Characterizing Effect Heterogeneity 1

Cyrus Samii

2019-06-03

Contents

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 Measuring 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,

```
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:

```
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.

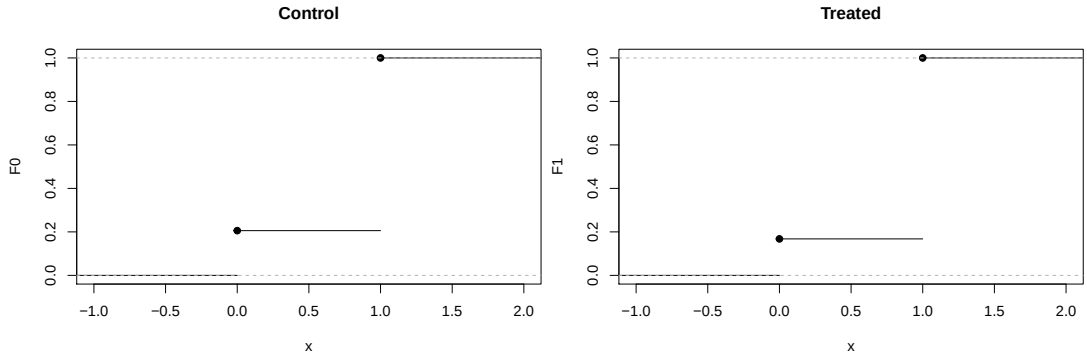
Such monotonicity is a strong assumption, however, and so we now turn to more general bounds.

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, $F_0(0) = P_{\cdot N}$ and $F_0(1) = 1$, while $F_1(0) = P_{N\cdot}$ 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:

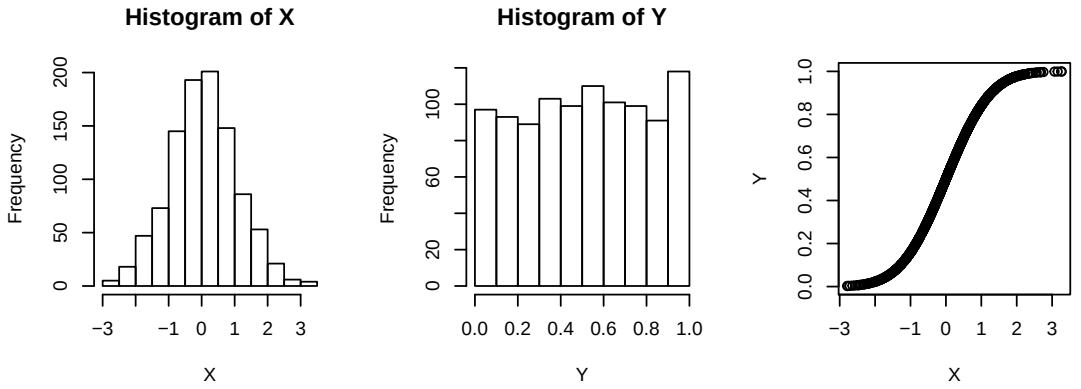
$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\},$$

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 $P_{\cdot N} < u < 1$. 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:

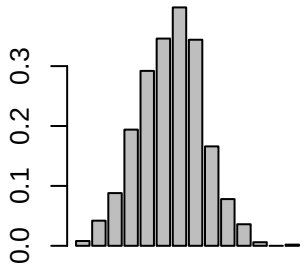
> If F is a bivariate CDF with marginals F_0 and F_1, then there exists a copula C such that Sklar I holds. Moreover, if the margins are continuous, then C is unique and equals,
>
> $$C(u_0, u_1) = F(F_0^{-1}(u_0), F_1^{-1}(u_1)).$$

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.

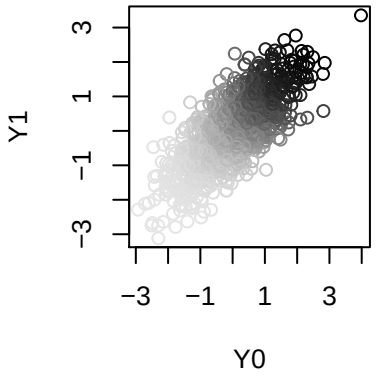
All of these properties hold for discrete variables as well, although an issue with discrete variables is that the copula for a given joint distribution may not be unique.

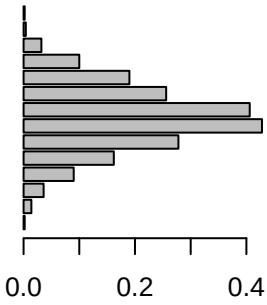
See for example a bivariate normal case.

2.1.1.1 Bivariate normal distribution

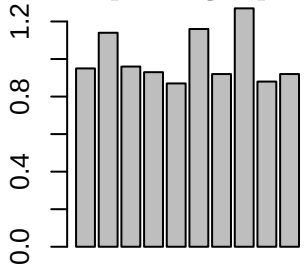


Bivariate Normal.
Shading is joint
CDF values.

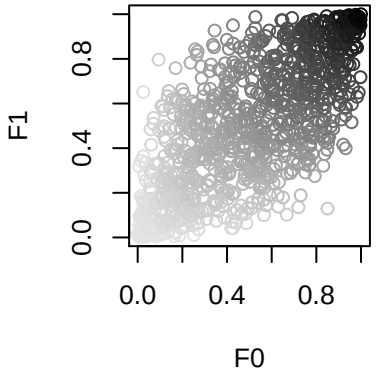


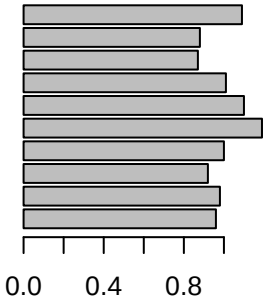


2.1.1.2 Corresponding copula

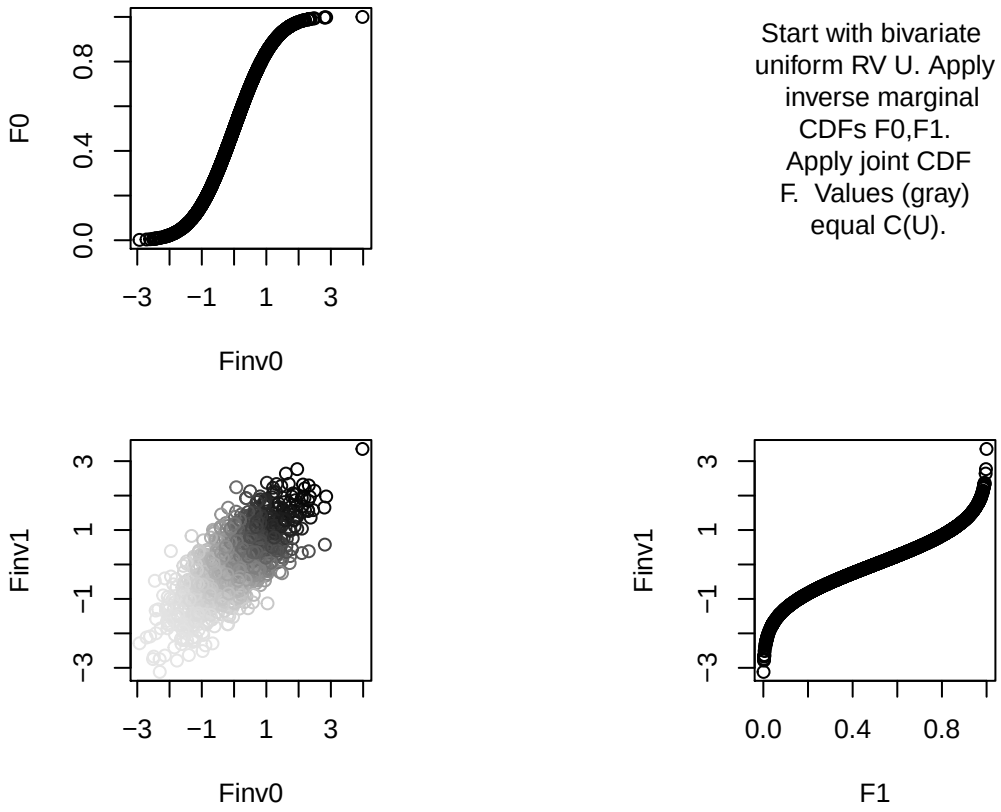


Joint distn of U.
Shading is copula
values.

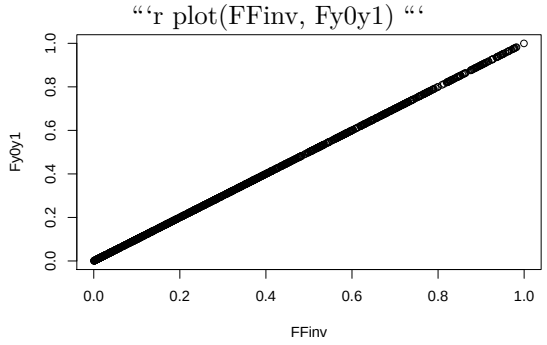




2.1.1.3 Illustrating of Sklar II, from CDF to copula



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



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

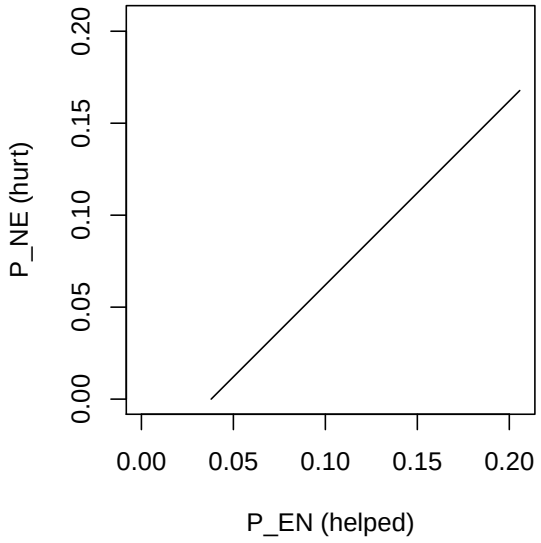
Returning to our example, recall that we are interested in the joint distribution of outcomes under treatment and control, where the outcome are each binary (either enrolled or not enrolled). Recall as well that we defined, e.g., P_{EE} to refer to the share of individuals for which potential outcomes 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 ($P_{E.}$ and $P_{.E}$) in each stratum.

Age	Male	Control	Treated	CATE	PEN_ub	PNE_ub
6	0	0.91	0.91	0.00	0.09	0.09
7	0	0.96	0.96	0.00	0.04	0.04
8	0	0.96	0.97	0.01	0.04	0.03
9	0	0.96	0.97	0.01	0.04	0.03
10	0	0.96	0.96	0.01	0.04	0.04
11	0	0.93	0.95	0.03	0.07	0.05
12	0	0.77	0.88	0.11	0.23	0.12
13	0	0.68	0.75	0.07	0.32	0.25
14	0	0.51	0.64	0.13	0.49	0.36
15	0	0.34	0.40	0.07	0.40	0.34
16	0	0.25	0.32	0.06	0.32	0.25
6	1	0.91	0.92	0.01	0.09	0.08
7	1	0.96	0.96	0.00	0.04	0.04
8	1	0.97	0.97	0.01	0.03	0.03
9	1	0.95	0.98	0.03	0.05	0.02
10	1	0.96	0.97	0.01	0.04	0.03
11	1	0.92	0.96	0.04	0.08	0.04
12	1	0.84	0.90	0.06	0.16	0.10
13	1	0.79	0.84	0.05	0.21	0.16
14	1	0.57	0.72	0.16	0.43	0.28
15	1	0.49	0.52	0.04	0.51	0.48
16	1	0.31	0.33	0.02	0.33	0.31

The table above shows outcome means, the conditional average treatment effect (CATE), and then the upper bounds on the share of those induced to enroll (EN) and those induced not to enroll NE) for strata

that distinguish between girls and boys aged between 6 and 16. The EN and NE shares are the upper bounds of the shares of units within the given stratum for which there may be a treatment effect. If effect monotonicity holds such that the treatment never induces anyone not to enroll, then the share of people within the given stratum that might be induced to enroll is equal to the value of the CATE.

When thinking about optimal treatment regimes, if the sum of EN and NE is small, then the potential for different treatment regimes to produce different outcome distributions is quite limited. In our application, for younger children, such limits are quite apparent. It is only for older youth, above the age of 12, where there is even the potential for there to be effect heterogeneity that is at all substantial. Under effect monotonicity the limits are even tighter.

2.2 Decomposing systematic and idiosyncratic treatment variation

Ding, Feller, and Miratrix (2019, *JASA*) consider the following decomposition of treatment effects:

$$\tau_i = Y_i(1) - Y_i(0) = X_i\beta + \epsilon_i,$$

where β is defined as the finite population OLS coefficient of the regression of τ_i on X_i. Then $X_i\beta$ is the systematic treatment effect variation explained by X and ϵ_i is the idiosyncratic variation. Note that if we could observe the τ_i, then we would have that

$$\beta = \gamma_1 - \gamma_0,$$

where γ_1 and γ_0 are the finite population coefficients from the regressions of the potential outcomes on the covariates, in which case ϵ_i is the differences in residuals from those two regressions. We cannot compute these quantities directly, but we can estimate them without bias (assuming random assignment) by working with regression fits to the treatment and control observations separately. These results motivate a test whereby on fits an interacted regression and then does a joint test on the coefficients for the terms that interact the treatment with covariates. This is all very natural – what is distinct in Ding et al.'s contribution is that their inferential methods are based solely on the randomization distribution. They also bound the extent of idiosyncratic variation using FH bounds like above. This is like what Djebbari and Smith (2008) do, essentially apply FH bounds to the residuals from an interacted regression.

I won't go into more detail on this now.