STATISTICS 536B, Lecture #9

March 26, 2015

Propensity scores - What is the high level idea?

Have $(Y, X, C_1, \ldots, C_p)$ data, interested in the association between Y and X given C.

Direct route: study this via regression of Y on X and C.

Indirect route: consider $Z = Pr(X = 1|C) = \pi(C)$ (in theory), or $\hat{Z} = \hat{\pi}(C)$ (in practice). Then focus on the association between Y and X given Z.

The underlying mathematics validates this approach.

Mongelluzzo et. al. - corticosteroids and mortality from bacterial meningitis

Outcome Y is time-to-event (time from hospitalization for bacterial meningitis to death, or time from hospitalization to discharge)

Binary exposure X is *adjuvant* use of corticosteroids

Potential confounders (C) include sex, race, vancomycin use within 24 hours, etc,...

Traditional analysis might involve proportional hazards regression model for Y using X and $C_1, \ldots, C_p$ as explanatory variables.

Instead, these authors use X and $\hat{Z} = \hat{\pi}(C)$ as the explanatory variables.

Some discussion points

Fitted propensity model for $(X|C)$ model gives AUC=0.74 ... "better than chance,' but "little concern about nonoverlapping propensity score distributions" ???

Discussion points, continued

But then: "The propensity scores were not equally distributed. When the propensity scores were stratified by quintiles, a greater proportion of X=1 patients were in the highest quintile and a greater proportion of $X = 0$ patients were in the lowest quintile. To address this imbalance..."

PUZZLING!!!

Discussion points, continued

'Residual confounding by indication' concern.

Often plausible that sicker patients more likely to get the intervention ($X = 1$) being studied. (So a crude two group comparison would be 'unfair' on $X = 1$).

Not a problem if 'sicker' is completely captured by C.

Otherwise, can make an intervention appear less efficacious than it really is. E.g., say that (C, C^*) completely capture 'sicker', but C^* is unmeasured.

Results

Table 3: no evidence for a (Y, X) association given C - for either Y.

Table 4: no evidence for a (Cost, X) association given C. Suggestive of (or at least consistent with) C being 'good enough.'

Plausible that if C wasn't fully capturing disease severity and $X = 1$ was being preferentially offered to those with more severe disease, then we would see a positive association between X and Cost given C.

Back to simpler framework of continuous outcome Y. Where are we at?

Trying to estimate

$$\Delta \quad = \quad E\{E(Y|X=1,C) - E(Y|X=0,C)\}.$$

If we are confident in our ability to model Y given X and C:

Could fit a $(Y|X,C)$ outcome model, to estimate
$m_x(C) = E(Y|X=x,C)$, then

$$\hat{\Delta}_R \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(c_i) - \hat{m}_0(c_i)$$

is a consistent estimator, if the form of the outcome model is right.

Or the propensity route

If we are confident in our ability to model X given C:

Recall (last time) we can rewrite the target parameter as

$$\Delta = E\left\{Y\left(\frac{X}{\pi(C)} - \frac{1-X}{1-\pi(C)}\right)\right\}$$

Could fit a $(X|C)$ propensity model, to estimate $\pi(C) = Pr(X = 1|C)$, then

$$\hat{\Delta}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} y_i\left(\frac{x_i}{\hat{\pi}(c_i)} - \frac{1-x_i}{1-\hat{\pi}(c_i)}\right).$$

is a consistent estimator if form of propensity model is right.

Back to nasty dataset from last time

```
### outcome model and fitted values
outmod <- lm(y~x+cnf)
m0 <- cbind(1,0,cnf)%*%coef(outmod)
m1 <- cbind(1,1,cnf)%*%coef(outmod)

### propensity model and fitted values
promod <- glm(x~cnf, family=binomial)
prpns <- fitted(promod, response=T)

### regression estimate
mean(m1-m0)
[1] 1.23

### IPW estimate
mean(y*(x/prpns - (1-x)/(1-prpns)))
[1] 1.14

### Double-robust estimate
 mean((y*x - (x-prpns)*m1)/prpns) - mean((y*(1-x) + (x-prpns)*m0)/(1-prpns))
[1] 1.16
```

Standard errors for these estimates?

All three estimates are means of n values, but ...

So bootstrap...

```
ests.bb <- matrix(NA,200,3)
for (i in 1:200) {
  smp <- sample(1:n, replace=T)

  ### outcome model
  outmod <- lm(y[smp]~x[smp]+cnf[smp,])
  ...
  ### propensity model
  promod <- glm(x[smp]~cnf[smp,], family=binomial)
  ...
  ests.bb[i,] <- c(mean(m1-m0), ...)
}

sqrt(apply(ests.bb,2,var))
[1] 0.12 0.12 0.12
```