Quantitative Research Methods IV - 17.806

Recitation, Week 7.

Topic: Causal Machine Learning II and Longitudinal Data.

Benjamín Muñoz

March 24, 2023

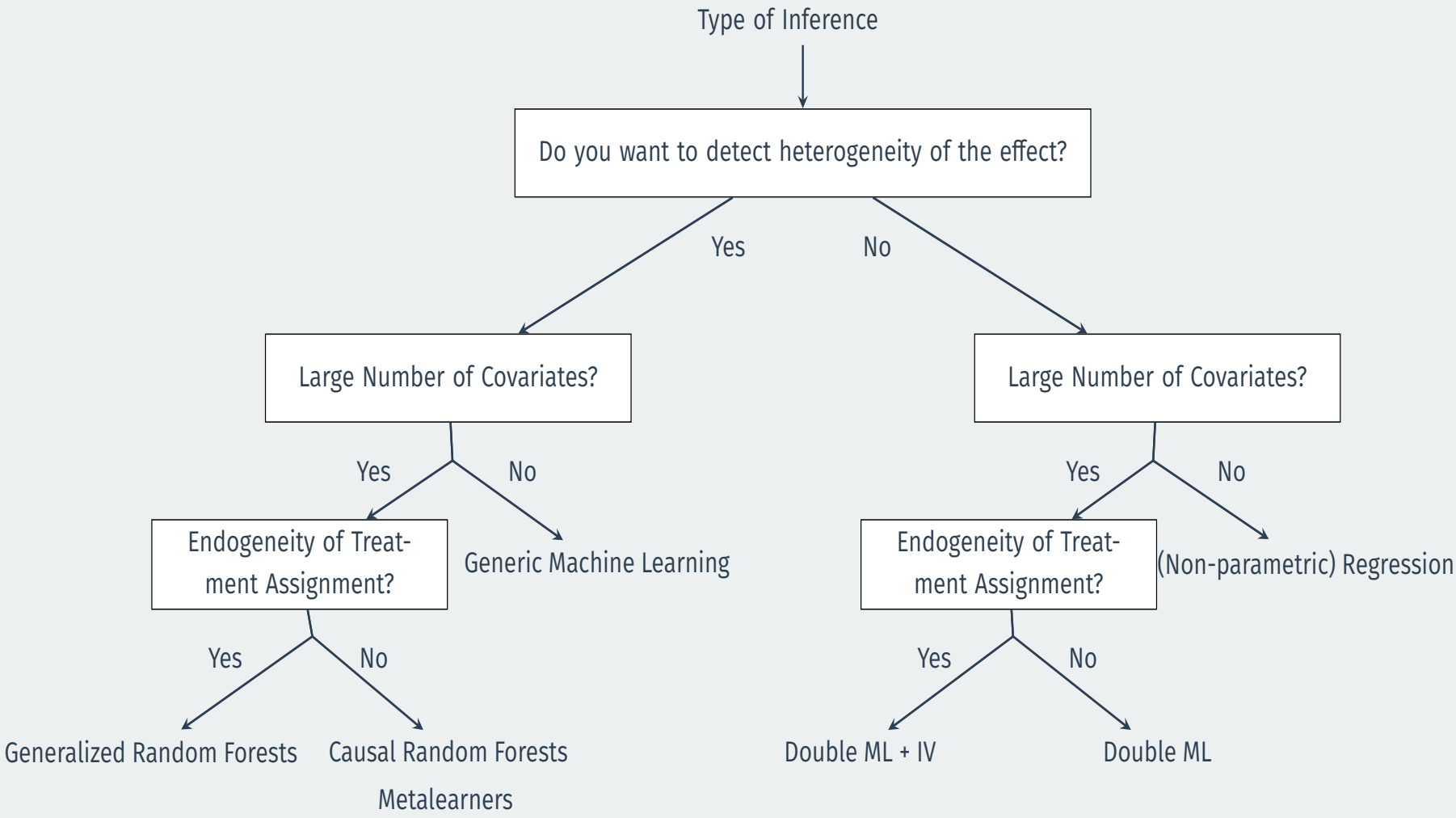
MIT

Massachusetts Institute of Technology

Table of contents

1/ Introduction

Causal Machine Learning



Heterogeneous Treatment Effects

- Degree to which different treatments have differential causal effects on each unit.
- Multiple justifications:
  1. Alternative Quantities of Interest.
  2. Selecting the most effective treatment (optimal allocation).
  3. Avoid strong functional form assumptions.
  4. Subgroup Analysis: identifying subgroups of observations for which the treatment in particularly efficacious or deleterious.
- **Risk of Multiple Testing Problem**:
  1. Ex Ante: Preregistration Plan of Study (specify relevant subgruops).
  2. Ex Post: Statistical Adjustment and Specific Techniques.

Multiple testing

- Suppose you have 20 hypotheses in your mind.
- Set $\alpha = 0.05$ (significance level)
- What is the probability of getting at least one significant result just by chance?

$$\begin{aligned}\Pr(\text{at least one significant result}) &= 1 - \Pr(\text{no significant result}) \\ &= 1 - (1 - 0.05)^{20} \approx 0.64\end{aligned}$$

- This shows that the probability of getting significant result is 60% even if all of the tests are actually not significant, much higher than $\alpha = 0.05$ (our procedure is not achieving the theoretical error rate).

Multiple testing

R Code

```
#A) When you run the test independently
pval <- rep(NA, 1000)

for(k in 1:100) {
        x       <- rnorm(n = 1000, mean = 0, sd = 1)
        pval[k] <- t.test(x = x, mu = 0)$p.value
}
mean(pval < 0.05)
## [1] 0.052

#B) When you conduct tests for 20 variables simultaneously
n    <- 20
pval <- matrix(NA, nrow = 1000, ncol = n)

for(k in 1:1000) {
          X <- replicate(n = n, expr = rnorm(100))
  pval[k, ] <- apply(X, 2, function(x)t.test(x=x, mu=0)$p.value)
}
# for each row check if there is any statistically significant results
mean(sapply(1:1000, function(i)any(pval[i, ] < 0.05)))
## [1] 0.631
```

Potential Solutions

1. Pre-registering Plan (including all the details, not just the study design).

2. Statistical adjustments:
   - Bonferroni correction: set cut-off at α / n.
   - Multiple options for adjustment (see **Ratkovic, 2021**).

3. **Honest** Procedures: Machine Learning techniques with sampling splitting steps.

R Code

```
#B) When you conduct tests for 20 variables simultaneously
# Correction of P-Values
mean(sapply(1:1000, function(i)any(pval[i, ] < 0.05 / n)))
## [1] 0.049
```

2/ Heterogeneous Treatment Effects

Conditional Average Treatment Effects

- CATE: $\tau(D; X) = \mathbb{E}(Y_i^1 - Y_i^0 | X_i = x)$

1. Linear Interactive Model

$$Y = \beta_0 + \beta_1 D + \beta_2 M + \beta_3 D \times M + \gamma X + \epsilon$$

   - **Extra Assumptions:** the treatment effect varies linearly with M, and for each value of M, the treatment and control groups should have a sufficient number of overlapping cases (see **Hainmueller, Mummolo, and Xu, 2019**).

2. Machine learning tools to estimate the heterogeneous treatment effects:
   - Loss function approach: Squared Loss SVM with separate LASSO constraints (**Imai and Ratkovic, 2013**), R-learner (**Nie and Wager, 2021**).
   - Construct potential outcomes: X-learner (**Künzel et al., 2019**), Causal Forests (**Wager and Athey, 2018**).

Causal Forests

- **Tree:** method that recursively partitions the high-dimensional covariate space into smaller units.
- Prediction ≠ Detection of Heterogeneity.
  - Minimization of within-leaf variance of Y (all the cases belonging to the same leaf should be homogeneous in terms of the outcome) $\rightsquigarrow$ Minimization of the within-leaf variance of the estimated treatment effects (inter-leaf variation should be large).
  - New Splitting Rule:

$$-\widehat{\mathrm{EMSE}_\tau}(\mathrm{S}^{\mathrm{Tr}}, \mathrm{N}^{\mathrm{est}}, \sqcap) = \underbrace{\frac{1}{N^{Tr}} \sum_{i \in S^{Tr}} \hat{\tau}^2(X_i | S^{Tr}, \sqcap)}_{\substack{\text{Variance of Treatment Effects} \\ \text{across Leaves} \\ \text{Prefer leaves with} \\ \text{Heterogeneous Effects}}} \underbrace{- \left(\frac{1}{N^{Tr}} + \frac{1}{N^{Est}}\right) \sum_{l \in \sqcap} \left(\frac{S^2_{S^{Tr}_{Treat}}(l)}{p} + \frac{S^2_{S^{Tr}_{Control}}(l)}{1-p}\right)}_{\substack{\text{Uncertainty about Leaf} \\ \text{Treatment Effects} \\ \text{Prefer leaves with Good Fit} \\ \text{(Leaf-Specific Effects estimated} \\ \text{Precisely)}}}$$

Causal Forests

- **Honest Procedure:** not use the same information for selecting the model structure as for estimation given a model structure.
    - One (independent) split of the data is used to learn the tree structure/-partition, and the second split of the data is used to conduct inference (estimation of treatment effects).

R Code

```
# Run Causal Forests (Basic Algorithm)
library(grf)

# Estimate Causal Forest
cf <- causal_forest(X = X, Y = Y, W = W, num.trees = 10000,
                    honesty = TRUE, honesty.fraction = 0.5,
                    tune.parameters = "all", seed = 17806)

# Estimate Predicted Values (CATEs)
pred <- predict(object = cf, newdata = newX,
                estimate.variance = TRUE)$predictions

# Use expanded algorithm for Problem 3.
```

Metalearners

- Meta-aalgorithms decompose estimating the CATE into several subregression problems that can be solved with any supervised ML method.
- Combination of **base learners** in a specific manner while allowing the base learners to take any form.

1. **S-Learner** (Single) $\rightsquigarrow$ using all of the features and the treatment indicator (without giving to D a special role).

$$\mu(x) = \mathbb{E}[Y|X = x, D = d]$$
$$\hat{\tau}(x) = \hat{\mu}(x, D = 1) - \hat{\mu}(x, D = 0)$$

   - Risk of dropping the treatment.
   - Low statistical efficiency.

Metalearners

2. **T-Learner** (Two) $\rightsquigarrow$ use base learners to estimate the conditional expectations of the outcomes separately for control and treatment groups.

$$\mu_0(x) = \mathbb{E}[Y|D = 0, X = x]$$
$$\mu_1(x) = \mathbb{E}[Y|D = 1, X = x]$$
$$\hat{\tau}(x) = \hat{\mu_1}(x) - \hat{\mu_0}(x)$$

- Ignore group size.
- Prediction $\neq$ Heterogeneity.

Metalearners

3. **X-Learner** $\rightsquigarrow$ uses each observation in the training set in an X-like shape (Sample Splitting for Fundamental Problem of Causal Inference).

   3.1 Estimate the response functions:

$$\mu_0(x) = \mathbb{E}[Y|D = 0, X = x]$$
$$\mu_1(x) = \mathbb{E}[Y|D = 1, X = x]$$

   3.2 Impute the individual treatment effects (for the Treated group with the control-outcome estimator and for the Control group with the treatment-outcome estimator):

$$\tilde{D}_i^1 = Y_i^1 - \hat{\mu}_0(X_i^1)$$
$$\tilde{D}_i^0 = \hat{\mu}_1(X_i^0) - Y_i^0$$

Metalearners

3.3 Use any learner to estimate/predict the imputed treatment effects for each group:

$$\hat{\tau}_1(x) = \mathbb{E}[\tilde{D}_i^1 | D = 1, X = x]$$
$$\hat{\tau}_0(x) = \mathbb{E}[\tilde{D}_i^0 | D = 0, X = x]$$

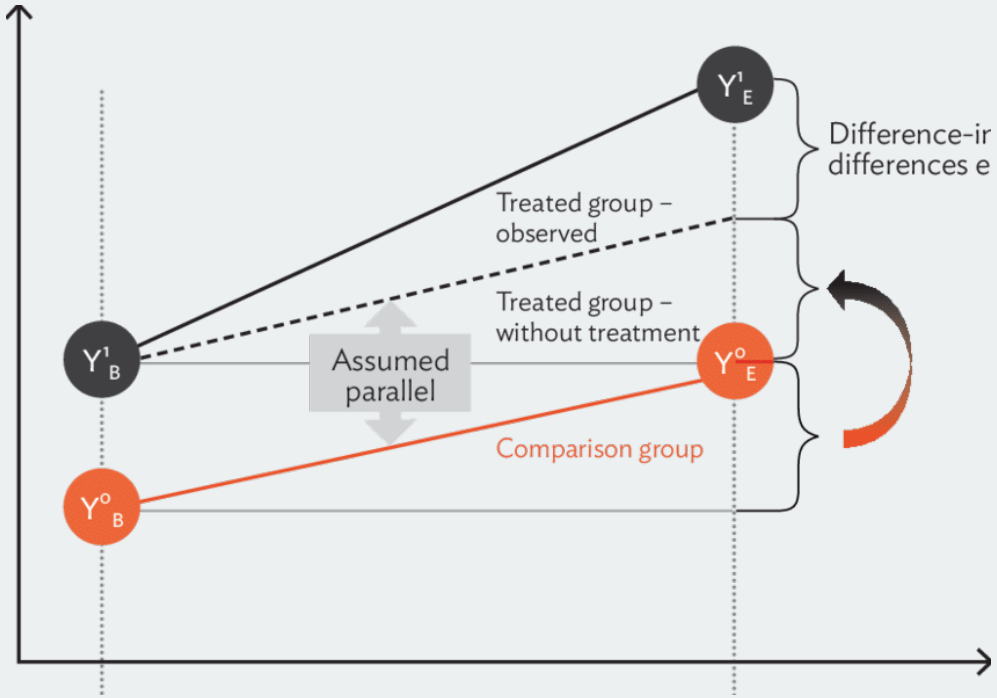
3.4 Define the CATE estimate by the weighted average of the two estimates in 3.3:

$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$$

- $g(x) \in [0, 1]$. Good option =Propensity Score.
- Efficient use of unbalanced design.

3/ Longitudinal Causal Inference

Difference-in-Differences



	Pre (t=0)	**Post** (t=1)	**Post - Pre**
Treated (d=1)	$\bar{y}_{1,0}$	$\bar{y}_{1,1}$	$\bar{y}_{1,1} - \bar{y}_{1,0}$
Control (d=0)	$\bar{y}_{0,0}$	$\bar{y}_{0,1}$	$\bar{y}_{0,1} - \bar{y}_{0,0}$
Tr - Ctr	$\bar{y}_{1,0} - \bar{y}_{0,0}$	$\bar{y}_{1,1} - \bar{y}_{0,1}$	$(\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{0,1} - \bar{y}_{0,0})$

Difference-in-Differences

- **Quantity of Interest:** Average Treatment Effect on the Treated (ATT).

$$Y = \beta_0 + \beta_1 Time + \beta_2 Group + \beta_3 Time \times Group + \epsilon$$

	Pre (t=0)	**Post** (t=1)	**Post - Pre**
Treated (d=1)	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Control (d=0)	β_0	$\beta_0 + \beta_2$	β_2
Tr - Ctr	β_1	$\beta_1 + \beta_3$	β_3

- In a 2×2, our estimate of the ATT is $\widehat{\beta_3}$.

- Extensions:
  - Correction of Standard Errors.
  - **Controls:** $Y_{idt} = \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Group} + \beta_3 \text{Time} \times \text{Group} + \gamma X + \epsilon$
  - Many Groups or Many Time Periods without Treatment Timing Variation.

Recent Advances

Mantra

- Multi-Period DID $\neq$ TWFE
- $Y_{it} = \theta_t + \eta_i + \alpha D_{it} + \beta X_{it} + \upsilon_{it}$
- TWFE only work if the Treatment Effect is α for all units (Strong Assumption).
- Multi-Period DID = Weighted TWFE
- Sometimes the weights are negative (Dynamic effects: ATT vary over time).

- **Typical Assumption** $\rightsquigarrow$ Staggered Adoption
  - Not all treated units are treated at the same time.
  - Once a unit participates in the treatment, they remain treated.

Goodman-Bacon Decomposition

- TWFE is a weighted average of all possible 2×2 DiD estimators that compare timing groups to each other.
- The weights are proportional to timing group sizes and the variance of the treatment dummy in each pair (highest for unit treated in the middle of the panel).
- $\text{plim}_{N \to \infty} \beta^{DD} = \beta^{DD} = \text{VWATT} + \text{VWCT} - \Delta\text{ATT}$
- Use of Already-treated groups as controls: substract average changes in their untreated outcomes and their treatment effects (DANGER!). This is the source of the negative weights (Bias). ΔATT equals zero if effects are constant (homogeneity assumption).

Goodman-Bacon Decomposition

R Code

```
### Load packages
library(bacondecomp)

### Run TWFE
fit_tw <- lm(l_homicide ~ post + factor(state) + factor(year),  data = bacondecomp::castle)

#         term    estimate std.error statistic   p.value
#         post      0.0818    0.0317     2.58  1.02e-  2

### Run Goodman-Bacon Decomposition
df_bacon <- bacon(formula = l_homicide ~ post,  data = bacondecomp::castle,  id_var = "state",
                  time_var = "year")

### Summary of the Decomposition (25 pairs)
bacon_summary(df_bacon)

#>                        type  weight  avg_est
#> 1 Earlier vs Later Treated 0.05976 -0.00554
#> 2 Later vs Earlier Treated 0.03190  0.07032
#> 3     Treated vs Untreated 0.90834  0.08796

### Weighted average of the Decomposition
sum(df_bacon$estimate * df_bacon$weight)
# 0.08181162
```

Goodman-Bacon Decomposition