ORIE 6745: Causality and Learning for Intelligent Decision Making	September 13, 2017
Lecture 7-8	
Lecturer: Nathan Kallus	*Scribe: Uzi Smadja*

1 Recap

Last lecture, we talked about controlled experiments setting where we completely control the assignment & application of treatment.

1.1 Randomization

We can actually estimate average causal effects even without observing counterfactuals. Specifically, we discussed one random design: **complete randomization-** inference with permutation test.
$H_0 : \forall i = 1, 2, ..., n \quad Y_i(0) = Y_i(1) \Leftrightarrow TE_i = 0$

1.1.1 Permutation test procedure

1. Fix significance level $\alpha \in (0, 1)$
2. Draw $z_{1:n} \in CR_n$
3. Assign & apply treatments
4. Measure $Y_{1:n}$
5. Compute $\hat{\tau}$:
   (a) For each $z'_{1:n} \in CR_n$ compute $\hat{\tau}_{z'_{1:n}} = \frac{2}{n} \sum_i (-1)^{1+z'_i} Y_i$
   (b) set $p = \frac{\sum_i z'_{1:n} \in CR_n \mathbb{I}[|\hat{\tau}_{z'_{1:n}}| \geq |\hat{\tau}|]}{|CR_n|}$. This p is Fisher's exact p-value.
   (c) If $p \leq \alpha$ then reject H_0.

2 Randomization tests

when n is big $\Rightarrow |CR_n| = \binom{n}{n/2}$ is really big $\Rightarrow$ permutation test is hard(infeasible).

1. Fix a large B
2. Draw $z'_{1,1:n}, ..., z'_{B,1:n} \in CR_n$
3. let $p = \frac{\sum_{i=1}^{B} 1 + \mathbb{I}[|\hat{\tau}_{z'_{1:n}}| \geq |\hat{\tau}|]}{1+B}$

2.1 Alternative statistics

1. $\hat{\tau}^{log} = \frac{2}{n}\sum_i (-1)^{1+z_i} log(Y_i)$ if Y_i is skewed or if effect is multiplicative
2. $\hat{\tau}^{alt} = \frac{2}{n}\sum_{i:z_i=0} Y_i$ would lead to same p as $\hat{\tau}$
3. $\hat{\tau}^{med} = median\{y_i : Z_i = 1\} - median\{Y_i : Z_i = 0\}$ - robust to outliers
4. Kolmogorov-Smirnov(KS) statistic - let $\hat{F}_z(y) = \frac{2}{n}\sum_{Z_i=z} \mathbb{I}[Y_i \leq y]$
   $\hat{\tau}^{KS} = \underset{y}{sup}|\hat{F}_1(y) - \hat{F}_0(y)|$

2.2 Distributional tests

1. Now consider a random sample drawn from a population, a weaker null hypothesis (Neymanian), no average effects:
   $H_0 : PATE = E[TE] = E[Y(1) - Y(0)] = 0$ ($\forall i \quad TE_i = 0$ as R.V $\Rightarrow TE = 0$)
   Suppose $Y_i(z) \sim \mathcal{N}(\mu_z, \sigma^2) \Rightarrow PATE = \mu_1 - \mu_0$
   Let $\hat{\mu_z} = \frac{2}{n}\sum_{Z_i=z} Y_i$

$$\begin{aligned} s_z^2 &= \frac{1}{\frac{n}{2} - 1} \cdot \sum_{Z_i=z} (Y_i - \hat{\mu_z})^2 \\ &= \frac{1}{n-2}(\frac{n}{2} - 1)(s_0^2 + s_1^2) \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\mu_{z_i}})^2 \end{aligned}$$

   The t-statistic:
   $t = \frac{\hat{\tau}}{s} = \frac{\hat{\mu_1} - -\hat{\mu_0}}{s}$
   under the sampling distribution assumption and the null hypothesis H_0: $t \sim Student - T(n-2)$
   To test H_0, let $p = 1 - F_{|student-T(n-2)|}(|t|) = 2 - 2F_{student-T(n-2)}(|t|)$
   reject H_0 if $p \leq \alpha$
   When n is sufficiently large, the test is robust to violation of normality

2. We know that:
   Fisherian Null: $H_0 : TE = 0 asr.v \Rightarrow$ Neymanian null: $H_0 : E[TE] = 0$
   We get p_1 from the permutation test with the Fisherian null hypothesis and p_2 with t-test for the Neymanian null hypothesis(assume normal so p is exact). Can we say that $p_1 \leq p_2$? Not always (Further reading in Ding '17 in the course website)

3 Decision making & Null hypothesis testing

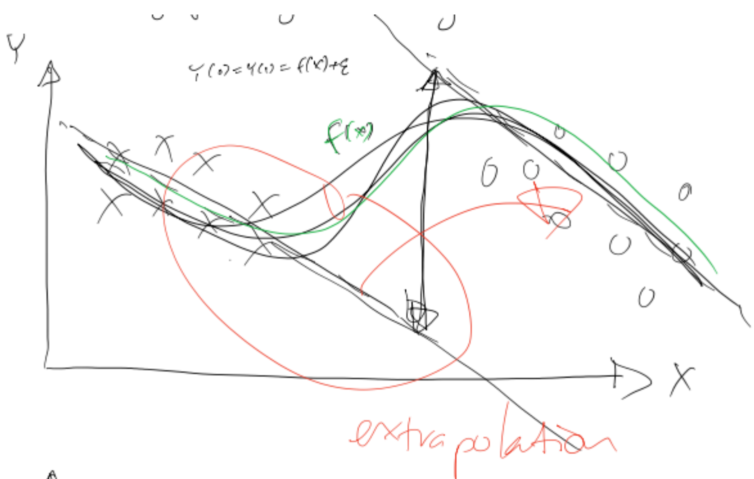
Suppose we're choosing between two alternatives: $z = 0$ always vs. $z = 1$ always
We conducted a controlled trial & measured $\hat{\tau}$ and note $sign(\hat{\tau})$ What to choose?

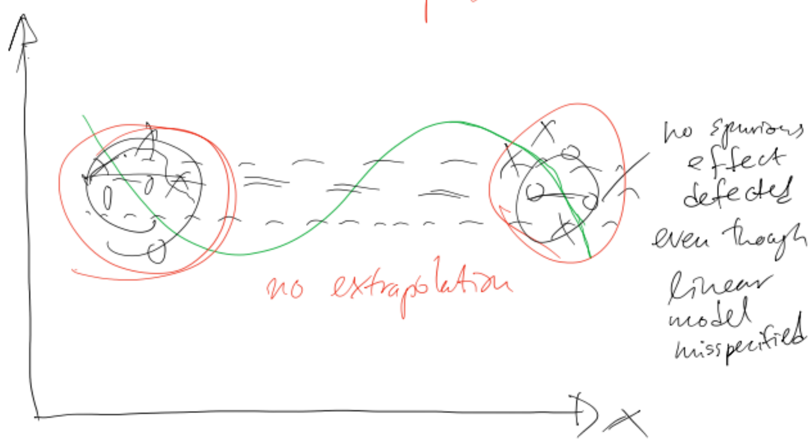
3.1 Regression adjustment

Fit a linear model $Y_i \sim \alpha + \tau z_i + \epsilon_i$
$\hat{\tau}^{OLS}, \hat{\alpha}^{OLS} = \underset{\alpha,\tau}{argmin} \sum_i (Y_i - \alpha - \tau z_i)^2$
$\hat{\tau}^{OLS} = \frac{\sum_i (z_i - \bar{z}_{1:n})(y_i - \bar{(y)}_{1:n})}{\sum_i (z_i - \bar{z}_{1:n})^2}$ where $\bar{z}_{1:n} = \frac{1}{n}\sum_i z_i = \frac{1}{2}$ for CR. also, $y_{1:n} = \frac{1}{n}\sum_i y_i$
$\hat{\alpha}^{OLS} = \bar{y}_{1:n} - \hat{\tau}^{OLS} \cdot \bar{z}_{1:n}$
With simple algebra we get $\hat{\tau}^{OLS} = \hat{\tau}$ and as before $\hat{\tau}$ is unbiased and consistent.

3.2 Regression adjustment with covariates

Observe x_i baseline covariates(pre-treatment)
Fit $Y_i \sim \alpha + \tau z_i + \beta^T x_i + \epsilon_i$
$\hat{\tau}^{OLS}, \hat{\alpha}^{OLS}, \hat{\beta}^{OLS} = \underset{\alpha,\tau,\beta}{argmin} \sum_i (Y_i - \alpha - \tau z_i - \beta^T x_i)^2$
$\Rightarrow \alpha^*, \tau^*, \beta^* = \underset{\alpha,\tau,\beta}{argmin} E[(Y - \alpha - \tau z - \beta^T x)]^2$
alternatively, we have: $Y_i \sim \alpha + \tau z_i + \beta^T x_i + \gamma^T (z_i(x_i - \bar{x})) + \epsilon_i$
If $y_i(z)$ is not linear then $\hat{\tau}^{OLS}$ is BIASED and inference may be invalid. but asymptotically we may be ok.





From Chapter 6 of Imbens & Rubin 15:
Under CR, $\tau^* = PATE$
$\sqrt{n}(\hat{\tau}^{OLS} - PATE) \sim \mathbb{N}(0, \frac{1}{4}E[(Y - \alpha^* - \tau^* z - \beta^{*T} x)^2])$

Asymptotically okay. Can reduce estimation but can still lead to bias and invalid inference in finite n if linear misspecified.

From Agnostic notes on regression adjustments to experimental data: Reexamining Freedmans critique (Lin '13):
It is better to add interaction terms:
$Y_i \sim \alpha + \tau z_i + \beta^T x_i + \gamma^T (z_i(x_i - \bar{x}_{1:n})) + \epsilon_i$
Alternative interpretation:

fit a linear predictor $\hat{f}_0(x) on (x_i, y_i) : z_i = 0$
fit a linear predictor $\hat{f}_1(x) on (x_i, y_i) : z_i = 1$
Measure $\hat{\tau} = \frac{1}{n}\sum_i(\hat{f}_1(x_i) - \hat{f}_0(x_i))$ where we impute potential outcomes using the predictive model.
$\Rightarrow$ Can actually use any ML predictive model - e.g. CART,BART,Neural Nets
proof:

$$\begin{aligned}
&\sum_i (Y_i - \alpha - \tau z_i - \beta^T x_i - \gamma^T (z_i(x_i - \bar{x}_{1:n})))^2 \\
= \sum_{i:z_i=0} (Y_i - \alpha - \beta^T x_i)^2 + &\sum_{i:z_i=1} (Y_i - (\alpha + \tau + \gamma^T \bar{x}_{1:n}) - (\beta + \gamma)^T x_i)^2 \\
&Where \quad \alpha' = \alpha + \tau + \gamma^T \bar{x}_{1:n} \\
&\beta' = \beta + \gamma \\
&\gamma, \tau \quad unrestricted
\end{aligned}$$

we get:
$\underset{\alpha,\beta,\alpha',\beta'}{argmin} \quad \sum$(squared error in control) + $\sum$(squared error in treatment)