Statistical Learning
Causal Inference for Complex Longitudinal Data

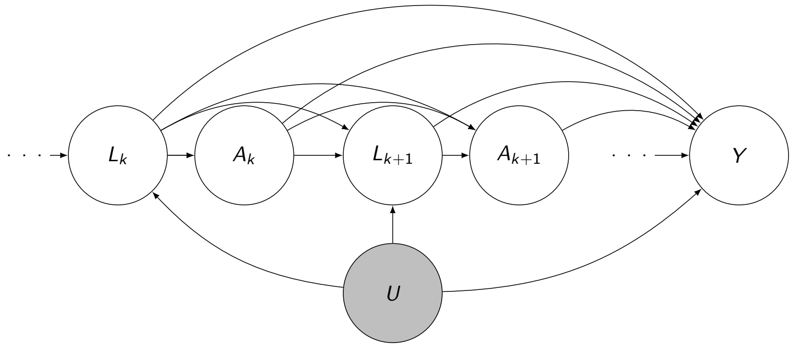
Setup

- Consider the effect of a time-varying dichotomous treatment on a continuous outcome Y measured at the end of study at time $K+1$.
- Treatment history $\overline{A} = \overline{A}_K$, where $\overline{A}_k = (A_0, A_1, \ldots, A_k)$.
- Covariates history $\overline{L} = \overline{L}_K$, where $\overline{L}_k = (L_0, L_1, \ldots, L_k)$.

Identifying assumptions

- **Consistency**: If $\overline{A} = \overline{a}$, then $Y = Y^{\overline{a}}$.
- **Sequential randomization**: $Y^{\overline{a}} \perp A_k | \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k$ for all possible values $\overline{a}$.
- **Positivity**: If $f_{\overline{A}_{k-1},\overline{L}_k}\left(\overline{a}_{k-1}, \overline{l}_k\right) \neq 0$, then we have $f_{A_k|\overline{A}_{k-1},\overline{L}_k}\left(a_k|\overline{a}_{k-1}, \overline{l}_k\right) \neq 0$ for all a_k.

Causal DAGs



g-formula

Under the three assumptions, with covariate L_t discrete, the g-formula for $E\left(Y^{\overline{a}}\right)$ is

$$E\left(Y^{\overline{a}}\right) = \sum_{\overline{l}} E\left(Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}\right) \prod_{k=0}^{K} f\left(l_k|\overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}\right)$$

When covariate L_t is continuous, the sum is replaced by integral in the g-formula.

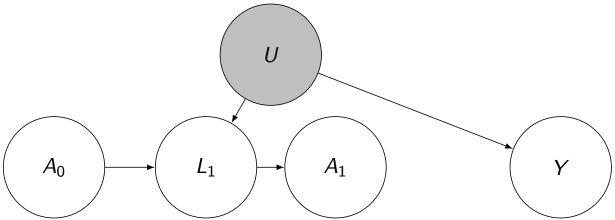
Proof

Take 2 stages as an example.

$$
\begin{aligned}
& E(Y^{a_0,a_1}) \\
= & \int E(Y^{a_0,a_1} | L_0 = l_0) f_{L_0}(l_0) \mathrm{d} l_0 \\
= & \int E(Y^{a_0,a_1} | L_0 = l_0, A_0 = a_0) f_{L_0}(l_0) \mathrm{d} l_0 \\
= & \int \int E(Y^{a_0,a_1} | L_0 = l_0, A_0 = a_0, L_1 = l_1) \times \\
& f_{L_1|A_0,L_0}(l_1 | A_0 = a_0, L_0 = l_0) f_{L_0}(l_0) \mathrm{d} l_0 \mathrm{d} l_1 \\
= & \int \int E(Y | L_0 = l_0, A_0 = a_0, L_1 = l_1, A_1 = a_1) \times \\
& f_{L_1|A_0,L_0}(l_1 | A_0 = a_0, L_0 = l_0) f_{L_0}(l_0) \mathrm{d} l_0 \mathrm{d} l_1
\end{aligned}
$$

Null paradox

The main issue of specifying models of g-formula is that under standard parametrization, there is no parameter to encode the null hypothesis of no joint effect of (a_0, a_1).



Null paradox

- Suppose that L_1 is binary and Y is continuous, so that the g-formula in this graph gives

$$E(Y^{a_0,a_1}) = \sum_{l_1=0}^{1} E(Y|A_0 = a_0, A_1 = a_1, L_1 = l_1) \times f_{L_1|A_0}(l_1|A_0 = a_0)$$

- A standard modeling approach would fit a linear regression

$$E(Y|A_0 = a_0, A_1 = a_1, L_1 = l_1; \gamma) = (1, a_0, a_1, l_1)\gamma$$

and a logistic regression

$$\text{logit}\{f(L_1 = 1|A_0 = a_0; \alpha)\} = (1, a_0)\alpha$$

where $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)^\top$ and $\alpha = (\alpha_0, \alpha_1)^\top$.

Null paradox

- The counterfactual mean is

$$
\begin{aligned}
& E(Y^{a_0,a_1}) \\
& = \sum_{l_1=0}^{1} E(Y|A_0 = a_0, A_1 = a_1, L_1 = l_1; \gamma) \times f(l_1|A_0 = a_0; \alpha) \\
& = \left(1, a_0, a_1, \frac{\exp((1, a_0)\alpha)}{1 + \exp((1, a_0)\alpha)}\right) \gamma
\end{aligned}
$$

Null paradox

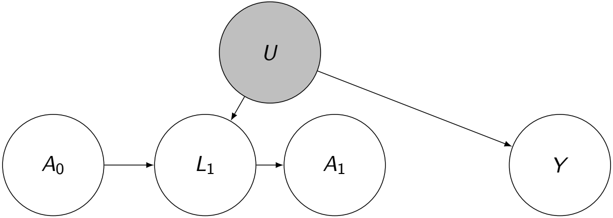
- The counterfactual mean is

$$
\begin{aligned}
&E(Y^{a_0,a_1}) \\
&= \sum_{l_1=0}^{1} E(Y|A_0 = a_0, A_1 = a_1, L_1 = l_1; \gamma) \times f(l_1|A_0 = a_0; \alpha) \\
&= \left(1, a_0, a_1, \frac{\exp((1, a_0)\alpha)}{1 + \exp((1, a_0)\alpha)}\right) \gamma
\end{aligned}
$$

- $E(Y^{a_0,a_1})$ does not depend on (a_0, a_1) if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_1 = 0$

Null paradox

- Recall that



$$E(Y|A_0 = a_0, A_1 = a_1, L_1 = l_1; \gamma) = (1, a_0, a_1, l_1)\gamma$$
$$\text{logit}\{f(L_1 = 1|A_0 = a_0; \alpha)\} = (1, a_0)\alpha$$

- We have

$$\gamma_3 \neq 0$$
$$\alpha_1 \neq 0$$

Marginal structural models

- An alternative is to directly specify a model for the marginal mean $E(Y^{a_0,a_1};\psi)$ with finite dimensional parameter ψ.
  For example, we could specify

$$
\begin{aligned}
E(Y^{a_0,a_1};\psi) &= \psi_0 + \psi_1 a_0 + \psi_2 a_1 \text{ or}\\
E(Y^{a_0,a_1};\psi) &= \psi_0 + \psi_1 (a_0 + a_1) \text{ or}\\
E(Y^{a_0,a_1};\psi) &= \psi_0 + \psi_1 a_1
\end{aligned}
$$

- Note here that ψ has a causal interpretation, it is the parameter of a Marginal Structural Mean Model (MSMM)

Marginal structural models

- MSMs focus on $E\left(Y^{\overline{A}=\overline{a}}\right)$ for all possible values $\overline{a}$.
- It is said that the time-varying treatment has a causal effect on the average value of Y if $E\left(Y^{\overline{a}}\right) - E\left(Y^{\overline{a}'}\right) \neq 0$ for at least two values $\overline{a}, \overline{a}'$.
- For the first example on the previous page, $E(Y^{a_0,a_1};\psi) = E(Y^{0,0};\psi) \Longleftrightarrow \psi_1 = \psi_2 = 0$.

Mean MSMs

Under the three assumptions, for the following mean MSM,

$$E(Y^{\overline{a}}) = \mu_\psi(\overline{a}),$$

we have

$$E\left[h(\overline{A})(Y - \mu_\psi(\overline{A}))/W\right] = 0,$$

where

$$W = \prod_{k=0}^{K} W_k = \prod_{k=0}^{K} \frac{f\left(A_k|\overline{L}_k, \overline{A}_{k-1}\right)}{f^\star\left(A_k|\overline{A}_{k-1}\right)}.$$

IPTW

Under the three assumptions, the IPTW formula for $E\left(Y^{\overline{a}}\right)$ is the mean of Y among the subset $(\overline{A} = \overline{a})$ in a pseudo-population, constructed by weighting each subject by their subject-specific weights

$$W = \prod_{k=0}^{K} f\left(A_k | \overline{A}_{k-1}, \overline{L}_k\right)$$

or stabilized weights

$$SW = \prod_{k=0}^{K} \frac{f\left(A_k | \overline{A}_{k-1}, \overline{L}_k\right)}{f^{\star}\left(A_k | \overline{A}_{k-1}\right)}$$

IPTW

IPTW creates a pseudo-population, in which

- the mean of $Y^{\overline{a}}$ is identical to that in the true population
- the treatment at each time t depends at most on past treatment history.

The difference is that in the unstabilized pseudo-population $P_{ps}\left(A_k = 1|\overline{A}_{k-1}, \overline{L}_k\right) = \frac{1}{2}$, while in the stabilized pseudo-population $P_{ps}\left(A_k = 1|\overline{A}_{k-1}, \overline{L}_k\right)$ is equal to $P\left(A_k = 1|\overline{A}_{k-1}\right)$ in the true population, where the subscript ps refers to the pseudo-population. Hence, $E\left(Y^{\overline{a}}\right)$ in the true population is $E_{ps}\left(Y|\overline{A} = \overline{a}\right)$.

Procedure

- Specify models for $f(A_0|L_0)$ and $f(A_1|L_1, A_0, L_0)$; say logistic regressions and obtain the MLEs $\hat{\alpha}_0$ and $\hat{\alpha}_1$

- For each person in the study, compute the weight $\hat{W} = \hat{f}(A_0|L_0; \hat{\alpha}_0)\hat{f}(A_1|L_1, A_0, L_0; \hat{\alpha}_1)$ which corresponds to the estimated probability of receiving the treatment you did indeed receive

- Regress Y on A_0 and A_1 using weighted least-squares with weights $\hat{W}^{-1}$

Proof sketch

$$
\begin{aligned}
& E\left(SW^{-1}h\left(A_0, A_1\right)Y\right) \\
& = E\left(SW^{-1}h\left(A_0, A_1\right)E\left(Y \mid A_0, A_1, L_0, L_1\right)\right) \\
& = \sum_{l_0, l_1, a_0, a_1} SW^{-1}h\left(a_0, a_1\right)E\left(Y \mid a_0, a_1, l_0, l_1\right) \\
& \times f\left(a_1 \mid l_1, a_0, l_0\right) f\left(l_1 \mid l_0, a_0\right) f\left(a_0 \mid l_0\right) f\left(l_0\right) \\
& = \sum_{l_0, l_1, a_0, a_1} \frac{f^\star\left(a_0\right) f^\star\left(a_1 \mid a_0\right)}{f\left(a_0 \mid l_0\right) f\left(a_1 \mid l_1, a_0, l_0\right)} h\left(a_0, a_1\right) E\left(Y \mid a_0, a_1, l_0, l_1\right) \\
& \times f\left(a_1 \mid l_1, a_0, l_0\right) f\left(l_1 \mid l_0, a_0\right) f\left(a_0 \mid l_0\right) f\left(l_0\right) \\
& = \sum_{a_0, a_1} f^\star\left(a_0\right) f^\star\left(a_1 \mid a_0\right) \sum_{l_0, l_1} E\left(Y \mid a_0, a_1, l_0, l_1\right) h\left(a_0, a_1\right) \\
& \times f\left(l_1 \mid l_0, a_0\right) f\left(l_0\right) \\
& = \sum_{a_0, a_1} f^\star\left(a_0\right) f^\star\left(a_1 \mid a_0\right) E\left(Y^{a_0, a_1}\right) h\left(a_0, a_1\right) \\
& = E(SW^{-1}h(A_0, A_1)E[Y^{A_0, A_1}])
\end{aligned}
$$

Censoring

For time k, denote lost-to-follow-up by $C_k = 0$, follow-up by $C_k = 1$. The stabilized weight for each subject is $SW^+ \times SW$, where

$$SW^+ = \prod_{k=0}^{K} \frac{Pr\left(C_k = 1 | \overline{C}_{k-1} = \overline{1}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \overline{l}_k\right)}{Pr\left(C_k = 1 | \overline{C}_{k-1} = \overline{1}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}\right)}$$