LONGITUDINAL CAUSAL INFERENCE

Consider a sequence of treatments over time. The goal is to compare the effect of different sequences on the outcomes of interest. See Hernan and Robins (2018) and the chapter by Hernan and Robins (2009).

e.g.

1. in neuroimaging, measurements Y of neural activity are taken at periods $t = 1, \ldots, T$. During each period, a stimulus Z may or may not be given.
2. in medicine, a researcher may wish to test if a medicine Z with nasty side effects administered every day t yields better outcomes Y than administering the medication every other day. Or the researcher might want to know if administering a medication only days when a patient's symptoms were worse than those of the previous day produces the same outcome the day after as versus administering the medication every day.
3. a company that sends its clients promotional material might want to know whether more frequent mailings are more effective than less frequent advertising.

LONGITUDINAL CAUSAL INFERENCE

We consider one subject and suppress the subscript i. The subject is observed at time periods $t = 1, \ldots, T$.

Let

1. $Z_t = 1$ if the subject is assigned to treatment in period t, 0 otherwise.
2. $\bar{Z}_T = (Z_1, \ldots, Z_T)$ denote the assigned treatment regimen.
3. $\bar{Z}_t = (Z_1, \ldots, Z_t)$ denote the sub-regimen of $\bar{Z}_T$ consisting of assignments through period t, $\bar{z}_T = (z_1, \ldots, z_T) \in \Omega_z \subseteq \{0, 1\}^T$.
4. $Y_t(\bar{z}_T)$ the potential outcome under treatment regimen $\bar{z}_T$, measured at the end of period t.

We assume $Y_t(\bar{z}_T)$ depends only on treatments administered prior to the time of measurement and write $Y_t(\bar{z}_T)$ as $Y_t(\bar{z}_t)$. We will focus on the case where Y is continuous.

Primary interest is in averages of unit differences: $Y_t(\bar{z}_t) - Y_t(\bar{z}_t^*)$.

For each t, we consider the average effect of treatment regimen $\bar{z}_T$ vs. $\bar{z}_T^*$, given covariates $\mathbf{X}_1$ observed prior to period 1 treatment,

$$E(Y_t(\bar{z}_t) - Y_t(\bar{z}_t^*) \mid \mathbf{X}_1)$$

and the average effect,

$$E(Y_t(\bar{z}_t) - Y_t(\bar{z}_t^*)).$$

LONGITUDINAL CAUSAL INFERENCE

e.g. in neuroimaging research, subjects are assigned to treatment regimens in Ω_z. The question is whether the same outcome can be attained when patients are given medication every day vs. every other day.

The researcher might randomly assign half the subjects to receive daily treatment and half to receive the treatment every other day. In this case, inference proceeds as in previous lessons.

However, suppose the assignment of medication on day t depends on day t symptoms. If the outcomes, $Y_t(\bar{z}_t)$, also depend on day t symptoms, day t symptoms are a confounder. Adjustment is necessary.

Day t symptoms are also an outcome of previous assignments and symptoms. Adjusting for these intermediate outcomes will generally create comparisons that are not causal.

LONGITUDINAL CAUSAL INFERENCE

Fortunately, it is possible to extend the unconfoundedness conditions developed in part one to cover "time varying confounding".

Let $\mathbf{X}_t$ denote measurements of the time varying confounders taken in period t. $\mathbf{X}_t$ may or may not include the outcome Y_{t-1} in period $t-1$.

Assume that in each period t, measurements $\mathbf{X}_t$ are first recorded, then treatment Z_t is given, followed by measurement of Y_t. The cumulative history is denoted $(\mathbf{X}_1, ...\mathbf{X}_t) \equiv \bar{\mathbf{X}}_t$.

LONGITUDINAL CAUSAL INFERENCE

Identification requires extending the assumption (Rosenbaum and Rubin 1983) that treatment assignment is strongly ignorable, given covariates.

$$Y(0), Y(1) \perp\!\!\!\perp Z \mid \mathbf{X}_1, \quad 0 < \Pr(Z = 1 \mid \mathbf{X}_1) < 1.$$

Assume (Robins and Hernan 2009) for all regimens $\bar{z}_T$ of interest, for all values $\bar{\mathbf{X}}_T$ and for all t,

$$Y_t(\bar{z}_t), \ldots, Y_T(\bar{z}_T) \perp\!\!\!\perp Z_t \mid \bar{\mathbf{X}}_t = \bar{\mathbf{x}}_t, \bar{Z}_{t-1} = \bar{z}_{t-1}, \tag{1}$$

and, for values $\bar{z}_{t-1}, \bar{\mathbf{x}}_t$ with probability function $f_{\bar{Z}_{t-1}, \bar{\mathbf{X}}_t}(\bar{z}_{t-1}, \bar{\mathbf{x}}_t) > 0$,

$$0 < \Pr(Z_t = 1 \mid \bar{Z}_{t-1} = \bar{z}_{t-1}, \bar{\mathbf{X}}_{t-1} = \bar{\mathbf{x}}_{t-1}) < 1. \tag{2}$$

The first condition is "sequential randomization" or "conditional exchangeability". The second condition is "positivity". Under these conditions, the distributions $f(y_t(\bar{z}_t))$, $f(y_t(\bar{z}_t) \mid \mathbf{x}_1)$ are identified.

e.g. if the probability function f is discrete:

$$f(y_t(\bar{z}_t)) = \sum_{\bar{x}_t} f(y_t \mid \bar{z}_t, \bar{\mathbf{x}}_t) \prod_{\ell=1}^{t} f(\mathbf{x}_\ell \mid \bar{\mathbf{x}}_{\ell-1}, \bar{z}_{\ell-1}),$$

where $f(\mathbf{x}_1 \mid \mathbf{x}_0, z_0) \equiv f(\mathbf{x}_1)$; thus,

$$E(Y_t(\bar{z}_t)) = \sum_{\bar{x}_t} E(Y_t \mid \bar{z}_t, \bar{\mathbf{x}}_t) \prod_{\ell=1}^{t} f(\mathbf{x}_\ell \mid \bar{\mathbf{x}}_{\ell-1}, \bar{z}_{\ell-1})$$

is also identified.

Robins and Hernan (2009) call this result the g-formula.

It is not difficult to establish and is evident from the following derivation for the case $t = 2$:

$$
\begin{aligned}
f(y_2(z_1, z_2)) &= \sum_{\mathbf{x}_1} f(y_2(z_1, z_2) \mid \mathbf{x}_1) f(\mathbf{x}_1) \\
&= \sum_{\mathbf{x}_1} f(y_2(z_1, z_2) \mid z_1, \mathbf{x}_1) f(\mathbf{x}_1) \\
&= \sum_{\mathbf{x}_2, \mathbf{x}_1} f(y_2(z_1, z_2) \mid z_1, \mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2 \mid \mathbf{x}_1, z_1) f(\mathbf{x}_1) \\
&= \sum_{\bar{\mathbf{x}}_2} f(y_2(z_1, z_2) \mid z_1, z_2, \mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2 \mid \mathbf{x}_1, z_1) f(\mathbf{x}_1) \\
&= \sum_{\bar{\mathbf{x}}_2} f(y_2 \mid z_1, z_2, \mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2 \mid \mathbf{x}_1, z_1) f(\mathbf{x}_1).
\end{aligned}
$$

To use the g-formula to compare different regimens and sub-regimens, it is necessary to estimate the expectations $E(Y_t(\bar{z}_t))$ above. This requires estimation of $E(Y_t \mid \bar{z}_t, \bar{\mathbf{x}}_t)$ for all $\bar{\mathbf{x}}_t$ realized under the sequence $\mathbf{z}_{t-1}$ and the probability functions $f(\mathbf{x}_\ell \mid \bar{\mathbf{x}}_{\ell-1}, \bar{z}_{\ell-1})$ for $\ell = 1, ..., t$.

Misspecification generally leads to biased estimates. As t increases, misspecification becomes more likely.

Additionally, as t increases, the number of sequences grows exponentially.

e.g. Robins, Greenland and Hsu (1999) analyze data from an observational study of 167 children, observed for 30 days. On each day, the child may be ill or not (the outcome) and the mother may be experiencing stress or not (the treatment). There are 2^{30} possible treatment regimens.

Robins and Hernan (2009) also point out a problem that is called the "null paradox of g-estimation". Under general conditions, using parametric models to compare the means under different regimens leads to falsely rejecting the null hypothesis of no effect, even when this is true.

This is one reason the g-formula has been less used than marginal structural models. In addition, commercial software (in SAS) has only recently become available for implementing this approach.