STATISTICS 536B, Lecture #13

April 9, 2015

Causal inference and data over time - yikes!

Apologies for the notation switch - not my fault!

Consider five binary variables: X_1, Z_1, X_2, Z_2, Y

Say $Z_t = 1$ indicates "on-drug" in month t

Say $X_t = 1$ indicates some manifestation of "worse symptoms" in month t

Say $Y = 1$ indicates bad end-of-study outcome

Worry that maybe X_t is both a confounder and a mediator? How to untangle?

Toy dataset, 5000 subjects

X_1	Z_1	X_2	Z_2	$Y = 0$	$Y = 1$
0	0	0	0	1364	744
0	0	0	1	355	185
0	0	1	0	371	335
0	0	1	1	104	88
0	1	0	0	8	5
0	1	0	1	91	45
0	1	1	0	1	2
0	1	1	1	25	11
1	0	0	0	18	13
1	0	0	1	4	3
1	0	1	0	279	389
1	0	1	1	71	100
1	1	0	0	2	1
1	1	0	1	11	11
1	1	1	0	19	19
1	1	1	1	141	185

Start with a completely general factorization for the joint pmf (for a given g)

$$
\begin{aligned}
f(x_{1:2}, z^*_{1:2}, z_{1:2}, y_{1:2}) \quad &= \\
f(x_1) f(z^*_1|x_1) g(z_1|z^*_1, x_1) f(x_2|x_1, z_1, z^*_1) & \\
f(z^*_2|x_{1:2}, z_1, z^*_1) & \\
g(z_2|z^*_2, x_{1:2}, z_1, z^*_1) & \\
f(y|x_{1:2}, z_{1:2}, z^*_{1:2}) &
\end{aligned}
$$

To help keep track: $g()$ denotes (potentially probabilistic) **intervention**. Distinguish between the exposure status that would arise **without any intervention** (Z^*, think 'patient's wish') and Z that arises from the intervention g being considered. Examples?

Example interventions

What intervention is this?

- $g(z_1|z_1^*, x_1) = 0.5$
- $g(z_2|z_2^*, x_{1:2}, z_1, z_1^*) = I\{z_2 = z_2^*\}$

What about this one?

- $g(z_1|z_1^*, x_1) = z_1$
- $g(z_2|z_2^*, x_{1:2}, z_1, z_1^*) = z_2$

Fundamental assumptions one can/must make

Immediately **after** intervention, the 'patient's wish' becomes irrelevant. (knock out all the terms in red on the previous slide)

- Only what actually happens can influence the future

All the $f()$ terms are unaffected by the choice of interventions $g()$. Really an assumption about controlling for confounding!

- For instance, the same $f(y \mid x_{1:2}, z_{1:2})$ describes randomized trials and observational cohorts - usual idea that x controls for confounding.

Data and estimation

We have in mind the situation that **we only have observational data**, i.e., just following a cohort over time [data arise from $g(z_1|x_1, z_1^*) = I\{z_1 = z_1^*\}$, $g(z_2|z_2^*, x_{1:2}, z_1) = I\{z_2 = z_2^*\}$].

That's OK, the data still gives us $\hat{f}()$ for all the $f()$ terms. So for any choice of interventions g of interest to us, we can, in principle, estimate the joint distribution of $(X_{1:2}, Z_{1:2}^*, Z_{1:2}, Y)$.

For instance, we could use data collected under
$g \equiv$ no intervention to estimate $Pr(Y = 1)$ under both
$g \equiv$ always treat and $g \equiv$ never treat.

In our simple set-up, the g-formula gives the following "bridge"

$$Pr_{\text{always}}(Y=1) = \sum_{x_1=0}^{1}\sum_{x_2=0}^{1}[Pr(X_1=x_1)Pr(X_2=x_2|X_1=x_1,Z_1=1)\times Pr\{Y=1|X=(x_1,x_2),Z=(1,1)\}],$$

with an analogous expression for $Pr_{\text{never}}(Y=1)$.

Full disclosure, our toy dataset was simulated under very specific conditions

At each timepoint, those who are sicker are more likely to start treatment (hallmark of "counfounding by indication).

There is a dual benefit of treatment:

- Direct effect of Z_1 and Z_2 on Y
- Indirect effect whereby $Z_1 = 1$ induces lower chance of undesirable transition from $X_1 = 0$ to $X_2 = 1$, in turn $X_2 = 1$ raises risk of $Y = 1$.

Analysis of the toy data

Naive target: $Pr(Y = 1|Z = (1, 1)) - Pr(Y = 1|Z = (0, 0))$

Estimated via sample proportions as 0.06 (SE=0.02).

g-formula analysis for: $Pr_{always}(Y = 1) - Pr_{never}(Y = 1)$

Estimated as -0.043 (SE=0.028).