Characterizing Effect Heterogeneity

Cyrus Samii

2019-05-23

Contents

1 Introduction 2

2 Bounding the Extent of Effect Heterogeneity 2

2.1 Frechet-Hoeffding bounds . . . 3

2.1.1 A digression on copulas . . . 3

2.1.2 Applying Frechet-Hoeffding Bounds . . . 7

3 Estimating Conditional Treatment Effects (CATES) 9

3.1 Overview and Goals . . . 9

3.2 Algorithms . . . 9

4 Analyzing Features of CATES (Chernozhukov et al, 2017) 9

4.1 Goals . . . 9

4.2 Setting . . . 9

4.3 Methods . . . 10

4.4 Choosing an ML method . . . 11

4.5 Applications . . . 11

5 Wager and Athey (2018) 11

6 Notes 11

6.1 Low dimensional features versus CATEs . . . 11

1 Introduction

Below are notes on approaches to characterizing effect heterogeneity. The methods are reviewed in light of the goals in the paper, Gechter et al. (2019, arxiv). It would be worth becoming familiar with that paper before reviewing what I have below.

2 Bounding the Extent of Effect Heterogeneity

We will start with Heckman, Smith, and Clements (1997) examination of the binary outcomes case. Let's bring in the data from the Progresa to illustrate. We have the marginal enrollment distributions under treatment and control,

```r
progd <- as.data.frame(import("for_analysis_el.dta"))
progd$treat <- 1-progd$control
enTab <- t(crossTab(progd$enrolled, progd$treat, "Enrolled", "Treated")[[2]])
colnames(enTab) <- rep("", ncol(enTab))
kable(enTab, align="r")
```

		Enrolled				
		0		1		
Treated	0	2131	0.21	8230	0.79	10361
	1	2866	0.17	14217	0.83	17083
		4997		22447		27444

and the corresponding treatment effect:

```r
ATE <- mean(progd$enrolled[progd$treat==1]) - mean(progd$enrolled[progd$treat==0])
ATE
```

```
## [1] 0.037906
```

Our aim is to fill in a contingency table for principal strata, which are defined in terms of whether units are enrolled (E) or not enrolled (N) under control and then under treatment. The share of units enrolled under both is P_{EE} and so on. The full contingency table looks like the following:

		Control		
		E	N	
Treated	E	P_{EE}	P_{EN}	$P_{E\cdot}$
	N	P_{NE}	P_{NN}	$P_{N\cdot}$
		$P_{\cdot E}$	$P_{\cdot N}$	$P_{N\cdot}$

The ATE is equivalent to,

$$\begin{aligned} ATE &= P_{E\cdot} - P_{\cdot E} \\ &= (P_{EE} + P_{EN}) - (P_{EE} + P_{NE}) \\ &= P_{EN} - P_{NE} \end{aligned}$$

The question is whether the experimental data allow us to identify these principal strata shares. Generally speaking, with no assumptions, the answer is no.

So let's start to consider some assumptions. Suppose, for example, that the effects of the intervention are weakly monotonic in that they either increase or do not change enrollment. Then, $P_{NE} = 0$, in which case $ATE = P_{EN}$. As such, for the share given by ATE, the intervention has an effect of 1, and for the share given by $1 - ATE$, the intervention has no effect. In our application, under such weakly monotonic beneficial effects, 3.8% of the population would have effects of 1, and 96.2% would have effects of 0.

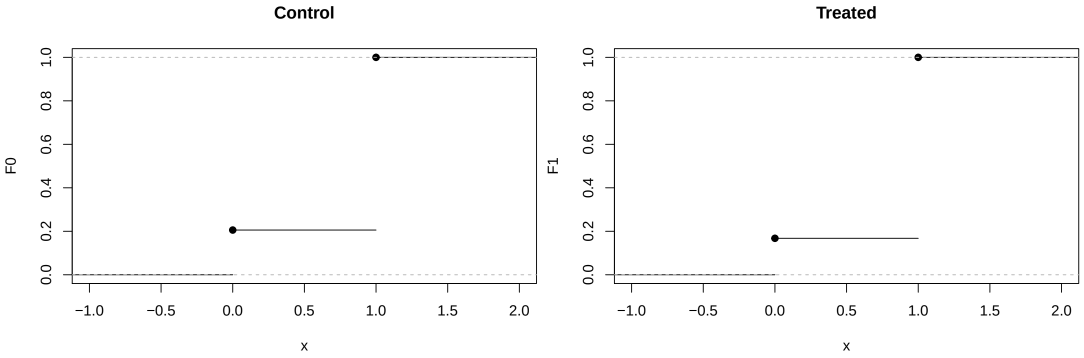
Such monotonicity is a strong assumption, however, and so we now turn to more general bounds.

2.1 Frechet-Hoeffding bounds

A general bound can be derived a la Frechet-Hoeffding. We explain this by way of a lesson on copulas, upon which Frechet-Hoeffding bounds are based.

2.1.1 A digression on copulas

Suppose two potential outcomes $Y(0), Y(1)$ with a joint distribution $G(y_0, y_1)$ and marginal distributions $F_0(y)$ and $F_1(y)$. As such, $F_0(0) = P_{.N}$ and $F_0(1) = 1$, while $F_1(0) = P_{N.}$ and $F_1(1) = 1$. We can draw these for our data:



Note that given the discrete random variables, we use the generalized inverse of the CDF:

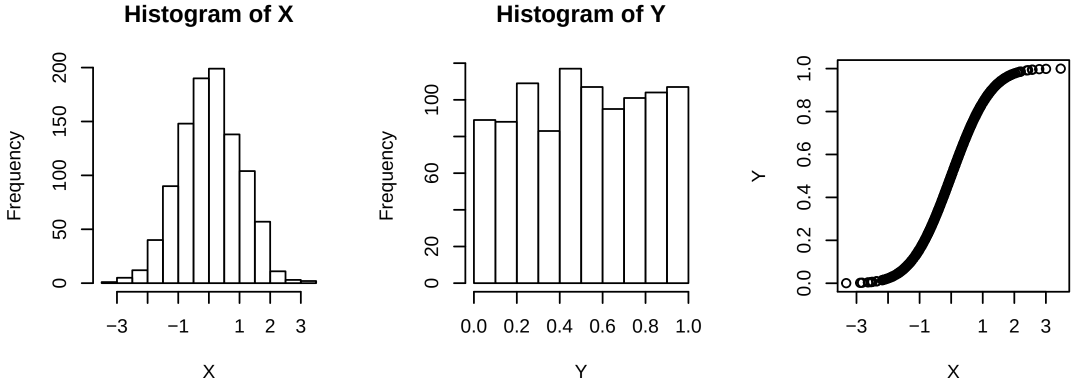
$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\},$$

that is, it returns the smallest x that returns $F(x) \geq p$. So, $F_0^{-1}(1) = 1$, we choose $y_0 = 1$. Similarly, $F_0^{-1}(u) = 1$ for $P_{.N} < u < 1$. More generally for generalized inverse function,

- $F(F^{-1}(u)) \geq u$.
- $F(x) \geq u$ iff $x \geq F^{-1}(u)$.
- For $U \sim U[0,1]$, $X = F^{-1}(U)$ has CDF F.

Note how a CDF transforms a random variable. E.g,

```
X <- rnorm(1000)
Y <- pnorm(X)
par(mfrow=c(1,3), pty="s")
hist(X)
hist(Y)
plot(X, Y)
```



Now let us define a copula and study its properties. For our case, we define it as mapping from two marginal distributions to range of a joint distribution. Generically, a copula for a bivariate distribution can be defined as $C : [0,1]^2 \to [0,1]$ such that for (U_1, U_2) with margins that are distributed $Unif[0,1]$, we have,

$$C(u_1, u_2) = \Pr[U_1 \leq u_1, U_2 \leq u_2].$$

Given this formulation, a few things follow:

- If $u_1 = 0$ or $u_2 = 0$ then $C(u_1, u_2) = 0$. The reason is that fixing one of the arguments to zero sets us at the edge of the cdf $C(\cdot)$ and so the mass is zero along this edge.
- $C(1, u_2) = u_2$ and $C(u_1, 1) = u_1$. The reason is that being at the 1-edge for u_j of the cdf $C(\cdot)$ implies that we have already incorporated all of the mass in the j dimension, meaning that we are just varying mass in the i dimension.
- If $a_j \leq b_j$, then
$$(C(b_1, b_2) - c(a_1, b_2)) - (C(b_1, a_2) - C(a_1, a_2)) \geq 0.$$
This is a sort of monotonicity.
- C is non-decreasing in its arguments.
- C is cts (because it is Lipschitz).

Sklar's theorem I says that we can construct multivariate CDFs using copulas. Here we state it for the bivariate case.

> Let C be a bivariate copula, and suppose univariate CDFs F_0 and F_1. Then,
>
> $$F(y_0, y_1) = C(F_0(y_0), F_1(y_1))$$
>
> is a bivariate CDF with margins F_0 and F_1.

Thus, the function C gives rise to a bivariate CDF with marginals F_0 and F_1. The proof is very simple. Suppose $(U_0, U_1) \sim C$, the function defined in the proof. What are the marginals of this distribution when the arguments are defined as in the proof? Well, if $U_0 = F_0(Y_0)$, say, then $Y_0 = F_0^{-1}(U_0) \sim F_0$; similarly $Y_1 = F_1^{-1}(U_1) \sim F_1$.

Sklar's theorem II says that any multivariate CDF has a copula. Again we state for the bivariate case:

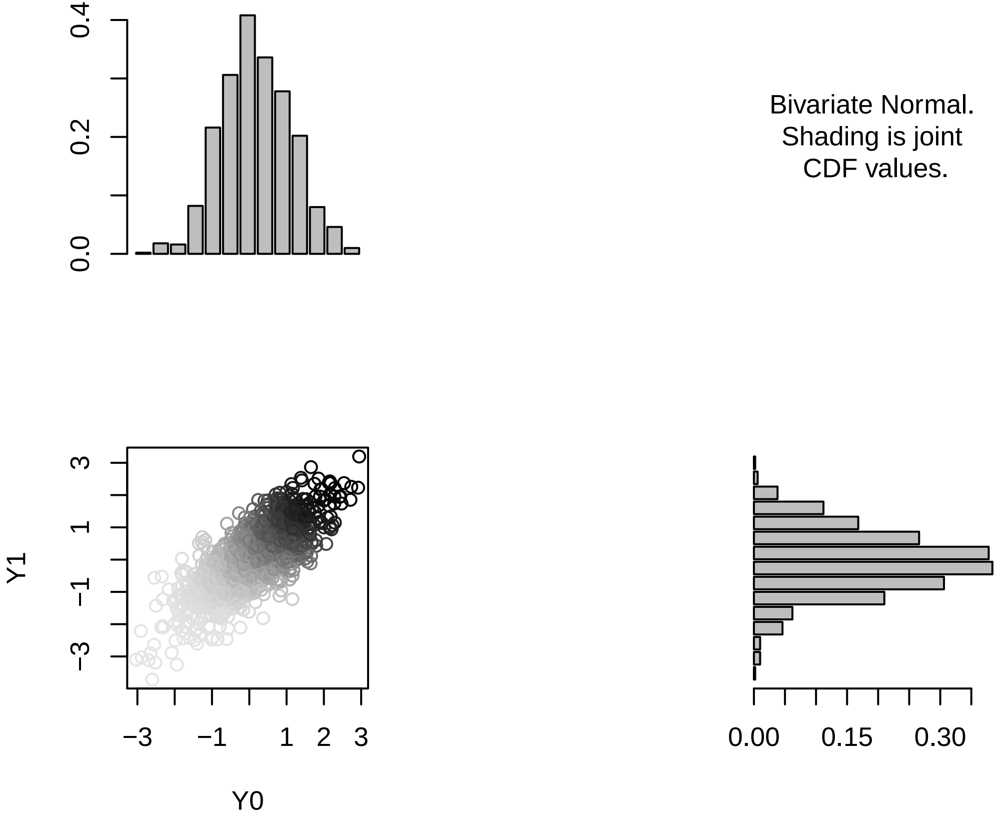
> If F is a bivariate CDF with marginals F_0 and F_1, then there exists a copula C such that Sklar I holds. Moreover, if the margins are continuous, then C is unique and equals,
>
> $$C(u_0, u_1) = F(F_0^{-1}(u_0), F_1^{-1}(u_1)).$$

The proof is as follows. Suppose the margins are continuous. Let $(Y_0, Y_1) \sim F$. Now, $U_j = F_j(Y_j) \sim Unif[0,1]$. Then, $(U_0, U_1) \sim C$ as defined in Sklar I holds.

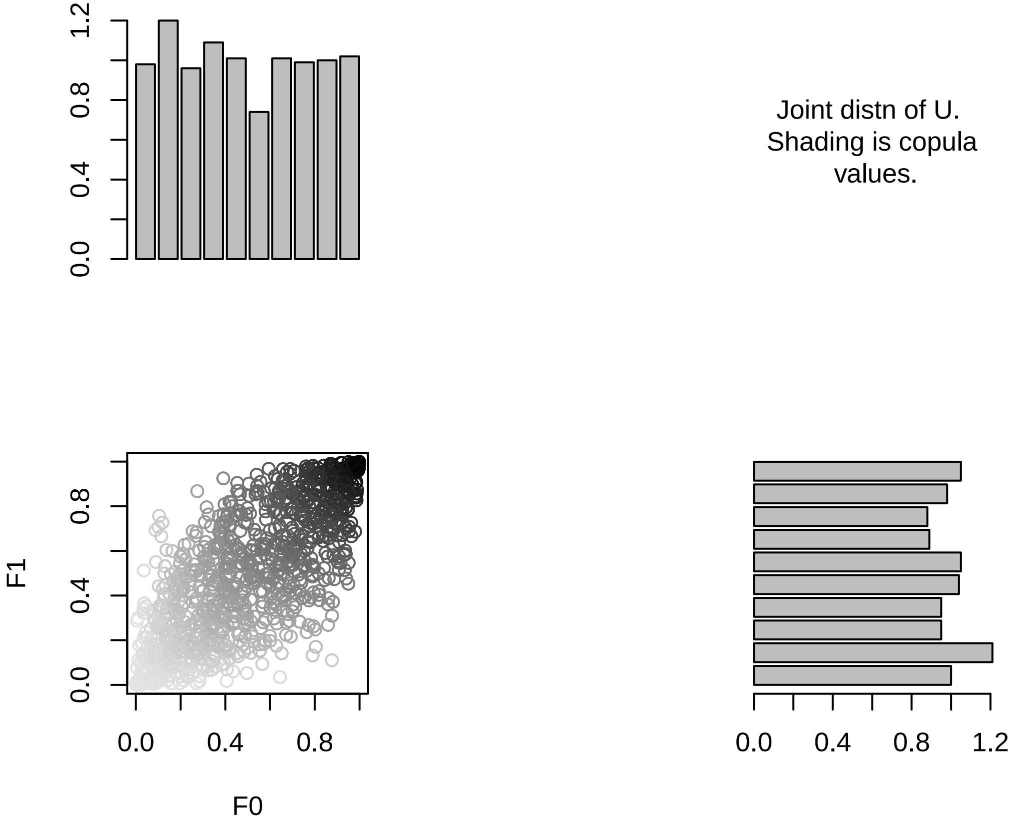
All of these properties hold for discrete variables as well, although an issue with discrete variables is that the copula for a given joint distribution may not be unique.

See for example a bivariate normal case.

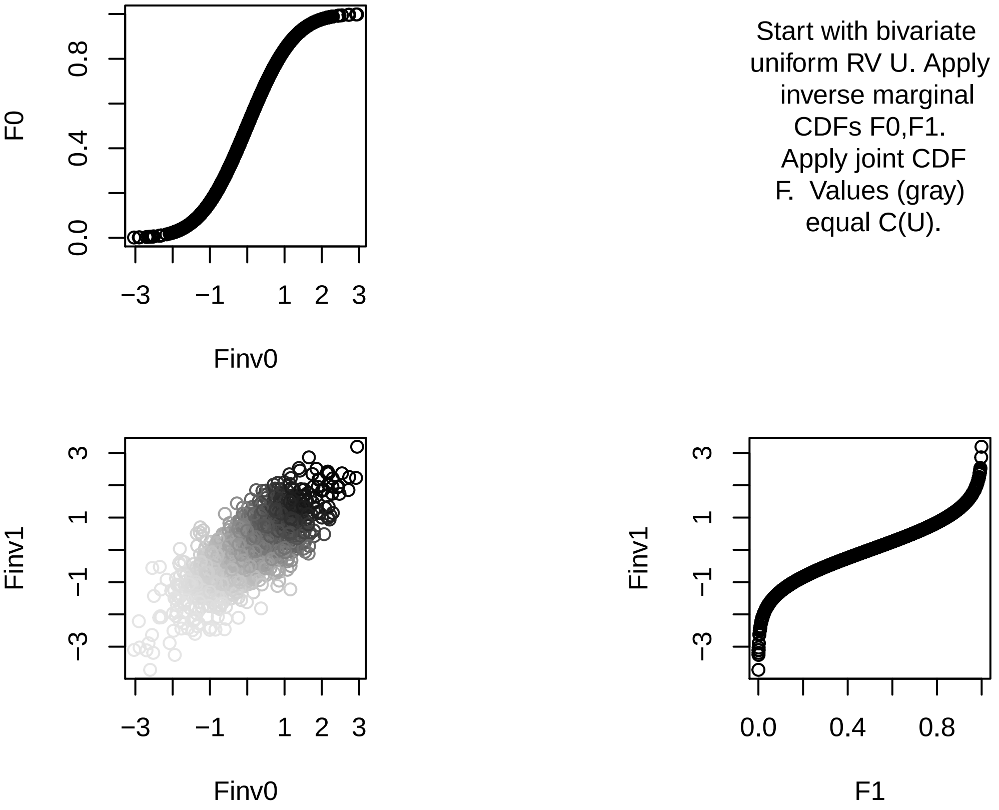
2.1.1.1 Bivariate normal distribution



2.1.1.2 Corresponding copula



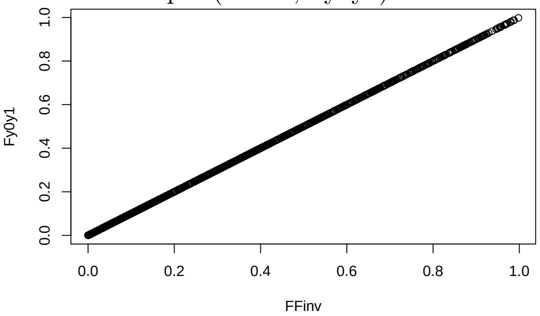
2.1.1.3 Illustrating of Sklar II, from CDF to copula



Start with bivariate uniform RV U. Apply inverse marginal CDFs F0,F1. Apply joint CDF F. Values (gray) equal C(U).

We can thus see that the C and F values are the same:

“‘r plot(FFinv, Fy0y1) “‘



2.1.1.4 Bounds

Okay, now we have a sense of copulas and how they relate to CDFs. Here is the statement for the Frechet-Hoeffding bounds:

> Any bivariate copula $C(u, v)$ verifies,
>
> $$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v).$$

The proof is as follows. By the properties of copulas, we have

$$C(u, v) \leq C(u, 1) \leq u$$

and

$$C(u, v) \leq C(1, v) \leq v,$$

which establishes the upper bound, $\min(u, v)$. Then, by the "monotonicity" property described above, defining $a_1 = u$,$b_1 = 1$, $a_2 = v$, and $b_2 = 1$, we have

$$(C(1, 1) - C(u, 1)) - (C(1, v) - C(u, v)) = 1 - u - v + C(u, v) \geq 0,$$

which yields the lower bound.

For intuition, recall

$$C(u, v) = \Pr[U \leq u, V \leq v]$$

for U, V with uniform marginals on $[0, 1]$. Then, the Frechet-Hoeffding bound is based on situations where U and V are perfectly correlated and perfectly anti-correlated. Suppose they are perfectly correlated. Then, $V = U$ and $C(u, v) = \Pr[U \leq u, U \leq v] = \min(u, v)$. Now suppose they are perfectly anti-correlated. Then, $V = 1 - U$ and $C(u, v) = \Pr[U \leq u, U \geq 1 - v]$. If $u < 1 - v \Rightarrow u + v - 1 > 0$, then this equals 0. Otherwise, it equals the space between u and $1 - v$, which is $u - (1 - v) = u + v - 1$.

Now, the idea is to go from this result on copulas to a result on joint distributions. This is where Sklar I comes into play. Recall that it states that

$$F(y_0, y_1) = C(F_0(y_0), F_1(y_1))$$

is a bivariate CDF with marginals F_0 and F_1. As such, we substitute the arguments $F_0(y_0)$ and $F_1(y_1)$ for u and v and then $F(y_0, y_1)$ for $C(u, v)$ n the statement of the bounds to obtain,

$$\max(F_0(y_0) + F_1(y_1) - 1, 0) \leq F(y_0, y_1) \leq \min(F_0(y_0), F_1(y_1)).$$

Relating the correlation and anti-correlation from the copulas to this result, we have that the upper bound is reached when Y_0 and Y_1 are comonotonic (i.e., Y_1 is a deterministic non-decreasing function of Y_0, implying a rank correlation of 1) and the lower bound then they are countermonotonic (i.e., Y_1 is a deterministic non-increasing function of Y_0, implying a rank correlation of -1).

2.1.2 Applying Frechet-Hoeffding Bounds

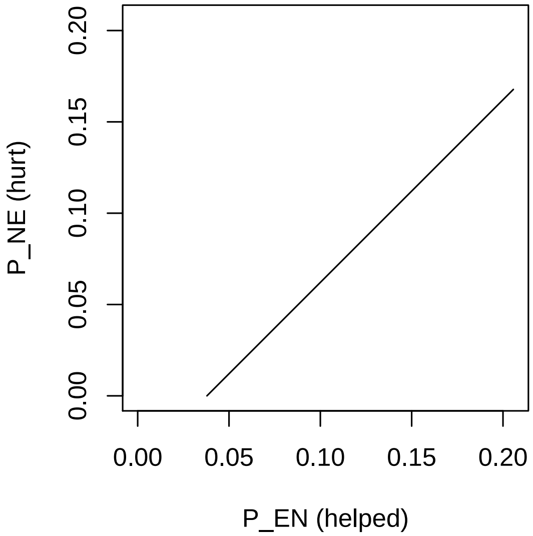
Returning to our example, recall that we are interested in the joint distribution of outcomes under treatment and control, where the outcome are each binary (either enrolled or not enrolled). Recall as well that we defined, e.g., P_{EE} to refer to the share of individuals for which potential outcomes under both control and treatment are "enrolled." Let us define two indicator variables, E_0 and E_1, for whether an individual is enrolled in treatment and control, respectively. Then P_{EE} is the joint distribution for E_1 and E_0, while $P_{\cdot E}$ is the marginal distribution of E_0 and $P_{E\cdot}$ is the marginal distribution of E_1. We can proceed similarly for P_{EN}, P_{NE}, and P_{NN}. Then, applying the Frechet-Hoeffding bounds we have, following Heckman et al. (1997), we have

$$\begin{aligned}
\max[P_{E\cdot} + P_{\cdot E} - 1, 0] &\leq P_{EE} \leq \min[P_{E\cdot}, P_{\cdot E}] \\
\max[P_{E\cdot} - P_{\cdot E}, 0] &\leq P_{EN} \leq \min[P_{E\cdot}, 1 - P_{\cdot E}] \\
\max[-P_{E\cdot} + P_{\cdot E}, 0] &\leq P_{NE} \leq \min[1 - P_{E\cdot}, P_{\cdot E}] \\
\max[1 - P_{E\cdot} - P_{\cdot E}, 0] &\leq P_{NN} \leq \min[1 - P_{E\cdot}, 1 - P_{\cdot E}].
\end{aligned}$$

From the table above, we see that in our application we have the following:

	Lower	Upper
EE	0.6265557	0.7943249
EN	0.0379060	0.2056751
NE	0.0000000	0.1677691
NN	0.0000000	0.1677691

So as many as 20.5% and as few as 3.8% may have been induced to enroll while as many as 16.8% and as few as 0% may have been induced not to enroll. Now, we have the identity $ATE = P_{EN} - P_{NE}$, so high P_{EN} require high P_{NE} and vice versa. Given $ATE = 0.038$, we can construct the range of joint (P_{EN}, P_{NE}) values consistent with our ATE estimate:



Based on these results, the maximal effect heterogeneity would be where 20.6% of the population would be helped, 16.8% would be hurt, and then the remainder (62.7%) would not be affected.

We can also compute these upper bounds on the extent of effect heterogeneity by strata. The terms that we need to compute are the treatment and control enrollment indicator means ($P_{E\cdot}$ and $P_{\cdot E}$) in each stratum.

```r
fit_agesex <- lm(enrolled~treat*as.factor(age)*male,
                  data=progd)
ageVec <- 6:16
predMat1 <- data.frame(age=c(ageVec, ageVec),
                       male=c(rep(0, length(ageVec)),
                              rep(1, length(ageVec))),
                    treat=rep(1, 2*length(ageVec)))
predMat0 <- predMat1
predMat0$treat <- 0
agesexMeans <-cbind(predMat1[,c("age","male")],
              predict(fit_agesex, newdata=predMat0),
              predict(fit_agesex, newdata=predMat1),
              predict(fit_agesex, newdata=predMat1)-predict(fit_agesex, newdata=predMat0))
colnames(agesexMeans) <- c("Age","Male","Control","Treated","CATE")
```

3 Estimating Conditional Treatment Effects (CATES)

3.1 Overview and Goals

We first consider off-the-shel methods for estimating heterogenous treatment effects. The focus initially will be on point predictions, rather than inference on such predictions. This is because the applications that we use work with summaries of the point predictions rather than the point predictions in and of themselves. See the section below on "features of CATES", referencing Chernozhukov et al. (2017, arxiv), for more on this point.

We are also interested in methods that work well when we entertain a high-dimensional covariate vector. I say "entertain" because it may be that, in fact, the covariates that predict heterogeneity are few, but this is something that we do not know *a priori,* and rather we have at our disposal many covariates that we want to consider as candidates for predicting effect heterogeneity. This leads us to machine learning approaches that use regularization to balance that ability to make very fine grained predictions with penalties for overfitting.

3.2 Algorithms

We consider the following algorithms:

- BART
- Generalized Random Forests
- Elastic Net on Minimum gCV Basis Expansion

4 Analyzing Features of CATES (Chernozhukov et al, 2017)

Link to the paper: arxiv

4.1 Goals

The goal is to learn about effect heterogeneity in a way that does not overfit and to provide "uniformly valid inference" on conditional average treatment effects (CATEs), or at least *features* of such CATEs. The approach is two-step: (i) build a machine learning (ML) proxy predictor of the CATE, then (ii) do inference on features of this proxy predictor. The features considered are first, best linear predictor (BLP), second, sorted group average effects (GATES), which are ATEs by heterogeneity groups, and third classification analysis (CLAN), which are the average effects of the most and least affected units.

4.2 Setting

Some notation:

- Potential outcomes, $Y(1), Y(0)$.
- Covariates Z.
- Baseline conditional average (BCA): $b_0(Z) = E[Y(0)|Z]$.
- CATE: $s_0(Z) = E[Y(1)|Z] - E[Y(0)|Z]$.
- CIA holds: $D \perp\!\!\!\perp (Y(1), Y(0))|Z$.
- Subvector of stratifying covariates, $Z_1 \subseteq Z$, such that
- propensity score is $p(Z) = P[D = 1|Z_1]$, with $0 < p_0 \leq p(Z) \leq p_1 < 1$.
- Observe $Y = DY(1) + (1 - D)Y(0)$, in which case,
- $Y = b_0(Z) + Ds_0(Z) + U$, with $E[U|Z, D] = 0$, where
- $b_0(Z) = E[Y|D = 0, Z]$ and $s_0(Z) = E[Y|D = 1, Z] - E[Y|D = 0, Z]$. CIA allows us to write b_0 and s_0 in terms of observables.
- Observe N iid draws of (Y, Z, D) with law P. Draws are indexed by $i = 1, ..., N$.

A result to keep in mind (already noted in Athey and Imbens PNAS, I believe):

Horvitz-Thompson scaling Define

$$H = H(D, Z) = \frac{D - p(Z)}{p(Z)(1 - p(Z))}.$$

Given the assumptions above, we have,

$$\begin{aligned} E[YH|Z] &= E\left[\frac{Y(D - p(Z))}{p(Z)(1 - p(Z))}\middle|Z\right] \\ &= \frac{1}{p(Z)(1 - p(Z))} E\left[D(D - p(Z))Y(1) + (1 - D)(D - p(Z))Y(0)|Z\right] \\ &= E[Y(1)|Z] - E[Y(0)|Z] \\ &= s_0(Z). \end{aligned}$$

We can use YH as an "unbiased signal" about $s_0(Z)$. It is, however, a noisy signal, and so in using this, Chernozhukov et al. make adjustments (e.g., for their second BLP method). Nonetheless, it does provide a target that one can use in trying to tune methods for estimating CATEs (see below the section on choosing an ML method).

4.3 Methods

For Chernozhukov et al., directly learning about s_0 in a generic way seems impossible at the moment When P is high dimensional, ML methods will not be consistent. Moreover, inference is complicated for adaptive estimators, particularly in trying to bound biases. Finally, tuning ML models is sort of free for all at the moment, with little to guide.

For these reasons Chernozhukov et al. focus on inference for low dimensional features of s_0. First, partition the indices $\{1, ..., N\}$ into two sets, M and A for "main" and "auxiliary", respectively, to yield a main sample, Data_M, and auxiliary sample, Data_A. Suppose for the time being that the two sets are about the same size. From sample A obtain ML estimates of b_0 and s_0, to be called "proxy predictors":

$$z \mapsto B(z) = B(z; \text{Data}_A) \text{ and } z \mapsto S(z) = S(z; \text{Data}_A).$$

These proxies need not be consistent. Then construct the features of interest (BLP, GATES, CLAN) in sample M. Then do inference in a way that accounts for estimation uncertainty given sample A and splitting uncertainty given the split into M and A. Their precise method is to use many splits and then take the median of the feature estimates and then the medians of the random conditional confidence sets, adjusting them to account for splitting uncertainty. They refer to this inferential approach as "variational estimation and inference" (VEIN).

BLP method 1 involves the following:

Consider the weighted linear projection:

$$Y = \alpha_1 + \alpha_2 B(Z) + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S(Z) - E[S(Z)]) + \epsilon,$$

with weights,

$$w(Z) = \frac{1}{p(Z)(1 - p(Z))}.$$

Chernozhulov et al. explain that "the interaction $(D - p(Z))(S - E[S])$ is orthogonal to $D - p(Z)$ under the weight $w(Z)$". This is due to LIE – cf. p. 34 of the paper.

I am going to leave the BLP alone for now. I can see how it may provide for a test of heterogeneity, but not sure how useful it is for our purposes.

Sorted group ATE is quite relevant to us. We want to put units into groups based on their predicted treatment effects. E.g., suppose we want a partition,

$$G - k = \{S \in I_k\}, k = 1, ..., K,$$

where $I_k = [\ell_{k-1}, \ell_k)$ are intervals that divide the support of S. Then, we want the conditional average effects within these bins,

$$E[s_0(Z)|G_k], k = 1, ..., K.$$

given the monotonicity restriction,

$$E[s_0(Z)|G_1] \leq ... \leq E[s_0(Z)|G_K].$$

4.4 Choosing an ML method

Options that they cover: (i) "ability to predict YH using BH and S" or (ii) "ability to predict Y using B and $(D - p(Z))(S - E[S])$ given $w(Z)$.

4.5 Applications

The application that is most relevant for us is the analysis of effect heterogeneity in the Morocco microfinance study. In this case, $p(Z) = 1/2$ for everyone. They try Random Forest, Elastic Net, Boosted Tree, and Neural Network, finding that the first two do best when judged using metrics geared toward the BLP and GATES analyses.

5 Wager and Athey (2018)

Link to the paper: JASA

They start with the same HT scaling result, expressing it in a different (but equivalent) manner (keeping with the notation from above):

$$E\left[Y\left(\frac{D}{p(Z)} - \frac{1-D}{1-p(Z)}\right)\middle|Z\right] = s_0(Z).$$

Now I am going to move over to Aaron's code.

6 Notes

6.1 Low dimensional features versus CATEs

In our applications, we may find ourselves doing something similar to Chernozhulov et al., in that we may have the ability to produce proxies for the target setting, but then have to take a low dimensional summary, insofar as policy makers may not be able to target on the basis of all of the covariate information available. We could adopt their approach to inference, then.

I should say that this feature of what Chernozhukov et al. are doing is very reminiscent of the "targeted learning" work of Van der Laan and coauthors.