# Data Science Lab - 5

## Academic year 2019-2020

**Lecturer Falco J. Bargagli-Stoffi**

IMT School for Advanced Studies Lucca & KU Leuven

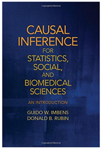
# *Causal Inference - Introduction*

# Why Causal Inference?

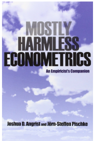
- Correlation is not causation: simple correlations can lead to misguided policies
- Among many different options, important to choose the *most effective* intervention
- Accurate cost-benefit analysis

# Causality Frameworks

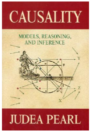
- Rubin Causal Model (Imbens & Rubin, 2015)




- Angrist & Pischke (2009)




- Pearl (2000)



# RCM (1980): Potential Outcomes Framework

# RCM (1980): Set Up

- Rubin's potential outcome framework (1974):
  - Given a set of $N$ units, indexed by $i = 1, ..., N$. Let $W_i$ be the binary indicator of the reception of the treatment:

$$W_i \in \{0, 1\}$$

  - Given this notation and SUTVA we can postulate the existence of a pair of potential outcomes for each unit:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) & if\ W_i = 0 \\ Y_i(1) & if\ W_i = 1 \end{cases}$$

  - We can define the Causal Effect as a simple difference between the potential outcome under treatment and under control:

$$\tau_i = Y_i(1) - Y_i(0)$$

# RCM (1974): Science World

- Imagine that we want to assess the effect (*causal effect*) of a job training (*treatment*) on a pool of students (*units*)

| ID | Education $X_i$ | Treated $W_i$ | No job training $Y_i(0)$ | Job training $Y_i(1)$ | Treatment effect $\tau_i = Y_i(1) - Y_i(0)$ |
|---|---|---|---|---|---|
| 1 | High school | 0 | 0 | 1 | 1 |
| 2 | High school | 1 | 0 | 1 | 1 |
| 3 | High school | 1 | 1 | 1 | 0 |
| 4 | College | 1 | 1 | 1 | 0 |
| 5 | College | 0 | 1 | 1 | 0 |
| 6 | College | 0 | 0 | 1 | 1 |

- Average Treatment Effect (ATE):

$$\begin{aligned}\bar{\tau} &= \bar{Y}(1) - \bar{Y}(0)\\ &= 1 - 0.5\\ &= 0.5\end{aligned}$$

# RCM (1974): Real World

| ID | Education $X_i$ | Treated $W_i$ | No job training $Y_i(0)$ | Job training $Y_i(1)$ | Treatment effect $\tau_i = Y_i(1) - Y_i(0)$ |
|---|---|---|---|---|---|
| 1 | High school | 0 | 0 | ? | ? |
| 2 | High school | 1 | ? | 1 | ? |
| 3 | High school | 1 | ? | 1 | ? |
| 4 | College | 1 | ? | 1 | ? |
| 5 | College | 0 | 1 | ? | ? |
| 6 | College | 0 | 0 | ? | ? |

- Average Treatment Effect:

$$\bar{\tau} = 0.66$$

$$\downarrow$$

32% bigger: why this bias?

# Selection Bias (intuition)

- People do not randomly select into various programs which we would like to evaluate

| ID | Education $X_i$ | Treated $W_i$ | No job training $Y_i(0)$ | Job training $Y_i(1)$ | Treatment effect $\tau_i = Y_i(1) - Y_i(0)$ |
|---|---|---|---|---|---|
| 1 | High school | 0 | 0 | ? | ? |
| 2 | High school | 1 | ? | 1 | ? |
| 3 | High school | 1 | ? | 1 | ? |
| 4 | College | 1 | ? | 1 | ? |
| 5 | College | 0 | 1 | ? | ? |
| 6 | College | 0 | 0 | ? | ? |

Higher treatment rate & higher treatment effects: $W_i \not\perp Y_i(0), Y_i(1)$

# Selection Bias (mathematical intuition)

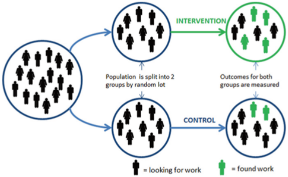
- As noted above, simply comparing those who are and are not treated may provide a misleading estimate of a treatment effect
- This problem can be efficiently described by using mathematical expectation notation to denote population averages:

$$\begin{aligned}\bar{\tau} &= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]}_{\text{Average Treatment Effect on the Treated}} + \underbrace{\left[\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]\right]}_{\text{Selection bias}}\end{aligned}$$

- Thus, the naive contrast can be written as the sum of two components, ATET, plus Selection Bias
- Average earnings of non-trainees, $\mathbb{E}[Y_i(0)|W_i = 0]$, may not be a good standing for the earnings of trainees had they not been trained, $\mathbb{E}[Y_i(0)|W_i = 1]$

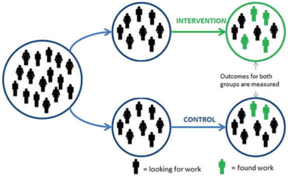
# Possible solutions

- The problem of selection bias motivates the use of:
  1. Random assignment (ex-ante) $\rightarrow$ experimental set-up




The basic design of a randomised controlled trial (RCT)

  2. Unconfoundedness (ex-post) $\rightarrow$ observational studies




  3. Instrumental variable (ex-post) $\rightarrow$ observational studies

# Random Assignment

- Random assignment ensures that the potential earnings of trainees had they not been trained are well-represented by the randomly selected control group
- Formally, when $W_i$ is randomly assigned, then:

$$\mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0] = [Y_i(1) - Y_i(0)|W_i = 1] = E[Y_i(1) - Y_i(0)]$$

- Replacing $E[Y_i|W_i = 1]$ and $E[Y_i|W_i = 0]$ with the corresponding sample analog provides a consistent estimate of ATE

# Unconfoundedness (or CIA)

- The unconfoundedness assumption states that conditional on observed characteristics, the selection bias disappears
- Formally, we overcome the problem that we have seen at slide 9, because: $W_i \perp\!\!\!\perp Y_i(0), Y_i(1) | X_i$
  This holds true even if conditioning just on:
  $e(x) = P(W = 1 | X_i = x)$
- Given unconfoundedness, comparison of average effects of job training have a causal interpretation:

$$\bar{\tau} = \mathbb{E}[Y_i(1)|W_i = 1, X_i] - \mathbb{E}[Y_i(0)|W_i = 0, X_i] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$$

- This can be generalized to the case of a continuous treatment variable (i.e effects of education on employment): $s_i \perp\!\!\!\perp Y_{s_i} | X_i$
- Conditional on $X_i$, what is the average causal effect of a one-year increase in collage attendance?

$$\mathbb{E}[Y_i|s_i = s, X_i] - \mathbb{E}[Y_i|s_i = s-1, X_i] = \mathbb{E}[f_i(s) - f_i(s-1)|X_i]$$

# *Machine Learning and Causality*

## *Using machine learning to estimate heterogeneous causal effect*

# Machine Learning and Causality

- Econometrics/ Statistics/ Social Science
  - Formal theory of causality
    - Potential outcomes methods (Rubin) maps onto economic approaches
  - Well-developed and widely used tools for estimation and inference of causal effect in experimental and observational studies
    - Used by social science, policy-makers, development organizations, medicine, business, experimentation
  - Weaknesses
    - Non-parametric approaches fail with many covariates
    - Model selection unprincipled

# Motivations

- Experiments and Data-Mining
  - Concerns about ex-post "data-mining"
    - In medicine, scholars are required to pre-specify analysis plan (similar in economic field experiments)
- How is it possible to deal with sets of treatment effects among subsets of the entire population?
- Estimate of treatment effect heterogeneity needed for optimal decision-making

**Definition 1 (Athey and Imbens, 2015; 2016)**

1. Estimating heterogeneity by features in causal effects in experimental or observational studies
2. Conduct inference about the magnitude of the differences in the treatment effects across subsets of the population

# Causal Inference Framework

- Causal inference in observational studies:
  - As we saw previously, assuming unconfoundedness to hold, we can treat observations as having come from a randomized experiment
  - Therefore we can define the conditional average treatment effect (CATE) as follows:

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$$

  - The population average treatment effect then is:

$$\tau^p = E[Y_i(1) - Y_i(0)] = E[\tau(X_i)]$$

# Why is CATE important?

- There are a variety of reasons that researchers wish to conduct estimation and inference on $\tau(x)$:

  1. It my be used to assign future units to their optimal treatment (in presence of different levels of the treatment):

  $$W_i^{opt} = max\, \tau(X_i)$$

  2. If we don't pre-specify the sub-populations it can be the case that the overall effect is negative, but it can be positive on subpopulations, then:

  $$W_i^{PTE} = \mathbf{1}_{\tau(X_i)\geq 0}$$

  e.g.: treatment is a drug $\rightarrow$ prescribe it just to those who benefit from it

# Using Trees to Estimate Causal Effects

Athey and Imbens (2015; 2016) propose 3 different approaches:

- Approach I: Analyze two groups separately:
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i$=1
  - Estimate $\hat{\mu}(0, x)$ using dataset where $W_i$=0
  - Preform within group cross-validation to choose tuning parameters
  - Predict $\hat{\tau} = \hat{\mu}(1, x) - \hat{\mu}(0, x)$
- Approach II: Estimate $\mu(w, x)$ using just one tree:
  - Estimate $\hat{\mu}(1, x)$ and $\hat{\mu}(0, x)$ using just one tree
  - Preform within tree cross-validation to choose tuning parameters
  - Predict $\hat{\tau} = \hat{\mu}(1, x) - \hat{\mu}(0, x)$
  - Estimate is zero for $x$ where tree does not split on $w$

# The CATE Transformation of the Outcome

- The authors' goal is to develop an algorithm that generally leads to an accurate approximation of $\hat{\tau}$ the Conditional Average Treatment Effect.
  1. Ideally we would measure the quality of the approximation in terms of goodness of fit using the MSE:

$$Q^{infeas} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2$$

  2. We can address this problem of infeasibiliy by transforming the outcome using the treatment indicator $W_i$ and $e(X)$:

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - e(X_i)}{(1 - e(X_i)) \cdot e(X_i)}$$

  3. Then:

$$E[Y_i^* | X_I = x] = \tau(x)$$

# How to estimate the In-Sample Goodness of fit?

- The ideal goodness of fit measure would be:

$$Q^{infeas}(\hat{\tau}) = \mathbb{E}[(\tau_i - \hat{\tau}(X_i))^2].$$

- A useful proxy that can be used for the goodness of fit measure is:

$$\mathbb{E}[\tau_i^2 | X_i \in S_j] = \frac{1}{N} \sum_i \hat{\tau}(x_i)^2.$$

This leads to our In-sample goodness of fit function:

$$Q^{is} = -\frac{1}{N} \sum_i \hat{\tau}(x_i)^2.$$

# Transformed Outcome Tree Model

- Approach 3:

  1. Model and Estimation
     - Model Type: Tree structure
     - Estimator $\hat{\tau}_i^{TOT}$: sample average treatment effect within leaf
  2. Criterion function (for fixed tuning parameter $\lambda$)
     - In-sample Goodness-of-fit function:

$$Q^{is} = -MSE = -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{TOT})^2$$

     - Structure and use of criterion:

$$Q^{crit} = Q^{is} - \lambda \times leaves$$

     - Select member of set of candidate estimators that maximizes $Q^{crit}$, given $\lambda$
  3. Cross-validation approach
     - Out-of-Sample Goodness-of-fit function:

$$Q^{oos} = -MSE = -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{TOT} - Y_i^*)^2$$

     - Approach: select tuning parameter $\lambda$ with highest $Q^{os}$

# Critique to the TOT approach

- Transformation of the Outcome in a randomized set-up:

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - p}{(1-p) \cdot p} = \begin{cases} \frac{1}{p} \cdot Y_i^{obs} & if \;\; W_i = 1 \\ -\frac{1}{1-p} \cdot Y_i^{obs} & if \;\; W_i = 0 \end{cases}$$

- Within a leaf the sample average of $Y_i^*$ is not the most efficient estimator of treatment effect

- The proportion of treated units within the leaf is not the same as the overall sample proportion

- We use a weighted estimator similar to the Hirano, Imbens and Ridder (2003) estimator

# Causal Tree Approach

- In details the Treatment Effect in a generic leaf $\mathbb{X}_j$ is:

$$\tau^{CT}(X_i) = \frac{\sum_{j:X_j \in \mathbb{X}_j} Y_i^{obs} \cdot \frac{W_i}{\hat{e}(X_i)}}{\sum_{j:X_j \in \mathbb{X}_j} \frac{W_i}{\hat{e}(X_i)}} - \frac{\sum_{j:X_j \in \mathbb{X}_j} Y_i^{obs} \cdot \frac{(1-W_i)}{(1-\hat{e}(X_i))}}{\sum_{j:X_j \in \mathbb{X}_j} \frac{(1-W_i)}{(1-\hat{e}(X_i))}}$$

- This estimator is a consistent estimator of:

$$\tau_{\mathbb{X}_j} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathbb{X}_j]$$

- The variance can be estimated the Neyman estimator:

$$\hat{\mathbb{V}}_{Neyman} = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$$

These two quantities can be estimated as:

$$s_{t,j}^{te,2} = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i(1) - \overline{Y}_t^{obs})^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i - \overline{Y}_t^{obs})^2$$

$$s_{c,j}^{te,2} = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i(0) - \overline{Y}_c^{obs})^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i - \overline{Y}_c^{obs})^2$$

# Attractive features of Causal trees

1. Can easily separate tree construction from treatment effect estimation
2. Tree constructed on training sample is independent of sampling variation in the test sample
3. Holding tree from training sample fixed, can use standard methods to conduct inference within each leaf of the tree on test sample
4. Can use any valid method for treatment effect estimation, not just the methods used in training
5. Simulations run by the authors show that the Causal Tree Algorithm outperforms the ST, TT and TOT approaches

# Case Study

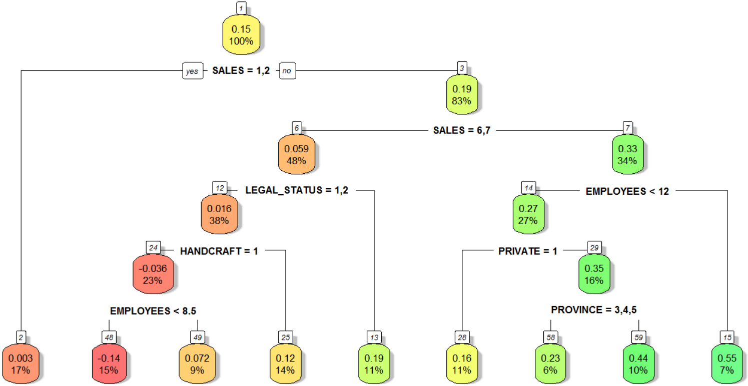



Figure: Bargagli-Stoffi & Gnecco (2020)

# Causal Forests

An individual tree can be *noisy* as we saw in the last lecture $\rightarrow$ instead, fit a causal forest

1. Draw a sample of size $s$
2. Split into a $\mathcal{D}$ and $\mathcal{I}$ sample
3. Grow a tree on $\mathcal{D}$
4. Estimate the effects on $\mathcal{I}$

Repeat many times

- Pros:
    1. Consistency for true $t(x)$
    2. Asymptotic normality
    3. Asymptotic variance is estimable
- Cons:
    1. Require sample splitting
    2. Large samples for asymptotic properties
    3. Not interpretable

# Bayesian Causal Forest (BCF)

- BCF were introduced by Hahn et al. (2020)
- BCF is a causal version of BART that:
  - has a similar priors of BART (higher probability of smaller trees and *stumps*, different hyper-priors to scale the leaves distribution of $\tau_i$)
  - accounts for *measure* confounding through the inclusion of the propensity score in the model
- Model parametrization:

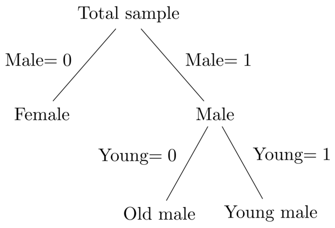
$$y_i = \mu(x_i, \hat{\pi}(x_i)) + \tau(x_i)w_i + \epsilon_i$$

Direct effects of $x_i$ and $\hat{\pi}(x_i)$ on $y_i$

Heterogeneous causal effects

# Causal rules and interpretability

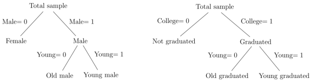
- In a causal scenario, interpretability can be defined as the ability of the algorithm to identify the subgroups where the effects are heterogeneous
- Decision rules are simple *if-then* statements regarding several conditions
- Rule-based learning improves interpretability



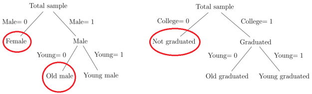

- Causal Rule Ensemble (CRE) algorithm (Lee, Bargagli-Stoffi and Dominici, 2020)

# Intuition on CRE

- Intuition on the CRE algorithm (5 steps):
  1. Divide the overall sample into a *discovery* and *estimation* sample
  2. Estimate the unit-level treatment effect $\tau^d(x)$ (where $X_i = x$)
  3. On the *discovery* build a series of causal rules by regressing $\tau^d(x)$ on $X_i$ using random forest (Breiman, 2001) and gradient trees (Friedman, 2001)




  4. Select the *most important* rules using stability selection (Meinshausen and Bühlmann, 2010)




  5. On the *estimation* sample estimate the treatment effects by regressing the estimated unit level treatment effects $\tau^e(x)$ on the selected rules

# Conclusions

1. The main problem to face is the absence of a *ground truth* when we deal with causal inference problems
2. The approaches developed are strongly data-driven: selection of subpopulation is optimized by the algorithm
3. Work well with randomized experiments and some techniques (i.e., BCF, CRE) control for potential confounding bias
4. The approaches are tailored for applications where:
    1. there may be many attribute relative to the number of units observed (*fat-data*)
    2. the functional form of the relationship between treatment effects and the attributes of units ins not known

# Further Readings


- S.Athey, G.Imbens *Machine learning methods for estimating heterogeneous causal effects,* 2015
- S.Athey, S.Wager *Estimation and Inference of Heterogeneous Treatment Effects using Random Forest*, 2015
- L. Breiman. *Random Forest,* Machine learning, 24:123-140, 2001
- L. Breiman, J.H. Olshen, C.J. Stone. *Classification and Regression Trees,* CRC press, 1984
- T.J. Hastie, R.J. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning.* Speringer, New York, 2009
- K.P. Murphy. *Machine Learning. A Probabilistic Perspective.* The MIT Press, Cambridge, Massachusetts, 2012
- K. Lee, F. J. Bargagli-Stoffi, F. Dominici, *Causal Rule Ensemble: Interpretable Inference of Heterogeneous Treatment Effects,* forthcoming, 2020