Introduction to Causal Inference with Longitudinal Data

Kuan Liu

June 7, 2022

ICES Western Research Seminar

Institute of Health Policy, Management and Evaluation
UNIVERSITY OF TORONTO

Dalla Lana
School of Public Health

Outline

Fundamentals of Causal Inference

Causal inference with longitudinal data

Implementation in R with Simulated Data - Live Demo

Additional Topics

Fundamentals of Causal Inference

Observational Study

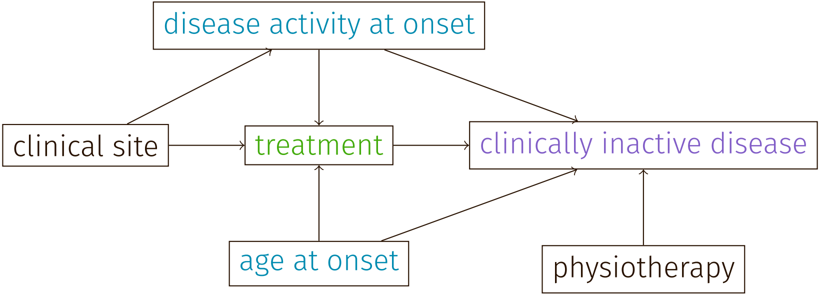
- Clinicians and other care providers need to make informed health care decisions with causal knowledge about treatment/intervention effectiveness and safety.
- Randomized controlled trials are the gold standard approach but not always feasible.

Observational Study

- Clinicians and other care providers need to make informed health care decisions with causal knowledge about treatment/intervention effectiveness and safety.
- Randomized controlled trials are the gold standard approach but not always feasible.
- Observational studies becomes invaluable!
- Under an observational setting, treatment assignment mechanism is unknown and key issues like confounding by indication often arise.
- Statistical causal inference methods are required to gather evidence from observational study.

Fundamentals of Causal inference

- Causal diagram is an essential graphical tool to specify the hypothesized causal mechanism for causal modelling
- Directed Acyclic Graph (DAG)[1] and Single World Intervention Graph (SWIG)[2]
- minimally sufficient set of confounders



[1] ▸ Greenland et al (1999)

[2] ▸ Richardson and Robins (2013)

Causal Inference Frameworks

- Potential outcomes[3]
- Decision-theoretic data-generating framework [4]
- Graphic models [5]
- Structural equation modelling [6]

[3] ▸ Neyman (1923) & Rubin (1974, 1977, 1978, 1983, 2005)

[4] ▸ Dawid et al (2010)

[5] ▸ Pearl (2014)

[6] ▸ Bollen and Pearl (2013)

Potential Outcome Framework

- Post-treatment outcomes under all possible treatment options, although only one of these is observed.
- Most widely used: adopted for both frequentist and Bayesian approaches
- Treatment options indexed by a; The potential outcomes is denoted as Y^a for treatment $Z = a$.

Assumptions

- Stable unit treatment value assumption (no interference & consistency)
- Strongly ignorable treatment assignment (no unmeasured confounding & positivity)

Causal Estimands

- Average treatment effect (ATE)

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

- Average treatment effect for the treated (ATT)

$$\tau^{ATT} = E[Y_i(1) - Y_i(0) \mid Z_i = 1]$$

- Conditional average treatment effect (CATE)

$$\tau^{CATE} = E[Y_i(1) - Y_i(0) \mid X_i = x]$$

Causal Methods

Some frequentist methods

- Propensity score analysis (PSA) ▸ Rosenbaum and Rubin (1983)
- Marginal structural models (MSM) ▸ Robins et al (2000)
- Generalized PSA ▸ Hirano and Imbens (2004)
- G-computation ▸ Robins (1986)
- Targeted maximum likelihood estimation (TMLE) ▸ Van Der Laan and Rubin (2006)

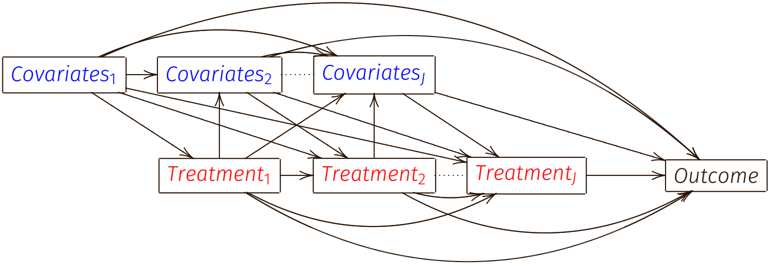
Some Bayesian methods

- Bayesian PSA ▸ McCandless et al (2009) ▸ Zigler (2016)
- Bayesian MSM ▸ Saarela et al (2015) ▸ Liu et al (2020)
- Bayesian g-computation ▸ Keil et al (2018)
- Bayesian non-parametric methods (BART, Dirichlet & Gaussian process) ▸ Hill (2011) ▸ Roy et al (2018)

Causal inference with longitudinal data

Longitudinal DAG

Figure 1: Longitudinal DAG with visits indexed by j, $j = 1, ..., J$.



Notations

- A longitudinal observational study with n subjects indexed by i, $i = 1, \ldots, n$ and J number of visits indexed by j, $j = 1, \ldots, J$.
- Y_i, X_{ij} and Z_{ij} are random variables representing an end-of-study response, covariates and the treatment for individual i at visit j.
- History up to visit j are denoted as $\bar{X}_{ij} = \{X_{i1}, \ldots, X_{ij}\}$ and $\bar{Z}_{ij} = \{Z_{i1}, \ldots, Z_{ij}\}$.
- We assume at each visit j, X_{ij} are measured first and treatment assignment Z_{ij} is decided after Figure 1.

Dataset visualization

Example longitudinal dataset with two visits and a binary time-varying treatment.

i	x_1	z_1	x_2	z_2	y	y^{00}	y^{01}	y^{10}	y^{11}
1	1	1	0	0	25	-	-	25	-
2	0	0	0	1	23	-	23	-	-
...	...								
n	0	1	1	1	15	-	-	-	15

Causal Framework and Assumptions

- For each unique treatment sequence $\bar{a} = (a_1, \ldots, a_k)$ in set $\mathcal{A}$, we have a corresponding potential outcome for subject i, $Y_i^{\bar{a}}$.
- consistency: $Y_i = \sum_{\bar{a} \in \mathcal{A}} I_{\{\bar{Z}_i = \bar{a}\}} Y^{\bar{a}}$
- "no unmeasured confounders" (sequential randomization given measured covariates):

$$Y^{\bar{a}} \perp Z_j \mid \bar{X}_j, \bar{Z}_{j-1}, \text{for } j = 1, \ldots, J.$$

- positivity: at each visit, any treatment sequence that is compatible with the complete treatment history has a non-zero probability of occurring.
- causal parameter of interest: Average potential outcome, $E(Y_i^{\bar{a}})$

Frequentist Methods for Longitudinal Data

Methods	*Estimation*
MSMs	Feature time-dependent treatment assignment model and IPTW, relay on the correct specification of the treatment model
G-computation	Feature outcome and covariates models, relay on the correct specification of the outcome model
TMLE	Doubly robust method which considers all three models, potential estimation issues with small sample

TMLE: targeted maximum likelihood estimation

MSMs: marginal structural models

MSMs

- It's called "marginal" since the method models the expectation of the potential outcome, $E(Y_i^{\bar{a}})$, without directly conditioning on any covariates!

$$g_Y(E[Y_i^{\bar{a}}]) = h_Z(\bar{Z}_i = \bar{a})\theta,$$

  where $g_y(\cdot)$ and $h_z(\cdot)$ represent the canonical link function and the treatment history function, respectively.
- In practice, we often use cumulative treatment exposure to formulate the marginal model,

$$g_Y(E[Y_i^{\bar{a}}]) = \theta_0 + \sum_j \theta_j Z_j.$$

MSMs

- It's called "marginal" since the method models the expectation of the potential outcome, $E(Y_i^{\bar{a}})$, without directly conditioning on any covariates!

$$g_Y(E[Y_i^{\bar{a}}]) = h_Z(\bar{Z}_i = \bar{a})\theta,$$

where $g_y(\cdot)$ and $h_z(\cdot)$ represent the canonical link function and the treatment history function, respectively.

- In practice, we often use cumulative treatment exposure to formulate the marginal model,

$$g_Y(E[Y_i^{\bar{a}}]) = \theta_0 + \sum_j \theta_j Z_j.$$

- **what about time-dependent confounders?**

MSMs

- MSMs use inverse-probability-of-treatment weighting (IPTW) to adjust for time-dependent confounding.
- The conventional IPT weight (unstabilized) is defined as,

$$w_i = \left\{ \prod_{j=1}^{J} P(Z_{ij} = a_j \mid \bar{Z}_{ij-1} = \bar{a}_{j-1}, \bar{X}_{ij} = \bar{X}_{ij}). \right\}^{-1},$$

  and the corresponding IPT stabilized weight is defined as

$$sw_i = w_i \times \prod_{j=1}^{J} P(Z_{ij} = a_j \mid \bar{Z}_{ij-1} = \bar{a}_{j-1}).$$

- Weights are calculated by fitting a series of time-specific treatment models (i.e., propensity score models).
- Given the estimated weights, a weighted glm is then used to fit the marginal model to obtain θ and $E[Y_i^{\bar{a}}]$.

MSMs

- Stabilized weights are less variable and favoured in the longitudinal cases
- Weight truncation is a common approach to deal with extreme stabilized weights, however, be aware of the ”variance-bias” trade-off.

MSMs

- Stabilized weights are less variable and favoured in the longitudinal cases
- Weight truncation is a common approach to deal with extreme stabilized weights, however, be aware of the ”variance-bias” trade-off.
- Many authors have pointed out that MSM is highly sensitive to misspecified treatment model.
- “Kitchen-sink” is not always the answer, e.g., risk of efficiency lost due to inclusion of instrumental variables.
- Nonetheless, due to its resemblance to standard regression models, MSM is a widely used.

G-computation

- The distribution of $Y^{\bar{a}}$ can be identified via the g-computation algorithm (Robins, 1986)

$$
\begin{aligned}
E[Y_i^{\bar{a}}] &= \int_{X_{i1}} E[Y_i^{\bar{a}} \mid X_{i1}] p(X_{i1}) dX_{i1} \\
&= \int_{X_{i1}} E[Y_i^{\bar{a}} \mid X_{i1}, Z_{i1} = a_1] p(X_{i1}) dX_{i1}, \text{ by sequential randomization} \\
&= \int_{X_{i2}} \int_{X_{i1}} E[Y_i^{\bar{a}} \mid X_{i1}, X_{i2}, Z_1 = a_1, Z_2 = a_2] p(X_{i1}) p(X_{i2} \mid X_{i1}, Z_1 = a_1) dX_{i1} dX_{i2} \\
&= \ldots \\
&= \int_{X_{i1}} \ldots \int_{X_{ik}} E[Y_i^{\bar{a}} \mid \bar{X}_i, \bar{Z}_i = \bar{a}] p(X_{ik} \mid \bar{X}_{ik-1}, \bar{Z}_{ik-1} = \bar{a}_{k-1}) \\
&\qquad \ldots p(X_{i2} \mid \bar{X}_{i1}, Z_{i1} = a_1) p(X_{i1})\; dX_{ik} \ldots dX_{i1} \\
&= \int_{X_{i1}} \ldots \int_{X_{ik}} E[Y_i \mid \bar{X}_i, \bar{Z}_i = \bar{a}] \prod_{j=1}^{k} p(X_{ij} \mid \bar{X}_{ij-1}, \bar{Z}_{ij-1} = \bar{a}_{k-1})\; dX_{ik} \ldots dX_{i1}
\end{aligned}
$$

G-computation

- Time-dependent treatment models are not featured!
- We estimate $P(Y_i \mid \bar{X}_i, \bar{Z}_i = \bar{a})$ by specifying a parametric outcome model and use an iterative Monte Carlo simulation approach to estimate $P(X_{ij} \mid \cdot)$
- Computation steps
    1. For a given treatment sequence of interest $\bar{a}$, we first generate $X_{i1} \ldots X_{i_k}, i = 1, \ldots, n$ in visit order (can be parametrically or non-parametrically).
    2. Simulate $Y_i, i = 1, \ldots, n$ given the generated $\bar{X}_i$ and $\bar{a}$ following the pre-specified parametric outcome model.
    3. Taking the mean across the simulated outcomes yields an estimate for $E[Y_i^{\bar{a}}]$.

G-computation

- “g-null paradox”: if there are unmeasured variables that affect a time-dependent confounder and the outcome, then it’s increasingly probable as sample size increases that the null hypothesis of no treatment effect is rejected, even when the null hypothesis holds (Robins, 2013)
- G-computation is computationally expensive compared to other methods.
- Many Bayesian causal methods adopt G-computation algorithm - very useful in Bayesian causal literature.

TMLE

- Targeted estimation is a semiparametric statistical approach to estimate parameters with favourable asymptotic properties.
- TMLE proceeds in two steps:
  (i) direct initial estimation of the distribution of the data (e.g., via maximum likelihood, or a machine learning algorithm),
  (ii) and a bias-reducing/targeting step in which the initial fit is updated to produce a substitution estimator with reduced bias for the parameter of interest (Van der Laan, 2006 & Van der Laan 2011)

TMLE

- A “doubly robust” estimator such that consistent estimation of the average potential outcome can be achieved if at least one of the outcome model or the treatment assignment model is correctly specified.
- TMLE has the advantage over MSMs and g-computation due to its double robustness.

TMLE - simple demo (1)

- Consider a simple causal study with a binary, time-varying treatment assignment
- we are interested in estimating the causal contrast between always treated, $\bar{a}^1 = (1, \ldots, 1)$, and never treated, $\bar{a}^0 = (0, \ldots, 0)$.
- The initial causal estimator is defined as

$$\hat{\phi} = \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[Y_i \mid \bar{Z}_i = \bar{a}^1, \bar{X}_i] - \hat{E}[Y_i \mid \bar{Z}_i = \bar{a}^0, \bar{X}_i]).$$

TMLE - simple demo (2)

1. getting an initial fit for $E[Y_i \mid \bar{Z}_i = \bar{a}, \bar{X}_i]$ by minimizing some global loss function w.r.t $P(Y_i \mid \bar{Z}_i, \bar{X}_i)$
   - not restricted to use parametric models, data-adaptive learning methods can be applied (i.e., Super Learner)
2. we tune the initial fit $\hat{E}[Y_i \mid \bar{Z}_i = \bar{a}, \bar{X}_i]$ by an estimated factor $\hat{\epsilon}\,\hat{f}(\bar{Z}_i = \bar{a}, \bar{X}_i)$, where $\hat{f}(\bar{Z}_i = \bar{a}, \bar{X}_i)$ depends on the treatment model and is related to the efficient influence function of ϕ.
   - ϵ ("error" term), can be interpreted as the amount of residual confounding after initial fit.
   - we can fit $y_i = \hat{E}[Y_i \mid \bar{Z}_i = \bar{a}, \bar{X}_i] + \hat{\epsilon}\,\hat{f}(\bar{Z}_i = \bar{a}, \bar{X}_i)$ to obtain ϵ
3. through an iterative estimation process, at convergence of ϵ at 0, we solve for $\hat{\phi}$, the causal parameter.

TMLE

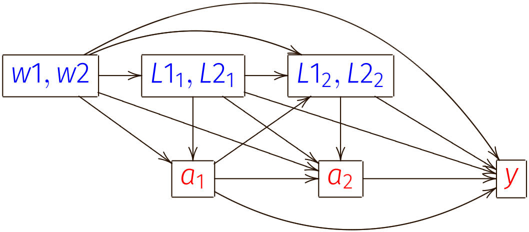
- Longitudinal TMLE is also computationally intensive.
- A small sample may limit the adoption of learning-based approaches
    - when faced with multiple time-dependent confounders, convergence of the fitting algorithms may not work!
- The use of TMLE with longitudinal data in applications is growing.

Implementation in R with Simulated Data - Live Demo

Simulated Dataset

- 1000 patients and 3 visits
- y, an end-of-study continuous outcome;
- A, a binary treatment assignment;
- w1 and w2 are two baseline covariates (one continuous, one binary) i.e., age, sex;
- L1 and L2 are two time-dependent covariates (one continuous, one binary);
- no missing data

Simulated DAG



Summary

- All three methods returned similar estimates.
- ltmle package can be used to fit g-computation, however, we suggest using the gfoRmula package for g-computation.
- Although there are potential computational efficiency issues when using "kitchen-sink" approach to specific the treatment model and the conditional outcome model, it's still preferred when there are manageable number of time-varying covariates.
- If you suspect non-linear relationships between regressors and predictor and the sample size is sufficient ($n \geq 500$), ltmle with SuperLearn is the method to use.

Additional Topics

Missing data

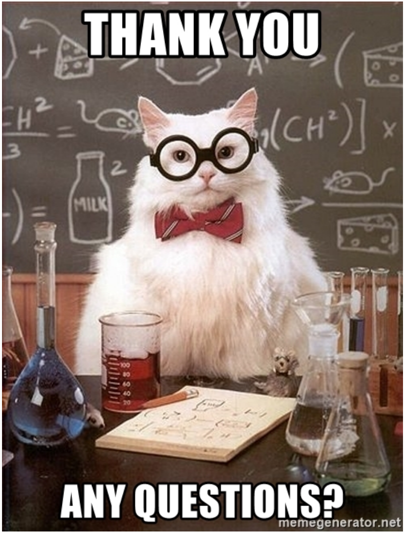
- Right-censoring
    - Under right-censoring at random conditionally on the observed history, we can model the time-specific probability of right-censoring via logistic regressions (upon creating visit-specific censoring indicators).
    - The estimated time-specific censoring probability can be incorporated in the IPTW weights and used in both MSMs and TMLE.

Missing data

- Right-censoring
  - Under right-censoring at random conditionally on the observed history, we can model the time-specific probability of right-censoring via logistic regressions (upon creating visit-specific censoring indicators).
  - The estimated time-specific censoring probability can be incorporated in the IPTW weights and used in both MSMs and TMLE.
- Intermittent censoring
  - Under censoring at random given observed data, we can use multiple imputation to predict missing data.
  - We estimate average potential outcome for each imputed dataset and use Rubin's formula to pool these estimates.
  - Past simulation study has shown promising performance if imputation model include both treatment and outcome.

Bayesian methods

- Bayesian MSMs and Bayesian g-computation are two existing Bayesian methods that can handle longitudinal data
- Bayesian causal methods can provide probabilistic summary that aids scientific conclusion (over P-value)
  1. Actionable knowledge for decision-marking
  2. ”Absence of evidence is not evidence of absence”
- A flexible modelling framework to accommodate unmeasured confounding (i.e., adding bias parameter as a random variable into the outcome model)
- and small sample size using weakly informative priors on model parameters to improve precision.


THANK YOU
ANY QUESTIONS?
memegenerator.net