※ Slides are uploaded on the website - materials!

Section 5

Observational Studies 1

Sooahn Shin

GOV 2003

Oct 7, 2021

Overview

- Logistics:
  - **No pset this week!**

- Today's topics:
  1. Review session
  2. No unmeasured confounding + regression

What we have learned so far?

- Fisher's approach to inference: randomization inference
- Neyman's approach to inference for the ATE: diff-in-means estimator

What we have learned so far?

- Fisher's approach to inference: randomization inference
- Neyman's approach to inference for the ATE: diff-in-means estimator

  $\hat{\tau}_{diff} \to V(\hat{\tau}) \to \hat{V}(\hat{\tau})$

- Analyzing experiments with regression
  - Simple OLS estimator + robust variance estimator
  - + Covariates
  - + Block design
  - + Cluster design

* unbiased
consistent
asymp. norm.

variance estimator
- conservative = positive bias
- efficient = small var

What we have learned so far?

- Fisher's approach to inference: randomization inference
- Neyman's approach to inference for the ATE: diff-in-means estimator
- Analyzing experiments with regression
  - Simple OLS estimator + robust variance estimator
  - + Covariates
  - + Block design
  - + Cluster design
- This week: observational studies
  - Before we move on, let's quickly review experimental designs!

Experimental design

- Types of experiments by their assignment mechanism
  - **Bernoulli randomization**: Each unit is assigned $D_i = 1$ with prob. p independently (coin flips)
  - **Completely randomized experiment**: Randomly sample n_1 units from the population to be treated
  - **Block/stratified randomized experiment**: Completely randomized experiment in each block $\rightsquigarrow$ always efficient for PATE
  - **Cluster randomized experiment**: Treatment assignment at a higher level $\rightsquigarrow$ allows for interference within clusters

Experimental design

- Types of experiments by their assignment mechanism
  - **Bernoulli randomization**: Each unit is assigned $D_i = 1$ with prob. p independently (coin flips)
  - **Completely randomized experiment**: Randomly sample n_1 units from the population to be treated
  - **Block/stratified randomized experiment**: Completely randomized experiment in each block $\rightsquigarrow$ always efficient for PATE
  - **Cluster randomized experiment**: Treatment assignment at a higher level $\rightsquigarrow$ allows for interference within clusters

- Exercise: comparing experimental designs through simulation
  1. Assume true potential outcomes
  2. Select one assignment mechanism
  3. Randomly generate treatment assignment
  4. Estimate SATE (using diff-in-means estimator)
  5. Repeat 3-4 multiple times
  6. Draw a distribution of estimates

Experimental design

- Setup:
  - SATE = $\frac{1}{16}\sum_{i=1}^{16}\tau_i = 8.5$
  - Design is balanced (except for Bernoulli)

Unit	$Y_i(0)$	$Y_i(1)$	τ_i	Block/Cluster
1	0	1	1	A
2	0	2	2	A
3	0	3	3	A
4	0	4	4	A
5	0	5	5	B
⋮	⋮	⋮	⋮	⋮
16	0	16	16	D

Experimental design

- Setup:
  - SATE = $\frac{1}{16}\sum_{i=1}^{16}\tau_i = 8.5$
  - Design is balanced (except for Bernoulli)

Unit	$Y_i(0)$	$Y_i(1)$	τ_i	Block/Cluster
1	0	1	1	A
2	0	2	2	A
3	0	3	3	A
4	0	4	4	A
5	0	5	5	B
⋮	⋮	⋮	⋮	⋮
16	0	16	16	D

- Q: Which design would have the largest (smallest) variance?

- Check the results here: https://twitter.com/aecoppock/status/1442545254423486465?s=21

Observational studies

- Problem:
  - Non-randomized treatment
  - $\rightsquigarrow \{Y_i(1), Y_i(0)\} \not\perp D_i$
  - $\rightsquigarrow$ selection bias = unidentified ATT

$$\underbrace{\mathbb{E}[Y_i|D_i=1] - \mathbb{E}[Y_i|D_i=0]}_{\text{diff-in-means}} = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i=1] - \mathbb{E}[Y_i(0)|D_i=0]}_{\text{selection bias}}$$

① consistency

② $+ E[Y_i(0) | D_i=1] - E[Y_i(0) | D_i=1]$

Observational studies

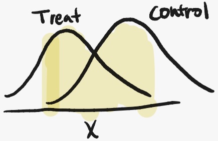
- Problem:
  - Non-randomized treatment
  - $\rightsquigarrow \{Y_i(1), Y_i(0)\} \not\perp\!\!\!\perp D_i$
  - $\rightsquigarrow$ selection bias $=$ unidentified ATT

$$\underbrace{\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]}_{\text{diff-in-means}} = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- What can we do for the **identification**?
  - Assume no unmeasured confounding with positivity
  - Partial identification: analysis of bounds for the ATE
  - Sensitivity analysis ...

Identification: No unmeasured confounding

- Identification
    - Let's begin with most common set of assumptions:
        1. **Overlap**/Positivity: $0 < \Pr[D_i = 1|\mathbf{X}_i] < 1$
        2. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid X_i$



Identification: No unmeasured confounding

- Identification
  - Let's begin with most common set of assumptions:
    1. **Overlap**/Positivity: $0 < \Pr[D_i = 1|\mathbf{X}_i] < 1$
    2. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid X_i$
  - This will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i(1) - Y_i(0) \mid X_i]\} \quad \text{iter.} \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(0) \mid X_i]\} \quad \text{linear.} \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i(1) \mid D_i = 1, X_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, X_i]\} \quad \text{n.u.c.} \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i]\} \quad \text{consistency.}
\end{aligned}
$$

Identification: No unmeasured confounding

- Identification
  - Let's begin with most common set of assumptions:
    1. **Overlap**/Positivity: $0 < \Pr[D_i = 1|\mathbf{X}_i] < 1$
    2. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid X_i$
  - This will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i(1) - Y_i(0) \mid X_i]\} \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(0) \mid X_i]\} \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i(1) \mid D_i = 1, X_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, X_i]\} \\
&= \mathbb{E}_{\mathbf{X}}\{E[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i]\}
\end{aligned}
$$

- Estimation
  - Regression
  - Matching/Weighting (Module 7)

Estimation: Regression-based estimators

- Treated and control conditional expectation functions (CEFs):

$$\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$$

- By consistency and no unmeasured confounding:

$$\underbrace{\mu_1(\mathbf{x})}_{\text{counterfactual}} = \underbrace{\mathbb{E}[Y_i \mid D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\text{observational}}, \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i = \mathbf{x}]$$

Estimation: Regression-based estimators

- Treated and control conditional expectation functions (CEFs):

$$\mu_1(\mathbf{x}) = \mathbb{E}[\,Y_i(1) \mid \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[\,Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$$

- By consistency and no unmeasured confounding:

$$\underbrace{\mu_1(\mathbf{x})}_{\text{counterfactual}} = \underbrace{\mathbb{E}[\,Y_i \mid D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\text{observational}}, \qquad \mu_0(\mathbf{x}) = \mathbb{E}[\,Y_i \mid D_i = 0, \mathbf{X}_i = \mathbf{x}]$$

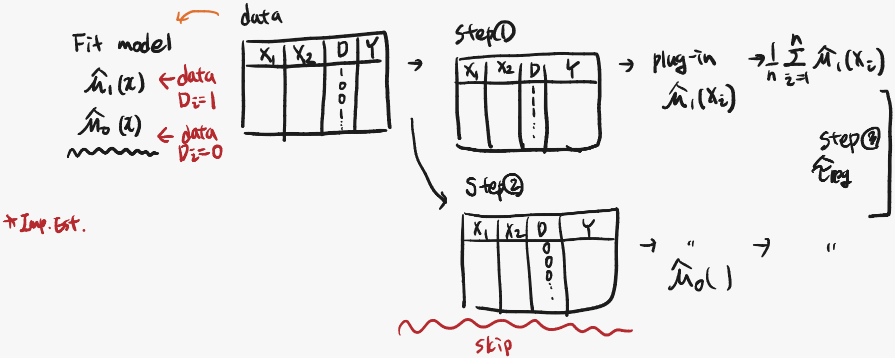
- Estimate CEFs using regression estimators $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$.
  - Might be linear or nonlinear models (e.g., GAMs)
  - $\rightsquigarrow$ Regression estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

Estimation: Regression-based estimators

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- General procedure:
  - Obtain predicted values for all units when $D_i = 1$.
  - Obtain predicted values for all units when $D_i = 0$.
  - Take the average difference between these predicted values.



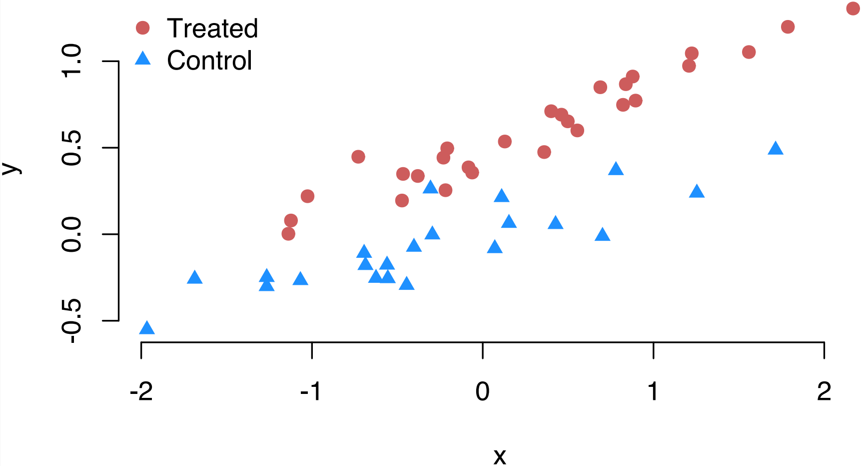
Estimation: Regression-based estimators

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- General procedure:
  - Obtain predicted values for all units when $D_i = 1$.
  - Obtain predicted values for all units when $D_i = 0$.
  - Take the average difference between these predicted values.
- Safest practice:
  - Estimate separate regression in each treatment group.
  - Sometimes called an imputation estimator.
  - Procedure:
    - Regress Y_i on X_i in the treatment group and get predicted values for all units (treated or control).
    - Regress Y_i on X_i in the control group and get predicted values for all units (treated or control).
    - Take the average difference between these predicted values.

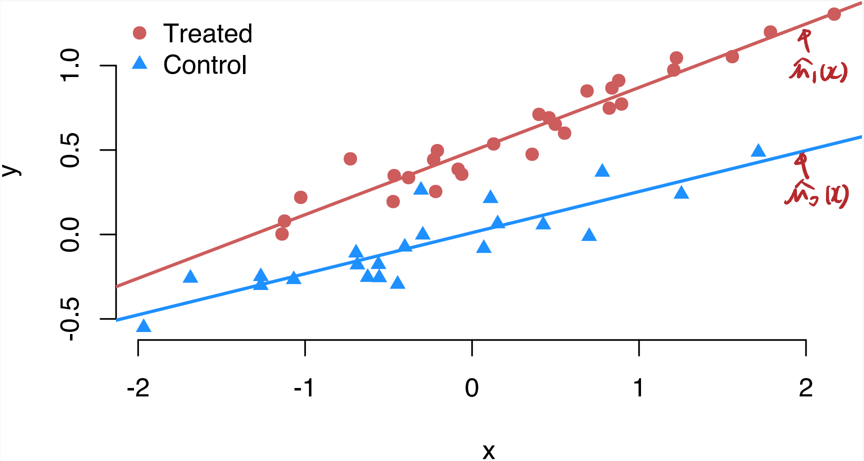
Toy example

- Data is as follows and we will use linear regression to estimate CEFs



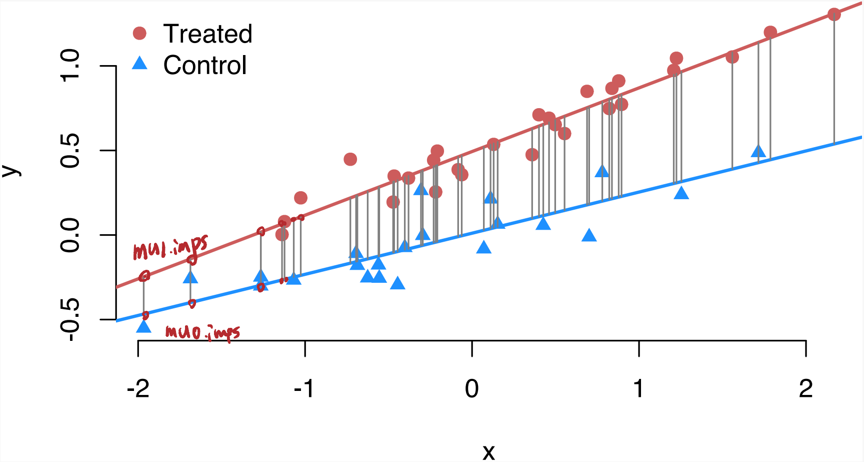
Imputation estimator visualization

```
mod0 <- lm(y~x, data = toy_data, subset = d==0)
mod1 <- lm(y~x, data = toy_data, subset = d==1)
```



```r
mu0.imps = predict(mod0, toy_data); mu1.imps = predict(mod1, toy_data)
cat("Estimate of ATE:", mean(mu1.imps - mu0.imps))
```

```
## Estimate of ATE: 0.4873975
```



Fully interacted OLS visualization

- What if $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$ are from fully interacted OLS with centered covariates?
  - Equivalent to running separate models for $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
    - Recall: Under linear models, $\widehat{\tau}_{\text{reg}}$ is **sometimes** equivalent to a coefficient.

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau} D_i + \tilde{X}_i \hat{\beta} + \tilde{X}_i D_i \hat{\gamma}$$

$$\hat{\mu}_1(X_i) = (\hat{\alpha} + \hat{\tau}) + \tilde{X}_i (\hat{\beta} + \hat{\gamma})$$

$$\hat{\mu}_0(X_i) = \hat{\alpha} + \tilde{X}_i \hat{\beta}$$

$$(\hat{\alpha}, \hat{\tau}, \hat{\beta}, \hat{\gamma}) = \underset{(\alpha, \tau, \beta, \gamma)}{\arg\min} \sum_i (Y_i - \alpha - \tau D_i - \tilde{X}_i \beta - \tilde{X}_i D_i \gamma)^2$$

$$= \underset{(\alpha, \tau, \beta, \gamma)}{\arg\min} \sum_{i: D_i = 0} (Y_i - \alpha - \tilde{X}_i \beta)^2 + \sum_{i: D_i = 1} (Y_i - \underbrace{(\alpha + \tau)}_{\alpha'} - \tilde{X}_i \underbrace{(\beta + \gamma)}_{\beta'})^2$$

$$= \underset{(\alpha, \beta)}{\arg\min} \sum_{i: D_i = 0} \text{''} + \underset{(\alpha', \beta')}{\arg\min} \sum_{i: D_i = 1} \text{''}$$

$$\hat{\tau}_{reg} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\tau} + \tilde{X}_i \hat{\gamma})$$

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i = 0$$

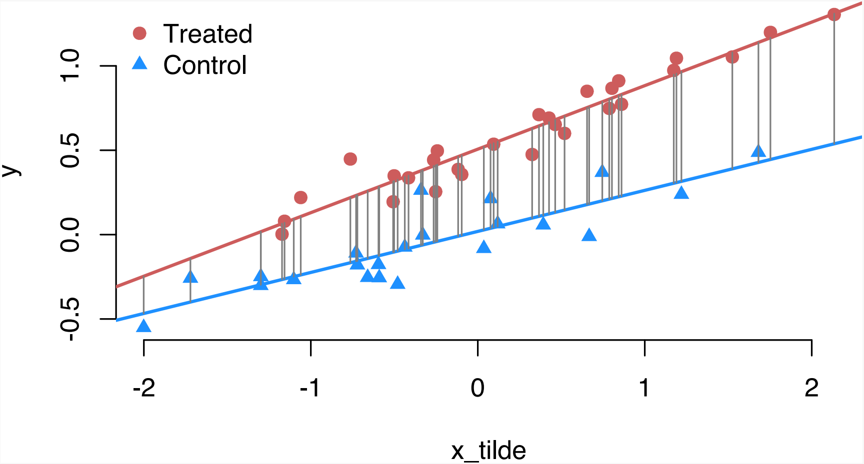
$$= \frac{1}{n} \sum_{i=1}^{n} \hat{\tau} = \hat{\tau}$$

Fully interacted OLS visualization

- What if $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$ are from fully interacted OLS with centered covariates?
  - Equivalent to running separate models for $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
    - Recall: Under linear models, $\widehat{\tau}_{\text{reg}}$ is **sometimes** equivalent to a coefficient.

```
toy_data$x_tilde <- toy_data$x - mean(toy_data$x)
mod_full <- lm(y~d+x_tilde+d*x_tilde, data = toy_data)
dat0 <- toy_data %>% mutate(d = 0); dat1 <- toy_data %>% mutate(d = 1)
mu0.full = predict(mod_full, dat0); mu1.full = predict(mod_full, dat1)
cat("Estimate of ATE (Fully interacted):", mean(mu1.full - mu0.full),
    "\nEstimate of ATE (Imputation):", mean(mu1.imps - mu0.imps),
    "\nEstimated coefficient on Di", mod_full$coefficients["d"])
```

```
## Estimate of ATE (Fully interacted): 0.4873975
## Estimate of ATE (Imputation): 0.4873975
## Estimated coefficient on Di 0.4873975
```



Treated
Control
y
1.0
0.5
0.0
-0.5
-2
-1
0
1
2
x_tilde

Uninteracted OLS visualization

- What if $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`?
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i

Uninteracted OLS visualization

- What if $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`?
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i

$Y_i = \hat{\tau} D_i + X_i'\beta + \varepsilon_i$

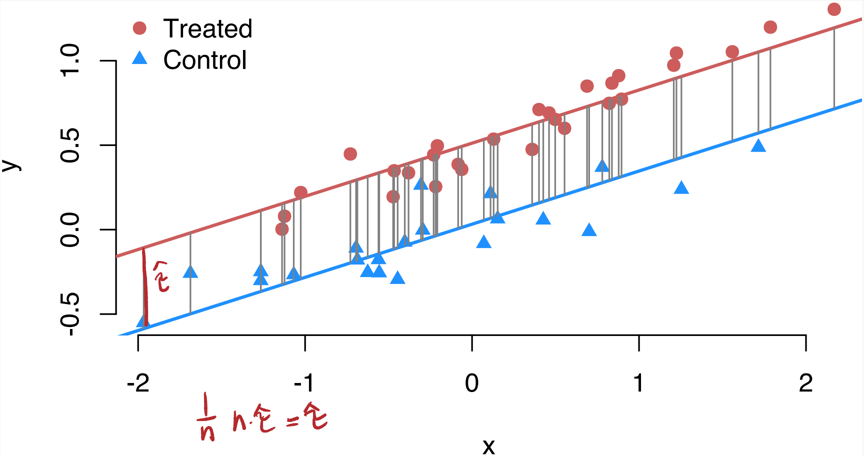
$\hat{\mu}_1(X) = \hat{\tau} + X_i'\hat{\beta}$

$\hat{\mu}_0(X_i) = X_i'\hat{\beta}$

```r
mod <- lm(y~d+x, data = toy_data)
mu0 = predict(mod, dat0); mu1 = predict(mod, dat1)
cat("Estimate of ATE (Uninteracted):", mean(mu1 - mu0),
    "\nEstimated coefficient on Di", mod$coefficients["d"],
    "\nEstimate of ATE (Fully interacted):", mean(mu1.full - mu0.full),
    "\nEstimate of ATE (Imputation):", mean(mu1.imps - mu0.imps))
```

```
## Estimate of ATE (Uninteracted): 0.479676
## Estimated coefficient on Di 0.479676
## Estimate of ATE (Fully interacted): 0.4873975
## Estimate of ATE (Imputation): 0.4873975
```

$$\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau} = \frac{1}{n}\cdot n\hat{\tau} = \hat{\tau}$$


Treated
Control
1.0
0.5
0.0
-0.5
y
-2
-1
0
1
2
x

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- **Nonparametric bootstrap**
  - Recall: Source of variance is due to **sampling**
  - Idea: View sample (data) as "population" → in-sample "sampling"

$$F \to \text{data} \to \hat{\theta}$$

$$\hat{F}_n \to \text{data}_{(s)} \to \hat{\theta}_{(s)}$$

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- **Nonparametric bootstrap**
  - Recall: Source of variance is due to **sampling**
  - Idea: View sample (data) as "population" → in-sample "sampling"
- Procedure:
  - Randomly resample n rows of the data with replacement
  - Refit the regressions on the bootstrapped data.
  - Calculate $\widehat{\tau}_{\text{reg}}$ in each bootstrap
  - Repeat several times and use empirical variance of the bootstraps

Bootstrap sample codes

```
set.seed(02138); sims<-500; tau_hat_draws<-rep(NA, sims)
for (i in 1:sims) { # Repeat the following several times
  # 1. Randomly resample n rows of the data with replacement
  sample_boot <- dplyr::slice_sample(toy_data, n = nrow(toy_data),
                                     replace = TRUE)
  # 2. Refit the regressions on the bootstrapped data
  model <- lm(y ~ d + x_tilde + d*x_tilde, data = toy_data)
  dat1 <- sample_boot; dat1$d <- 1
  dat0 <- sample_boot; dat0$d <- 0
  mu1_hat  <- predict(model, newdata = dat1)
  mu0_hat  <- predict(model, newdata = dat0)
  # 3. Calculate tau_hat in each bootstrap
  tau_hat_draws[i]  <- mean(mu1_hat - mu0_hat)
}
# 4. Use empirical variance of the bootstraps
var(tau_hat_draws)
```

```
## [1] 0.0003247686
```

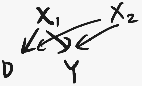
DAG

- How do we know if no unmeasured confounders holds?
    - One way: use DAGs and look at back-door paths.

DAG

- How do we know if no unmeasured confounders holds?
  - One way: use DAGs and look at back-door paths.
- **D-separation**
  - Can we determine conditional independence from our causal DAG?
  - Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
    1. Find all paths between A and B.
    2. Check if each path is **blocked**.
    3. If all paths are blocked, then A is **d-separated** from B by C

DAG



- How do we know if no unmeasured confounders holds?
  - One way: use DAGs and look at back-door paths.
- **D-separation**
  - Can we determine conditional independence from our causal DAG?
  - Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
    1. Find all paths between A and B.
    2. Check if each path is **blocked**.
    3. If all paths are blocked, then A is **d-separated** from B by C
- Ways to block $A \to B$ (each is a node):
  1. $A \to C \to B$, C is observed (conditioned)
  2. $A \leftarrow C \leftarrow B$, C is observed
  3. $A \leftarrow C \to B$, C is observed
  4. $A \to C \leftarrow B$, C is **not** observed
  - If C observed $\rightsquigarrow$ collider bias
  - e.g., A=bicycle accident, B=stomachache, C=hospitalization; Sackett (1979)