Causal Inference – Jennifer Hill

See bartCause R package

Other BART packages in R:

BayesTree

dbarts is better than BayesTree

# Motivation

Observational studies didn’t support effectiveness of Salk vaccine at preventing polio. Randomized experiments did!

Early days of search engine marketing wasted tons of money by assuming clicking and buying was very closely related. In fact, people found most of the things they bought anyway, unrelated to the ads (big quasi-experimental research at ebay showed this)

Women’s Health Initiative and Nurses Health Study found opposite results re HRT and coronary heart disease. Hernan and Robins (Epidemiology, 2008) found how to reconcile results w/ causal analysis.

# Example

Effect of an enrichment program on subsequent test scores

Suppose that being admitted to the program is

-determined based on one pre-test score

-is probabilistic, not deterministic

And treatment effect of the enrichment program varies across students as a function of pre-test scores

So, exposure to the program depends on pre-test score, and so does effect of the program, but in different ways.

Parametric assumptions: implications of non-linearity and lack of overlap

Our data will have two groups of students, ones who got the enrichment program and ones who didn’t, who vary based on pretest scores

Linear regression’s not an appropriate assumption here. (In this example, non-program students have logarithmic relationship b/w pre- and post-test scores, while program students have exponential relationship).

Here, lack of overlap in pretest scores b/w the two groups exacerbates the problem by forcing model extrapolation outside the range of pretest scores the program was actually used in

When there are 30, 50, 100 covariates, it’s a lot harder to go “oh, let me try things other than linreg” than when there are few.

# Causal inference from observational studies requires strong assumptions:

-Structural: all confounders measured

Everything that affects *both* whether you get the treatment and the outcome

-Parametric: the model used to adjust for all those confounders is correct

And if you’re measuring a ton of possible confounders, this is harder!

Multi lin regression on a ton of different things may be an issue.

## Notation

X is (vector of) observed covariates

Z is binary treatment variable

Y(0), Y(1) are potential outcomes

Y is the observed outcome

Individual level causal effects compare potential outcomes, e.g.

Y1(1) – Y1(0)

With the goal of estimating something like

E[Y(1) – Y(0)]

Expected value of potential outcomes

or

E[Y(1) – Y(0) | Z=1]

Expected value of potential outcomes given that treatment variable = 1

Structural assumption: the key assumption in most obs studies is that **all confounders have been measured** (ignorability, selection on observables, conditional independence, etc**.) Formally, this implies:**

**Y(0), Y(1) are orthogonal to Z|X**

This assumption is untestable and hard to satisfy.

Parametric assumptions are where most of the action is. Re structural assumption, can only fix by collecting more data or doing a randomized experiment instead. More we can fix on parametric side.

Parametric assumptions are that we can tie our potential outcomes to X through a model. E.g. if we assumed a linear model:

E[Y(0) } X] = X\*beta^gamma

E[Y(1)|X] = X\*beta^gamma + tau

The more covariates we include – e.g. to satisfy the “all cofounders measured” assumption – the bigger trouble we’re in if our parametric assumptions aren’t right.

The massive literature on propensity score matching is primarily about reducing our reliance on these parametric assumptions.

# Roadmap

Bayesian additive regression trees (BART) are a useful ML approach.

Can use BART to fit the response surface, which is the E[Y|Z,X]. The expected value of the outcome given treatment and covariates.

Can use BART automatic uncertainty quantification to understand when don’t have sufficient common support or overlap across the groups.

Heterogeneity, generalizability

**bartCause R package**

BART works really well in a lot of cases.

Bayesian version of boosted regression trees. Bayesian inference/ MCMC.

Regression trees progressively split the data into more and more homogeneous subsets. Within each of these subsets the mean of the response variable can be calculated.

Hard to decide when to stop splitting

So, boosted regression trees! First, fit a little tiny tree to your data.

Then, fit a little tiny tree to the residuals.

Then, fit a little tiny tree to…

On each step, try to explain what the previous step didn’t explain.

Ensemble of weak learners. Sum of trees. Additive model.

Fit using a back-fitting algorithm.

Vulnerable to overfitting. Adjust by multiplying by a tuning variable, cross-validate, use bootstrapping to quantify uncertainty… gets complicated.

Boosting is great for prediction, but:

Reqs ad-hoc choice of tuning parameters (number of trees, depths of trees, shrinkage for the fit of each tree)

Estimating uncertainty is hard. Generally use booststrapping; can be cumbersome and time-consuming, and really you should bootstrap your cross-validation *too*, and that’s a pain.

# So, BART!

A stochastic alternative to boosting algorithms, for fitting a sum of trees model.

Is within a likelihood framework

F(x,z) is a random var sampled using MCMC

Treesare exchangeable

**Avoids overfitting by the prior specification that shrinks toward a simple fit**

Priors tend toward small trees (“weak learners”)

Fitted values from each tree are shrunk using priors

The priors are extremely clever, and are explained in the 2007 paper.

Distribution for the treatment effect is created by differencing the (posterior predictive) distributions for Y(1) and Y(0)

Make predictions for if everyone got the treatment and if nobody did

Natural way to quantify uncertainty

Linreg will often be strongly and wrongly confident in places where treatment and non-treatment groups follow diff distribs.

BART has great way of showing confidence intervals where are bad – good at telling you where data is lacking, w/ confidence interval lines.

Sd discard rule – can get rid of obs where SD of y distribution is too wide, to avoid uninformative data.

Can often use ratios of SDs to discard what to use.

# BART vs propensity scores

Propensity scores run into trouble when there *isn’t* a good match for someone in one group in the other group – that’ll keep you from making an inference for that person.

If you use propensity score on overlap, can get situations where you’re actually predicting causes of treatment rather than effect of treatment.

Propensity scores would drop a lot of obs that shouldn’t really drop, when there’s diff in covariates b/w groups

Good for overlap because:

BART identifies neighborhoods that lack common support and thus gets better treatment effect estimates in settings that lack common support, compared to propensity scores, bc prop scores ignore outcome info

Including the outcome var is curcial for understanding common causal support

Heterogeneous treatment effects – BART outperforms prop score approaches when these happen. Also, better generalizability of average treatment effects to different/ broader populations, which can be tricky when treatment effects are heterogeneous across observations.

# The hard part

So, what are the assumptions for BART?

Sensitivity to unobserved confounders

Need to care about your data generating process!

Our model of the data generating process has sensitivity parameters that let us estimate this.

MCMC algo: specify a pair of sensitivity parameters, fit a model for treatment assignment and a model for the response surface and get a treatment effect estimate

Alternate b/w estimating a parameter, drawing from distrib given that parameter, drawing from parameter given that distrib