathbb{E}[Y^a|V] = \beta_0 + (\beta_1 + \beta_2 V)a + \beta_3 V.$$

Now, $SW^A(V) = \frac{f(A|V)}{f(A|L)}$ generally results in narrower CI compared to using SW^A.

Standardization via modelling

TU/e

$$
\begin{aligned}
\mathbb{E}[Y^a] &= \mathbb{E}[\mathbb{E}[Y^a|L]] \\
&\overset{\text{CE}}{=} \mathbb{E}[\mathbb{E}[Y^a|A=a,L]] \\
&\overset{\text{C}}{=} \mathbb{E}[\mathbb{E}[Y|A=a,L]] \\
&= \int_{l\in\mathcal{L}} \mathbb{E}[Y|A=a,L=l]dF_L(l)
\end{aligned}
$$

In the case of a large number of strata, we cannot obtain meaningful nonparametric stratum-specific estimates of the mean outcome in the treated. We thus model,

$$\mathbb{E}[Y|A=a,L=l],$$

e.g. with a regression model.

Standardization via modelling

TU/e

To estimate $\mathbb{E}[Y^a]$ Robins and Hernán suggest to compute the average of $n^{-1}\hat{\mathbb{E}}[Y|A=a,L]$ empirically from the data.

Necessary additional assumption Chapter 13.3

$$\mathbb{E}[\hat{f}_n(l)] = f(l)$$

If $\mathbb{E}[\hat{f}_n(l)] = \tilde{f}(l) \neq f(l)$, then

$$\begin{aligned}\mathbb{E}[n^{-1}\hat{\mathbb{E}}[Y|A=a,L]] &= \int_{l\in\mathcal{L}} \mathbb{E}[Y|A=a,L=l]\tilde{f}(l)dl \\ &\neq \int_{l\in\mathcal{L}} \mathbb{E}[Y|A=a,L=l]f(l)dl = \mathbb{E}[Y^a]\square.\end{aligned}$$

IP weighting or standardization?

TU/e

IP weighting

- Models $f(a|L)$.

Standardization

- Models $\mathbb{E}[Y|L, A]$.

Doubly robust methods exists that requires a correct model for either treatment $A|L$ or $Y|L, A$, see Technical Point 13.2.

Intermezzo - Parametric g-formula

TU/e

Computing the standardized mean outcome with parametrically estimated conditional means is a particular case of the *parametric g-formula*.

The parametric g-formula density, $f^{G,g=a}(y)$ equals the density factorization of the causal DAG we intervened on. So,

- Leave out terms for treatment variables given their parents (e.g. $\mathbb{P}(A = a|L = l)$).
- Whenever a treatment variable appears as a parent, set it equal to the value specified by a.

Under (C), (CE) and (P) the g-formula has the desired causal interpretation.

Intermezzo - Parametric g-formula - Example

TU/e

Let us consider the causal DAG (applying lemma previous meeting)

$$
\begin{aligned}
L &= f_1(\epsilon_1) \\
A &= f_2(L, \epsilon_2) \\
Y &= f_y(L, A, \epsilon_y)
\end{aligned}
\qquad\qquad
\begin{aligned}
L &= f_1(\epsilon_1) \\
A &= f_2(L, \epsilon_2) \\
Y^a &= f_y(L, a, \epsilon_y)
\end{aligned}
$$

$$
\begin{aligned}
f(y) &= \sum_{l \in \mathcal{L}} f(y, l) &&= \sum_{l \in \mathcal{L}} \sum_{a \in \mathcal{A}} f(y|a, l) f(a|l) f(l) \\
f(y|A = a) &= \sum_{l \in \mathcal{L}} f(y, l|a) &&= \sum_{l \in \mathcal{L}} f(y|a, l) f(l|a) \\
f^{G, g=a}(y) &= \sum_{l \in \mathcal{L}} f^{G, g=a}(y, l) &&= \sum_{l \in \mathcal{L}} f(y|a, l) f(l)
\end{aligned}
$$

Intermezzo - Parametric g-formula - Example

TU/e

Under (C), (CE) and (P) $f^{G,g=a}(y) = f_{Y^a}(y)$.

Proof:
$$
\begin{aligned}
f(Y^a) &= \sum_{L \in \mathcal{L}} f(Y^a, L) \\
&= \sum_{L \in \mathcal{L}} f(Y^a | L) f(L) \\
&\overset{\text{CE}}{=} \sum_{L \in \mathcal{L}} f(Y^a, A = a | L) f(L) \\
&\overset{\text{C}}{=} \sum_{L \in \mathcal{L}} f(Y | A = a, L) f(L) \\
&= f^{G,g=a}(y) \ \square
\end{aligned}
$$

Taking the g-formula mean results in standardization $\mathbb{E}[Y^a] = \mathbb{E}^{G,g=a}[Y]$.

Stratification (Chapter 15.1)

TU/e

Assume

$$\mathbb{E}[Y^a|L] = \beta_0 + \beta_1 a + \beta_2 aL + \beta_3 L.$$

Under (C), (CE) and (P) the parameters of the above structural model can be estimated by fitting the outcome regression model

$$\mathbb{E}[Y|A, L] = \alpha_0 + \alpha_1 a + \alpha_2 aL + \alpha_3 L.$$

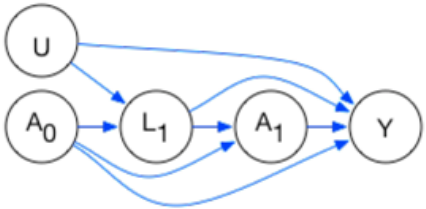
If $\beta_2 = 0$, then $\hat{\beta}_1 = \hat{\alpha}_1$ is an estimate of both the conditional and marginal ACE. In the point-exposure case stratification and standardization are equivalent.

Robins came up with the G-methods (IPTW, G-formula, G-estimation) to deal with time-dependent confounding where stratification fails.

Time-dependent confounding

TU/e

Consider the causal graph:



The difficulty with standard methods (stratification, matching, propensity regression) is that:

1. In order to estimate the joint effects of A_0 and A_1, we must adjust for the confounding effect of L_1 in order to consistently estimate the effect of A_1 on Y.
2. However, if we adjust for confounding by standard methods, we cannot consistently estimate the effect of A_0, because the association of L_1 with A_0 results in selection bias (conditioning on L_1 opens a collider $A_0 \rightarrow L_1 \leftarrow U \rightarrow Y$).

Example

TU/e

A_0	L_1	A_1	N	$\mathbb{E}[Y\|A_0, L_1, A_1]$	SW^A	N_{pseudo}
0	0	0	6000	50	$\frac{1}{2}\frac{8}{6} = \frac{2}{3}$	4000
0	0	1	2000	70	$\frac{1}{2}\frac{8}{2} = 2$	4000
0	1	0	2000	200	$\frac{1}{2}\frac{8}{2} = 2$	4000
0	1	1	6000	220	$\frac{1}{2}\frac{8}{6} = \frac{2}{3}$	4000
1	0	0	3000	230	$\frac{6}{16}\frac{4}{3} = \frac{1}{2}$	1500
1	0	1	1000	250	$\frac{10}{16}\frac{4}{1} = \frac{5}{2}$	2500
1	1	0	3000	130	$\frac{6}{16}\frac{12}{3} = \frac{3}{2}$	4500
1	1	1	9000	110	$\frac{10}{16}\frac{12}{9} = \frac{5}{6}$	7500
			32000			32000

Example

TU/e

What is the ACE of $A_0 = 1$ additional to $A_1 = 1$, i.e. $\mathbb{E}[Y^{(1,1)} - Y^{(0,1)}]$?

Association

$$\begin{aligned}\mathbb{E}[Y^{(1,1)} - Y^{(0,1)}] &= \mathbb{E}[Y^{(1,1)}] - \mathbb{E}[Y^{(0,1)}] \\ &= \frac{1 \cdot 250 + 9 \cdot 110}{1 + 9} - \frac{2 \cdot 70 + 6 \cdot 220}{2 + 6} \\ &= 124 - 182.5 \\ &= -58.5\end{aligned}$$

Stratification

$$\begin{aligned}\mathbb{E}[Y^{(1,1)} - Y^{(0,1)}] &= \sum_{l \in \{0,1\}} \mathbb{E}[Y^{(1,1)} - Y^{(0,1)} | L = l] \cdot \mathbb{P}(L = l) \\ &= (250 - 70)\frac{12}{32} + (110 - 220)\frac{20}{32} \\ &= -1.25\end{aligned}$$

Example

TU/e

What is the ACE of $A_0 = 1$ additional to $A_1 = 1$, i.e. $\mathbb{E}[Y^{(1,1)} - Y^{(0,1)}]$?

Standardization

$$\sum_{l \in \{0,1\}} \mathbb{E}[Y^{(1,1)}|L = l]\mathbb{P}(L = l|A_0 = 1) - \mathbb{E}[Y^{(0,1)}|L = l]\mathbb{P}(L = l|A_0 = 0)$$
$$= (110 \cdot \tfrac{3}{4} + 250 \cdot \tfrac{1}{4}) - (220 \cdot \tfrac{1}{2} + 70 \cdot \tfrac{1}{2})$$
$$= 145 - 145$$

IP Weighting, $SW^A = \frac{P(A_0)}{P(A_0)} \frac{P(A_1|A_0)}{P(A_1|A_0,L)}$, see (Robins et al., 2000)

$$\begin{aligned}\mathbb{E}[Y^{(1,1)} - Y^{(0,1)}] &= \mathbb{E}[Y^{(1,1)}] - \mathbb{E}[Y^{(1,0)}] \\ &= \frac{2.5 \cdot 250 + 7.5 \cdot 110}{2.5 + 7.5} - \frac{4 \cdot 70 + 4 \cdot 220}{4 + 4} \\ &= 145 - 145\end{aligned}$$

Structural nested models (SNM)

Under (CE) we can define the *structural nested mean model*

$$\mathbb{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aL$$

Structural nested models are semiparametric because they are agnostic about both the intercept and the main effect of L, i.e. we are not interested in estimating $\mathbb{E}[Y^{a=0}|L = l] = \beta_0 + \beta_3 L$ (the mean counterfactual outcome under no treatment in stratum $L = l$).

Rank preservation

When the effect of treatment A on the outcome Y is exactly the same, on the additive scale, for all individuals with the same values of L, in the study population, we say that *conditional additive rank preservation* holds. An example is

$$Y_i^a - Y_i^{a=0} = \psi_1 a + \psi_2 a L_i, \text{ for all subjects } i.$$

The assumption of constant treatment effect for all individuals with the same value L is not applicable in practice, but makes it easier to introduce *g-estimation*. The procedure of g-estimation is actually the same for rank-preserving and non-rank-preserving models.

G-estimation

Let

$$\mathbb{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a.$$

More specifically, let $\psi_1 = \beta_1$ and assume a rank-preserving model

$$Y^{a=0} = Y^a - \psi_1 a \overset{\text{C}}{=} Y - \psi_1 A.$$

Define

$$H(\psi^\dagger) = Y - \psi^\dagger A.$$

Note $H(\psi_1) = Y^{a=0}$. To estimate ψ_1 in practice we fit separate logistic regression models

$$\text{logit}\left(\mathbb{P}(A = 1|H(\psi^\dagger), L)\right) = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 L,$$

for a grid of values of $\psi^\dagger$. Under (CE), $\mathbb{P}(A = 1|Y^0, L) = \mathbb{P}(A = 1|L)$, thus the candidate $H(\psi^\dagger)$ with $\alpha_1 = 0$ is the counterfactual $Y^{a=0}$, and $\psi^\dagger$ is the estimate for ψ_1.

G-estimation

The g-estimation algorithm for ψ_1 produces a consistent estimate of the parameter β_1 of the mean model, assuming the mean model is correctly specified (that is, if the average treatment effect is equal in all levels of L).
Let

$$\mathbb{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a.$$

This is true regardless of whether the individual treatment effect is constant (rank-preserving model). In other words, the validity of the g-estimation algorithm does not actually require that $H(\beta_1) = Y^{a=0}$ for all subjects, but requires that $\mathbb{E}[H(\beta_1)|L] = \mathbb{E}[Y^{a=0}|L]$.

G-estimation + discussion

TU/e

If the ACE vary across strata of L the SNM will be misspecified and causal inference will be wrong. This is in contrast with MSMs, which are not misspecified if we fail to add terms $\beta_2 aV$ and $\beta_3 V$ even if there is effect modification by V, because MSMs that do not condition on V estimate the ACE in the population. SNMs estimate, by definition, the ACE within levels of the confounders L.

Now assume

$$\mathbb{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aV, \text{ where V} \subseteq \text{L}.$$

$$H(\psi_1^\dagger, \psi_2^\dagger) = Y - \psi_1^\dagger A - \psi_2^\dagger AV$$

$$\text{logit}\left(\mathbb{P}(A = 1|H(\psi_1^\dagger, \psi_2^\dagger), L)\right) = \alpha_0 + \alpha_1 H(\psi_1^\dagger, \psi_2^\dagger) + \alpha_2 H(\psi_1^\dagger, \psi_2^\dagger)V + \alpha_3 L,$$

Now, we need to search for a combination of values $\psi_1^\dagger$ and $\psi_2^\dagger$ that make both α_1 and α_2 equal to zero. If $\mathbb{E}[H(\beta_1, \beta_2)|L] = \mathbb{E}[Y^{a=0}|L]$, then $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are consistent estimates for β_1 and β_2.